Eduardo de Mattos Pinto Coelho

Ontologias Difusas no Suporte à Mineração de Dados: aplicações na Secretaria de Finanças da Prefeitura Municipal de Belo Horizonte

Belo Horizonte
Escola de Ciência da Informação
2012

Eduardo de Mattos Pinto Coelho

Ontologias Difusas no Suporte à Mineração de Dados: aplicações na Secretaria de Finanças da Prefeitura Municipal de Belo Horizonte

Tese apresentada ao curso de Doutorado do Programa de Pós Graduação da Escola de Ciência da Informação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Ciência da Informação.

Área de concentração: Produção, Organização e Utilização da Informação.

Linha de pesquisa: Gestão da Informação e do Conhecimento.

Orientador: Prof. Marcello Peixoto Bax

Coorientador: Prof. Wagner Meira Júnior

Belo Horizonte
Escola de Ciência da Informação da UFMG
2012

Coelho, Eduardo de Mattos Pinto.

C672o Ontologias difusas no suporte à mineração de dados [manuscrito] : aplicações na Secretaria de Finanças da Prefeitura Municipal de Belo Horizonte/ Eduardo de Mattos Pinto Coelho. – 2012.

232 f. : il., enc.

Orientador: Marcello Peixoto Bax.
Coorientador: Wagner Meira Júnior.
Tese (doutorado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências: f. 205-222
Anexos: f. 223-232

1. Ciência da informação – Teses. 2. Representação do conhecimento (Teoria da informação) – Teses. 3. Ontologias (Recuperação da informação) – Teses. 4. Mineração de dados (Computação) – Teses. I. Título. II. Bax, Marcello Peixoto. III. Meira Júnior, Wagner. IV. Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

CDU: 025.4.03

Ficha catalográfica: Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG

UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

ATA DA DEFESA DE TESE DE **EDUARDO DE MATTOS PINTO COELHO**, matrícula: 2007671365

Às 9:00 horas do dia 05 de novembro de 2012, reuniu-se na Escola de Ciência da Informação da UFMG a Comissão Examinadora aprovada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação em 04/10/2012, para julgar, em exame final, o trabalho intitulado ***Ontologias difusas no suporte à mineração de dados: aplicações na Secretaria de Finanças da Prefeitura Municipal de Belo Horizonte***, requisito final para obtenção do Grau de DOUTOR em CIÊNCIA DA INFORMAÇÃO, Área de Concentração: Produção, Organização e Utilização da Informação, Linha de Pesquisa: Gestão da Informação e do Conhecimento - GIC. Abrindo a sessão, o Presidente da Comissão, Prof. Dr. Marcello Peixoto Bax, após dar conhecimento aos presentes do teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a argüição pelos examinadores com a respectiva defesa do candidato. Logo após, a Comissão se reuniu sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foram atribuídas as seguintes indicações:

Prof. Dr. Marcello Peixoto Bax - Orientador	APROVADO
Prof. Dr. Wagner Meira Júnior - Co-orientador	APROVADO
Prof. Dr. Frederico Luiz Gonçalves de Freitas	APROVADO
Prof. Dr. Alberto Henrique Frade Laender	APROVADO
Prof. Dr. Maurício Barcellos Almeida	APROVADO
Prof. Dr. Fernando Silva Parreiras	APROVADO
Profa. Dra. Renata Maria Abrantes Baracho Porto	APROVADO

Pelas indicações, o candidato foi considerado APROVADO.

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a sessão, da qual foi lavrada a presente ATA que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 05 de novembro de 2012

Prof. Dr. Marcello Peixoto Bax
ECI/UFMG (Orientador)

Prof. Dr. Wagner Meira Júnior
DCC/ICEX/UFMG (Co-orientador)

Prof. Dr. Frederico Luiz Gonçalves de Freitas
UFPE

Prof. Dr. Fernando Silva Parreiras
FUMEC

Prof. Dr. Alberto Henrique Frade Laender
DCC/ICEX/UFMG

Prof. Dr. Maurício Barcellos Almeida
ECI/UFMG

Profa. Dra. Renata Maria Abrantes Baracho Porto
ECI/UFMG

Profa. Renata Maria Abrantes Baracho
coordenadora do Programa de Pós-Graduação em Ciência da Informação - ECI / UFMG

Obs: Este documento não terá validade sem a assinatura e carimbo da Coordenadora.

UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

FOLHA DE APROVAÇÃO

"ONTOLOGIAS DIFUSAS NO SUPORTE À MINERAÇÃO DE DADOS: APLICAÇÕES NA SECRETARIA DE FINANÇAS DA PREFEITURA MUNICIPAL DE BELO HORIZONTE"

Eduardo de Mattos Pinto Coelho

Tese submetida à Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais, como parte dos requisitos à obtenção do título de **"Doutor em Ciência da Informação"**, Linha de Pesquisa: **"Gestão da Informação e do Conhecimento - GIC"**.

Tese aprovada em: 05 de novembro de 2012.

Por:

Prof. Dr. Marcello Peixoto Bax - ECI/UFMG (Orientador)

Prof. Dr. Wagner Meira Júnior - DCC/ICEX/UFMG (Co-orientador)

Prof. Dr. Frederico Luiz Gonçalves de Freitas - UFPE

Prof. Dr. Fernando Silva Parreiras- FUMEC

Prof. Dr. Alberto Henrique Frade Laender - DCC/ICEX/UFMG

Prof. Dr. Maurício Barcellos Almeida - ECI/UFMG

Profa.Dra. Renata Maria Abrantes Baracho Porto – ECI/UFMG

Aprovada pelo Colegiado do PPGCI

Profa. Renata Maria Abrantes Baracho Porto
Coordenadora Pró-Tempore

Versão final Aprovada por

Prof. Marcello Peixoto Bax
Orientador

Dedicatórias

Sílvia

Uma dádiva! Parece-me estranho
só há pouco conhecer-te, pois
parece que sempre nos tivemos:
"eu te gosto, você me gosta
desde tempos imemoriais".[1]

Carolina

Repara um pouquinho nesta,
no queixo, no olhar, no gesto,
e na consciência profunda
e na graça menineira,
e dize, depois de tudo,
se não é, entre meus erros,
uma imprevista verdade.
Esta é minha explicação,
meu verso melhor ou único,
meu tudo enchendo meu nada.[2]

Camilla

A gente se apega, se achega, e se engrandece no afeto

Sylvia, Lígia e Eliana

"Tudo se realizará"
dizem as ondas ao duro litoral[3]

---

[1] Drummond: *Balada do Amor Através das Idades.*
[2] Drummond: *A mesa.*
[3] Neruda*: Ode à Esperança*.

Agradecimentos

À Escola de Ciência da Informação da UFMG (ECI/UFMG) pela oportunidade de desenvolver este projeto.

Agradecimento também em especial ao Prof. Marcello Peixoto Bax pela valiosa aprendizagem, orientação, paciência e amizade.

Ao Prof. Wagner Meira Júnior, amigo, orientador e incentivador de primeira hora.

Ao Prof. Maurício Barcellos de Almeida também pela amizade e valiosa aprendizagem nas duas disciplinas ministradas sobre ontologias aplicadas e no Estudo Especial sobre Causalidade.

Ao Auditor Milton Rodrigues Alves, gerente da Gerência de Inteligência e Fiscalização Estratégica (GEIFE), pela confiança e pelo apoio do início, ao último momento, sem o qual a conclusão do projeto não teria sido possível.

Ao Auditor João Bráulio Cruz de Vilhena, gerente da Gerência de Inteligência Fiscal (GEINFE) pelo apoio recente e crucial nesses últimos momentos.

Aos meus familiares e amigos pela tolerância e compreensão com o meu afastamento do convívio.

Ao Celso Antônio Alves Neto pelo apoio, discussão e revisão de algumas questões filosóficas abordadas na tese.

À funcionária da Secretaria da ECI/UFMG, Gisele da Silva Reis, pela gentileza e profissionalismo.

Agradecimento especialíssimo à minha mãe, Da. Sylvia que sempre valorizou e estimulou a formação dos filhos, pelo contínuo apoio e confiança.

SUMÁRIO

FIGURAS

TABELAS

ABREVIAÇÕES E SIGLAS

AKD	*Actionable Knowledge Discovery and Delivery*
CMC	Cadastro Municipal de Contribuintes de Tributos Mobiliários
DDDM	*Domain Driven Data Mining*
D^3M	*Domain Driven Data Mining*
DIG	*Description-Logic Implementation Group*
DL	Lógica Descritiva (tradução de DL - *Description Logic)*
IBPT	Instituto Brasileiro de Planejamento Tributário
ICMS	Imposto Sobre a Circulação de Mercadorias e Serviços
ISSQN	Imposto Sobre Serviços de Qualquer Natureza
JDBC	Java Database Connectivity
ODBC	Open Database Connectivity
OIL	Ontology Interchange Language
OWL	Web Ontology Language
PBH	Prefeitura Municipal de Belo Horizonte
RDF	Resource Description Framework
RDF-S	Resource Description Framework Schema
RuleML	Rule Markup Language
SWRL	*Semantic Web Rule Language*
SQWRL	*Semantic Query Web Rule Language*
W3	*World Wide Web*
W3C	*W3 Consortioum*
WWW	*World Wide Web*
XML	*EXtended Markup Language*

RESUMO

Este projeto de pesquisa tem como objetivo o aprimoramento de tecnologia de suporte à detecção, análise e combate a fraudes fiscais no Imposto Sobre Serviços de Qualquer Natureza – ISSQN – a partir da recuperação e processamento de grande volume de dados. A hipótese que lançamos é que *a associação de metodologias e técnicas de ontologias e sistemas difusos pode auxiliar e mesmo viabilizar o sucesso da mineração de dados na recuperação destes grandes volumes de dados*. Esta hipótese fundamenta-se em três premissas. A primeira premissa é que a solução passa pela necessidade da captura, tratamento, modelagem, representação formal e incorporação do conhecimento de domínio, em especial, o constituído pelo conhecimento coletivo de especialistas. A segunda premissa é que ontologias adequam-se perfeitamente à tarefa dessa incorporação de conhecimento. Essa adequação se faz em vista das características das ontologias em explicitar, formalizar, verificar, e consolidar conhecimentos, tornando-os compartilháveis, reutilizáveis e interoperáveis. Além disso, o conhecimento representado e incorporado é naturalmente utilizado para fins de classificação, a partir dos mecanismos de inferências inerentes às ferramentas de desenvolvimento de ontologias, e aos mecanismos de inferências que podem ser agregados a elas. A terceira premissa é que, considerando-se que esse conhecimento coletivo de especialistas é de natureza vaga e subjetiva, as metodologias e técnicas da área de sistemas difusos são adequadas para capturar, tratar e modelar esse conhecimento. Com isto, desenvolvemos uma solução para o pré e pós-processamento de dados na mineração, focada na modelagem de conhecimento especialista de natureza vaga e subjetiva. Este conhecimento vago e subjetivo é modelado em atributos com técnicas de sistemas difusos, guiando o processo de mineração de dados, e gerando uma medida subjetiva que suporta a análise e interpretação de resultados, que de outro modo, seria mais laboriosa, difícil ou mesmo impossível de ser realizada. Desse modo, apresentamos uma solução efetiva para o aumento da arrecadação tributária, através da identificação de evidências de fraudes e evasão tributária em serviços importados por empresas de Belo Horizonte. Com base nesta abordagem, nos testes e simulações realizados, pudemos reduzir o número de regras de interesse geradas pela mineração de dados em 42%. Comparando-se o número de registros detectados na mineração de dados por esta abordagem, com o número de registros conhecidos envolvendo fraudes comprovadas, obtivemos uma taxa de sucesso global de 95,88%. Esta solução tem o potencial para ser aplicada em outras situações, e em amplas áreas de aplicação na esfera pública e privada. O projeto explora a convergência das habilidades desenvolvidas por três diferentes instituições: a Escola de Ciência da Informação da UFMG, o Departamento de Ciência da Computação da UFMG e a Secretaria Municipal de Finanças da Prefeitura de Belo Horizonte.

ABSTRACT

This research project aims at the improvement of technology support for detecting, analyzing and fighting tax evasion in Service of Any Kind Tax- ISSQN - from recovery and processing large volumes of data. The hypothesis is that the association of methodologies and techniques of ontologies and fuzzy systems, can even help facilitate the success of data mining in the recovery of these large volumes of data. This hypothesis is based on three premises. The first premise is that the solution uses the capture, processing, modeling, representation and formal incorporation of domain knowledge, in particular, constituted by the collective knowledge of experts. The second premise is that ontologies are perfectly suited to the task of knowledge incorporation. This adequacy is obtained in view of the characteristics of ontologies in explicit, formalize, verify and consolidate the knowledge, making them sharable, reusable and interoperable. In addition, the knowledge represented and incorporated is naturally used for classification, from the inherent inference mechanisms of development tools ontologies, and the mechanisms of inferences that can be added to them. The third premise is that, considering that the collective knowledge of experts is vague and subjective in nature, methodologies and techniques in the area of fuzzy systems are adequate to capture, treat and model this knowledge. Then, we have developed a solution for the pre-and post-processing of data mining, focused on modeling expert knowledge of nature vague and subjective. This vague and subjective knowledge is modeled to attributes with fuzzy systems techniques, guiding the process of data mining, and generating a subjective measure that supports the analysis and interpretation of results that otherwise would be more laborious, difficult or even impossible to be performed. Thus, we present an effective solution for increasing tax revenue through the identification of evidence of fraud and tax evasion on imported services for companies of Belo Horizonte. Based on this approach, tests and simulations conducted, we can reduce the number of rules of interest generated by data mining by 42%. Comparing the number of records in data mining detected by this approach, with the number of known records involving fraud proven, we obtained an overall success rate of 95.88%. This solution has the potential to be applied in other situations, and in wide areas of application in the public and private sphere. The project explores the convergence of skills developed by three different institutions: the School of Information Science at UFMG, the Department of Computer Science at UFMG and the Finance City Office of Belo Horizonte.

1 Introdução

1.1 Escopo

As organizações responsáveis pela fiscalização e controle de setores econômicos desenvolvem um alto grau de especialização na detecção de fraudes e irregularidades. Além de serem desafiadas pela diversidade e dinamismo das práticas ilícitas adotadas, frequentemente elas são sobrepujadas pelo volume gigantesco e crescente de dados que afluem às suas bases.

Em vista disso, são crescentes as demandas por soluções originárias da Tecnologia da Informação amparadas em desenvolvimentos oriundos da Ciência da Computação e da Ciência da Informação. A Ciência da Computação dá o principal suporte ao desenvolvimento de ferramentas de *hardware* e *software* em Tecnologia da Informação que prometem amenizar os problemas e aprimorar as soluções, sob o critério do desempenho, da eficiência e da eficácia no processamento dos dados. Entretanto, muitas vezes a adoção dessas ferramentas falha completamente, ou possui seu alcance limitado, por não considerar o problema em sua concepção mais ampla, onde aspectos sócio-humanos e culturais são definidores e decisivos para o efetivo sucesso das soluções adotadas.

Nesse aspecto, a Ciência da Informação, como ciência social aplicada, é capaz de levar a consideração do problema para uma perspectiva mais elaborada, multidisciplinar e multifacetada. Além disso, oferece um elenco relevante de abordagens e soluções que permite atingir seus objetivos com um enfoque muito mais abrangente do que um mero suporte tecnológico, fazendo a ponte e integração das várias áreas de saberes e levando em consideração não somente o dado, mas a inserção da informação e do conhecimento na organização.

O problema técnico-científico com que trabalharemos é o da representação do conhecimento, e de como a constituição de ontologias podem auxiliar o processo de mineração de dados.

As ontologias construídas são integradas à ferramenta de mineração de dados, nas etapas de pré e pós-processamento, de modo a permitir a recuperação de

informações relevantes que se constituam em indícios de fraude e sonegação a partir de excessiva massa de dados. Desse modo, temos uma efetiva ferramenta para auxiliar o processo de decisão sobre quais setores, empresas, ou transações merecerão uma auditoria específica.

Considerando que os conhecimentos a serem incorporados às ontologias são de natureza sócio-humana, onde a definição de descrições, limiares e critérios de classificação envolvem alto grau de imprecisão, incerteza, vaguidade e considerações de ordem subjetiva alguns aspectos lógico-ontológicos serão considerados sob a ótica de sistemas difusos.

Ontologias difusas[4] apresentam a vantagem de permitir uma elaboração e representação de aspectos conceituais, simbólicos, qualitativos e abstratos de uma área de domínio. Também apresentam a vantagem de lidar, simultaneamente, com aspectos quantitativos, concretos ou numéricos dessa área de domínio. O sistema resultante foi submetido a simulações, testes e mensurações dos resultados obtidos. Assim as ontologias de aplicação incorporarão metaclasses, regras e mecanismos de inferência que darão suporte à classificação de instâncias difusas.

1.2 Relevância Sócio-Econômica da Pesquisa

A pesquisa insere-se no trabalho de inteligência fiscal desenvolvido na Gerência de Inteligência e Fiscalização Estratégica (GEIFE) da auditoria tributária da Secretaria de Finanças da Prefeitura de Belo Horizonte. A Auditoria Tributária possui em seu elenco de atribuições e deveres a contribuição para o aperfeiçoamento do sistema fiscal, a prevenção e o combate à fraude e à evasão fiscal.

Seu programa de inteligência consiste, em especial, do acompanhamento, monitoramento e supervisão de empresas e de eventos e segmentos econômicos de grande porte, com foco nos principais contribuintes e responsáveis tributários do Imposto Sobre Serviços de Qualquer Natureza, o ISSQN.

---

[4] O termo difuso está associado às noções de imprecisão, incerteza, vaguidade e nebulosidade.

Este trabalho de inteligência é fundamentado na busca, coleta, recepção, seleção e análise de dados a fim de gerar informação e conhecimento, permitindo a detecção de padrões já previamente conhecidos, ou de novidades, permitindo construir e atualizar perfis e cenários envolvendo os contribuintes, suas práticas e situações de mercado e de panorama econômico. Utilizam-se técnicas de cruzamentos e conciliações dos dados obtidos, de modo a detectar e tratar práticas eventuais e sistêmicas de evasão, sonegação, elisão e fraude fiscal.

De um lado, há dados estruturados e semiestruturados advindos de inúmeras fontes. De outro lado, há o conhecimento dos auditores estabelecido em anos de experiência que fazem a análise das informações disponíveis para a tomada de decisão. Com base nessas informações, as ações fiscais são direcionadas, facilitando e otimizando o trabalho, onde o rol de informações é filtrado, a base de conhecimento é ampliada, e as ações fiscais podem ser mais bem planejadas e orientadas (SILVA, 2006a).

Um estudo realizado pelo Instituto Brasileiro de Planejamento Tributário (IBPT, 2008) informa que a sonegação das empresas brasileiras corresponde a 25% do seu faturamento, e alcança um faturamento não declarado de R$1,32 trilhão, com tributos sonegados na ordem de 200 bilhões por ano, atingido, em conjunto com os tributos sonegados pelas pessoas físicas a 9% do PIB. De acordo com este estudo, considerando-se o porte das empresas, o percentual de sonegação estimado ocorre nos seguintes níveis.

TABELA 1.1 Indícios de sonegação por porte das empresas

Porte	Percentual de sonegação
Pequenas	64,65
Médias	48,94
Grandes	26,78

Fonte: IBPC, 2008, p. 4.

No setor de serviços, setor alcançado pela incidência do ISSQN, o índice de sonegação é da ordem de 25,02%, e foram encontrados fortes indícios de sonegação fiscal em aproximadamente 26,84% das empresas pesquisadas (TABELA 1.1). A TABELA 1.2 indica valores associados a setores de atividade econômica, e a TABELA 1.3 destaca esses valores para o setor de serviços.

TABELA 1.2 Sonegação por setor de atividade econômica em 2008

N°	PESSOA JURÍDICA – Setor Econômico	Crédito (R$)
1.	Indústria	78.772.920.287
2.	Comércio	74.146.288.051
3.	Transportes e serviços relacionados	4.666.689.659
4.	Construção civil	3.965.684.774
5.	Serviços de comunicação, energia e água	5.588.010.364
6.	Serviços financeiros	8.432.087.323
7.	Outros Serviços	24.715.429.350
	TOTAL SONEGADO	**200.287.109.807**

TABELA 1.3 Valores e índices de sonegação em setores de serviços

N°	PESSOA JURÍDICA – Setor Econômico	Crédito (R$)	PERCENTUAL
1.	Transportes e serviços relacionados (1)	4.666.689.659	9,8520
2.	Construção civil (2)	3.965.684.774	**8,3720**
3.	Serviços de comunicação, energia e água (1)	5.588.010.364	11,7970
4.	Serviços financeiros (2)	8.432.087.323	**17,8013**
5.	Outros Serviços (2)	24.715.429.350	52,1775
	TOTAL SONEGADO	**47.367.901.470**	

(1) Sujeito ao ICMS (Importo de Circulação de Mercadorias e Serviços), ou ao ISSQN, dependendo o caso. Transporte intramunicipal, ISSQN; Transporte intermunicipal, ICMS. Serviços diretos de comunicação, energia e água, ICMS. Serviços acessórios, ISSQN.
(2) Sujeito ao ISSQN.

Fonte: Elaborado a partir da TABELA 1.2 (IBPT, 2008, p. 6), considerando-se apenas serviços (sujeitos à incidência do ICMS e ISS).

O uso da mineração de dados na busca do fato gerador do tributo que por fraude ou ocultação, simulação e dissimulação foge ao alcance tributário permitirá a sua devida inclusão na esfera da incidência tributária. Além disso, a eficiente ação antievasiva e antielisiva do Fisco alcança, e ajuda a combater a atuação de organizações criminosas, a corrupção, os financiamentos ilícitos de campanhas eleitorais, o contrabando, a lesão aos direitos, às finanças e saúde de consumidores.

A despeito de sua potencialidade, relatos recentes de projetos de mineração de dados nos EUA informam que 51% dos projetos são mal sucedidos[5]. No Brasil, a pesquisadora Lóren Gonçalves (GONÇALVES, 2004 e 2002) relatou estudos de casos e as inúmeras dificuldades na introdução com sucesso da mineração de dados em empresas do varejo.

Constata-se que boa parte da responsabilidade pelas dificuldades e fracassos em mineração de dados advém do mau conhecimento do negócio e dos dados submetidos à mineração de dados e à análise subsequente.

Diante dessas dificuldades, e justamente por sua capacidade de tratar, incorporar e formalizar o conhecimento do negócio e do entendimento dos dados, as ontologias possuem o potencial de auxiliar de forma decisiva na operacionalização e inserção da mineração de dados em uma organização.

Com o suporte de ontologias pode-se sobrepujar as limitações de autonomia e independência, e as limitações de utilidade dos sistemas de mineração de dados apontadas em Gonçalves (2004 e 2002). Com este suporte, a mineração de dados passa a possibilitar o alcance de uma solução, em situações onde seu sucesso seria pouco promissor.

Além disso, cabe lembrar que essa solução usufrui das vantagens usuais associadas a ontologias. Isto é, ela permite agregar ao sistema de mineração de dados uma disciplina metodológica na modelagem do conhecimento, buscando explicitá-lo e validá-lo do ponto de vista da consistência, completude e expressividade lógica. Tornando-o ainda compartilhável, reutilizável, e mais facilmente interoperável.

Esse projeto é capaz de despertar um amplo interesse por parte de organizações públicas e privadas.

Na esfera pública, a partir de onde o projeto foi desenvolvido, podemos citar como potenciais interessados as fiscalizações tributárias municipais dos mais de cinco mil

---

[5] Vide: http://www.kdnuggets.com/2012/05/tma-webinar-data-mining-failure-to-launch.html.

municípios brasileiros, as, fiscalizações tributárias estaduais e da União, Banco Central, Polícia Federal, Comissão de Valores Mobiliários, Ministério Público, Tribunais Eleitorais, e toda a gama de órgãos fiscalizadores e repressores de crimes contra a ordem tributária, crimes contra a ordem econômica, e de repressão ao crime organizado. Na esfera privada, não só em contextos de auditoria e controle, as metodologias e técnicas apresentadas podem ser úteis no apoio a decisões associadas ao negócio fim da organização.

1.3 Objetivo da Tese

Neste projeto de tese, partindo-se destas considerações e motivações, busca-se a construção de uma solução metodológica e técnica que permita aumentar a eficácia e a eficiência das organizações em detectar fraudes e irregularidades em registros econômico-financeiros para fins contábeis e fiscais. Trata-se de uma solução que envolve a recuperação de informações para apoio a decisões.

O problema estrito consiste na questão: *dado um conjunto de dados referentes a transações econômico-financeiras, como se pode aprimorar a recuperação de transações que indiquem ocorrências de irregularidades?*

Uma possível solução óbvia, e já plenamente defendida e justificada, consiste na adoção de metodologias e técnicas de mineração de dados para detectar essas irregularidades. Entretanto, a despeito de utilizarmos a mineração de dados, ela não é o foco de nossos estudos.

A **hipótese** que lançamos é que *a associação de metodologias e técnicas de ontologias e sistemas difusos pode auxiliar e mesmo viabilizar o sucesso da mineração de dados*.

Esta hipótese fundamenta-se em três premissas. A **primeira premissa** é que a solução passa pela necessidade da captura, tratamento, modelagem, representação formal e incorporação do conhecimento de domínio, em especial, o constituído pelo conhecimento coletivo de especialistas.

A **segunda premissa** é que ontologias, adequam-se perfeitamente à tarefa dessa incorporação de conhecimento. Essa adequação se faz em vista das características

das ontologias em explicitar, formalizar, verificar, e consolidar conhecimentos, tornando-os compartilháveis, reutilizáveis e interoperáveis. Além disso, o conhecimento representado e incorporado é naturalmente utilizado para fins de classificação, a partir dos mecanismos de inferências inerentes às ferramentas de desenvolvimento de ontologias, e aos mecanismos de inferências que podem ser agregados a elas.

A **terceira premissa** é que, considerando-se que esse conhecimento coletivo de especialistas é de natureza vaga e subjetiva, as metodologias e técnicas da área de sistemas difusos são adequadas para capturar, tratar e modelar esse conhecimento.

Assim, a segunda e terceira premissa atendem à demanda identificada na primeira premissa e sustentam a hipótese de solução: o desenvolvimento de ontologias difusas no suporte à mineração de dados.

1.4 Alcances e Limitações das Contribuições da Tese

No percurso do alcance desses objetivos, é feita uma apuração do estado da arte na construção de ontologias difusas, e do uso de linguagens de regras e de mecanismos de inferência baseados em regras. Desenvolvemos regras declarativas nas linguagens SWRL e SQWRL para expandir as capacidades expressivas da ontologia, e usufruir de sua capacidade de manipular cláusulas Horn, e de sua melhor capacidade de manipular operadores aritméticos necessários para definir os limites máximo e mínimo de conjuntos difusos.

Em meio a este levantamento, também é feita uma ligeira apresentação de linguagens desenvolvidas a partir de lógicas descritivas difusas, construídas como extensões de versões de linguagens suportadas pela lógica descritiva. Apuram-se as linguagens e raciocinadores difusos que vem sendo desenvolvidos, mas, considerando-se que ainda não se encontravam disponíveis para uso, não puderam ser efetivamente testados e avaliados.

Esclarecemos que não trazemos contribuições específicas para a área de mineração de dados, ou ontologias. O que é novo em nossa solução é a abordagem inter e

multidisciplinar que busca recursos na área de mineração de dados, ontologias, sistemas difusos, e de análise de inteligência.

O uso de noções, metodologias e técnicas advindas da área de sistemas difusos permeia toda esta pesquisa, e é o que a caracteriza. A engenharia de sistemas difusos permite capturar, incorporar e formalizar conhecimento especialista de natureza vaga e subjetiva, comum em domínios sócio-humanos, tais como economia, administração, contabilidade, direito, política, dentre outros domínios. No caso de serviços de inteligência do Estado, o conhecimento especialista vago e subjetivo é quase todo o seu material de trabalho.

Há uma grande dificuldade por parte dos profissionais com formação pura na área de ciências exatas em reconhecer a ocorrência dos fenômenos da vaguidade e da subjetividade na análise de fatos empíricos. Há ainda maior dificuldade em admitir a possibilidade de representar e incorporar estes fenômenos em uma modelagem lógico-matemática. Em vista disto, uma das contribuições da tese é a discussão do fenômeno da vaguidade. Mostra como a qualidade da informação é prejudicada por este fenômeno, considerando-se as dificuldades que ele traz relacionadas à representação, à recuperação, validação e avaliação da informação. Discute-se como ele é confundido com fenômenos similares, apresentam-se critérios de distinção e discutem-se suas várias correntes de abordagem, dando destaque às correntes usadas em sistemas formais que, por sua vez, permitem sua utilização na construção de ontologias. As abordagens destacadas são a da lógica sobreatribuitiva e da lógica difusa.

Também em vista destas dificuldades, esta tese busca contribuir com a discussão das especificidades envolvidas no tratamento de bases de conhecimento empíricas em domínios sócio-humanos. Mostra como as tarefas de descrição, prescrição, explicação, interpretação, previsão de fatos empíricos vinculam-se com ontologias, mineração de dados e, mesmo, a análise de inteligência. Destacam-se as especificidades do domínio de ciências humanas que exige na construção de suas teorias as noções de possibilidade, subjetividade e interpretação, contrastantes com as noções de necessidade, objetividade e explicação usadas em domínios de ciências naturais. Tais exigências demandam novas perspectivas em ontologias e mineração de dados. Tais demandas encontram guarida no paradigma emergente

da computação suave e do raciocínio aproximado, onde as abordagens baseadas em sistemas difusos se inserem.

Há metodologias e técnicas de análise de inteligência que buscam capturar e tratar o conhecimento coletivo a partir de opiniões de especialistas, tentando objetivar o que é intrinsecamente subjetivo e intersubjetivo. Essas metodologias e técnicas permitem refinar sistemática e paulatinamente esse conhecimento coletivo. Elas associam-se automaticamente à abordagem desenvolvida e são citadas ao longo da tese, mas não foram diretamente trabalhadas neste projeto. O conhecimento coletivo é apenas simulado, em vista da indisponibilidade de uma equipe de especialistas. Isto é, as opiniões usadas para fins de testes foram emitidas por um único especialista. A acoplagem efetiva dessas metodologias e técnicas ficou para as pesquisas futuras decorrentes.

1.5 Contribuições Específicas e Resultados obtidos

Além das contribuições gerais já citadas podemos destacar algumas contribuições mais específicas, mas que são técnicas que podem ser aplicadas em distintos problemas envolvendo mineração de dados.

O desenvolvimento do projeto envolveu a construção de duas ontologias e respectivos mecanismos de inferência. Uma ontologia para atuar na etapa prévia à mineração de dados, na categorização dos atributos a serem minerados. Outra ontologia para atuar na etapa posterior à mineração de dados, na classificação dos resultados gerados.

Um dos principais problemas da mineração de dados envolve o grande número de atributos que podem ser submetidos à mineração de dados. Primeiro, um número grande de atributos pode comprometer a performance do minerador, em função do aumento significativo das combinações possíveis. Segundo, e em consequência da multiplicidade combinatória, pode contribuir para aumentar significativamente o número de resultados da mineração de dados, dificultando, ou inviabilizando a análise dos resultados.

Desenvolvemos uma modelagem de modo a ser possível utilizar um operador de agregação que permitiu a transformação de um conjunto de atributos simples, em um único atributo que expressa a influência desses atributos simples sobre um determinado atributo composto. No caso de uso estudado, vários atributos simples concorrem para influenciar o *nível de suspeição* de transação econômica-financeira realizada. O peso da influência de cada um desses atributos foi capturado e ponderado com base em opiniões de especialistas, e o operador difuso de agregação utilizado permitiu que esses distintos atributos fossem resumidos em um único atributo. No caso em questão, transformamos dez atributos simples em um único atributo composto. Com isso, conseguimos a redução das regras resultantes da mineração de dados, propiciada pela redução das possibilidades combinatórias dos atributos.

Uma segunda contribuição específica do uso da modelagem, baseada em sistemas difusos na construção desse atributo composto, é que ele traz um nível de significação aos resultados capaz de orientar seu processo de análise. Nesse aspecto, esse atributo exerce a função de uma métrica subjetiva dos resultados obtidos, sobrepujando as dificuldades e limitações das métricas objetivas, já amplamente discutidas na literatura.

Com isso conseguimos a redução do número de regras de interesse, o que vai além da “mera” redução do número de regras obtida através da redução das possibilidades combinatórias. A construção e uso desse atributo que atua como uma métrica subjetiva permite-nos focar nas regras de relevância. Elas serão aquelas que tiverem esse atributo no consequente. Essa métrica associada com métricas objetivas e a sumarização de dados pertinentes permitem-nos construir critérios de seleção e ordenamentos dos resultados. Esses critérios são incorporados na ontologia de pós-mineração de dados. Desse modo, esse atributo conduz e facilita o processo de mineração de dados e a análise de resultados.

Com base nesta abordagem, nos testes e simulações realizados, pudemos reduzir o número de regras de interesse em 42%. Comparando-se o número de registros detectados na mineração de dados por esta abordagem, com o número de registros conhecidos envolvendo fraudes comprovadas, obtivemos uma taxa de sucesso global de 95,88%.

Considerando-se essas contribuições e os resultados obtidos pudemos concluir que a abordagem difusa é adequada para capturar e modelar conhecimento vago e subjetivo, fazendo uso de atributos simples que expressam esse conhecimento, e que podem ser agregados em um atributo composto que aprimora a mineração de dados e facilita o processo de análise de seus resultados.

O processo desenvolvido é semi-automático, pois as etapas de acoplamento e passagem de dados entre ontologias, bancos de dados, inserção e extração de dados das ontologias são ativadas manualmente. Entretanto, uma vez obtendo-se uma versão estável e funcional para determinado caso, ele pode ser plenamente automatizado.

1.6 Estrutura da Tese

A tese está estruturada em sete capítulos. Além dessa parte introdutória, divide-se em mais duas partes, com três capítulos cada.

Primeira Parte – Revisão da Literatura

A primeira parte apresenta trabalhos similares desenvolvidos em ontologias, especificamente ontologias e lógica difusa. Além disso, apresenta o tratamento que vem sendo dado ao suporte de ontologias à mineração de dados, e as especificidades em se trabalhar o conhecimento na área de ciências humanas que resvala nas organizações, onde o caráter aplicado dessa pesquisa se insere.

No **Capítulo 2**, *Vaguidade e a Importância de Sua Consideração para a Construção de Ontologias Aplicadas*, tratamos do fenômeno da vaguidade, ou vagueza e de seus efeitos de incerteza, imprecisão, inexatidão, ou indeterminação que gera dificuldades na representação, recuperação, validação e análise da informação. A vaguidade é conceituada buscando situá-la, ou distingui-la de conceitos e fenômenos relacionados, com as quais, normalmente, é confundida. Apresentam-se distintas correntes lógico-filosóficas que interpretam e lidam diferentemente com o fenômeno e suas decorrentes abordagens em ontologias aplicadas.

Há uma ampla discussão com destaque na consideração da lógica difusa e da lógica sobreatribuitiva. Estas duas lógicas vêm se sobressaindo no desenvolvimento de soluções em ontologias, em domínios onde o fenômeno da vaguidade se impõe.

A discussão da vaguidade é uma das grandes contribuições dessa tese e irá respaldar a adoção de soluções baseadas em conjuntos e lógica difusa.

No **Capítulo 3**, *Suporte de Ontologias Aplicadas á Mineração de Dados*, argumentamos que a mineração tende a gerar um número intratável de regras prejudicando a sua aplicabilidade. Para solucionar este problema, propõe-se o uso de ontologias nas etapas de pré e pós-processamento no suporte à mineração.

Além disso, o capítulo ressalta que organizações humanas exigem as noções de possibilidade, subjetividade e interpretação, contrastantes com as noções de necessidade, objetividade e explicação, úteis em domínios de ciências naturais. Tais exigências demandam novas perspectivas em ontologias e mineração de dados. Uma área emergente que vem apresentando e ampliando novas perspectivas nessa área é a de computação suave (*soft computing*) que engloba raciocínio aproximado e sistemas difusos.

A principal contribuição deste capítulo é mostrar a inserção de ontologias e mineração de dados na articulação da tarefa de descrição, onde a mineração de dados cumpre um papel importante, e na articulação das tarefas de predição, prescrição, explicação e interpretação. Através dessas tarefas, as ontologias dinâmica e dialeticamente são nutridas e dão suporte a essas tarefas.

No **Capítulo 4**, *Ontologias Difusas*, é apresentado o estado da arte no desenvolvimento de ontologias que incorporam componentes e mecanismos de inferência difusos. Além das iniciativas de extensão das ontologias clássicas pela reestruturação ontológica, são apresentadas iniciativas de extensão de lógicas descritivas para incorporar a lógica difusa e dar suporte a mecanismos de inferência difusos. As linguagens lógicas são comentadas considerando-se os critérios de decidibilidade, consistência, completude e expressividade.

Segunda Parte – Estudo de Caso e Pesquisas Futuras.

A segunda parte fica por conta do detalhamento da metodologia, modelagem, desenvolvimento da solução e discussão de resultados realizados no Capítulo 5 e 6. E, por fim, indicações de pesquisas futuras realizadas no Capítulo 7, a partir da abertura de campo propiciada pela pesquisa realizada.

No **Capítulo 5**, Metodologia, *Modelagem, Implementação de Classes e Regras de Classificação a partir de Estudo de Caso,* é desenvolvida uma solução de modelagem com a finalidade de ilustrar aspectos conceituais, metodologias e técnicas discutidas nos capítulos anteriores. São destacadas a construção de classes nas ontologias e o desenvolvimento de regras em SWRL e SQWRL para o mecanismo de inferência de classificação de instâncias das ontologias

A maior contribuição consiste em mostrar como se pode incorporar e tratar com ontologias conhecimentos especialistas de natureza vaga, típico de domínios sócio-humanos, como o de auditoria tributária, de modo a auxiliar tarefas de tomada de decisão.

No **Capítulo 6,** *Testes, Simulações* e Resultados relata-se os testes e simulações realizados, e os principais problemas enfrentados. Confrontam-se os resultados da utilização de uma “abordagem clássica” em mineração de dados, com os resultados da “abordagem difusa”. Apresenta-se os resultados, discutindo-os, esclarecendo seus significados e alcances, e destacando algumas vantagens, e melhorias conseguidas pela abordagem difusa. Conclui-se pela adequação do uso da abordagem difusa em domínios sócio-humanos.

No Capítulo 7, *Conclusões*, apresentamos as conclusões gerais do trabalho e, indicamos possibilidades de pesquisas futuras. As perspectivas futuras vão desde a aplicação da solução em outros casos de uso; o uso de modelagens geométricas baseadas em espaços conceituais; considerações de ontologias no contexto de governo eletrônico; a agregação de técnicas de análise de inteligência na análise das regras, e no desenvolvimento da modelagem ontológica; e o estudo de relações de natureza causal.

2 Vaguidade e a Importância de sua Consideração para a Construção de Ontologias Aplicadas

2.1 Introdução[6]

A vaguidade, ou vagueza é um fenômeno fonte de incerteza, inexatidão, indefinição e indeterminação para a informação. Portanto, no âmbito da Ciência da Informação deve ser compreendida e tratada, considerando-se as dificuldades que ela traz relacionadas à representação, à recuperação, validação e avaliação da informação oriunda de diversos meios tais como textos, sons, ou imagens.

Em ontologias aplicadas a questão vem sendo tratada com ênfase em duas abordagens distintas. De um lado, a abordagem sobreatribuitiva, destacando-se os trabalhos desenvolvidos por Thomas Bittner, Barry Smith, Mauren Donnelly, e Berit Brogaard em ontologias aplicadas através de proposição de partições granulares (BITTNER e SMITH, 2001a 2001b, 2003a, 2003b, 2003c; SMITH e BROGAARD, 2001; BITTNER, DONNELLY e SMITH, 2010, 2004).

De outro lado, há a abordagem difusa já com significativo volume de trabalhos abordando diversos aspectos da representação e recuperação de informação. Em ontologias aplicadas, podemos destacar os trabalhos de Calegari e Ciucci (CALEGARI e CIUCCI, 2008, 2007a, 2007b, 2007c e 2006) e (CALEGARI e SANCHEZ, 2007). No Brasil, Pereira, Ricarte e Gomide (PEREIRA, 2004; PEREIRA, RICARTE e GOMIDE, 2006 e 2005a, 2005b) e, mais recentemente, Yaguinuma, Biajiz, Santos e Escovar (YAGUINUMA *et al.* 2007a e 2007b , e ESCOVAR *et al.* 2007), e Ferraz (FERRAZ *et al.*, 2010). Esses trabalhos serão abordados com maiores detalhes no Capítulo 4 que tratará do desenvolvimento de ontologias difusas.

---

[6] O tema deste capítulo foi conteúdo da seguinte publicação: COELHO, Eduardo de Mattos Pinto Coelho; BAX, Marcello Peixoto e MEIRA JÚNIOR. Wagner. Vaguidade e a importância de sua consideração para a Ciência da Informação e a construção de ontologias aplicadas.Trabalho aprovado para ser apresentado no *XIII Encontro Nacional da Associação Nacional de Pesquisa em Ciência da Informação* (Enancib XIII). Rio de Janeiro, 28 e 31 de outubro de 2012.

Nas seções seguintes elucidaremos noções, peculiaridades e conceitos associados ao fenômeno da vaguidade. Apresentaremos distintas perspectivas lógico-filosóficas sobre a vaguidade e, por fim, as propostas de extensões feitas em ontologias aplicadas para lidar com o fenômeno.

2.2 Definindo, Caracterizando e Distinguindo Vaguidade

A qualidade da informação é prejudicada pelos fenômenos da vaguidade e da ambiguidade. Ambas geram incerteza, imprecisão e são fontes de indeterminação, mas por envolverem conceitos, influenciarem a qualidade da informação em modos distintos, e exigirem tratamentos bem diferenciados devem ser claramente distinguidas.

2.2.1 Ambiguidade

Na Ciência da Informação, o fenômeno da ambiguidade já vem obtendo um tratamento pertinente no âmbito da língua portuguesa. Os professores Marisa Bräsher e Mamede Lima-Marques, da Universidade de Brasília (Unb), vêm desenvolvendo e orientando pesquisas que traçam um amplo panorama teórico e esclarecem as várias nuanças da ambiguidade, discriminando seus vários tipos, suas fontes e procurando aprimorar a recuperação da informação em textos através de processos de desambiguação (BRÄSCHER, 2002; SANTOS, 2006 e SILVA, 2006b).

No fenômeno da ambiguidade, uma palavra, ou frase possui vários significados distintos e podem gerar mais de uma interpretação de significado. (BRÄSCHER, 2002, p. 3). Por exemplo, ao se dizer de “banco” pode-se estar querendo dizer de banco de sangue, banco instituição financeira, banco mobiliário, banco de areia, etc.

Segundo Bräscher (2002, p.4), a ambiguidade causa ruído na recuperação da informação, pois sob um mesmo termo, o usuário encontrará informação relevante e irrelevante. Em vista disso, destaca-se a importância da desambiguação para as ferramentas de busca:

> os trabalhos mais recentes na área baseiam-se na premissa de que ferramentas de busca, ao fazerem uso da linguagem natural, necessitam de conhecimento sobre o significado das expressões que são tratadas e das

> relações que se estabelecem entre elas. Essas ferramentas devem, ainda, ser capazes de tratar determinados fenômenos linguísticos que afetam a qualidade da recuperação, como o da ambiguidade. (BRÄSCHER, 2002, p.2)

Max Black, autor de um dos artigos clássicos que reacenderam o debate sobre a vaguidade (BLACK, 1937), identifica a ambiguidade com generalidade, e define-a como a associação de um número finito de distintos significados tendo a mesma forma fonética. Black foca no signo, a palavra, e, portanto, seu foco é a ambiguidade lexical.

Estudos e classificações mais recentes discriminam vários tipos e origens de ambiguidades. Santos (2006, p. 24) analisa as classificações apresentadas e propostas em Bräscher (2002), e Silva (2006b) elaboradas a partir dos fatores causadores de ambiguidades. Considerando como critério de seleção a estabilidade, isto é, a capacidade de uma classificação de permitir que uma única palavra ou oração ambígua seja classificada em um único tipo de ambiguidade na taxonomia, ele opta pela classificação de Fuchs[7] apresentada por Bräscher (1999)[8].

O discernimento sobre o que é ambiguidade e seus tipos permite-nos perceber que tal informação imperfeita é tratável. Ambiguidade pode ser resolvida apresentando-se uma palavra alternativa, vaguidade não (WILLIAMSON, 1994, p. 73).

Também, a desambiguação pode ser alcançada através da relativização. Isto é, é possível descobrir o sentido conhecendo-se como o léxico, ou a frase associa-se ao cotexto, ou ao contexto. O cotexto é o texto ao redor, o que está escrito antes ou após um enunciado. O contexto é o conjunto de condições de uso da língua, que envolve, simultaneamente, o comportamento linguístico e o social, e é constituído de dados comuns ao emissor e ao receptor (HOUAISS, 2007). Ambos fornecem elementos para a compreensão unívoca do sentido. Então, no caso da ambiguidade, a incerteza, ou indeterminação gerada por ela vaporiza-se mediante mais e melhor informação que restaura o sentido pretendido. No entanto, o mesmo ocorre com a vaguidade? Tal imperfeição é resolvida com mais informação?

---

[7] Fuchs, C. *Les ambiguïtés du français*. Paris: Orphys, 1996. 183 p. Nessa classificação, a ambiguidade é classificada em seis tipos: Ambiguidade Morfológica, Lexical, Sintática, Predicativa, Semântica e Pragmática.

[8] Bräscher, Marisa. *Tratamento automático de ambiguidades na recuperação da informação*. 286 p. Tese, Universidade de Brasília, Brasília, 1999.

2.2.2 Vaguidade

Já a vaguidade, apresenta-se como um fenômeno ainda mais espinhoso. O termo "vaguidade" foi popularizado por Russell, em mais um de seus artigos seminais, *Vagueness* de 1923. Em 1937, Max Black trouxe novas perspectivas sob a questão, envolveu-se em um famoso debate com Hempel (1939) sobre suas posições e, desde então, tal assunto tornou-se emergente em semântica e filosofia, desafiando os pressupostos teóricos da lógica e da semântica clássica.

Com a ambiguidade, a vaguidade compartilha a falta de definição, a consequente falta de certeza, de precisão. Assim como a ambiguidade, a vaguidade é constituída por uma relação um-muitos, entre simbolização e sistemas simbolizados, e isto leva a vaguidade a também ser confundida com generalidade, como confundido em Russel (1923) e Black (1937). Williamson (1994, p. 73) esclarece que vaguidade diz respeito a limites indeterminados, enquanto generalidade diz respeito ao alargamento de sentidos entre limites bem definidos. Entretanto, diferentemente da ambiguidade, a vaguidade não se dissolve com a elucidação do cotexto, ou do contexto: se uma palavra é ambígua, o falante pode resolver a ambiguidade sem partir do uso literal; se a palavra é vaga, o falante não pode resolver este caso limite.

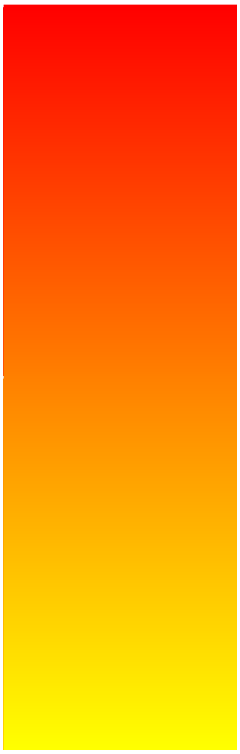

FIGURA 2.1
Degradê do vermelho para o amarelo

Para Pierce, um dos primeiros a retomar a discussão da vaguidade:

> Uma proposição é vaga quando há possíveis estados de coisas que são intrinsecamente incertos se, tendo eles sido contemplados pelo falante, ele os tiver considerado como excluídos ou permitidos pela proposição. Por intrinsecamente incerta, nós significamos que não há incerteza em consequência de qualquer ignorância do intérprete, mas pela indeterminação dos hábitos do falante que são indeterminados (PEIRCE, 1902[9], *apud* KHATIB, 2008, p.5).

Também contribuindo para o esclarecimento da questão, Kit Fine (1975), de modo rigoroso, estipula que vaguidade é deficiência de significado. Como essa deficiência

[9] Peirce, Charles Sander. Vague. *In*: *Dictionary of Philosophy and Psychology*, BALDWIN, J. M. (ed.), New York: Macmillan, 1902, p. 748.

é compartilhada por outros fenômenos, deve ser distinguida, em especial, de generalidade, incapacidade de decisão ("*undecidability*") e ambiguidade que são, respectivamente, falta de conteúdo, falta de conhecimento possível, e falta de significado unívoco, como em RUSSEL (1923).

Em princípio, vaguidade é uma questão epistêmica associada à falta de estabilidade de nossos juízos e de nossas percepções. Apresenta-se como um problema em se estabelecer limites, fronteiras, ou termos. Apresenta-se também como um problema em que este limite pode ser estabelecido de múltiplas formas. Ainda, emerge a partir de uma série paradoxal.

Emergindo do paradoxo, a vaguidade fica intimamente associada a predicados. Desse modo, de uma questão epistêmica, a vaguidade tende a se tornar eminentemente uma questão semântica, assim como a ambiguidade.

Este paradoxo nos é legado pela antiga filosofia grega. Atribui-se a Eubulides de Mileto (4º séc. a.C.) o chamado paradoxo do monte (*sorites*) que surge do seguinte enigma. Um único grão de areia, certamente, não é um monte. Nem a adição de um único grão de areia é bastante para transformar um não monte em um monte: quando temos uma coleção de grãos de areia que não é um monte, então adicionando somente um grão de areia, não será criado um monte. Entretanto, sabemos que em algum momento teremos um monte. Que momento é este?

Este paradoxo pode ser traduzido pelo encadeamento das seguintes proposições:

1. Um grão de areia não faz um monte.
2. Se 1 grão de areia não faz um monte, então dois grãos de areia não o fazem.
3. Se dois grãos de areia não fazem um monte, então três grãos de poeira não o fazem
4. ...
5. Se 9.999 grãos de areia não faz um monte então 10.000 grãos de areia não o fazem
6. 10.000 grãos de areia não fazem um monte.

Assim, empregando-se apenas *modus ponens*, rigorosa regra de inferência endossada tanto pela lógica estoica, quanto pela lógica clássica moderna (HYDE,

2005), chegamos a uma série de premissas verdadeiras que levam a uma conclusão absurda, aparentemente falsa.

Tal paradoxo elucida bem o caráter da vaguidade: a falta de limite dos predicados envolvidos. O predicado da proposição revela sua vaguidade e, por derivação, nomes, adjetivos, advérbios, e, assim por diante, são alcançados por essa vaguidade (HYDE, 2005). Isto é, as palavras vagas somente são vagas indiretamente, em virtude de se ter um sentido do que é vago (SORENSEN, 2006).

> (...) as regras que governam os predicados imprecisos ordinários, simplesmente não permitem refinar e precisar linhas dividindo objetos para os quais os predicados se aplicam a partir de objetos de qualquer outro tipo (...). (SOAMES, 2005, *apud* RORTY, 2005)

Outro exemplo bem didático para elucidar a vaguidade é o clássico exemplo da suave e gradual transição de uma cor para a outra em um espectro, como vemos em imagem elaborada por John Slaney[10]. Nesta imagem, FIGURA 2.1, vai-se do claramente vermelho (RGB 255,0,0) para o claramente amarelo (RGB 255, 255, 0). Em que ponto, em que momento nesta faixa deixa-se de ser vermelho e passa-se a ser amarelo? Ou, ainda, partindo do amarelo, deixa-se de ser amarelo e passa-se a ser vermelho? Esses dois limites tendem a ser diferentes, considerando-se diferentes direções de análise? A histerese, a persistência dos efeitos da observação precedente, afeta a determinação (vide ÉGRÉ, 2009)? Em diferentes tentativas, ou, ainda, diferentes pessoas chegam ao mesmo resultado?

2.2.3 Enfrentado o Paradoxo

Roy Sorensen (2003, p.*xi*), tem uma consideração interessante sobre paradoxos. Para este autor, assim como matemáticos consideram os números primos como sendo átomos da matemática, os paradoxos são como átomos da filosofia. Os paradoxos constituem os pontos básicos de partida para a especulação disciplinada. Assim, chegando ao paradoxo, chegamos àquilo que começa.

Penso que o paradoxo leva-nos ao limite da cognição humana. Na medida em que se furta ao senso comum, maravilha-nos, dá-nos vertigens, subverte o pensamento e exige de nós a invenção de novos métodos e formas de pensamento. Podemos

---

[10] Fonte:<http://users.cecs.anu.edu.au/~jks/sorites.html>, acesso em 01-02-2010.

querer ignorá-lo, fingir que não nos diz nada de relevante, ou transmutá-lo de modo a se tornar algo mais palatável, mas inadvertidamente o paradoxo nos aguarda e manifesta-se em sua plena integridade.

Também segundo Sorensen, a filosofia se mantém coesa mais por suas questões do que por suas respostas. As questões filosóficas básicas surgem a partir de problemas dentro de nosso esquema conceitual ordinário, e esses paradoxos aproximam gerações com problemas comuns e geram um repositório acumulativo de respostas. De fato, o paradoxo *sorites* foi descoberto no século IV a.C., na escola megárica, foi amplamente levado a sério e debatido pelos estoicos e, desde então, ficou na surdina durante muitos séculos para emergir novamente no final do século XIX quando a lógica formal uma vez mais assumiu um papel central na filosofia (HYDE, 2005).

Na tentativa de se lidar com o paradoxo, considera-se que há quatro atitudes possíveis (RESTALL, 2004[11]; exposições similares são feitas em SLANEY, 1988; e HYDE, 2005)[12]:

1) Negar que o problema seja legitimamente construído. Isto é, defender que a lógica não se aplica a expressões vagas.
2) Aceitar que a lógica aplica-se legitimamente aqui, mas defender que este argumento em particular é inválido.
3) Aceitar ambas, isto é, a lógica aplica-se em tais casos e o argumento é válido, mas negar a validade de uma das premissas.
4) Aceitar o argumento e as premissas e, assim, aceitar também a conclusão.

---

[11] Na obra citada, Restall atribui a Slaney a classificação adotada por ele.

[12] Mora, 2004, p. 2964, expõe uma alternativa a essas atitudes apresentadas por I.M. Copi (Copilowish). Dentre as três atitudes apresentadas, Copi opta pela que busca manter a integridade das leis da lógica. Curioso é que, para tal, ele reduz a vaguidade a um caso especial de ambiguidade. Essa redução, parece-me, levaria a uma deformação do sentido do que seja vaguidade. Vide: Copilowish, I. M. Border-Line Cases: Vagueness and Ambiguity, *Philosophy of Science*, Vol. 6, 1939, p. 181-195. Acrescente-se que Copilowish não está só. Também D. Lewis entende a vaguidade como um fenômeno semântico em forma especial de ambiguidade, a hiperambiguidade. A ideia de Lewis que é afirmativas ambíguas são verdadeiras quando elas tornam-se verdadeiras sob todas desambiguações. Como veremos, isto é uma forma de supervaloracionismo.

A atitude (1) é a de Frege e Russell. Para ambos, a rigorosa delimitação de cada conceito era uma das premissas para a possibilidade de aplicar as regras da lógica.

> "Uma definição de um conceito (de um predicado possível) deve... sem ambiguidade determinar, em relação a qualquer objeto, se ele cai, ou não sob o conceito (se, ou não o predicado é verdadeiramente assertivo a partir dele). Então não deve haver qualquer objeto a ser considerado em que a definição deixa em dúvida se ela cai sob o conceito.... Nós devemos expressar isto metaforicamente como segue: o conceito deve ter uma fronteira aguda."(FREGE, 1903, § 56, *apud* VARZI, 2001a, p. 3)

Então, começando com Frege e repercutindo ao longo do século XX, vários lógicos tomam a vaguidade como uma porta aberta para o desastre lógico, e, como tal, defendem que ela deva ser eliminada da lógica. Por exemplo, a despeito de Russell, 1923, admitir que toda linguagem natural seja vaga e que a vaguidade se infiltra nos termos (é penetrante, "pervasive")[13] atingindo praticamente tudo[14], ele defende que uma linguagem formal, e a lógica enquanto disciplina normativa, deverá ser expurgada de toda vaguidade.

Atualmente, de um modo geral, tal posição é considerada inaplicável por restringir extremamente o campo da lógica. Se a vaguidade é incompatível com a lógica clássica, a lógica clássica é inaplicável aos cada vez mais relevantes linguagem e pensamento ordinários.

A atitude (4) não é levada a sério por ser difícil sustentar que o paradoxo é verdadeiro e o senso comum, que admite que em algum momento a adição de um grão constituirá um monte, não deva ser considerado.

Pelo contrário, o senso comum, muitas vezes referenciado como conhecimento pré-filosófico, tendo sua defesa já em Thomas Reid[15] no século XVIII (SORENSEN,

---

[13] Michel Dummett chega a dizer que a vaguidade é penetrante tal qual poeira, (DUMMET, 1995, p. 207, *apud* VARZI, 2001a, p. 3)

[14] Russel argumenta que até os nomes próprios são vagos, pois, quando dizemos, por exemplo, Paulo não se sabe a que estado de Paulo se está referindo. Por exemplo, se ao Paulo de há 10 segundos, se ao Paulo vestindo uma camisa azul, se ao Paulo professor, e por ai vai...

[15] Ver interessante discussão sobre senso comum em SORENSEN, 2003, p. 268-283, em que é debatida a posição de Thomas Reid em "Inquiry into the Human Mind: On the Principles of Common Sense" de 1764. Mais próxima à Ciência da Informação, ver Barry Smith em Formal Ontology,Commonsense and Cognitive Science. *International Journal of Human Computer Studies*, 43 (5/6): 626-640.

2003, p. 268-283) vem sendo considerado com crescente relevância em áreas como ciência cognitiva e filosofia da mente (associada à área da "folk psychology").

Mesmo na moderna filosofia, George Moore[16], um dos fundadores da filosofia analítica, defende a importância do senso comum para os fundamentos do conhecimento e, contemporaneamente, Saul Kripke, o grande expoente da filosofia analítica a partir dos anos 70 reforça indiretamente este entendimento (RORTY, 2005).

Na Ciência da Computação e Ciência da Informação, mais especificamente em inteligência artificial e ontologias, vem sendo desenvolvidos trabalhos com base no conhecimento de senso comum. Podemos destacar o famoso projeto *Cyc*[17] da *Cyc Corporation* e os projetos do MIT (Massachusetts Institute of Technology): *Using Common Sense Reasoning to Enable Semantic Web*[18], *Open Mind Common Sense* (OMCS)[19], *Concept Net*[20]; o projeto *Epilog* da *University of Rochester* e *University of Alberta*. O pioneiro da Inteligência Artificial, Marvin Minsky (MINSKY, 2006), destaca-se como um pesquisador relevante da área, e também se pode considerar que grande parte das iniciativas em lógica e sistemas difusos, por sua especial preocupação em tratar a linguagem natural, também lida direta, ou indiretamente com a questão do senso comum.

Voltando à questão do paradoxo, a atitude (3) é a opção mais aceita nos círculos filosóficos, conforme nos esclarece Greg Restall (2004).

> (3) é, talvez, a opção ortodoxa nos círculos filosóficos. Ela tem a vantagem de não ter que modificar suas teorias lógicas, mas tem a desvantagem de nos exigir o reconhecimento preciso da premissa falsa em um grupo (de premissas) aparentemente plausível. Atualmente, existem duas maneiras principais de desenvolver esta opção. Uma dessas maneiras é chamada de método de sobreatribuições[21] (devido a originalmente a Bas van Fraassen (...). De acordo com essa abordagem, nosso conceito de "vermelho" não

---

[16] Moore, George (1925/1959). Uma Defesa do Senso Comum, in Escritos Filosóficos, p.243 ; trad. Paulo R. Mariconda. - São Paulo: Nova Cultural, 1989. Vide também: Coates, John. *The claims of common sense: Moore, Wittgenstein, Keynes and the social sciences*. Cambridge University Press, 1996, 176 p.

[17] **Cyc** é um projeto de inteligência artificial que tenta montar uma ontologia e uma base de conhecimento do senso comum usado no dia a dia. Vide: http://cyc.com/cyc/technology/whatiscyc

[18] http://agents.media.mit.edu/projects/semanticweb_old/

[19] http://csc.media.mit.edu/

[20] Rede semântica de fatos e relações de senso comum. http://web.media.mit.edu/~hugo/conceptnet/.

[21] Nota minha: "*supervaluations", supervalorações.* Outros, traduzem como *sobreatribuições*, como na tradução feita em Branquinho *et al*., 2006. Outros traduzem com sobreatribuições, etc.

> fornece detalhes claros da localização da fronteira entre vermelho e amarelo. Existe uma classe de fronteiras aceitáveis, cada uma tão boa quanto outra. Testamos nosso argumento muitas vezes, e cada vez demarcamos a fronteira em lugares diferentes. Se o argumento é válido em todas estas demarcações, ele funciona, e se ele é inválido em alguma demarcação, ele falha. Além disso, se um enunciado é verdadeiro em todas as demarcações, ele é verdadeiro[22]; se é falso em todas as demarcações, ele é falso[23]; e se ele for verdadeiro em algumas e falso em outras, trata-se de uma "lacuna de valor-verdade". Isso porque nosso conceito original "vermelho" é vago. Ele não é capaz de decidir sobre se esse enunciado é verdadeiro ou não (RESTALL, 2004).

É a opção mais aceita, pois, mantém-se a integridade dos princípios da lógica clássica, conforme veremos na seção seguinte, essa abordagem também tem restrições no que diz respeito à própria lógica e limitações no alcance de soluções práticas.

A atitude (2) é plausível (RESTALL, 2004), os argumentos *sorites* parecem totalmente suspeitos e é atraente dizer que são inválidos. Entretanto, é acompanhada por renitentes críticas nesses mesmos círculos filosóficos. Teme-se jogar fora o bebê lógico junto com a água do banho *sorites* (RESTALL, 2004). Isto é, flexibilizando-se os princípios da lógica clássica, em especial, os princípios da Bivalência, do Terceiro Excluído e da Não Contradição, aparentemente estaremos alcançando soluções, mas ao custo de afastarmo-nos de concepções simples, bem estabelecidas e amplamente aceitas.

Na seção seguinte, iremos apresentar os argumentos básicos das distintas correntes filosóficas que sustentam as atitudes (2) e (3), em especial as correntes que se baseiam na lógica sobreatribuitiva e em lógicas multivaloradas, em especial, a difusa. Na seção subsequente, discutimos as soluções adotas em ontologias aplicadas, desenvolvidas a partir das abordagens de sobreatribuição e de lógicas difusas.

2.3 Correntes Lógico-Filosóficas da Vaguidade

Nos últimos trinta e cinco anos, a questão da vaguidade veio à baila gerando uma profusão de artigos e grandes debates, tão polêmicos quanto irresolutos, em vista

---

[22] Diz-se que o enunciado é superverdadeiro.
[23] Diz-se que o enunciado é superfalso.

dos relevantes e impactantes problemas filosóficos que ela levanta e repercutem numa variedade de campos.

A despeito de ser uma questão eminentemente filosófica ela nos alcança em questões que afetam diretamente nosso dia a dia. Para vinculá-la de imediato a questões práticas, consideremos as questões do Direito onde a vaguidade se instaura tais como as envolvendo a vida, quando ela surge e quando ela se extingue, a responsabilidade penal em vista da idade, ou a identidade de raça para usufruir de uma cota universitária. A vaguidade e suas dificuldades inerentes certamente surgirão quando pretendermos fazer interpretações, ou afirmativas discretas sobre fenômenos contínuos (cor, comprimento, vida, idade, velocidade evolução das espécies, etc.).

As áreas de inteligência artificial, ontologias aplicadas e linguagens formais computacionais lidam com o que as máquinas precisam para poder reconhecer. A máquina submetida a estímulos, os dados de entrada, parte de um modelo de representação do mundo e respectivos mecanismos de inferência para adotar determinados procedimentos (manipulação dos dados, e/ou entidades, classificação, tomadas de decisão, etc.). Em vista disso, esses debates filosóficos também alcançam em cheio a área da semântica, da lógica, das ontologias e as linguagens formais de interesse direto em Ciência da Informação e Ciência da Computação.

> Em filosofia, discute-se amplamente se a vaguidade surge:
>
> (1) das próprias coisas;
> (2) de nosso modo de nos expressarmos sobre as coisas, isto é, da linguagem;
> (3) da limitação de nossos sentidos, ou da limitação de nossa capacidade de entendimento: isto é, por mais que tenhamos dados, não seremos capazes de compreendê-los, distingui-los, discerni-los.

Em (1) temos uma **vaguidade ontológica**, ou, segundo alguns, ôntica. As coisas, os objetos, em suma, a realidade é, por si só, vaga. De imediato, isso nos remete a sérios problemas quanto à identidade e individuação das coisas em si. Os recentes debates a respeito dessa questão foram deflagrados por um artigo de uma página de Garreth Evans, *Can There Be Vague Objects*? (EVANS, 1978). Não iremos entrar

nesse debate, mas um bom panorama da vaguidade ôntica é apresentado em Chibeni (2003). Uma da várias alternativas a Evans em defesa dos objetos vagos pode ser vista em Unger (1979), Burgess (1989), Tye (1990) e Prinz (1998). Reelaborando o argumento de Evans, vide Lewis (1988).

Em (2) temos uma **vaguidade semântica**. A linguagem ordinária é vaga, nosso aparato linguístico ordinário não é suficientemente adequado para tornarmos as coisas exatamente distintas. Isso nos remete a sérias dificuldades no falarmos sobre, ou referirmo-nos às coisas, e, consequentemente, ao falar sobre a identidade e individuação das coisas.

Em (3) temos uma **vaguidade epistêmica**.Com ela somos levados ao niilismo, ou ao ceticismo que leva-nos a inúmeros outros problemas, e não leva-nos a novas soluções, propriamente ditas, em ciências aplicadas, mas confere-nos excepcional capacidade de criticarmos as abordagens oriundas de (1) e de (2).

O velho entrelaçamento entre realidade, linguagem e pensamento tendo no foco o homem, levanta questões filosóficas que persistem sem resposta: por exemplo, o pensamento faz contato direto com a realidade, ou sempre é mediado pela linguagem? Entretanto, em nossa área, as questões do nível do pensamento (*de cogitate*) situam-se no que este nível interfere no fazermos a ciência a que nos propomos, mas, pelo menos no que diz respeito à capacidade que precisamos dotar a máquina para reconhecer, podemos nos deter ao relacionamento entre coisas e linguagem. Mais precisamente, podemos nos deter no relacionamento entre e entidades e linguagens formais. Portanto, a discussão alcançar-nos-á no que diz respeito a se a vaguidade se dá no nível das coisas (*de re*), ou no nível semântico[24], da referência (*de dicto*), qual a lógica mais adequada para tratá-la e quais as condições de verdade corretas para uma linguagem vaga (FINE, 1975).

Na abordagem da vaguidade, há textos filosóficos de relevância histórica que esclarecem o significado e o sentido desse fenômeno, inclusive no nível em que ela se manifesta. Alguns desses esclarecimentos já foram feitos na seção anterior.

[24] Fontes da vaguidade semântica: predicados, nomes, quantificadores e, mesmo, operadores.

Russel (1923), Black (1937), Hempel (1939), e Dummett (1975) são referências sempre citadas.

Já no que diz respeito à abordagem da lógica e de métodos mais adequados para tratá-la, teremos autores identificados a uma série de correntes de pensamento como o indeterminismo (RUSSELL, 1923; BURGESS, 1998), epistemicismo (WILLIAMSON, 1994; SORENSEN, 1988), sobreatribuicionismo[25] (MEHLBERG, 1958; FINE, 1975; VARZI, 2001a, 2001b, 2002 e 2007) e polivalência (EDGINGTON, 1997 e GOTTWALD, 2000)[26], em especial a abordagem baseada na lógica difusa (ZADEH, 1965 e 1975, MACHINA, 1976 e HÁJEK, 2009).Há ainda as ditas correntes menores tais como pragmatismo/contextualismo (VAN KERKHOVE, 2001 e 2002 e RAFFMAN, 1996) intuicionismo (WRIGHT, 2003), niilismo, paraconsistência, incoerentismo, relevantismo, subvaloracionismo, transvaloracionismo, etc.

Uma excelente referência para uma ampla perspectiva histórica e crítica das principais correntes dentre as citadas é o livro de Timothy Williamson, *Vagueness* de 1994. Williamson, adepto do epistemicismo, faz uma exaustiva análise crítica dessas principais correntes desmantelando seus principais argumentos. Nesse percurso de desmantelamento geral, abre-se o campo para a ocupação do niilismo. Entretanto, Williamson também o rebate, e leva-nos a uma formulação epistemicista, que é uma variante limitada de ceticismo, em que nossa ignorância não é absoluta, é relativa e pode ser saneada. Também Sorensen, 1988, outro adepto do epistemicismo, e Hyde (2005) traçam um bom panorama da vaguidade e do paradoxo *sorites*.

Estas correntes de interpelação e interpretação da vaguidade, por sua vez, podem ser classificadas em dois grandes ramos: as que **consideram que a vaguidade não pode, em rigor, ser eliminada**; e as que **consideram que a vaguidade pode, rigorosamente, ser eliminada**.

Em ciências aplicadas é possível identificar uma atitude distinta. Em certo sentido as preocupações, ou pretensões em se eliminar a vaguidade são flexibilizadas. Na verdade, não importará se a vaguidade é, ou não eliminável, o que importa é que ela, como uma das fontes de imprecisão e indeterminação, seja explicitada,

---

[25] "Supervaluotionism". Santos, 2006a, traduz como sobreatribuição.
[26] Também referenciada como lógica multivalorada, ou de vários graus de verdade.

modelada, gerenciável, contida por meio da lógica e das linguagens formais, mesmo que sofra um claro processo reducionista em atenção e em coerência à demanda de soluções de determinadas questões práticas.

Com efeito, em seu conhecido texto *Two Kinds of Definitions*, Popper (1945), relativiza a importância das “definições” (no sentido aristotélico) para as ciências empíricas (exclui-se a matemática) e argumenta que é muito plausível que a precisão da linguagem dependa da precisão de seus termos, embora também possa ser um mero preconceito. Popper afirma que a exatidão de uma língua depende, antes, do fato de ela tomar o cuidado de não sobrecarregar seus termos com a tarefa de serem exatos. Os termos “Duna” ou "vento" são muito vagos, mas para muitos dos objetivos do geólogo, eles têm um grau suficiente de exatidão; e, para outros fins, quando é necessário um maior grau de diferenciação, ele sempre pode dizer "vento de uma velocidade entre 20 e 40 quilômetros por hora" ou “dunas entre 4 e 30 metros de altura". Ou seja, a exatidão não consiste em tentar reduzir essa margem a zero nem em fingir que ela não existe, mas sim em reconhecê-la explicitamente. Nesse sentido Ramsey, abaixo, ilustra bem a questão.

> "O principal perigo para nossa filosofia, à parte a preguiça e a obscuridade, é o *escolasticismo* [...], que equivale a tratar o que é vago como se fosse exato." (F.P. RAMSEY *apud* POPPER, 1945).

Vale notar que ao relativizar a importância das definições, Popper (1945) se referia a uma ciência totalmente feita por humanos. Entretanto, ao lançar mão de ontologias informacionais, a ciência contemporânea objetiva obter o auxílio de máquinas na verificação da coerência de nossos modelos. Encontramo-nos assim em um outro cenário, onde a explicitação de definições, mesmo que flexibilizadas para termos vagos, é crucial.

Essa atitude que consiste em explicitar, modelar e gerenciar a vaguidade será explicitada quando formos expor na seção seguinte as extensões de ontologias aplicadas baseadas em lógica sobreatribuitiva e lógica difusa.

A típica corrente do primeiro ramo é o indeterminismo que considera não poder eliminar a vaguidade. Correntes do segundo ramo (que a eliminam) são o

epistemicismo, o sobreatribuicionismo e a lógica difusa rigorosa[27]. Comentaremos um pouco dessas quatro abordagens introduzidas, a seguir.

O indeterminismo é abordagem originalmente atribuída a Russell e, possivelmente, é a mais aceita. A vaguidade advém da indeterminação nas condições reais de conhecimento, e considera a vaguidade como sendo, principalmente, um problema de linguagem. Desse modo, a principal questão é determinar o limite entre propriedades opostas, semanticamente falando, entre sua extensão e sua contra extensão, como no caso do vermelho e do amarelo, quente e frio, jovem e velho, nascido e não nascido, sóbrio e embriagado, alto e baixo, lento e veloz, etc. Entretanto, essa dificuldade é admitida por várias correntes, o aspecto que a distingue é que se argumenta que mesmo com uma linguagem logicamente estruturada e muito rigorosa, ainda teremos vaguidade. Em suma, não há método para eliminar a vaguidade, poderemos fingir que a ignoramos, mas ela continuará lá. O que se pode fazer é procurar expurgá-la das linguagens formais, conforme Russell e Frege, mas como já comentamos, a despeito de se considerar o indeterminismo do fenômeno, a atitude de expurgá-lo da lógica, hoje, não recebe amparo.

Já o epistemicismo apregoa que a indeterminação associada às frases com predicados vagos resulta não de qualquer indeterminação no mundo refletido pelo nosso conhecimento, ou sobre o mundo e a linguagem que usamos para falar dele (SANTOS, 2006 p. 787). Os predicados imprecisos são de fato perfeitamente precisos, no sentido de que há finas e precisas linhas dividindo, de um lado, objetos aos quais eles verdadeiramente se aplicam, e de outro, objetos aos quais eles verdadeiramente não se aplicam. As fronteiras estão lá e as discerniríamos se pudéssemos conhecer melhor as coisas, ou o domínio de aplicação dos predicados referentes às coisas. A premissa maior do *sorites* é plenamente falsa. Há uma fronteira real entre o vermelho e o amarelo, entre a água e o óleo, entre a vida e a não vida. Em vista disso, a lógica clássica é uma boa ferramenta para lidar com ela, não havendo qualquer questionamento ao princípio da bivalência[28].

---

[27] Lógica difusa em sentido amplo, vide Seção 3.5.

[28] A vaguidade é um apenas um dos problemas que atingem o Princípio da Bivalência. Há ainda outras situações onde atribuir um valor de verdade, ou falsidade são problemáticas (RESTALL, 2004, p. 84-85): paradoxos, como o Paradoxo do Mentiroso; a ocorrência de termos não denotativos (o

Na abordagem epistemicista a vaguidade é puramente semântica, diferente da niilista que coloca em xeque a existência das próprias coisas, mas, assim como a niilista, é contra intuitiva e foge ao senso comum (para uma distinção entre as duas correntes além de WILLIAMSON, 1994, vide ENOCH, 2007). Há de se desenvolver métodos que permitam precisar esses limites, essas fronteiras. Entretanto, não há métodos gerais e, num viés cético, pensamos que muitos jamais estarão ao nosso alcance. Concluindo, o epistemicismo, assim como o niilismo permitem excepcionais elaborações argumentativas, **mas não encontram terreno fértil no campo das ciências sociais aplicadas**.

A corrente do sobreatribuicionismo defende a atribuição do valor verdade (superverdade) para afirmativas verdadeiras em todas as avaliações possíveis, admissíveis e completas; ou falsidade (superfalsidade) para afirmativas falsas em todas as avaliações possíveis, admissíveis e completas, e de nenhum dos dois valores para o resto. Uma precisão admissível é um modelo clássico para a linguagem, onde se atribuem valores semânticos aos termos de maneira tal que a classe completa das precisificações constitua uma elaboração apropriada e suficiente de seu significado.

O resto alcança justamente o caso limite, o caso de fronteira ao qual estará vinculada uma zona cinzenta, indeterminada, de penumbra, ou ponto cego, sobre o qual o falso e o verdadeiro não poderão ser exatamente atribuídos.

As origens do sobreatribuicionismo remontam ao trabalho de Mehlberg, 1958, que tratou de uma lógica de tal natureza, embora não usasse o termo. Depois, remontam a Bas van Frassen (1968 e 1969[29]) que utilizou essas noções para buscar um tratamento semântico de nomes que não tinham referência e de sentenças auto referenciais tais como as envolvidas no Paradoxo do Mentiroso, e introduziu o termo "supervaluotionism" (WILLIAMSON, 1994, p. 146). O sobreatribuicionismo foi usado para abordar a vaguidade em Kit Fine, 1975, e os trabalhos de Smith e Brogaard

---

objeto ao qual se refere não existe); falha de pressuposição (pressupomos que nossos termos denotam objetos; quando essa pressuposição falha, temos problemas em interpretar afirmações que envolvem tais termos) e futuros contingentes ("haverá uma batalha naval amanhã").

[29] van Frassen, B. C. "Presuppositions, Implications and Self-Reference", *Journal of Philosophy*, 65, 1968.
van Frassen, B.C. "Presuppositions, Supervaluations and Free Logic", in *The Logical Way of Doing Things*. LAMBERT, K. (ed.), Yale University Press, 1969.

(2000) e Achille Varzi (2001a) são relevantes por aproximarem este tratamento da área de Ciência da Informação. Os que defendem, ou simpatizam com essa abordagem consideram que ela reflete uma intuição profunda, pré-analítica referente à vaguidade conforme ela surge na linguagem ordinária (VARZI, 2001a, p. 11).

O sobreatribuicionismo propõe que verdade é superverdade e falsidade é superfalsidade. Por exemplo, podemos querer definir o que é o Saara, ou o Everest. Haverá uma região em que poderemos alcançar uma unanimidade que diz que é o Everest. Também, uma outra região em que poderemos alcançar uma unanimidade que diz que não é o Everest. Desse modo, necessitamos de um método que colete muitas descrições semânticas da linguagem que se pretende fazer precisa. Para isso, consideram-se todas as precisificações destes predicados como sendo relevantes. Toda precisificação destes predicados outorga um valor de verdade definido (verdadeiro ou falso) a cada aplicação possível do termo. Se em toda precisificação relevante recebe o mesmo valor de verdade, tem este valor de verdade de modo claro. Se não ocorre isto, não o tem.

A região onde houver contestações quanto a ser, ou não ser, constitui a zona de penumbra. As zonas de penumbra, isto é, as lacunas podem ser acomodadas de acordo com as necessidades, e as precisificações admissíveis são configuradas em uma linguagem convenientemente formalizada: precisificar um termo é tornar um termo preciso, e uma linguagem vaga pode se tornar precisa de mais de um modo. Considerando-se a zona de penumbra, o sobreatribuicionismo transgride a bivalência, mas, dentro de cada uma das precisificações, uma propriedade é bivalente.

Kit Fine (1975) propõe quatro condições (Condição de Fidelidade-F; Condição de Estabilidade-S; Condição de Completabilidade-C e Condição de Resolução-R) que levadas em consideração, e devidamente trabalhadas em função do domínio de aplicação e de considerações de otimização, reduzem a zona de penumbra de modo a torná-la irrelevante para o domínio de aplicação visado. Essas zonas penumbrais, distintas de deficiências, podem ser “fechadas”. Obtendo-se este fechamento, a lógica sobreatribuitiva recuperaria, para o domínio em questão, a consistência, a simplicidade e funcionalidade da lógica clássica.

A quarta corrente é a das lógicas polivalentes, ou multivaloradas. Também aqui é feito a transgressão da bivalência. Normalmente, a origem da lógica polivalente é atribuída a Jan Lukaziewicz que formalizou e debateu uma proposta de lógica trivalente para poder expressar o futuro contingente (por exemplo, “a batalha naval de amanhã” de Aristóteles), mas há quem identifique precedentes na Idade Média, e mesmo em Aristóteles, em vista dessa famosa proposição futuro-contingente (MORA, 2004c, p. 2313). Nessa lógica, uma proposição deixa de poder ter tão somente um valor verdadeiro, ou um valor falso. Podemos ter três (no caso de Lukaziewicz, 0, 1 e ½ para o valor indeterminado), quatro, cinco, ou infinitos valores verdade para uma dada proposição.

A lógica polivalente mais aceita é a que admite a assunção de infinitos valores verdade entre 0 e 1, onde a transição da verdade cheia (1) para a falsidade cheia (0) pode ser feita de modo gradual, não abrupto. Diferente da lógica sobreatribuitiva, aqui não há lacunas.

Dentre as lógicas que adotam infinitos valores para uma proposição, destaca-se a lógica difusa criada por Lotfi Zadeh a partir de artigo de 1965 que trata de conjuntos difusos (*Fuzzy Sets*). Petr Hájek atribui a Goguen (1969) o primeiro uso do termo lógica difusa (*fuzzy logic*) no artigo *The Logic Of Inexact Concepts*.

A abordagem difusa busca graduar uma medida do fenômeno em vista da dificuldade de se estimar experimentalmente o grau de verdade. Consideram-na intuitiva por começar a análise com rótulos de linguagem natural (variáveis que instanciam palavras e não números tais como alto, baixo, calvo, cabeludo), e poder atribuir valores aproximativos a esses rótulos.

Princípios lógicos, validade e verofuncionalidade

As duas primeiras correntes citadas até o momento, a indeterminista e a epistemicista, preservam o Princípio da Bivalência (há exatamente e tão somente dois valores de verdade).

Além do Princípio da Bivalência, são importantes também o Princípio da Não-contradição (um proposição não pode “ser” e “não ser” ao mesmo tempo), e o Princípio do Meio Excluído (ou do Terceiro Excluído, uma proposição só pode “ser”

ou “não ser”). O Princípio da Não-contradição estabelece que uma **conjunção** entre uma proposição e sua negativa é sempre falsa (*Não(P* e *Não-P)* é *V,* é uma tautologia). O Princípio do Meio Excluído estabelece que uma **disjunção** entre uma proposição e sua negativa é sempre verdadeira (*P* ou *Não-P* é *V,* também é uma tautologia).

Há de se observar que na lógica clássica o Princípio da Bivalência, normalmente, é confundido com estes dois outros princípios, pois, são logicamente equivalentes. Entretanto, há lógicas não clássicas em que não há essa equivalência lógica, como acontece em formulações das duas últimas correntes consideradas, a sobreatribuitiva e a polivalente.

Outros aspectos a serem considerados é o da verofuncionalidade e da validade lógica. Um operador ou conectivo lógico é verofuncional se o valor de verdade de uma proposição com dado operador depende inteiramente do valor de verdade dessa frase sem o operador. Por exemplo, “não chove” é totalmente determinado pelo valor verdade da proposição “chove”. Já, por exemplo, uma proposição de crença “acho que não chove” não tem seu valor verdade determinada pela proposição “acho que chove” (MURCHO, 2006). Uma proposição composta com conectivos é verofuncional, se o valor verdade da proposição composta é determinado pelo valor de verdade dos elementos, isto é, das proposições que as constitui. A lógica clássica é verofuncional, mas não o são, por exemplo, a lógica modal que trabalha com os valores verdade “Necessário”, “Possível” e “Contingente” e lógica deôntica “Obrigatório”, “Proibido” e “Indiferente”

Já a validade lógica diz respeito a um argumento, ou uma proposição ser aceita como verdadeira. As premissas e a conclusão não são elas próprias, válidas ou inválidas, mas verdadeiras ou falsas, então, a validade lógica lida com a questão de se o valor verdade da conclusão é decorrência das premissas. Deverá haver uma preservação dos valores verdade designados nas premissas para a conclusão, ou seja, as premissas derivam a conclusão. É inválido, por exemplo, um argumento, ou proposição que deriva uma conclusão falsa a partir de premissas verdadeiras.

Uma importante distinção diz respeito também ao alcance da validade. Há dois tipos de validade: validade local e validade global.

Na lógica sobreatribuitiva, um argumento é válido localmente se e somente se as premissas são verdadeiras para uma precisificação, tendo a conclusão verdadeira naquela precisificação. Um argumento é globalmente válido se e somente se as premissas são verdadeiras para todas as precisificações, tendo a conclusão verdadeira para todas as precisificações. Validade global é preservação de superverdade, enquanto que validade local é preservação de verdade (WILLIAMSON, 1994, p. 147-148). Como a lógica sobreatribuitiva tenta identificar superverdade com verdade, a questão do alcance da validade é um critério crucial para o entendimento do alcance da abordagem sobreatribuitiva.

Considerando as lógica sobreatribuitiva e a lógica difusa, obtemos os seguintes resultados para o operador de negação e para a avaliação dos princípios do Meio Excluído e da Não Contradição.

TABELA 2.1 Comparação entre lógica sobreatribuitiva e lógica difusa.

	Regra em Lógica sobreatribuitiva	**Regra em Lógica difusa**
Proposição	*P*	*P*
Negação	1. *Não-P* = 1 – *P* 2. *Não-*P_n = 1 – P_n = indefinido, para *n* pertencente à zona de indeterminação de P	*Não-P* = 1 – *P*
Terceiro Excluído	1. *P* ou *Não-P* = 1 2. P_n ou *Não-*P_n= indefinido, para *n* pertencente à zona de indeterminação de P	*P* ou *Q* = Máx(*P*,*Q*)
Não contradição	1. *P e Não-P* = 0 2. P_n e *Não-*P_n= indefinido, para *n* pertencente à zona de indeterminação de P	*P* e *Q* = Mín(*P*,*Q*)

Observemos na TABELA 2.1 que a lógica sobreatribuitiva não é verofuncional com respeito aos três valores verdade na negação, na conjunção e na disjunção. Ela permite instâncias de verdades conjuntivas/disjuntivas, em que nenhuma destas conjunções/disjunções é verdadeira em vista das lacunas de atribuição de valor verdade. Em vista disso, ela não valida os princípios do Meio Excluído e da Não

Contradição. A condicional (Se *P*, então *Q*) exibe características não clássica análogas. Como consequência, lógica e semântica sobreatribuitiva não são verofuncionais.

De outro lado, a lógica difusa, parcialmente, nega o Princípio do Meio Excluído e da Não Contradição. Parcialmente, pois, uma proposição difusa não é definitivamente verdadeira e nem definitivamente falsa, mas há a preservação do valor verdade na disjunção e na conjunção, em função dos componentes da proposição. De fato, na disjunção difusa a validade é definida como preservação dos valores atribuídos, e somente o máximo valor é designado. Na conjunção difusa a validade também é definida como a preservação dos valores atribuídos, e, nesse caso, somente o mínimo valor é designado. Como consequência, lógica e semântica difusa são verofuncionais.

As regras e exemplos desta tabela, assim como os esclarecimentos acima, serão importantes referências para compreendermos as avaliações e comparações entre essas abordagens a serem tratadas a seguir.

Avaliações sobre aspectos semânticos e lógicos de lógica sobreatribuitiva e lógica difusa

Significados vagos são concebidos a partir de especificações incompletas da referência. Para obtermos uma linguagem precisa temos que completar estas especificações sem contradizer qualquer coisa de seu conteúdo original. Entretanto, toda abordagem corre o risco de, no esforço de eliminar a vaguidade, alterar o significado. Assim, as abordagens, ao invés de efetivamente resolverem o fenômeno, correm o risco de falsificar o que elas supõem resolver. Na análise crítica dessas duas últimas abordagens que transgridem a lógica e a semântica clássica, vamos destacar alguns aspectos de natureza semântica e outros de natureza lógica que revelam limitações e vantagens de uma sobre a outra.

Nenhuma das duas abordagens está isenta de fortes críticas e não há unanimidades, mas de um modo geral, a abordagem sobreatribuitiva está mais ao gosto de lógicos e linguistas de distintas orientações. Essa receptividade se explica em vista de um aparente maior potencial da lógica de sobreatribuições em recuperar a consistência e simplicidade da lógica e semântica clássicas.

Na análise que Williamson (1994, p. 142) faz das abordagens que tentam eliminar a vaguidade, ele constata que elas padecem potencialmente de pelo menos umas das limitações:

> 1. "Uma linguagem vaga pode ser feita precisa em mais de um modo.
> 2. Pode somente em princípio; na prática não podemos fazer nossa linguagem vaga completamente precisa, mesmo de um único modo.
> 3. Se uma linguagem vaga é feita completamente precisa, suas expressões mudam de significado, de tal modo que uma descrição semântica apurada da linguagem precisa é uma descrição não apurada da correspondente linguagem vaga."

Ainda, de acordo com Williamson (1994, p. 143), o sobreatribuicionismo responde às objeções da seguinte forma. A primeira objeção acima seria respondida pela consideração de todos os modos de tornar a linguagem vaga precisa. A segunda objeção seria respondida considerando-as coletivamente, sem uma tentativa fútil de especificá-las individualmente. A terceira objeção seria respondida por que a incompletude de um significado vago é espelhada em uma variedade de suas completudes. Como veremos adiante, há controvérsias quanto a esta avaliação.

No que diz respeito à lógica difusa, as limitações potenciais apontadas por Williamson não conseguiriam o mesmo nível de solução. É ponto comum apontar que a abordagem difusa apenas substitui uma vaguidade por várias, criando uma precisão arbitrária e artificial (FINE, 1975; HAACK, 1994 e 1998; VARZI**,** 2001a).

Para Haack, 1994, a abordagem difusa não é imprecisa, mas rigorosa demais. Ao invés de modelar o modo que pessoas falam e pensam sobre vaguidade, lógica difusa força uma quantificação da vaguidade sem qualquer garantia, além disso, não há porque alegar que linguagem natural justifica ou demanda rótulos linguísticos associados a graus de verdade. Que sentido há em dizer que "verdadeiro = 0,3/0,6 + 0,5/0,7 +0,7/0,8+0,9/0,9 + 1/1"? (HAACK, 1998, p. 226).

Essas objeções já são contundentes, mas, o que costuma ser ainda mais realçado é o fato da lógica difusa poder levar a construções semânticas desprovidas de sentido, (WILLIAMSON, 1994, p. 146-147, SAUERLAND, 2009). Pensemos na frase que, em linguagem ordinária, é apreciada com naturalidade, tendo uma mensagem que é compreendida: "João não é alto, nem baixo". Traduzindo-a de modo a reduzir seus

termos, consideremos a frase logicamente equivalente: “João é alto e João não é alto”, “*P* e *Não-P*”. Consideremos também “João é alto e João é alto”, “*P* e *P*”.

Do ponto de vista semântico, as frases são indistinguíveis de seus componentes, ambas tem duas conjunções que não são nem verdadeiras e nem falsas, pois, o termo “alto” é vago. A partir desse exemplo, contrastando as duas abordagens sobreatribuitiva e por lógica difusa, teremos o seguinte:

Em lógica de sobreatribuições, conforme vimos na TABELA 2.1, o valor de “*P* e *Não-P”* será indefinido, e estará coerente com a apreensão cognitiva propiciada pela frase em linguagem ordinária: “João não é alto, nem baixo”.

Em lógica difusa, o grau de verdade da proposição ““*P* e *Não-P*”, supondo que o grau de verdade de *P*=0,5 (logo, “*Não-P* = 0,5”) será Max(0,5; 0,5) = 0,5. Mas considerando também o valor verdade de “*P* e *P*” será Max(0,5;0,5)=0,5. Ainda, considerando as disjunções “*P ou Não-P*” = Min(0,5;0,5) = 0,5, teremos frases de sentidos e significados extremamente distintos sendo consideradas com o mesmo valor semântico pela lógica difusa, o que soa como um grande absurdo[30], ao contrário do sobreatribuicionismo que seria sensível a distinções intuitivamente significantes obliterados pela funcionalidade da lógica polivalente. Na verdade, não somente a lógica polivalente, ou difusa, mas toda lógica verofuncional não clássica está sujeita a este tipo de deficiência semântica. Essa deficiência semântica foi apontada por Kamp[31] e Fine (1975).

Sauerland, no artigo *Vagueness in Language: The Case Against Fuzzy Logic Revisited*, de 2009, retoma essa objeção semântica de Fine e Kamp à abordagem difusa e pontua que linguistas estão interessados em formular modelos formais específicos que podem ser usados para lidar como o fenômeno linguístico, enquanto que lógicos exploram classes de modelos formais que possuem propriedades matemáticas interessantes, e, algumas vezes, também aplicações interessantes,

---

[30] Outro exemplo de limitação semântica é dado por Edgington, 1997, *apud*, Haggard, 2007: seja R_a, R_b e R_c as afirmativas que as bolas a, b e c são vermelhas, respectivamente, e A_s, S_b e S_c, as afirmativas que elas são pequenas. Suponhamos: (1) $|R_a| = 1$; $|S_a| = 0,5$; (2) $|R_b| = 0,5$; $|S_b| = 0,5$; (3) $|R_c|=0,5$; $|S_c| = 0$. Com base nisso temos: $|R_a$ e $A_s| = |R_b$ e $S_b| = 0,5$. Mas a bola “a” não é um caso melhor para “vermelho e pequeno” do que a bola “b”?

[31] Segundo Sauerland, 2009, Hans Kamp atribui o apontamento original a Nicholas Rescher (RESCHER, Nicholas. *Many-Valued Logic*. McGraw-Hill, New York, N.Y., 1969).

(SAUERLAND, 2009, p. 2). As intuições linguísticas frequentemente apresentam conflitos com a lógica, tanto a clássica, quanto as não clássicas, (SAUERLAND, 2009, p. 2).

Sauerland identifica três possibilidades de se lidar com essa deficiência. Primeiro, a exclusão na modelagem de aplicações técnicas desse tipo de situação ("*P* e *Não-P*"). Segundo, considerar que o operador "e" é ambíguo e tratá-lo com sentido diferente quando ocorrer "*P* e *Não-P*". Terceiro, avaliar o desenvolvimento axiomático da lógica difusa para verificar se tais problemas foram solucionados.

Nessa terceira via, Sauerland chega à conclusão que desde o primeiro uso semântico da lógica difusa por Lakoff (1973) essa inconsistência permanece, a despeito de todo o desenvolvimento e aparato formal e axiomático que vem sendo desenvolvido por diversos autores, em especial, por Petr Hájek[32]. Entretanto, Sauerland, vislumbra a adoção de estratégias alternativas, como o uso do "Argumento Dialético", ou dialeteísmo, (SAUERLAND, 2009, p. 7-12) que, embora não sobrepuje as objeções, obtém implicações positivas para se lidar com a vaguidade de modo paramétrico e que possa incluir tanto abordagens sobreatribuitivas, quanto abordagens baseadas em lógica difusa.

De outro lado, a abordagem sobreatribuitiva também apresenta algumas desvantagens assinaláveis que a tornam menos recomendável do que poderia parecer até o momento.

Santos (2006a, p. 718-719), discrimina quatro objeções à abordagem sobreatribuitiva. Primeiro, como a abordagem difusa, a sobreatribuitiva não está isenta de arbitrariedades:

> A tradução do comportamento semântico de um predicado vago em um conjunto de predicados precisos alternativos ignora o fato de que as zonas de aplicabilidade de um predicado vago não são determinadas arbitrariamente, sendo, portanto, dificilmente definíveis com o auxílio de uma variação arbitrária em um domínio de alternativas (precisas); não é arbitrário, p. ex., quais são os indivíduos aos quais "calvo" se aplica de forma correta, equívoca ou incorreta.

[32] HÁJEK, Petr. *Metamathematics of Fuzzy Logic*. Springer, 1998, *apud*, Sauerland, 2009.

Segundo,

> "a solução das sobreatribuições implica que disjunções da forma "P_n ou não P_n" (com *P* vago e *n* um número natural segundo a convenção mencionada acima) sejam sempre verdadeiras – mesmo que *n* pertença à zona de indeterminação de *P*. De fato, para cada versão precisa de *P*, P_n é ou verdadeira ou falsa; e, em cada um desses casos, não P_n é, respectivamente, ou falsa ou verdadeira. Logo, para cada versão precisa de *P*, exatamente um dos disjuntos de "P_n ou não P_n" é verdadeiro, o que torna a disjunção verdadeira em todas essas versões. Essa preservação do Terceiro Excluído mesmo no caso de frases com predicados vagos pode ser vista como uma vantagem (sobretudo para os adeptos da lógica clássica), mas tem o defeito sério de admitir que as disjunções da forma mencionada sejam verdadeiras até nos casos em que nenhum dos seus disjuntos o é; se *n* pertencer à zona de indeterminação de *P*, então nem P_n nem não $P_n$são verdadeiras (segundo a própria análise em termos de sobreatribuições), mas pelo raciocínio anterior, P_n, ou não $P_n$continua a ser.

Tal dificuldade já era apontada por Mehlberg em 1958 (*apud* WILLIAMSON, 1994, p. 145): "se verdade é superverdade, de acordo com Mehlberg, uma disjunção verdadeira pode ter disjunções, ambas, não verdadeiras". O exemplo dado é: "o número de árvores de Toronto pode ser par, ou ímpar". Esta proposição é verdadeira, mas a proposição "o número de árvores de Toronto pode ser par", assim como a proposição "o número de árvores de Toronto pode ser ímpar" pode não ser verdadeira e nem falsa, isto é, pela lógica de sobreatribuições pode ter valor de verdade indefinido. Essa situação leva Mehlberg a distinguir o princípio do meio-excluído como sendo lógico e meta-lógico (princípio da bivalência). A sobreatribuição nega o princípio meta-lógico do meio-excluído, e admite o princípio lógico do meio-excluído (WILLIAMSON, 1994, p. 145).

A terceira objeção apresentada por Santos, 2006a, é ainda mais significativa, e diz:

> O conceito de sobreatribuição implica que, dado um predicado vago *P*, existe um conjunto de versões precisas dele tais que: 1) são "adequadas", isto é, não "contradizem" o significado do predicado; 2) para cada uma dessas versões, existe um *n* tal que *Pn* é verdadeira e *Pn*+1 é falsa. Mas o traço distintivo de um predicado vago *P* (aquilo que o torna vago) é justamente o fato de que nenhum *n* na zona de indeterminação de *P* tem a característica 2) – a vagueza implica (por definição) a ausência de fronteiras distinguindo entre as várias zonas de aplicabilidade de um predicado. Logo, nenhuma das mencionadas versões precisas de *P* pode ser considerada "adequada" ou "consistente com o seu significado"; todas o contradizem. Logo, esse comportamento não pode ser definido por meio delas."

Em suma, precisificações mudam significados. Tal objeção também é amplamente explorada em Fodor e Lepore (1996).

Como vimos mais acima, considerações de domínio e de otimização com base em métodos de tratamento de opiniões sustentam a precisificação e o fechamento das zonas de penumbra. O sobreatribuicionismo também está sob pressão para rejeitar princípios semânticos que são intimamente associados com a aplicação de leis lógicas. Sobreatribuicionismo irá convergir com lógica clássica somente se cada palavra da sentença superavaliada é uniformemente interpretada, Hyde (2005). Entretanto, o que são essas considerações de domínio, otimização e opiniões, senão avaliações pragmáticas e arbitrárias para se tornar precisos predicados que, em si, não são precisos, como o fenômeno abordado diz, são vagos[33]?

A quarta objeção diz respeito à vaguidade de ordem superior, ou seja a vaguidade da vaguidade. Casos limites têm casos limites. Diz Santos, 2006.

> Objeção de caráter metodológico: diz respeito ao fato, já mencionado, de que a fronteira entre os casos de aplicação indeterminada de um predicado vago *P* e os casos inequívocos (de objetos que são inequivocamente *P* ou não *P*) é, ela própria, indeterminada. Nem sempre é inequívoco quando um objeto é indeterminadamente *P*; por outras palavras, o predicado "determinadamente *P*" é tão indeterminado como o próprio *P* – é chamada vagueza de segunda ordem. Por outras palavras, para *P* vago, a noção de *Pa* ser verdadeira é ele própria vaga; e a redução de semântica da vagueza à semântica da precisão por meio do método das sobreatribuições não é capaz de iludir esse fato.

A vaguidade de ordem superior também repudia a introdução de graus de verdade pela lógica polivalente. Se vaguidade já é uma questão inóspita, a vaguidade da vaguidade amplia por demais nosso horizonte, é desnecessária para nossa avaliação e, portanto, não será aprofundada aqui. Em todo o caso, o tema é tratado em Fine (1975); Wright (1992)[34]; Hyde (1994)[35] e Williamson (1999)[36].

---

[33] "Este espaço de vaguidade, essencial aos conceitos da linguagem comum, é o que Waismann chama de 'open texture' e Stegmüller de 'abertura de conceitos'. Nós, na realidade, podemos apenas, por meio de certas regras, diminuir o campo de vaguidade dos conceitos empíricos (na terminologia de Waismann, em contraposição aos conceitos matemáticos) ou dos conceitos de linguagem comum, mas afastar toda e qualquer vaguidade é impossível, pois isso pressupõe conceitos cuja significação está estabelecida de modo definitivo e não podemos, *a priori*, estabelecer regras para todos os casos. Reviravolta linguístico-Pragmática analítica.". OLIVEIRA, Manfredo A. de. *Reviravolta Linguístico-Pragmática na Filosofia Contemporânea*. Edições Loiola, 2006, p. 131.

[34] WRIGHT, Crispin. Is Higher Order Vagueness Coherent? *Analysis*, 52, 1992, p. 129-139.

[35] HYDE, Dominic. Why Higher-Order Vagueness is a Pseudo-Problem. *Mind*, Vol. 103, N° 409, 1994, p. 35-41.

[36] WILLIAMSON, Timothy. On the Structure of Higher-Order Vagueness. *Mind*, Volume 108, Número 429, 1999, p. 127-143.

Voltando à avaliação das abordagens, Machina (1976, p. 177-180) argumenta que enquanto o sobreatribuicionismo administra para preservar todos os teoremas logicamente verdadeiros da lógica clássica ele falha em preservar a validade de regras de inferência. Williamson (1994, p. 151-152) comenta no mesmo sentido: sobreatribuicionismo requer rejeição de regras de inferência tais como contraposição, argumentos por caso (ou eliminação), prova condicional e *reductio ad absurdum*. Em vista disso, sobreatribuicionismo invalida nosso modo natural de pensar dedutivamente. Tal característica, possivelmente, é um dos fatores do sucesso da lógica difusa junto a construtores de sistemas: engenheiros e analistas.

Em vista dessas objeções envolvendo as noções de princípios lógicos, validade, verofuncionalidade e consistência semântica, alguns defensores dessas abordagens ainda colocarão em xeque a importância dessas noções, amenizando sua importância, ou ainda descartando-as em nome de outros princípios lógicos que seriam mais adequados para serem considerados na abordagem da vaguidade. Ou, ainda, buscar-se-á que eles sejam parcialmente atendidos, na busca de se aproximar ao máximo de reintegração dos princípios lógicos, da sustentação da verofuncionalidade, da preservação da validade lógica e da capacidade inferencial da lógica adotada.

Desde 1970, por exemplo, Verma (1970, p. 74) propõe a inclusão do "Princípio do Quarto Excluído" para lidar com a vaguidade. Sanford (1976, p. 195) propõe a retenção da Lei Fraca do Meio Excluído, em conjunto com a rejeição do Princípio do Meio Excluído e da Não Contradição, a adoção do Princípio de Infinitos-Valores Semânticos, capacitando o tratamento lógico da vaguidade com alguns conectivos não definidos verofuncionalmente. Hoje em dia, alternativas continuam sendo apresentadas tanto na defesa da lógica sobreatribuitiva (VARZI, 2007), quanto no lado da lógica difusa (SMITH, 2009[37] *apud*. SAUERLAND, 2009 e SMITH, 2010). A variedade de busca de soluções inclui ainda a combinação das duas abordagens. Em Fermüller e Kosik (2006) a lógica trivalente de Lukasiewicz e lógica clássica são estendidas de modo a incorporar a modalidade sobreatribuitiva correspondente a "...verdadeiro em todas precisificações". Vide também uma proposição mais recente de conciliação em Fermüller (2009).

---

[37] SMITH, N.J.J. *Vagueness and Degrees of Truth*. Oxford University Press, 2009.

Na seção seguinte iremos verificar como a imposição do fenômeno da vaguidade se reflete na concepção e desenvolvimento de ontologias aplicadas.

2.4 Extensões em Ontologias Aplicadas para o Tratamento da Vaguidade

Em ontologias aplicadas a vaguidade vem tendo um tratamento relativamente recente e as proposições apresentadas baseiam-se, ou em lógica difusa, ou em lógica sobreatribuitiva.

2.4.1 Lógica Difusa

Teoria de conjuntos, lógica e sistemas difusos há muito já não são estranhos à Ciência da Informação e Ciência da Computação. Várias áreas de pesquisa buscaram subsídios nessa teoria. Dentre elas, podemos citar a representação e recuperação da informação, banco de dados, mineração de dados, reconhecimento de padrões, raciocínio aproximado, mecanismos de inferência, teorias da possibilidade e crença, estatística e probabilidade difusa, espaços métricos, medições, processamento de imagens, visão computacional, processamento de sinais, teoria dos jogos, controle e automação, sistemas híbridos, etc. Há uma vasta e diversificada teoria e bem sucedidos casos de aplicação comercial. Em vista da multiplicidade, seria contraproducente apresentar referências que traduzam este panorama, mas pode-se ter um panorama e um balanço unificado da área através do Volume 156 da revista *Fuzzy Sets and Systems*, publicada em 2005, em comemoração aos quarenta anos do artigo fundador *Fuzzy Sets* de Zadeh (1965).

As extensões de conjuntos e lógica difusa aplicada ao desenvolvimento de ontologias aplicadas são um dos pontos focais dessa tese e serão abordadas com maiores detalhes no Capítulo 4. As ideias fundamentais de conjuntos, lógica e sistemas difusos são apresentadas no Capítulo 4. Sua inserção na área de computação suave é comentada no Capítulo 3, Seção 3.5.2.

Na Seção 3.5 a seguir compararemos a abordagem realizada por lógica sobreatribuitiva com a de lógica difusa, e as noções fornecidas nas seções anteriores nesse capítulo serão suficientes para a compreensão dessa comparação.

2.4.2 Lógica Sobreatribuitiva

De modo distinto, Bittner e Smith (2001a, 2001b) desenvolvem a Teoria de Partições Granulares em sintonia com a lógica de sobreatribuições, baseada em uma abordagem *de dicto* da vaguidade: a vaguidade ocorre como uma propriedade semântica de nomes e predicados.

A Teoria de Partições Granulares fornece uma abordagem geral dentro da qual se pode compreender a relação entre termos e conceitos vagos de um lado e porções correlacionadas da realidade de outro lado. Entretanto, ao invés de estender a ontologia de modo a abrigar relações que busquem expressar a vaguidade, ela busca "reconceitualizar" as relações entre os termos e conceitos, através de projeções entre objetos e locações rigorosas que estão no mundo (BITTNER e SMITH, 2001b, p. 2).

Como vimos, a lógica sobreatribuitiva defende o ponto de vista de que não há a existência intrínseca de objetos, ou atributos vagos. Desse modo, em ontologias, abstém-se de compromisso ontológicos adicionais e da necessidade de investigar a questão se locação vaga de objetos vagos em regiões vagas é, ou não é a mesma relação da locação rígida. No caso rigoroso, cada partição é caracterizada por uma relação de projeção única e uma relação de localização única. A fim de acomodar a ideia de sobreatribuição, dá-se a restrição que cada partição seja associada com uma única relação de projeção, ou localização.

A Teoria das Partições Granulares considera todas as entidades como sendo rigorosas. Inclusive, no que diz respeito a nomes e predicados, argumenta-se que é insuficiente considerar nomes e predicados vagos do modo como eles ocorrem em sentenças consideradas em abstração. No caso, nessa teoria, os objetos, tomados como referências a serem consideradas, são os objetos normais do mundo, não os objetos extremos, normalmente considerados pela filosofia. Nessa teoria, o mundo é o mundo natural.

Além disso, na Teoria das Partições Granulares a abordagem de sobreatribuições é modificada de modo a tomar conta dos diferentes modos nos quais termos e conceitos projetam – vagamente, ou rigorosamente – em porções correspondentes da realidade em diferentes contextos. As partições vagas propostas por essa teoria

são partições contextualizadas, e este contexto é explicitado na ontologia. Dependendo do contexto, podemos ter uma partição com granularidade mais grosseira, ou mais refinada. Além disso, a avaliação semântica que se faz não se aplicam às sentenças tomadas da linguagem natural, mas dos juízos realizados a partir dessas sentenças (BITTNER e SMITH, 2001b, p. 3).

Uma terceira diferença em relação à abordagem de sobreatribuições é que a Teoria das Partições Granulares busca formular uma teoria da vaguidade que dispensa a noção de lacunas de valor verdade: mesmo juízos envolvendo termos vagos não são marcados pelo valor verdade da indeterminação.

Coerente com a abordagem de sobreatribuições, há para cada nome vago, múltiplas porções da realidade que são igualmente boas candidatas para serem seus referentes, e, para cada predicado vago, múltiplas classes de objetos que são, igualmente, boas candidatas para ser sua extensão. Entretanto, considerando-se o mundo natural e o contexto explicitado, cada nome, cada predicado terá sua porção da realidade determinada.

A teoria das partições granulares tem duas partes (BITTNER e SMITH, 2001b):

A) uma teoria de relações entre células e as partições na quais estão incluídas, e
B) uma teoria das relações entre células e objetos na realidade.

Um detalhamento mais formal é exposto em Bittner e Smith (2001b), mas, em linhas gerais, a parte (A) estuda as propriedades que as partições granulares têm em virtude das relações entre as operações realizadas sobre as células a partir da qual elas estão construídas. Tais partições envolvem células agrupadas em alguma estrutura em forma de grade. Esta estrutura é intrínseca para a própria partição; o que quer dizer, é que independentemente dos objetos em direção às quais deve ser projetada. Esta parte da teoria baliza igualmente bem, tanto partições rigorosas, como as partições vagas (BITTNER e SMITH, 2001b).

A Teoria (B) surge em vista do fato que partições são mais do que somente sistemas de células. Elas são construídas de modo a projetar sobre a realidade. Quando a

projeção é bem sucedida, diremos que o objeto direcionado pela célula pertinente é locado naquela célula. Locação pressupõe projeção.

FIGURA 2.2 Esquerda: uma partição com células Everest, Lhotse e Himalaias. Direita: Uma parte dos Himalaias vista do espaço com candidatos e referentes para o “Monte Lhotse” à esquerda e o “Monte Everest” à direita. (BITTNER e SMITH, 2003c, p. 3).

Em contraste com uma projeção rigorosa única do tipo indicado na parte direita da FIGURA 2.2, as partições vagas têm uma multiplicidade de projeções candidatas para suas células, indicadas pelas regiões limites que podem ser imaginadas como aglomerações de projeções em volta das duas montanhas à direita da FIGURA 2.3. Os limites dos candidatos atuais em direção as quais, as células “Lhotse” e “Everest” estão projetadas sobre os vários P_i em P^v estão incluídas em algum lugar dentro da aglomeração correspondente de regiões.

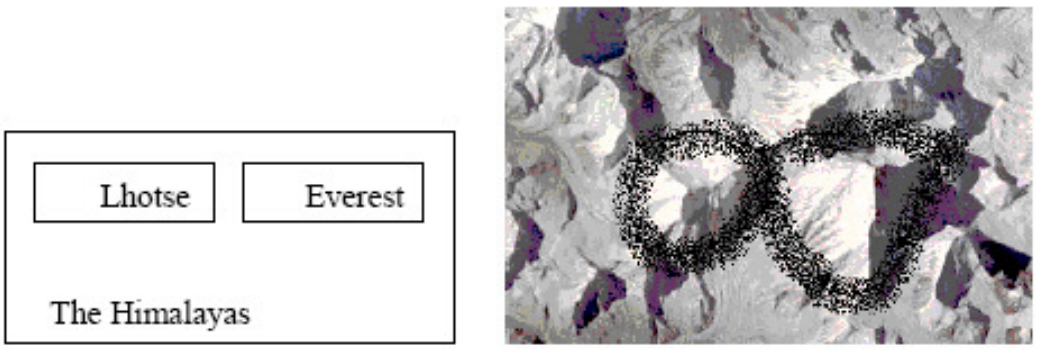

FIGURA 2.3 Esquerda: uma partição com células, Everest, Lhotse e Himalaias. Uma parte dos Himalaias vista do espaço com aglomeração de ovoides representando a família de referentes candidatos admissíveis para o “Monte Lhotse” à esquerda e o “Monte Everest” à direita

A Teoria (A) não é afetada pela rejeição de lacunas, mas a Teoria (B) demanda uma modificação da lógica sobreatribuitiva de modo que objetos rigorosos sejam incluídos como um caso especial. As relações de projeção e de locação são restringidas de modo que para cada projeção haja somente uma única locação, e que para toda célula projetada na realidade haja um objeto correspondente. A vaguidade *de dicto* é capturada no nível de partição via múltiplos modos de projetar rigorosamente (BITTNER e SMITH, 2001b, p. 8).

Na sua exclusão das lacunas, ou zonas de indeterminação, a Teoria das Partições Granulares busca alcançar um nível de precisão, de rigor que seja o bastante para o contexto avaliado, onde as fronteiras são estabelecidas mediante juízos feitos a partir de avaliações semânticas das sentenças que referenciam a realidade. Um grau último de precisão levaria à indeterminação, mas em contextos naturais, um de precisão adequado basta para os propósitos em vista.

A Teoria das Partições Granulares vem sendo utilizada em aplicações na área da saúde e em geografia.

2.5 Comentários Adicionais Sobre a Abordagem Difusa e a Abordagem Sobreatribuitiva

As abordagens apresentadas na seção anterior podem ser consideradas não somente concorrentes, mas também complementares. Também podem ser consideradas abordagens reducionistas e aproximativas.

De um lado, a transposição de questões da filosofia e da lógica, ou, de modo geral, das ciências teóricas para as ciências aplicadas, por vezes, exige a conciliação e harmonização de teorias, em teoria, incompatíveis (COELHO *et al.*, 2010). No caso da lógica sobreatribuitiva e da lógica difusa, exemplos de tal conciliação já são apresentadas nos trabalhos citados de Fermüller e Kosik (2006) e Fermüller (2009), e também sugerida por Bradley (2009a, p. 225).

De outro lado, é fato amplamente constatado a utilização de abordagem reducionista que descarta, ou flexibiliza um, ou alguns aspectos da proposição teórica mais

rigorosa e consistente em prol de uma viabilidade, ou maior alcance da aplicação da teoria a soluções práticas.

Na abordagem difusa, existe uma corrente rigorosa que acredita e busca realmente resolver a questão da vaguidade. Em vista dos progressos alcançados na teoria formal da lógica difusa (HAJÉK, 1998, dentre outros) e da crença na superação das dificuldades teóricas atuais. Vilém Novák (2005) é um dos pesquisadores que respondem positivamente à questão que ele mesmo apresenta: “Conjuntos nebulosos são uma ferramenta razoável para modelar fenômenos vagos?”.

Entretanto, dentro da própria comunidade de conjuntos e lógica difusa há outra corrente que admite uma abordagem mais suave (*soft*). Esta corrente contenta-se em não resolver de modo definitivo e exato as questões que lhe são apresentadas. Contentam-se com uma solução aproximativa e funcional da questão da vaguidade. A lógica difusa é considerada como suporte ao raciocínio aproximado (ZADEH, 1975).

Em sentido amplo, lógica difusa refere-se a um sistema de conceitos, princípios e métodos para lidar com problemas que envolvem classes com limites não agudos, isto é, sujeitas ao fenômeno da vaguidade. Em sentido estreito, lógica difusa refere-se ao cálculo lógico para raciocínio que envolve graus de verdade mais do que tão somente verdade e falsidade (BELOHLAVEK *et al.*, 2009, p. 26). Zadeh (1994a, 1994b) esclarece esses dois sentidos para se entender a lógica difusa.

> O termo lógica difusa é correntemente usado em dois sentidos diferentes. Em um sentido amplo, lógica difusa é um sistema lógico que objetiva uma formalização do raciocínio aproximado. Como tal, é enraizada na lógica multivalorada, mas sua agenda é nitidamente diferente dos sistemas lógicos multivaloradas tradicionais, por exemplo, a lógica de Lukasiewicz. Nesta conexão, o que poderia ser notado é que muitos dos conceitos cuja abordagem para a efetividade da lógica difusas como uma lógica de raciocínio aproximado não são uma parte dos sistemas lógicos multivalorados tradicionais. Entre eles, há o conceito de uma variável linguística, forma canônica, regra difusa Se-Então, quantificadores difusos, e tais modos de raciocínio como raciocínio interpolativo, raciocínio silogístico, e raciocínio disposicional. (ZADEH, 1994a, p. 78).

Shet, Ramakrishnan e Thomas (2005), no contexto da Web Semântica, apresentam três níveis de abordagens ontológicas para lidar com dados heterogêneos na Web: a Implícita, a Formal e a Poderosa. Esta última, também é chamada de Suave (*Soft*) e

baseia-se, justamente, na computação, ou raciocínio suave baseado em lógica difusa, conforme proposto por Zadeh (2002, 1994a e 1994b, 1979).

A divisão entre essas abordagens em lógica difusa, a suave e a rigorosa, a estreita e a ampla, e, ainda, a mais condizente com os problemas de engenharia, em contraposição aos problemas lógico-filosóficos também ficou bem perceptível no *Colóquio Internacional de Praga* de 2006, que tratou de raciocínio sobre probabilidade e vaguidade. Bradley (2009a e 2009b) propôs aos participantes quinze questões formuladas por Christian Fermüller que abordavam aspectos teóricos de natureza filosófica que demonstrariam as limitações da abordagem difusa para tratar a questão da vaguidade. O objetivo era apurar opiniões sobre se essas críticas teriam alguma relevância prática ou qualquer relevância em relação aos campos de aplicação. Se essas questões aumentariam ou diminuiriam a adequação da teoria para se lidar com a vaguidade.

Nas respostas, Bradley observou uma clara divisão entre o ramo técnico, e o ramo teórico da comunidade, este constituído pelos que possuem formação e interesses mais sólidos em matemática e filosofia. No ramo técnico, apenas algumas poucas questões foram consideradas relevantes. No ramo teórico, as questões foram consideradas altamente relevantes, e houve interesse no conhecimento das respostas a essas questões. Dentre desse ramo, o sucesso prático da teoria difusa era reconhecido como uma confirmação de que ela é uma excelente abstração da realidade, mas não implicava que ela é uma representação válida das muitas camadas de vaguidade encontradas na realidade (BRADLEY, 2009b, p. 7).

> "É ainda uma questão aberta se lógica difusa pode ser considerada uma teoria completa da vaguidade em relação às questões apresentadas. Seria definitivamente útil para a lógica difusa se certos princípios considerados elementares em lógica clássica pudessem ser geralmente assegurados em lógica difusa, e se não fosse necessário escolher entre varias alternativas de lógica difusa, algumas preservando certos princípios de lógicas clássicas e outras preservando outros aspectos. Isto leva a dar suporte a tentativas de combinar lógica difusa com outras teorias da vaguidade – como a abordagem sobreatribuitiva – em tentativas que utilizam as vantagens de ambas. (BRADLEY, 2009a, p. 225).

Os problemas enfrentados pela lógica difusa para, em última instância, tratar a vaguidade não são considerados empecilhos nem para um ramo, nem para o outro. Dubois e Prade (1994), dois dos mais antigos e importantes pesquisadores da área, declaram explicitamente que a lógica difusa é uma ficção conveniente. Se a

vaguidade apresenta-se como um fenômeno resistente ao tratamento, resta-nos gerenciá-la e manipulá-la dentro de determinados contextos e em vista de determinados fins para a solução de questões práticas.

Já os adeptos da abordagem sobreatribuitiva não chegam a assumir explicitamente o "caráter ficcional" do tratamento último da vaguidade e referem-se às suas soluções como sendo não só superiores, mas plenamente adequadas à solução da vaguidade. Entretanto, a versão da abordagem sobreatribuitiva adotada na Teoria das Partições Granulares é claramente uma redução do problema. Para eliminar boa parte das críticas à abordagem sobreatribuitiva e torná-la mais simples e funcional, Bittner e Smith introduzem e modelam a situação de contexto. A situação de contexto tira da abordagem sobreatribuitiva os problemas decorrentes da validade global, de modo que a Teoria de Partições Granulares tenha que lidar tão somente com a validade local e, desse modo, ao mesmo tempo em que recupera várias das funcionalidades e vantagens da lógica clássica afasta-se da real consideração do fenômeno que se pretendia resolver.

A consideração da modelagem da semântica de sentenças e o juízo aproximado também revelam o caráter aproximativo, ficcional da solução. Em verdade, parece-nos que tal qual para o ramo técnico da lógica difusa, as questões lógico-filosóficas da vaguidade são de somenos importância para os proponentes da Teoria das Partições Granulares.

Tal atitude, inclusive, é bastante coerente com a posição defendida por Barry Smith, conjuntamente com Kevin Mulligan e Peter Simons. No artigo *What is Wrong With Contemporary Philosophy?* (MULLIGAN, SIMONS e SMITH, 2006), onde, dentre outras, a Filosofia Analítica é colocada na berlinda, e eles elencam a vaguidade como sendo um dos assuntos inúteis. Consideram-na dentre os típicos assuntos da Filosofia Analítica que cogitam uma série de quebra-cabeças, inflam tendências, mas quedam-se sem soluções óbvias e sem deixar o mundo mais sábio.

A despeito das críticas à Filosofia Analítica, o próprio Smith ignorar suas críticas para tratar problemas específicos. A "Teoria da Projeção" apresentada pelo analítico Wittgenstein no *Tractatus Logico-Philosophicus* (1922) é a clara inspiração da

modelagem de projeção realizada na Teoria das Partições Granulares, vide Bittner, Donnelly e Smith (2004).

2.6 Conclusão

O fenômeno da vaguidade é um fenômeno resistente a abordagens. As soluções apresentadas como extensões de ontologias aplicadas que buscam tratá-la na área de Ciência da Informação e Ciência da Computação possuem caráter aproximativo e apenas “ficcionalmente” resolvem o fenômeno.

A despeito disso, abordagens que irão buscar o tratamento da vaguidade, mesmo que não possam eliminá-la, irão buscar gerenciá-la dentro de determinados contextos e para determinados fins.

Essas abordagens representam um grande avanço, seja pela consideração da relevância do fenômeno para a área, onde apresenta inúmeros campos de aplicação, seja pelas soluções já apresentadas no tratamento da vaguidade, ou do amplo campo aberto para o desenvolvimento e aprimoramento de novas soluções que, inclusive, podem se utilizar de abordagens mistas. Os campos de aplicação que se destacam são a recuperação de documentos, em especial no contexto da Web Semântica, ontologias geográficas e na área de saúde.

As abordagens teóricas que vem sendo incorporadas às soluções práticas envolvem lógica difusa e lógica sobreatribuitiva. Ambos os campos teóricos vem se revelando ativos e dinâmicos, avançando em suas formulações e axiomatizações, prometendo um contínuo aprimoramento nas soluções.

3 Suporte de Ontologias Aplicadas à Mineração de Dados

3.1 Introdução[38]

O presente capítulo trata da utilização de ontologias nas fases de pré e pós-processamento à mineração de dados por regras de associação. Este uso busca agregar o conhecimento de domínio e propiciar suporte semântico na fase de preparação de dados e na fase de pós processamento, auxiliando o analista a explicar e interpretar as regras obtidas.

Inicialmente, introduzimos noções básicas de mineração de dados e apresentamos uma taxonomia de métodos. Realçamos a divisão entre mineração de dados supervisionada e não supervisionada que, naturalmente, associam-se a tarefas de prescrição e descrição, respectivamente. Dentre a mineração de dados não supervisionada, destacamos a mineração de dados por regras de associação.

Em seguimento, é apresentado um levantamento dos estudos anteriormente realizados do uso de ontologias no suporte às fases de pré e pós-processamento em mineração de dados.

A partir da análise das características das tarefas de prescrição, predição, descrição, explicação e interpretação são discutidas e apresentadas indicações de conceitos e métodos que devem ser incorporados a metodologias de ontologias quando estas se associam a contextos humano-sociais, típicos de questões surgidas em economia, administração, contabilidade, direito, política, dentre outros domínios.

---

[38] O tema deste capítulo foi conteúdo das seguintes publicações:

COELHO, Eduardo de Mattos Pinto Coelho; BAX, Marcello Peixoto e MEIRA JÚNIOR. Wagner. Suporte de ontologias aplicadas à mineração de dados por regras de associação. *Proceedings of Joint IV Seminar On Ontology Research in Brazil and VI International Workshop on Metamodels, Ontologies, Semantic Technologies*, Gramado, Brazil, September 12-14, 2011, p. 171-176. Disponível em <http://ceur-ws.org/Vol-776/ontobras-most2011_paper20.pdf>, acesso em 28-12-2011.

COELHO, Eduardo de Mattos Pinto Coelho; BAX, Marcello Peixoto e MEIRA JÚNIOR. Wagner. "Ontologias nebulosas no suporte a mineração de dados". *International Conference on Information Systems and Technology Management, 7th CONTECSI. São Paulo, 2010.*

COELHO, Eduardo de Mattos Pinto Coelho; BAX, Marcello Peixoto e MEIRA JÚNIOR. Wagner. Ontologias nebulosas no suporte a mineração de dados. *X Encontro Nacional da Associação Nacional de Pesquisa em Ciência da Informação* (Enancib X). João Pessoa, Paraíba. Disponível em: <http://dci2.ccsa.ufpb.br:8080/jspui/handle/123456789/585>, acesso em 12-12-2009.

Tanto ontologias quanto mineração de dados são áreas prevalentemente construídas com base nas ciências naturais. Trabalham com as noções de necessidade, objetividade e explicação, associadas às ciências físicas, químicas e biológicas. Entretanto, em organizações humanas, os fenômenos humanos prevalecem, demandando metodologias específicas de ciências humanas que, por sua vez, exigem as noções de possibilidade, subjetividade e interpretação.

Argumenta-se que tais noções, conceitos e métodos, de um lado, já são parcialmente abrigados nas tecnologias humano-cêntricas desenvolvidas sob a perspectiva de computação suave e granular e, de outro lado, mesmo se não abrigadas, essas tecnologias são adequados à incorporação desses novos conceitos e métodos.

Com base na computação suave e granular, desenvolvida sob a perspectiva de sistemas de engenharia *difusa*, é proposta a utilização de ontologias difusas para o suporte à atividade de mineração de dados.

3.2 Mineração de Dados

A mineração de dados é uma das etapas do processo chamado Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* – KDD[39]). Fayyad *et al.* (1996) e Maimon e Rokach (2005, p. *i*) afirmam que o KDD consiste em um processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, em um conjunto de dados. Esses padrões podem se constituir, dentre outros, em regras de associação, ou sequências temporais que permitam revelar relacionamentos não aleatórios entre atributos de variáveis.

---

[39] No âmbito da indústria, há duas metodologias que são bastante usadas e tomadas como referências. Primeiro, a metodologia não proprietária "Cross Industry Standard Process for Data Mining", CRISP-DM, 1999,http://www.crisp-dm.org/. Segundo, a metodologia "Sample, Explore, Modify, Model", SEMMA, propriedade da SAS, empresa líder em BI (business inteligence) http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html. De acordo com a pesquisa de agosto de 2007 da KDnuggets (Knowledge Discovery Nuggets), registrada em http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm, o CRISP-DM é utilizado em 42%, a SEMMA em 13% e o processo KDD em 11% dos casos de aplicação de mineração de dados. Essas metodologias possuem vários aspectos comuns, e, para nossos comentários, tomaremos a metodologia KDD como referência.

Em especial, a complexidade e o volume excessivo de dados relevam a importância da mineração de dados para a descoberta de padrões que possam revelar informações úteis ao processo de tomada de decisão, tornando-a fundamental em contextos organizacionais.

Pode-se visualizar na FIGURA 3.1 as diversas etapas de transformação dos dados existentes neste processo, onde a Mineração de Dados é o núcleo do processo de descoberta de conhecimento em bases de dados de natureza diversas. Os dados podem ser estruturados, semi, ou mesmo não estruturados, podem se constituir em banco de dados, mas também de textos, sons, imagens, etc. como usualmente se encontram dispersas as informações nas organizações e na web.

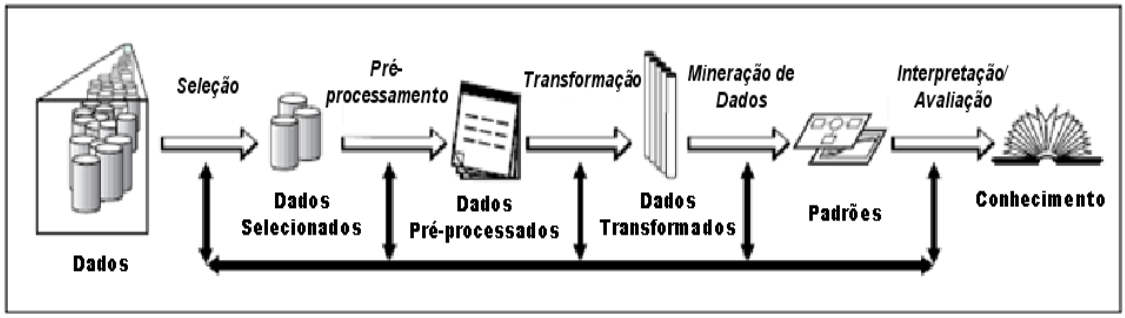

FIGURA 3.1 Passos constituintes do processo de descoberta de conhecimento em bases de dados, clássico diagrama de Fayyad (FAYYAD *et al* , 1996, p. 29).

A mineração de dados terá que lidar com questões de ordem computacional e de usabilidade (GONÇALVES, 2004 e 2002; SEIFERT, 2004). Para tornar-se efetivamente útil, terá que transpor problemas quanto à confiabilidade dos dados, a geração de resultados excessivos, as limitações de autonomia e independência que demandam o conhecimento do negócio e entendimento dos dados para a análise dos resultados, limitações de desempenho e produtividade, limitações de facilidade de uso no trabalho de rotina e limitações de utilidade.

3.2.1 Pré-processamento

Para buscar amenizar tais problemas e alcançar melhores níveis de eficácia, desempenho e produtividade, a mineração de dados adota métodos que aprimoram as etapas de seleção, pré-processamento e transformação dos dados.

Após a seleção e extração de características relevantes, os dados selecionados serão depurados de inconsistências (*cleansing*), valores de atributos ausentes serão tratados (MALETIC e MARCUS, 2005; PYLE, 1999), os dados supérfluos e redundantes serão cortados (*pruning*), e os dados sem interesse serão filtrados. Além disso, buscar-se-á a redução da dimensão dos dados, sua discretização e normalização

Para isso serão extraídos, filtrados, diferenciados, ordenados, redimensionados (BURGES, 2005; CHIZI e MAIMON, 2005), distribuídos, normalizados, suavizados, discretizados (PYLE, 1999; YANG, WEBB e WU, 2005) com nula, insignificante, ou mínima perda de informações potenciais. Também serão enriquecidos a partir da avaliação da ocorrência e potencial importância de conteúdos de variáveis vazios, ou perdidos (GRZYMALA-BUSSE e GRZYMALA-BUSSE, 2005).

Nesse pré-processamento, os dados serão avaliados quanto à sua consistência, integridade, nível de agregação, granularidade, monotonicidade, linearidade, dispersão, concorrência, etc. (PYLE, 1999).

Além disso, deverão ser considerados a ocorrência de valores muito desviantes (*outliers*), desagrupados, destoantes, ou isolados, no que se inclui a avaliação da ocorrência de valores raros e extremos (BEN-GAL, 2005 e WEISS, 2005).

O objetivo principal de toda essa etapa de pré-processamento é criar uma representação uniforme e confiável dos dados que serão entregues à mineração de dados.

3.2.2 Pós-processamento

A mineração de dados gera os padrões que, por sua vez, deverão também ser processados. No pós-processamento, as principais dificuldades relacionam-se ao excesso e a complexidade de padrões que normalmente são gerados, e de como esses padrões serão avaliados e interpretados para efetivamente se transformarem em informações úteis aos processos organizacionais.

Os padrões gerados serão avaliados quanto à qualidade e ao grau de interesse, e serão reduzidos por sumário, ou agrupamento. Recomenda-se, ainda, utilizar técnicas de visualização dos resultados (YANG, 2005, KEIM, 2002; THEARLING *et*

al., 2001 e KORPIPÄÄ, 2001 e KOHAVI 2000) de modo a se tentar aprimorar, ou até mesmo viabilizar a ação do intérprete, viabilizar a análise interpretativa dos dados resultantes da mineração. Um panorama dessas técnicas e as referências de seus desenvolvimentos originais é apresentada em Baesens *et al.*(2000), e Sahar (2005).

Dentre as medidas utilizadas para se avaliar os padrões resultantes da mineração de dados há as medidas de interesse objetivas e subjetivas que buscam avaliar a novidade do padrão descoberto (SOUZA e CARVALHO, 2007; e GONÇALVES, 2005).

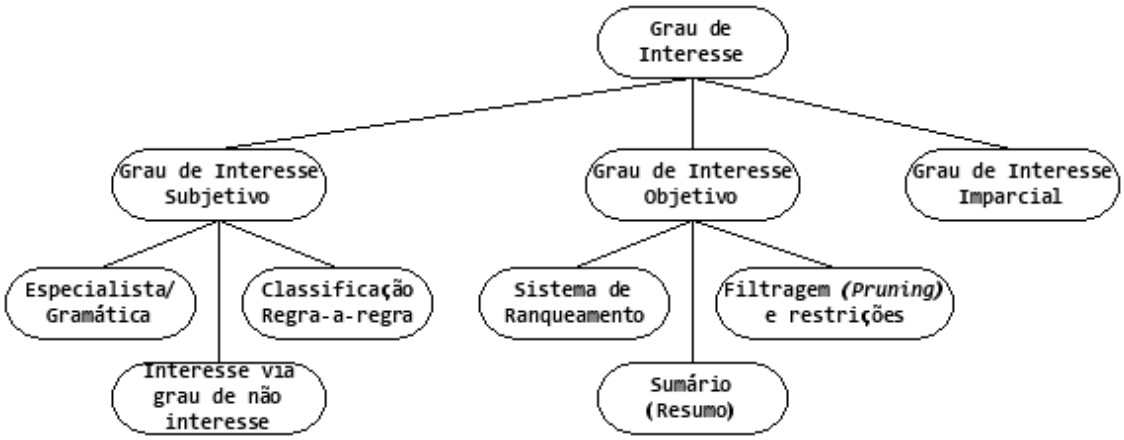

FIGURA 3.2 Tipos de abordagens de grau de interesse (SAHAR, 2005, p. 650).

A primeira, como o próprio nome informa, busca uma mensuração de critérios objetivos dos dados, independente do sujeito (YAO e ZHONG, 1999). São medidas de qualidade dos próprios dados. Já a segunda, busca uma mensuração de critérios dos dados dependente do sujeito (TAN *et al*, 2002 e SILBERCHATZ e TUSHILIN, 1996).

Nas referências citadas acima pode-se constatar a variedade de abordagens e medidas de graus de interesse que, por sua vez, podem ser interessantes em vista do caso a ser tratado. Entraremos em detalhes sobre algumas medidas quando formos tratar do pós-processamento de regras de associação.

3.2.3 Taxonomia de Paradigmas de Mineração de Dados

Há uma variedade de paradigmas de mineração de dados. Objetivando nossa análise, consideramos mais interessante a taxonomia que realça a existência de modelos supervisionados e não supervisionados.

Em modelos supervisionados, como o próprio nome já diz, o processo de detecção de padrões é supervisionado. As classes à qual cada padrão possa ser pertinente são pré-definidas. São modelos prescritivos. Além disso, são preditivos no sentido de serem mais adequados a revelar tendências.

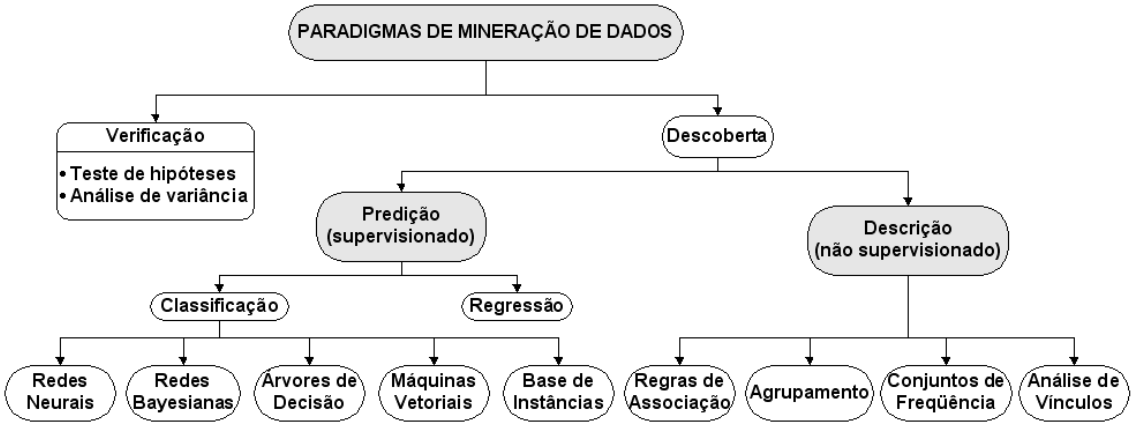

FIGURA 3.3 Taxonomia de paradigmas de mineração de dados (MAIMON e ROKACH, 2005, p. 7).

Já em modelos não supervisionados, não se conhecem a priori classes às quais os padrões possam ser pertinentes. São modelos de ênfase descritiva, no sentido de serem mais adequados a apresentarem descrições, relevando características dos dados minerados**.**

3.2.4 Conjunto de Itens Frequentes e Mineração de Dados Por Regras de Associação

Mineração de dados por regras de associação e por conjuntos de frequência são paradigmas amplamente utilizados na descoberta de conhecimento ainda oculto em grandes volumes de dados. Conforme a FIGURA 3.3 os situa, são paradigmas não supervisionados que permitem obter uma descrição dos dados minerados.

O exemplo clássico quase sempre utilizado é a descoberta de padrões de compras em supermercados. A identificação desses padrões de compras pode levar ao conhecimento de associações de itens de compras a partir de critérios de frequência de associação desses itens. A obtenção desse conhecimento influenciará estratégias mercadológicas que podem afetar da disposição dos produtos na prateleira até a elaboração da campanha publicitária.

Faremos a seguir uma apresentação mais formal das regras de associação e itens frequentes, conforme é apresentado em Simovice e Djeraba (2008, p. 273-293). Introduzimos apenas algumas passagens intermediárias na sequência de fórmulas, e alguns esclarecimentos adicionais, para tornar mais fácil seu acompanhamento.

Seja *I* um conjunto finito de *n* elementos, $I = \{i_1, \ldots, i_n\}$. Referimo-nos aos elementos de *I* como *itens*. Uma regra de associação é uma implicação da forma $X \rightarrow Y$, onde $X \subset I, Y \subset I \; e \; X \cap Y = \emptyset$. Ou seja, uma regra de associação sobre um conjunto de itens *I* é um par não vazio de conjuntos de itens disjuntos (*X*,*Y*).

Cada regra de associação é caracterizada por meio de seu **suporte** e **confiança** que, por sua vez, são definidos em termos de frequência de ocorrência de itens, com base nas seguintes definições:

1. Um **conjunto de transação de dados** sobre *I* é uma função $T:\{1,\ldots,n\} \rightarrow P(I)$. O conjunto $T(k)$ é a *k*-ésima transação de *T*. Os números 1,…,*n* são os identificadores da transação (*t*).
2. Seja $T:\{1,\ldots,n\} \rightarrow P(I)$ um conjunto de dados de transação sobre um conjunto de itens *I*. A **contagem do suporte** de um subconjunto *K* do conjunto de itens *I* em *T* é o número $cont\_suporte_T(K)$ dado por

$$cont\_suporte_T(K) = |\{k|\ 1 \leq k\ \leq n \; e \; K\ \subseteq T(k)\}\ |$$

O **suporte de um item** do conjunto *K* é o número:

$$suporte_T(K) = \frac{cont\_suporte_T(K)}{n}$$

3. Um item do conjunto *K* é ***μ-frequente*** relativo ao conjunto de transação de dados *T* se $suporte_T(K) \geq \mu$.

   Chamamos de $\mathcal{F}_T^{\mu}$ a coleção constituída pela união de todos os conjuntos de itens *μ-frequente* relativo ao conjunto de transação de dados *T* e por $\mathcal{F}_{T,r}^{\mu}$ a coleção de todos conjuntos de itens *μ-frequentes* que contém *r* itens para $r \geq 1$.

$$\mathcal{F}_T^{\mu} \bigcup_{r \geq 1} \mathcal{F}_{T,r}^{\mu}$$

Para compreendermos a dificuldade em lidarmos com o volume de regras de associação, observe que se $|I|=n$ e, conforme definido para $X \rightarrow Y$, onde $X \subset I, Y \subset I \; e \; X \cap Y = \emptyset$, temos que $|X|=k$, $|Y|=n-k$, logo: $x + y = n$.

O conjunto X de consequentes contem k elementos, e X é não vazio, logo k>0. Há $\binom{n}{k}$ modos de escolher X. Uma vez que X é escolhido, Y pode ser escolhido entre os subconjuntos não vazios remanescentes de $I-X$, ou seja, há $2^{n-k}-1$ modos de escolher Y.

$$\sum_{k=1}^{n} \binom{n}{k} (2^{n-k}-1) = \sum_{k=1}^{n} \binom{n}{k} 2^{n-k} - \sum_{k=1}^{n} \binom{n}{k}$$

Tomando-se $x=2$ na igualdade. Toma-se este valor considerando-se que na seleção de um item de frequência, o item pode estar presente, ou não na seleção:

$$(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k}$$

Temos:

$$(1+2)^n = \binom{n}{0} 2^{n-0} + \sum_{k=1}^{n} \binom{n}{k} 2^{n-k}$$

$$(3)^n = \frac{n!}{0!\,(n-0)!} \binom{n}{0} 2^n + \sum_{k=1}^{n} \binom{n}{k} 2^{n-k}$$

$$3^n = 2^n + \sum_{k=1}^{n} \binom{n}{k} 2^{n-k}$$

$$\sum_{k=1}^{n} \binom{n}{k} 2^{n-k} = 3^n - 2^n$$

Considerando-se que o conjunto *X* é não vazio, subtraímos o somatório de regras possíveis de 1, de tal modo que:

$$\sum_{k=1}^{n} \binom{n}{k} 2^{n-k} = 3^n - 2^n - 1$$

Podemos observar que o número de regras de associação pode ser um número considerável mesmo para pequenos valores de *n*. Por exemplo, para *n*=10,temos $3^{10} - 2^{11} + 1 = 57.002$ regras de associação (SIMOVICE e DJERABA, 2010, p. 282). Marinica e Guillet (2010, p. 1041) chegam a comentar sobre a existência de experimentos que mostram que um número de regras superior a 100 torna seu uso quase impossível. Isso basta para demonstrar que o pós-processamento das regras é totalmente imprescindível.

Uma vez que um item de conjunto *μ-frequente Z* esteja identificado, precisamos examinar os níveis de suporte dos subconjuntos *X* de *Z* para assegurar que uma regra de associação da forma *X*➔*Z* – *X* tem o nível de confiança suficiente, $confiança_T(X \rightarrow Y) = \frac{\mu}{suporte_T(X)}$. Observe que $suporte_T(X) \geq \mu$ porque *X* é um subconjunto de *Z*. Para obter um alto nível de confiança para *X*➔*Z-X*, o suporte de *X* deve ser tão pequeno quanto possível. Entretanto, quanto menor o suporte, maior o volume de regras geradas pela mineração.

4. Uma regra de associação insere-se em um conjunto *T* de transação de dados com suporte *μ* e confiança *c* se $\boldsymbol{suporte_T(X \rightarrow Y) \geq \mu}$ **e** $\boldsymbol{confiança_T(X \rightarrow Y) \geq c}$.

Claramente, se *X*➔*Z* – *X* não encontra o nível de confiança, então será inútil buscar por regras da forma *X'*➔*Z* – *X'* entre os subconjuntos *X'* de *X*.

Esclarecendo, a medida de confiança é a probabilidade condicional do consequente dado o antecedente. Mineração de dados por regra de associação seleciona as regras que possuem um suporte e uma confiança acima dos limites definidos pelo usuário, isto é, acima de um suporte mínimo e uma confiança mínima.

$$suporte\ (X \rightarrow Y) = \frac{número\ de\ transações\ suportadas\ X \cup Y}{total\ de\ número\ de\ transações}$$

$$\textit{confiança}\ (X \rightarrow Y) = \frac{suporte(X \cup Y)}{suporte(X)} = \frac{p(X \cup Y)}{p(X)} = p(X|Y)$$

O algoritmo mais utilizado para a determinação das regras de associação baseadas nas medidas de confiança e suporte, calculadas a partir dos itens frequentes, é o algoritmo Apriori de Agrawal e Srikant (1994).

Essa abordagem mostra-se limitada considerando-se a dificuldade em se determinar adequadamente os parâmetros de suporte e confiança mínima, pois casos de baixa frequência podem ser significativos. Além disso, a medida de confiança não é muito adequada para expressar a dependência do consequente em relação ao antecedente. De um modo geral, suporte e confiança seriam medidas insuficientes para a seleção das regras.

Além disso, a mera associação de regras com base nos itens frequentes não considera as forças das associações do ponto de vista semântico. As regras são dotadas de sentido e significado, e se apresentam com relevâncias distintas considerando-se o domínio em questão.

Em vista dessas limitações, inúmeras abordagens com métodos, técnicas e algoritmos alternativos, ou que agregam novas métricas e estratégias à abordagem original foram apresentados. Elas percorrem um amplo espectro que vão de puras técnicas estatísticas à crescente consideração do domínio.

Várias abordagens são citadas e comentadas em Baesens *et al. (*2000), Bruha e Famili (2000). Além dessas, consideramos interessantes Bayardo e Agrawal (1999) que incluem uma variedade de métricas de interesse incluindo *lift*, convicção, *chi*-quadrado, avaliação de entropia, laplace, dentre outras; e Melanda (2004) que agrega método de Pareto com associação de medidas para selecionar as regras; Neves (2002) que utiliza operadores que incidem sobre as regras para gerar novo conjunto de regras mais adequadas à análise e Chawla *et al.*(2004) que usam hipergrafos direcionados para filtrar regras.

Na via de consideração do conhecimento de domínio, consideramos interessantes os mecanismos de sumarização, a generalização de regras através de taxonomias (DOMINGUES, 2004; DOMINGUES e REZENDE, 2005; CARVALHO, 2007 e TSENG *et al.*, 2007), e também na construção de bases de meta-regras.

Avançando mais ainda na consideração do conhecimento do domínio de aplicação na mineração por regras de associação, veremos a seguir o uso de ontologias na fase de pré e pós-processamento.

3.3 O Uso de Ontologias no Suporte à Mineração de Dados

Como vimos, a despeito dos algoritmos e ferramentas de mineração de dados serem excepcionais nos critérios de desempenho e eficiência na criação de novos dados, frequentemente eles são insuficientes para garantir seu uso prático nos negócios onde buscam ser aplicados. O conhecimento do domínio é crucial para o sucesso, ou, mesmo, sua desconsideração inviabiliza o uso dos resultados da mineração de dados. Em vista disso, em contraposição ao paradigma de mineração de dados centrado no dado, e indo além das iniciativas isoladas, vem-se desenvolvendo todo um paradigma de mineração guiado pelo conhecimento de domínio (*domain-driven data mining*, D^3M) e pela praticabilidade de uso do conhecimento que é descoberto e entregue (*actionable knowledge discovery and delivery,* AKD). Um panorama crítico dessa evolução para o pode ser vislumbrado em Cao (2010). Em reforço a este paradigma recente, o uso de ontologias como elemento de expressar o conhecimento do domínio vem sendo utilizado já há alguns anos.

Uma ontologia é a especificação explícita e formal de uma conceituação compartilhada (GRUBER, 1993), fornecendo uma base conceitual para sistemas baseados em conhecimento. Elas buscam representar conhecimento consensual entre pessoas e aplicações, permitindo a interoperabilidade e a compatibilidade semântica entre distintos sistemas, e o reuso e compartilhamento de informações oriundas desses sistemas.

Para tal, relaciona conceitos a termos e relações em um dado domínio, fornecendo a base conceitual para a definição de elementos de dados, sua hierarquia e relacionamento na modelagem conceitual (SOERGEL, 1999). As definições e seus relacionamentos são formalizados em axiomas lógicos, o que permite agregar mecanismos de inferência lógica à ontologia e desenvolver mecanismos de classificação a partir desse mecanismo de inferência.

Essas características da ontologia que lhe permitem dar suporte a mecanismos de classificação baseados no conhecimento, tornam-nas candidatas naturais a propiciarem o suporte necessário aos mecanismos de mineração de dados.

Boas referências do uso de ontologias no pós-processamento de regras de associação geradas pela mineração de dados são Marinica, Guillet e Briand (2008), Marinica e Guillet (2010), e o recente Mansingh, Kweku-Muata e Reichgelt (2011). Esses trabalhos fazem um valioso levantamento e eventual discussão de trabalhos anteriores e relacionados a soluções baseadas em ontologias, e apresentam proposições com praticabilidade demonstrada em estudo de casos. Nessas proposições, como normalmente acontece, as soluções possuem caráter híbrido, onde ontologias estão associadas a esquemas de meta-regras, operadores de regras e métricas objetivas e subjetivas de modo a alcançarem seus objetivos de aprimorar a mineração de dados do ponto de vista de seu efetivo e eficiente uso (acionabilidade).

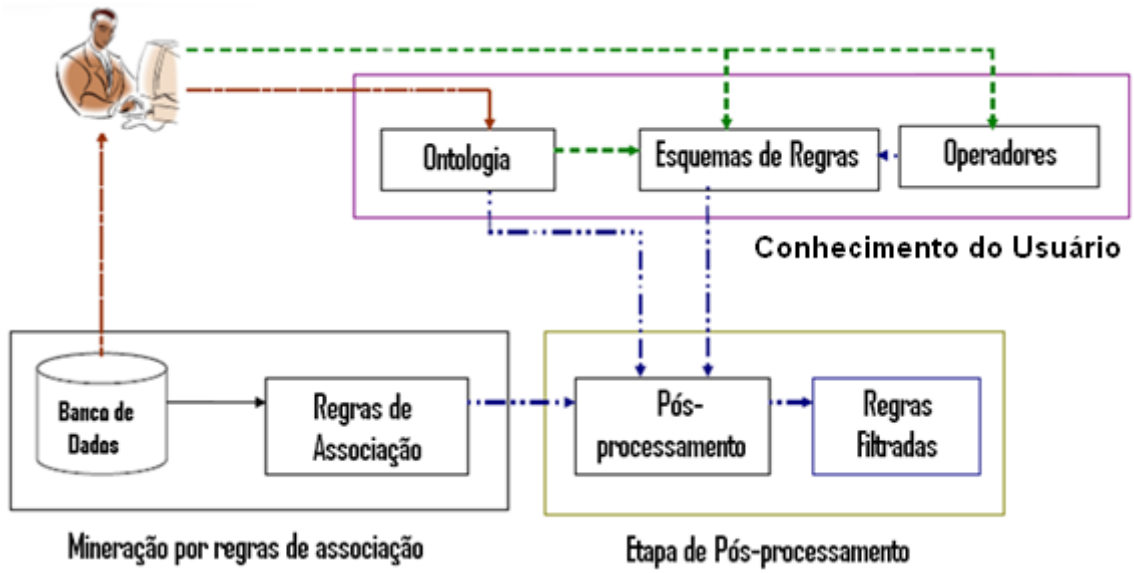

FIGURA 3.4 Mineração de dados integrada a ontologia, esquemas de regras e operadores. Marinica, a partir de slide de apresentação[40].

O trabalho de Mansingh, Kweku-Muata e Reichgelt (2011), propõe um método de depuração e filtragem híbrido que envolve o uso de análise objetiva e subjetiva,

[40]MARINICA, Claudia; GUILLET, Fabrice e BRIAND, Henri. Imagem extraída de slide projetado durante apresentação no *The Second International Workshop on Domain Driven Data Mining (DDDM 2008), IEEE International Conference on Data Mining (ICDM 2008),* Nantes, França, 2008.

onde uma ontologia dá suporte à análise subjetiva. O estudo de caso envolve um banco de dados médico. O conhecimento do domínio é requerido para determinar o que serão regras já conhecidas, ou que representem novidade, ou seja, realmente, novas associações descobertas.

Os trabalhos de Marinica *et al.* (2009) são de nosso especial interesse pela proximidade da abordagem que proporemos adiante. Eles usam ontologias para aprimorar a integração do conhecimento do usuário na tarefa de pós-processamento e esquema de regras (meta-regras) para especificar as expectativas do usuário.

Todos esses três trabalhos buscam tratar o conhecimento de caráter subjetivo a partir da abordagem apresentada por Liu *et al.*, 1999, ou Liu *et al.*, 2000. Nessa abordagem, cria-se uma linguagem de especificação simples que envolve três tipos de conhecimento prévio: *impressões gerais*, *conceitos razoavelmente precisos* e *conhecimento preciso*. *Impressões gerais* envolvem as crenças vagas dos usuários sobre as associações entre certos conceitos, *conceitos razoavelmente precisos* especificam tanto a associação entre conceitos como a direção da associação e *conhecimento preciso* inclui os valores de suporte e confiança. Eles também usam conhecimento de *impressão geral* e *conceito razoavelmente preciso* para categorizar as regras extraídas como conformes, *consequente inesperado*, *condição inesperada* e *ambos os termos das regras inesperados* (isto é, antecedente e consequente inesperados) (MANSINGH *et al.*, 2011, p. 420).

Em Marinica *et al.* (2009), a ontologia é criada a partir da generalização das regras de associações, e a definição de *impressões gerais*, com base nos objetivos, nas crenças e expectativas dos usuários, são utilizadas para constituir os esquemas de regra. Operadores de corte (*pruning*, *P*), e filtragem de acordo com a conformidade (*conform*, *C*) e de acordo com o inesperado (*unexpect*, *U*) são aplicados às regras a fim de reduzir o seu número. Em Marinica *et al.* (2010), tais procedimentos tornam-se interativos, permitindo aprimoramento via intervenção dos usuários.

Uma outra forma de agregar facilidades à mineração de dados com ontologias encontra-se em Wu *et al.*, 2009. Nesse trabalho, a ontologia de preferências do usuário mantém as consultas mais frequentes e representativas na história da mineração de dados facilitando e agilizando o processo iterativo de mineração de

dados. Os resultados das novas minerações são despachados automaticamente a usuários específicos de acordo com suas preferências.

Já no pré-processamento, há o grande potencial, pelo que sabemos ainda não explorado, de se usar as ontologias para auxiliar o processo de seleção de dados, seja no processo de consulta ao banco de dados, ou textos; seja no processo de corte e filtragem de dados pré-selecionados.

3.4 Peculiaridades de Construção de Bases de Conhecimento em Ciências Humanas

Métodos para empreender a análise científica dos fenômenos humano sociais, precisam articular os níveis descritivo, explicativo e interpretativo (DOMINGUES, 2004, p. 103). Estas tarefas se articulam dinâmica e dialeticamente para a constituição e no uso da base de conhecimento empírico de determinado domínio de interesse.

Ao longo dessa tese referimo-nos a bases de conhecimento, às vezes associando ao termo o adjetivo empírico, factual, difuso, em lógica descritiva, ontológica, especialista, ou do senso comum. Para não disseminar maiores ambiguidades, ou mal entendidos, esclarecemos que entendemos como sendo uma base de conhecimento empírico um conjunto arbitrário de registros de coisas, fatos, eventos, processos, ou acontecimentos, que foram registrados de alguma forma, após serem percebidos de algum modo. Essa base de conhecimento empírico pode ter o caráter de aglutinar conhecimento de natureza especialista ou do senso comum, conhecimento difuso (vago), conhecimento registrado em uma linguagem de lógica descritiva, ou expresso ontologicamente, em classes e propriedades.

Como vimos nas seções anteriores, os paradigmas de mineração de dados subdividem-se em dois grandes ramos, o de mineração supervisionada, e o de mineração não supervisionada.

A supervisão está associada à atividade de previsão a partir de conhecimentos prescritos isto é, explicitamente pré-estabelecidos (**prescrição**). Ela parte de conhecimentos prévios que, no processo de busca no banco de dados, requerem

reconhecimento, confirmação. Descobrem-se fatos novos, mas de algo (categorias, classes, atributos) que já se conhecia.

A não supervisão está associada à atividade de **descrição**. Não há conhecimento prévio a ser reconhecido, mas busca-se a descoberta de associações que se requerem serem descritas e que possam se constituir em conhecimento útil.

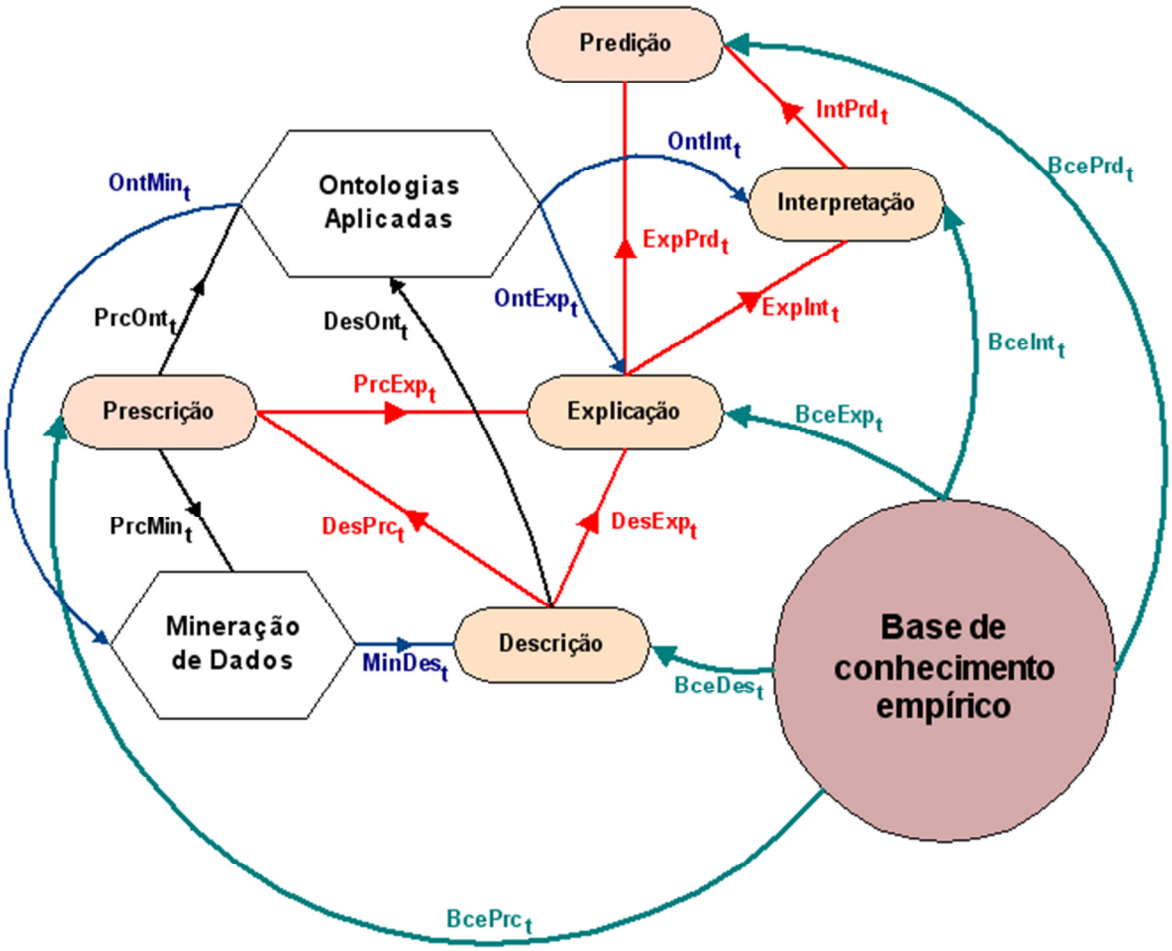

FIGURA 3.5 Suporte de ontologias e mineração de dados às tarefas que se articulam dinâmica e dialeticamente para a constituição e no uso da base de conhecimento empírico.

A consideração dessas tarefas permitirá mostrar que os métodos de constituição de bases empíricas de conhecimento, no âmbito das ciências humanas, possuem peculiaridades e especificidades próprias que os distinguem de métodos voltados às ciências naturais. Mais claramente, nas ciências naturais – em seus grandes ramos de ciências físicas, químicas e biológicas – as noções de objetividade, explicação e necessidade, são essenciais. Já, metodologias específicas de ciências humanas – tais como economia, contabilidade, administração, política e direito – por sua vez,

exigem as noções de subjetividade, interpretação e possibilidade. A ênfase na tarefa de explicação é típica das ciências naturais, enquanto que a ênfase na tarefa de interpretação é típica das ciências humanas (DOMINGUES, 2004, p. 103 a 135, em especial, p. 116 e p. 134-135).

A demanda de suporte de ontologias às tarefas de explicação e interpretação leva-nos à principal questão a ser desenvolvida nessa seção: o de que metodologias e métodos em ontologias aplicadas fundam-se principalmente em métodos de ciências naturais e visam a dar suporte a tarefas de explicação. Como tal, possuem alcances limitados em seus usos em organizações onde prevalecem fenômenos humano-sociais que demandam suporte à tarefa de interpretação. Portanto, ontologias aplicadas deveriam estender-se de modo a suportar tarefas de interpretação e, desse modo, dar o suporte adequado aos problemas atinentes a domínios das ciências humanas.

Na FIGURA 3.5, os vínculos da base de conhecimento empírico com as tarefas representam o fluxo de distintos subconjuntos da base de conhecimento empírico relevantes para cada uma das tarefas. Como essa base de conhecimento e, consequentemente, estes fluxos são dinâmicos, de conteúdo variável, associamos o índice *t* a cada um desses fluxos ($BceDes_t$, $BceExp_t$, $BcePrc_t$, $BcePrd_t$, e $BceInt_t$), e a cada um dos demais fluxos da figura.

No nosso caso, a base de conhecimento empírico será constituída por registros de transações envolvendo serviços prestados por empresas sediadas fora de Belo Horizonte, para empresas sediadas em Belo horizonte. O conteúdo dessa base de conhecimento muda ao longo do tempo. Uma descrição feita com um conteúdo anterior, assim como as explicações e interpretações realizadas a partir desse conteúdo, irá auxiliar na definição de prescrições e predições no tratamento de novos conteúdos da mesma base de conhecimento. Por exemplo, registros que levem a descrições que revelem novidade, ou inesperabilidade podem ser úteis para alimentar novas prescrições na avaliação futura da mesma base de conhecimento com novos registros.

Exemplificando de modo mais específico, em nossa ontologia incorporamos uma classe que prescreve que determinados municípios são paraísos fiscais. Em uma

nova mineração de dados, descobrimos mediante análise que um novo município também é paraíso fiscal, e esse novo município passará a constituir a classe de municípios paraíso fiscal para futuras minerações de dados. Isto é, esse município passará a ser prescrito como município paraíso fiscal.

Do mesmo modo, descrições obtidas dessa base de conhecimento empírico poderão revelar tendências. O conhecimento associado à nova essa tendência poderá ser incorporado à ontologia, e irá auxiliar a tarefa de mineração de dados e análise dos resultados.

Buscando esclarecer os fluxos associados à tarefa de prescrição, descrição, explicação e interpretação, podemos discriminar o seguinte:

$PrcOnt_t$: quando estivermos modelando os dados, ponderando como os diferentes atributos influenciam na constituição de um atributo composto nível de suspeição, estaremos prescrevendo a forma que se dará a mineração de dados. Por exemplo, podemos prescrever que determinados municípios do Colar Metropolitano possuem determinado peso na influência de um nível de suspeição; ou determinada faixa de valor de um atributo receberá o rótulo linguístico de “Irrisório”

$PrcExp_t$: quando estivermos descrevendo como os critérios das métricas e parâmetros de sumarização nos auxiliam na explicação, estaremos prescrevendo critérios potencialmente úteis para uma possível explicação e interpretação dos resultados.

$PrcMin_t$: quando estivermos determinando os parâmetros de configuração da mineração de dados (suporte, confiança, restrições em faixa de valores de atributos, seleção de atributos no antecedente, e no consequente) estaremos prescrevendo com base em experiências anteriores os critérios básicos de mineração de dados, potencialmente úteis para nossos objetivos.

$DesPrc_t$: descrição de coisas passadas que são norteadoras na detecção de novos eventos no processo de mineração de dados. A descrição que determinado município faz parte da região, ou do colar metropolitano, associado ao alto índice de empresas com domicílio simulado já detectado, leva à prescrição de que estes as transações oriundas destes municípios influenciam o nível de suspeição das transações relativas aos serviços prestados.

$DesExp_t$: descrições relevantes e úteis para a tarefa de explicação. a proximidade de municípios a BH, a afinidade dos donos de imóveis com os sócios das empresas, a debilidade econômica do município, a debilidade moral dos administradores públicos e contadores.

$DesOnt_t$: descrições de classes (atributos e propriedades) úteis para serem incorporados à ontologia e serem úteis nas tarefas de classificação de categorias na fase de pré mineração, e na fase de pós mineração

$ExpPrd_t$: explicações úteis na tarefa de predição (por exemplo, a explicação de que a fonte de simulação de domicílio fiscal, é incentivada por empresas de contabilidade, leva à predição que futuros clientes dessa contabilidade adotarão o mesmo comportamento (de simulação)

$ExpInt_t$ e $IntPrd_t$: a explicação de que empresas de contabilidade são sede de domicílio fiscal simulado, leva à interpretação de que seus contadores oferecem esse tipo de serviço a fim de captarem clientes. Ou ainda, transações suspeitas oriundas de cidades pertencentes ao Colar e Região Metropolitana de Belo Horizonte, sugerem que casas de campo, sítios, ou imóveis de lazer de contribuintes e seus familiares sejam utilizados para simular domicílio fiscal de contribuintes efetivamente domiciliados em Belo Horizonte.

Já no que diz respeito a ontologias, no suporte à mineração de dados, elas incorporam classes, atributos, axiomas de classes, e mecanismos de inferência úteis para a classificação (categorização) de atributos a serem minerados, ou classificação de regras resultantes da mineração, úteis para explicar e interpretar. Assim, elas podem ser vistas tanto como ferramentas para incorporar e classificar as instâncias que são o ponto de partida para a descrição (**$DescOnt_t$**) da base de conhecimento. Também representam o conhecimento previamente adquirido, necessário para o processamento da mineração (**$OntMin_t$**). Por fim, representam o conhecimento de referência útil à análise dos padrões descritivos obtidos no processo de mineração (**$MinDes_t$**) nas tarefas de explicação e interpretação (**$OntExp_t$** e **$OntInt_t$**).

3.4.1 A Tensão entre a Objetividade Inalienável *Versus* a Subjetividade Inexorável

O grande problema ao se buscar nas ciências humanas referências metodológicas para o desenvolvimento de abordagens ontológicas específicas para domínios humano sociais é que, de imediato, deparamo-nos com as questões associadas à sua própria fundamentação.

A questão do método em ciências humanas é uma questão ainda em aberto, seus padrões de cientificidade ainda encontram-se em busca de uma consolidação a partir de distintas vias[41], e seus fenômenos resistem às exigências da objetividade[42]. A fundamentação das ciências humanas apresenta-se inconclusa. Domingues (2004, p. 85), citando Granger (1994), expressa essas dificuldades:

> "os saberes sociológicos ou psicológicos, econômicos ou linguísticos não podem pretender, em seu estado presente e passado, ter a solidez e a fecundidade dos saberes físico-químicos, ou até biológicos"
>
> "tais fenômenos resistem "à sua transformação simples em objetos, ou seja, em esquemas abstratos lógica e matematicamente manipuláveis"".
>
> "as características cientificamente negativas dos fatos humanos, e em especial seus elementos de liberdade e de imprevisibilidade". Granger, 1994, p. 85.

No cerne dos dilemas do método, na busca inalienável da objetividade, encontra-se a dificuldade de objetivação da significação dos fatos sociais e de lidar com a subjetividade inerente aos fenômenos humano-sociais.

> A exigência de objetividade, por sua vez, se impõe, antes de mais nada, pela necessidade de o sujeito cognoscente adequar-se ao real, submeter-se a ele e ater-se aos indícios que ele lhe fornece, devendo para tanto livrar-se das paixões que estorvam, dos preconceitos que cegam e da vontade que precipita (objetividade do cientista – questão de atitude do sujeito). Impõe-se, por fim, pela necessidade de a ciência ater-se em suas formulações ao real e apoiar-se em indícios do próprio real em seus vaticínios, sob pena de, ao não o fazer, perder o real, evadir-se em abstrações e tomar uma quimera pela realidade (objetividade da ciência – questão de parâmetro aplicado à coisa). (DOMINGUES, 2004, p. 139).

---

[41] Vide, "Padrões de cientificidade nas ciências humanas – Formas de explicação (compreensão) da realidade humano-social". Domingues, 2004, p. 85-102.

[42] Vide, "As ciências humanas e a exigência de objetividade: as vias de Durkheim, Marx, Freud e Weber" (DOMINGUES, 2004, p. 137-166).

Aqui encontramos o grande ponto de tensão. Se a objetividade é a meta inalienável, a subjetividade é elemento inexorável que se imiscui nos fatos, nos fenômenos, e na práxis das ciências humanas.

À parte isso, outros aspectos das tarefas relacionadas apresentam particularidades e destacam as peculiaridades das ciências humanas.

3.4.2 Descrição

As maiores dificuldades envolvendo essa tarefa são que um mesmo fato, ou fenômeno pode ser descrito de diferentes maneiras e a descrição nunca será completa. A descrição depende da granularidade a ser observada (aspecto selecionado e nível de detalhamento), e da intencionalidade do observador (a forma em que sua consciência direciona seu "olhar"). Granularidade e intencionalidade estão relacionadas, pois, a intencionalidade é um dos condicionantes da granularidade. Outros condicionantes são o contexto, os recursos técnicos e tecnológicos disponíveis, etc.

Tais dificuldades não são exclusivas das ciências humanas, elas também pertencem às ciências naturais. Domingues (2004, p. 104) dá um interessante exemplo de como o som pode ser diferentemente descrito por um físico, um fisiólogo e um linguista.

As dificuldades peculiares das ciências humanas é que, além dos aspectos objetivos dos comportamentos humanos, a descrição deverá debruçar-se sobre um conjunto de elementos subjetivos (intenções, sentimentos, consciência, valores e fins visados pelos agentes humanos) (DOMINGUES, 2004, p. 107). Se as ciências naturais debruçam-se sobre as causas e a descrição e explicação causal dos fatos e fenômenos, as ciências humanas estendem este leque para buscar também a descrição e explicação das motivações, razões, crenças e esperanças.

Para tentar deixar isto mais claro, contrapomos as seguintes passagens. (1) Causalidade como típica relação de fenômenos naturais objetivos:

> No exemplo filosófico adotado à exaustão (e a recorrências desses mesmos exemplos na filosofia deveria despertar nossas suspeitas), a bola de bilhar *A* cumpre seu inevitável trajeto através do pano verde até colidir com a bola *B*, ponto em que *B* começa a mover-se e A imobiliza-se. Essa pequena

> cena, recontada vezes infindas, é o paradigma da causalidade: o evento de *A* colidindo com *B* causou o evento de *B* entrar em movimento. E, segundo a visão tradicional, quando testemunhamos essa cena não vemos realmente, nem observamos de forma alguma, quaisquer conexões causais entre o primeiro e o segundo evento. O que de fato observamos é um evento seguido de outro. Podemos, contudo observar a repetição de pares semelhantes de eventos e essa repetição constante autoriza-nos a dizer que os dois membros dos pares estão causalmente relacionados, mesmo que não possamos observar relação causal alguma. (SEARLE, *Intencionalidade*, 2002, p. 156)

Nessa passagem, ao mesmo tempo em que Searle aponta o caso típico da causalidade nos fenômenos naturais (no caso, físico-químicos), ele a coloca como uma noção insuficiente, mesmo nesse domínio.

A segunda passagem é de Elizabeth Anscombe (*Intensions,* 1957) citada pelo Prof. Ivan Domingues (2004, p. 107). Para nossos propósitos, ela ilustra (2) causalidade objetiva como insuficiente relação para descrever, explicar e interpretar fenômenos humano-sociais:

> Um homem está bombeando água potável para a cisterna de um edifício. Alguém descobriu uma maneira de contaminar sistematicamente o manancial com um veneno cumulativo mortal, cujos efeitos não são percebidos até que resultem incuráveis. O edifício é habitado por um pequeno grupo de dirigentes políticos e suas famílias, o qual controla uma grande nação; estão co-implicados no extermínio de judeus e possivelmente planejam uma guerra mundial. O homem que contaminou o manancial supõe que se essas pessoas forem destruídas indivíduos honestos assumirão o poder e governarão apropriadamente, ou inclusive poderão instaurar o reino dos céus na terra e assegurarão o bem-estar de todo o povo. Essa pessoa tem confessado suas suposições, além do assunto do veneno, ao homem que está bombeando. Demais, a morte dos habitantes do edifício acarretará muitos outros efeitos; entre eles, certo número de pessoas, desconhecidas para estes dois homens, receberão heranças (das vítimas), que não sabem de nada. (Anscombe, *Intensions*, 1957, *apud* DOMINGUES, 2004, p. 107)

A partir dessa situação hipotética, Anscombe tece uma série de descrições possíveis para todo o processo. Entretanto, toda e qualquer descrição interessante exige que se inclua o homem como sujeito. E para se chegar a uma descrição que possa revelar algo de interessante, Anscombe propõe que, ao invés de nos limitarmos ao "quê", perguntemos "por que" o sujeito se comporta de tal maneira (DOMINGUES, 2004, p. 108). Além disso, uma análise só seria mais aprofundada se ela envolvesse a correlação dos agentes, das intenções, dos atos, dos segmentos temporais e a consideração da realidade.

Ivan Domingues, indo além das conclusões da autora, segundo ele, restritas por seu empirismo, propõe a extensão do processo descritivo à interpretação de seus elementos descritivos, isto é, dos agentes, ações, atos, intenções, valores e fins que a integram (DOMINGUES, 2004, p. 113).

Aqui Ivan Domingues apresenta uma peculiaridade clara das ciências humanas diante das ciências naturais. A tarefa de descrição, mesmo quando considera os atos intencionais, não lança seus elementos descritivos apenas às tarefas de explicação, em especial com base nos aspectos causais, como veremos a seguir. A descrição dever-se-á, necessariamente, lançar seus elementos à tarefa interpretativa que, em ciências humanas, raramente já estará encapsulada pela tarefa explicativa.

3.4.3 Explicação

Com base nos elementos descritivos há distintas formas de explicação possíveis: genéticas (de origem), estruturais, funcionais, finais e, dentre outras, causais. **A explicação causal é considerada a forma de explicação por excelência** (DOMINGUES, 2004, p. 116). Entretanto, a questão da causalidade, assim como a questão da vaguidade discutida no Capítulo 2, é assunto de intensos debates nas últimas décadas e que apresenta uma série de dificuldades.

Russell (1912, p. 1)[43] chega a comparar a lei da causalidade com a monarquia: ambas seriam relíquias que somente sobrevivem até os dias atuais porque se supõe erroneamente que não causam danos. Somente em um sistema isolado seria possível estabelecer conexões causais, mas tal sistema não existe em realidade. Daí, se a noção de causalidade envolve uma falácia original e essencial, levando-nos a novas falácias, para quê recorrer à noção de causalidade?

Nesse aspecto, mais uma vez, a questão da causalidade aproxima-se da questão da vaguidade. Assim como ela, causalidade tem sido útil para fins de análise e, como a vaguidade, ao invés de abandonada, ressurge revigorada nas últimas décadas acompanhadas com distintas interpretações e métodos para tratá-la.

---

[43] Russel, Bertrand. *On the notion of cause*. 1912.

De fato, as questões envolvendo a causação tornam-se extremamente complexas. Além da dificuldade já há muito identificada de se distinguir entre “lei da natureza” e regularidades acidentais; de se distinguir a causa de seus efeitos, da impossibilidade de se fazer um número infinito de observações que sustentasse uma indução; ou de se fazer a regressão ao infinito (à causa primeira – Deus), e da dificuldade em lidar com possibilidades físicas ainda não realizadas (um corvo branco, por exemplo, considerando-se que o cisne negro, antes “impossível”, já foi encontrado em 1697 na Austrália), há uma série de outras dificuldades agudas.

Como apelaremos à indução diante dos eventos singulares? Além disso, uma mesma coisa pode ser causa de efeitos contrários; pode-se identificar uma causalidade recíproca ou circular; pode ocorrer dependência mútua, ação, ou influência de causa e efeito e, ainda, causalidade frequentemente é confundida com condicional lógica. Há as causas indiretas, concorrentes, espúrias[44], bidirecionais[45], retrocedentes. Há o problema dos epifenômenos, onde uma determinada causa pode ter dois ou mais efeitos e o problema da preempção, onde duas, ou mais causas coexistem, mas somente uma delas é eficaz. Também, é muito mais fácil trilhar o caminho de volta dos efeitos para suas causas (explicação) do que prever ou explicar eventos futuros.

Tal tema, a despeito de sua extrema importância não poderá ser realmente elucidado aqui. Primeiro, por exigir toda uma monografia para seu desenvolvimento. Segundo, para não desviarmo-nos de nossos objetivos principais. Em nosso percurso, prosseguiremos com a direção dada pelo Prof. Ivan Domingues. Das causas, ele ressalta três aspectos importantes (DOMINGUES, 2004, p. 117):

1. Identifica, em consonância com o que já dissemos, que causas não podem ser estabelecidas com base de conhecimento empírico e nem com base na razão (DOMINGUES, 2004, p. 117).
2. Em ciências humanas ocorre a dissociação da causalidade da lei. Citando Weber, ele informa que a causa dissocia-se da noção de necessidade, passando a ser tratada como uma possibilidade (possibilidade objetiva).

---

[44] Duas variáveis estão relacionadas porque elas compartilham uma causa comum, mas não porque uma delas causa a outra.
[45] Uma variável influencia a outra.

3. Identifica que a noção de causa enfatiza a dissociação da pergunta do "como" da pergunta do "por que". Isto prejudica a ênfase interpretativa que, segundo o autor, deve predominar em ciências humanas.

Além disso, se um mesmo fenômeno, ou fato pode levar a diferentes descrições, o mesmo ocorre com as explicações causais: a mesma base factual é compatível com mais de uma explicação causal. É comum respostas discreparem profundamente, cada um apontando a sua causa, sem chegar a um denominador comum[46]. A mesma base factual é compatível com mais de uma explicação causal, ficando a escolha da causa a depender de um contexto mais amplo que se decide em outro nível de análise.

Tal particularidade pode-se atribuir à insuficiência do princípio de análise, à deficiência do próprio analista, à confusão da causa superficial com a causa profunda, à incapacidade de combinar a pluralidade de causas para dar uma explicação total. Conclui Domingues (2004, p. 119) que o importante é a análise causal depender da consideração de um contexto mais amplo, que se decide em outro nível de análise: a interpretação (compreensão).

3.4.4 Interpretação (Compreensão)

Muitos pretendem que a interpretação já se decide no nível da explicação, e não é senão um de seus aspectos. Entretanto, para Domingues, mesmo reconhecendo que muitas vezes explicação e interpretação (compreensão) se encavalam e atuam num mesmo nível, é o caso de distinguir uma da outra, considerando que a explicação incide sobre os fatos, ou coisas. Já a interpretação envolve a significação, o sentido deles. Portanto, a interpretação irá introduzir as unidades significativas de análise, como as hipóteses, os modelos (tipos ideais), as postulações de sentido, e assim por diante (DOMINGUES, 2004, p. 119-120).

---

[46] Um dos exemplos dado por Domingues (2004, p. 118-119), onde cada analista aponta sua causa sem chegar a um denominador comum, é a decadência e ruína do Império Romano. Gibbon (triunfo do cristianismo e barbárie), Maquiavel (perda da virtú e da liberdade com o fim da República), Montesquieu (lei do ciclo/ambição desmesurada), Marx (luta de classes), Max Weber (declínio do escravismo e retorno à economia natural), Seeck (eliminação da elite), Kaphahan (degenerescência física), T. Frank (degenerescência racial), Huntington (seca prolongada), Liebig (degradação do solo) e Toynbee (invasões bárbaras conjugadas com a insatisfação das massas).

No contexto da discussão dos fundamentos das ciências humanas, Domingues conclui que dentre as tarefas consideradas, a candidata com maior chance de ter a primazia no método e de conduzir a análise é a tarefa interpretativa.

> Acreditamos que, entre os elementos envolvidos na abordagem dos fenômenos humano-sociais em seus aspectos descritivo, explicativo, interpretativo (compreensivo) e prescritivo, o candidato com maior chance de ter a primazia no método e de conduzir a análise, tendo por fio o sentido, é o elemento interpretativo. O argumento é ele ser de todos, o mais abrangente e o único auto-referente ou auto-aplicável. Podemos interpretar as prescrições, interpretar as explicações, interpretar as descrições e interpretar as próprias interpretações. Não podemos, diretamente, fazendo economia do sentido e da interpretação, prescrever, explicar e descrever interpretações, nem prescrever prescrições, explicar explicações e descrever descrições: simplesmente, os elementos sobre os quais operam – os dados, as normas, as relações: causais, funcionais, finalísticas – pressupõem o sentido, que é coextensivo à interpretação e é por ela decifrado, ao perguntar pelo sentido da norma, da relação, do dado e do próprio sentido, em suas mais variadas situações: pregnância do sentido, colapso do sentido, falta de sentido, etc. (DOMINGUES, 2004, p. 135-135)

Para esclarecermos e não perdermos o fio da meada, da discussão sobre as tarefas envolvendo a constituição e uso de uma base de conhecimentos, para nossos propósitos, é importante reter o seguinte.

TABELA 3.1 Tarefas relacionadas à base de conhecimento empírico e as peculiaridades das ciências humanas.

Tarefa	Foca no	Faz	Peculiaridades das Ciências Humanas
Descrição	"quê" "como"	Indaga o que são os fenômenos e como se comportam. *modus essendi* e *modus operandi*	Demanda a descrição dos atos intencionais que se iluminam à luz do "por que"
Explicação	"para quê"	Indaga como os fenômenos se comportam à luz de uma (aspectos): • origem (genéticas) • estrutura (estruturais) • função (funcionais) • fim (teleológicos) • **causa (causais)** *modus operandi*	Ênfase na possibilidade ao invés da necessidade. Indaga como os fenômenos se comportam, à luz de: • motivações, • razões, • esperanças e • crenças.
Interpretação	"por quê" "para quê" "qual o sentido"	Modo como significamos os fenômenos, bem como a forma como eles nos interpelam ou nos afetam *modus significandi*	Ênfase no significado, dando maior margem à teoria, ao subjetivo e ao intersubjetivo.

5.4.5. Intersubjetividade

Já temos estabelecido que a consideração da subjetividade, a despeito de se constituir em enorme problema e desafio, é crucial para o processo de análise e, consequentemente, para a representação do conhecimento em domínios de ciências humanas.

Indo ainda mais além, deve-se cogitar que tal consideração ainda deve ser ampliada. Como a própria Anscombe comenta, e é citado por Domingues (2004, p. 115), por vezes temos que considerar não só a subjetividade associada a um indivíduo, mas a coletividades inteiras. Entretanto, desafio ainda maior se instaura pela consideração não dessa coletividade que, a despeito de ampliada, ainda pode ser individualizada. Desafio ainda maior, é a consideração da inter-subjetividade, ou seja, a consideração da comunicação das consciências individuais.

No que diz respeito a metodologias de construção de ontologias, a intersubjetividade já vem sendo considerada e tratada. Observa-se a atenção com a intersubjetividade em dois aspectos. Primeiro, há a preocupação de que o conteúdo das ontologias expresse um conhecimento consensual. Deverá haver uma concordância com a especificação dos conceitos, relações, atributos e axiomas que a ontologia fornece. Segundo, é a viabilização do primeiro aspecto. Isto é, como construir uma ontologia com conteúdo consensual, sintonizando e consolidando os esforços conjuntos de diferentes especialistas distribuídos em diversos lugares. Para tal, foram desenvolvidas metodologias e protocolos. As principais referências são o CO4 e a $(KA)^2$. Outras iniciativas nesse sentido podem ser vislumbradas em Aschoff *et al.*(2004) e Kozaki *et al.*(2007).

O CO4 é um protocolo para alcançar consenso entre várias bases de conhecimento. Seu objetivo é que pessoas possam discutir e chegar a um acordo sobre o conhecimento introduzido nas bases de conhecimento do sistema. As bases de conhecimento são organizadas em árvore. As folhas são chamadas de usuários das bases de conhecimento, e os nodos intermediários, grupos de bases de conhecimento. De um lado, os usuários das bases de conhecimento não têm obrigatoriamente conhecimento consensual. De outro lado, cada grupo representa o conhecimento consensual entre seus filhos (chamados subscritores de bases de

conhecimento). Uma base de conhecimento pode estar subscrita a um grupo somente (CORCHO, LÓPEZ-FERNÁNDEZ e PÉREZ, 2001, p. 32)[47].

Já no $(KA)^2$, *Knowledge Annotation Intiative of the Knowledge Acquisition Community*, para facilitar o processo de construir a ontologia, há um agente de coordenação da ontologia que distribui modelos aos agentes coordenados (agentes ontópicos). É usado email em suas comunicações internas e também para enviar seus resultados para os agentes de coordenação (especialistas em diferentes tópicos). A ontologia é gerada a partir do conhecimento introduzido via modelo. Uma vez que agentes de coordenação da ontologia, obtém todas as porções das ontologias a partir dos agentes ontópicos, eles as integram. Neste caso, o processo de integração é facilitado em função dos agentes ontópicos usem o mesmo modelo (CORCHO, LÓPES-FERNÁNDEZ e PÉREZ, 2001, p. 32).

A despeito dessa atenção metodológica com o processo subjetivo e intersubjetivo de obtenção do consenso sobre o que se conhece, isso representa apenas um lado da questão, como foi dito, a incorporação de conhecimentos afetados por diferentes níveis de subjetividade na ontologia. Outro lado da questão é a adequação e uso de ontologias na etapa de decifração do sentido intersubjetivo dos atos humanos na tarefa da interpretação.

Haverá situações (fatos e coisas) em que devem ser consideradas não somente a subjetividade do observador, mas as subjetividades de distintos observadores que interagem entre si, ou não. Isso demandará a busca por métodos e ferramentas visando à modelagem e busca de opiniões qualificadas e majoritariamente consensuais. A verdade a ser considerada advirá da convicção de verdade recolhida pela análise dentro da comunidade.

Em ciência da informação e ciência da computação, a demanda pela consideração da intersubjetividade na análise da informação é eventual, a não ser, como veremos na seção 5.5.2, se ela for tratada dentro do paradigma de sistemas difusos, onde tal

---

[47] Para o CO4 (Corcho *et al*., 2001), citam as seguintes referências:

- EUZENAT, J. Corporative memory through cooperative creation of knowledge bses and hyper-documents. 1996, *Proceedings 10th KAW,* Banff (Canada).
- EUZENAT, J. Building consensual knowledge bases: context and architecture. *In*: MARS, N. (Ed.). *Building and sharing large knowledge bases*. IOS Press. Amsterdam (Netherlands). 1995, p. 143-155.

tratamento é quase inerente. Entretanto, tal demanda é agudamente sentida nas comunidades de inteligência. Nessas comunidades, a consideração da intersubjetividade é regra e não exceção.

Como dogma metodológico admite-se que as análises de distintos analistas devem ser confrontadas, e deve-se buscar metódica e sistematicamente uma análise conjunta que traduza a convergência das análises individuais, e ainda, considerando-se o contexto, o mérito e níveis de consenso.

Desse modo, ferramentas e métodos de análise de inteligência apresentam-se como suportes à atividade de interpretação a partir da consideração da intersubjetividade. Dentre os métodos mais aceitos pelas comunidades de inteligência, podemos citar o Método Delphi e o Método das Hipóteses Concorrentes.

O suporte de ontologias ao processo de mineração de dados pode, por via intermediária, dar suporte aos métodos das comunidades de inteligência citados acima e aos métodos de análise de consenso desenvolvidos no paradigma de sistemas difusos que serão citados na seção 3.5, a seguir.

3.5 As Peculiaridades das Ciências Humanas Demandam Abordagens Alternativas em Ontologias

Há pouco estipulamos as peculiaridades dos domínios das ciências humanas em relação ao domínio em ciências naturais: a primazia da interpretação sobre a explicação, a importância e necessidade de se considerar a subjetividade em conjunto com a objetividade, e a flexibilização da noção de necessidade (lei) para a noção de possibilidade.

Em vista disso, e considerando que nosso objetivo é agregar conhecimento de domínio, via ontologias, para dar suporte à mineração de dados em contextos organizacionais, onde prevalecem aspectos associados a ciências humanas, impõem-se a nós duas questões:

> 1) Estariam as abordagens e representações propiciadas pelas ontologias de fundamentação, ou de alto nível – que servem de referências para a

construção de novas ontologias – preparadas para lidar com peculiaridades atinentes ao domínio de ciências humanas?

2) Como ontologias poderiam incorporar as noções de subjetividade e possibilidade e, desse modo, dar um suporte mais adequado à tarefa de interpretação e, consequentemente, aos domínios de ciências humanas, onde prevalecem aspectos humano-sociais?

Tais questões exigem um exame minucioso que foge ao escopo de nossos objetivos principais. Portanto, não examinaremos essas questões exaustivamente, mas teceremos alguns comentários, indicações e avaliações iniciais que poderão nortear avaliações mais aprofundadas no futuro, e que, de imediato, já sejam suficientes para fundamentar ainda mais a estratégia de solução adotada ao longo de toda essa tese e que se alinha à abordagem de viés humano cêntrica, conforme preconizada pela computação suave, granular, em especial, a que adota metodologias, noções e técnicas de sistemas difusos.

Compartilhamos com o Prof. Domingues a visão de que em ciências humanas o conhecimento se faz e, consequentemente, se representa, a partir da inter-relação do sujeito e do objeto, em um processo circular, a despeito das ilusões do objetivismo puro.

> (…) “se o fundamento do conhecimento de si mesmo não pode ser encontrado do lado do objeto, como imaginava a ciência, nem do lado do sujeito, como queria a filosofia (a filosofia do sujeito), a saída será buscá-lo na interrelação dos dois. É então que o descobriremos, se não queremos encapsular o si no objeto, nem deixá-lo perder-se nos recessos do sujeito, a dupla dependência do conhecimento do sujeito e do objeto, o objeto pivoteado pelo real e arrancando o sujeito de si mesmo, o sujeito pivoteado pelo eu e mergulhando o objeto dentro de si mesmo ou na consciência de si. É então que descobriremos a estrutura reduplicada do processo de conhecimento, tendo por fundamento a estrutura reduplicada da reflexão, pivoteada pelo sujeito e pelo objeto, quando ficará evidenciado que o sujeito só pode ser apreendido no processo cognitivo como objeto ou ao modo de um objeto, enquanto o objeto só é apreensível segundo as condições ou as estruturas do sujeito. É então que descobriremos em epistemologia das ciências humanas aquilo que já sabíamos em gnosiologia ou teoria do conhecimento: a estrutura invencivelmente circular do processo cognitivo, e, como a serpente de Valéry, a epistemologia termina por engolir a própria cauda.” (DOMINGUES, 2004, p. 650).

Uma das vias de acesso ao sujeito parece-nos ser, conforme já comentada na seção anterior, o estudo dos atos intencionais, sua estrutura, suas categorias. No exemplo

citado de Anscombe, de todas as descrições e explicações possíveis, as que permitiam que o desvelamento de um ato criminoso, eram, justamente, as que se baseavam em intenções do sujeito guiadas pelo "por que" e "para quê". Ontologias de fundamentação que não propiciem condições para este tipo de representação parecem-nos contraindicadas para as ciências humanas.

A seguir comentaremos brevemente da emergência da abordagem por computação suave e granular e que vêm em reforço ao uso de ontologias no suporte a mineração de dados, em especial, em domínio sócio-humanos.

A insuficiência das soluções focadas exclusivamente em aspectos de detalhes técnicos tais como eficiência de algoritmos e módulos de programas, vem levando ao surgimento de abordagens que preconizam a importância da consideração de aspectos de usabilidade, conhecimento do domínio e capacidade de interpretação de resultados. No âmbito da mineração de dados e ontologias, citamos o surgimento da mineração de dados dirigida ao conhecimento do domínio ("*Domain-Driven Data Mining* - D^3M), e a crescente importância em se considerar aspectos cognitivos humanos, consideração da linguagem natural, do senso comum da subjetividade e intersubjetividade em domínios humano sociais.

Nesse contexto, a computação suave (*soft computing*) e computação granular, comentadas na Seção 2.5, vêm se associar a uma abordagem humano cêntrica flexível e com foco na interação do homem com a máquina, no desenvolvimento de projetos centrados no usuário. Isso, de modo a permitir que os sistemas desenvolvidos exibam flexibilidade, significativas habilidades de comunicação, e um adequado nível de adaptatividade.

Considerando-se essas três abordagens convergentes – computação suave, granular e humano cêntrica[48] – destacam-se as soluções surgidas no contexto de sistemas difusos[49], redes neurais[50] e algoritmos genéticos[51]. Em Mitra (2002) temos uma visão panorâmica da literatura disponível sobre mineração de dados

[48] Os trabalhos de um recente *workshop* dessa área podem ser verificados em: http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-696/
[49] Modelam incerteza, imprecisão, ambiguidade, vaguidade em dados.
[50] Redes neurais são robustas e exibem boas capacidades de aprendizagem e generalização.
[51] Algoritmos genéticos fornecem algoritmos de busca eficientes.

fornecendo uma categorização dessas ferramentas, utilidade das metodologias, seu uso na generalização de padrões, tratamento de dados, incompletos, ou com ruídos, e seu uso na interação humana e soluções híbridas.

Em especial, os sistemas difusos preocupam-se com a natureza amigável ao usuário dos sistemas resultantes manifesto através de um alto nível de cuidado com o contexto e obtenção de retorno que seja relevante. Além disso, enfatiza a formação de um balanceamento adequado entre acuidade e transparência, no sentido de capacidade de interpretação do que foi realizado, e, ainda, a incorporação de projetos amigáveis de mecanismos de desenvolvimento de sistemas facilitando uma agregação eficiente de várias fontes de informação disponível sendo apresentadas em vários níveis diferentes de abstração integração de conhecimento de domínio e dados numéricos experimentais (PEDRYCS e GOMIDE, 2007, p. xvii).

Sistemas difusos constituem o paradigma de computação suave mais antigo e reportado e podem ser integrados com outras ferramentas de computação suave levando à geração de sistemas mais poderosos, inteligentes e eficientes com aplicações em reconhecimento de padrões, processamento de imagens, e inteligência de máquina (MITRA e PAL, 2005; HÜLLERMEIER, 2005 e PETRATUS, 2006).

Nesse contexto, confirmando e reforçando a emergência dessas abordagens, vem surgindo estudos que aplicam tecnologias de sistemas difusos a mineração de dados e ontologias.

Nos próximos capítulos abordaremos em detalhes o uso do suporte de ontologias difusas no processo de mineração de dados por regras de associação em ambiente organizacional, em domínios sócio-humanos e, poderemos vislumbrar na prática algumas soluções propiciadas pelas abordagens citadas.

3.6 Conclusão

As poderosas ferramentas computacionais de mineração de dados, por si só, não garantem o sucesso de seu uso em ambientes organizacionais. Além disso, esses ambientes, normalmente, apresentam problemas surgidos em domínios sócio-

humanos que possuem peculiaridades e exigem abordagens que levem em consideração suas particularidades. De um modo geral, essas peculiaridades e particularidades demandam a consideração de noções distintas e, por vezes, antagônicas aos domínios associados às ciências naturais. Ao invés do caráter de lei associada à noção de necessidade, há a noção de possibilidade; ao invés da ênfase na explicação, há a ênfase na interpretação, e ao invés do expurgo da subjetividade, há a sua concomitante consideração, não em contraposição, mas em cooperação com a objetividade científica.

O surgimento de novas abordagens traz novas metodologias, métodos, técnicas e ferramental lógico-matemático que permitem tratar essas peculiaridades dos domínios sócio-humanos.

4 Ontologias Difusas

4.1 Introdução

Neste capítulo iremos tratar das ontologias difusas que são extensões das ontologias clássicas, com conjuntos rigorosos e lógica clássica. No Capítulo 2, ao discutirmos o fenômeno da vaguidade, foi amplamente discutida a inadequação da lógica clássica para lidar com conhecimento vago, ambíguo, impreciso, incerto, inexato, ou indeterminado, e a emergência da lógica difusa para lidar com esses fenômenos.

Nas ontologias difusas a extensão das ontologias clássicas se faz pela incorporação de recursos de sistemas difusos na definição de classes, propriedades, instâncias e diversos tipos de relacionamentos. Conjuntos difusos são incorporados mediante a construção de metaclasses e classes difusas. Mecanismos de inferência difusos são inseridos, ou mediante o uso de raciocinadores difusos, que incorporam linguagens e famílias de linguagens difusas, ou mediante a utilização de linguagens de regras que permitam o desenvolvimento de regras que implementam diretamente os mecanismos de inferência difusa.

Nesse âmbito, mostraremos como trabalhos correlatos vêm introduzindo a definição de ontologias difusas, a adaptação dos elementos ontológicos (classes, propriedades e instâncias) para incorporar características de sistemas difusos, e o desenvolvimento de extensões de linguagens baseadas em lógica descritiva difusa.

4.2 Formalização de Ontologias Difusas

O aprimoramento de estruturas ontológicas existentes para acomodar incerteza e vaguidade em descrições de conceitos e relações entre conceitos, além de outras aplicações, pode ser obtido evoluindo-se de uma estrutura ontológica clássica para uma estrutura ontológica difusa.

Silvia Calegari e Davide Ciucci (CALEGARI e CIUCCI, 2008) apresentam uma formalização simples de ontologia difusa que já permite vislumbrar o grau de pertinência regendo propriedades que relacionam conceitos a instâncias ontológicas e associações entre conceitos e instâncias.

> **Definição 1**. Uma ontologia difusa é uma ontologia estendida com valores difusos que são atribuídos mediante duas funções:
> g: (*Conceitos*∪*Instâncias*) × (*Propriedades*∪ Prop_val) → [0,1] e
> h: *Conceitos*∪*Instâncias* → [0, 1]

Aprofundando essa formalização simples, os trabalhos de Abulaish e Dey (2006) e de Calegari e Sanchez (2007) e Fúller (2008) apresentam um detalhamento dessa formalização, embora apresentem algumas distinções, enfatizando um, ou outro aspecto ao estender a estrutura ontológica clássica para a difusa conforme o interesse ditado pelo domínio de aplicação visado.

No trabalho de Abulaish e Dey (2006), esses autores partem da descrição tradicional de conceitos em uma ontologia usando uma estrutura *<propriedade, valor, restrição>*, onde um conceito descritor é representado como uma relação difusa que codifica o grau de valor de uma propriedade usando uma função de pertinência difusa. A estrutura ontológica difusa proposta armazena descrições em uma estrutura *<propriedade, valor, qualificador, restrição>,* onde o valor e o qualificador são ambos definidos como um conjunto difuso. Esta estrutura permite definir o valor-propriedade de um conceito com diferentes graus de incerteza, sem mudar o paradigma de descrição do conceito. Tais descrições de conceito podem ser finalizadas como descrições indeterminadas de conceitos.

Um interessante aspecto dos atributos modelados como conjuntos difusos é que com um conjunto subjacente de valores numéricos pode-se associar diferentes conjuntos difusos de quantificadores para representar aspectos diferentes do mesmo atributo.

Por exemplo, um valor de preço único pode ser interpretado como sendo "*próximo a*" ou "*longe de*" a partir de outro valor de preço, e ao mesmo tempo pode também pode ser interpretado como *"barato"* ou *"caro".* Além do mais, restrições podem também ser aplicadas para criar novos conjuntos difusos com diferentes significados. Então modelar um atributo como um conjunto difuso permite que este atributo único contribua para expressar diferentes tipos de indeterminação na descrição de um conceito.

Além de descritores de conceitos, outras relações na ontologia como "*É-um",* "*Tem-parte",* etc. são também vinculados a uma força de associação. Formalmente, uma

ontologia difusa (Θ_F) pode ser definida como segue, como o exposto em Abulaish e Dey (2006). Uma Ontologia Difusa Θ_F é uma quádrupla da forma:

$\Theta_F = (C, P_F, R_F, M)$, onde:

- C é um conjunto de conceitos definido para o domínio.
- P_F é um conjunto de propriedades difusas do conceito. Uma propriedade $p_f \in P_F$ é definido como um quádrupla da forma $p(c, v_f, q_f, f)$, onde $c \in C$ é um conceito ontológico, *'v_f'* representa valores difusos de atributos e pode ser tanto números difusos como quantificadores difusos, *'q_f'* modela qualificadores linguísticos e são restrições, que podem controlar ou alterar a força de um valor de atributo e f é a restrição considerada sobre v_f.
- R_F é um conjunto de relações entre conceitos. Como propriedades difusas de conceito, R_F é definido como um quádruplo da forma $R_F(c,c, t, q_f)$, onde $c \in C$ é um conceito ontológico, *'t'* representa o tipo de relação, e 'q_r' modela as forças de relação e são variáveis linguísticas (rótulos), que podem representar a força de associação entre pares de conceitos $<c,c>$.
- A escolha de números difusos ou quantificadores difusos para valores é ditado pela natureza do atributo subjacente e também suas restrições consideradas. A faixa completa de valores sobre a qual um atributo pode tomar valores define o universo do discurso *M.* O universo do discurso é decomposto em uma coleção de conjuntos difusos. Cada conjunto difuso é definido sobre um domínio que sobrepõe uma parte do universo do discurso.

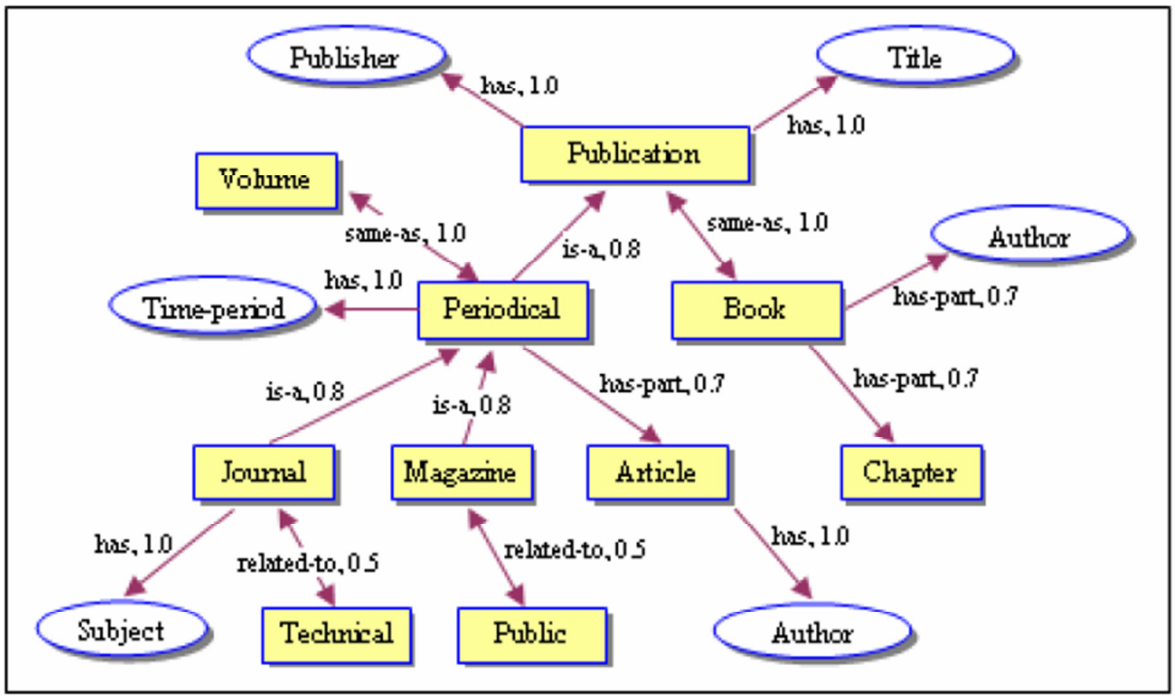

FIGURA 4.1 Uma ontologia derivada da WordNet que representa um domínio de publicações (ABULAISH e DEY, 2006).

Já baseando-nos em Calegari e Ciucci (2008), Calegari e Sanchez (2007) e Fullér (2008), podemos expressar uma ontologia difusa, como sendo uma quíntupla $O_d = <I,C,T,N,A>$, onde:

- I é o conjunto de indivíduos (objetos), também chamados de instâncias de conceitos.
- C é um conjunto de conceitos difusos (ou classes, como em OWL, de indivíduos, ou categorias, ou tipos). Cada conceito é um conjunto difuso sobre o domínio de instâncias.
- O conjunto de entidades da ontologia difusa é definido por $E = C \cup I$
- T indica as relações taxonômicas difusas entre os conjuntos de conceitos C. Elas organizam conceitos em estruturas em árvores de sub ou super conceitos. A relação taxonômica $T(i,j)$ indica que o filho j é uma especificação conceitual de um pai i com um certo grau.
- N indica as relações associativas difusas não taxonômicas que relacionam entidades através das estruturas em árvore, por exemplo: relações de nomeação, de localização, funcionais, subsunção, correlação, etc.

- *A* indica o conjunto de axiomas expresso em uma linguagem lógica apropriada, isto é, predicados que delimitam (restringem/ constrangem) o significado de conceitos, indivíduos, relacionamentos e funções.

Na próxima Seção veremos ilustrações de como essas extensões ontológicas vêm sendo empregadas com sucesso em várias aplicações no âmbito da modelagem de bases de conhecimento e sistemas de recuperação de informações.

4.3 Componentes Ontológicos Difusos

As primeiras abordagens em ontologias difusas inspiraram-se no tratamento difuso das estruturas hierárquicas em taxonomias e depois evoluíram para o tratamento de propriedades difusas consolidando uma estrutura ontológica difusa com vocabulários, relacionamentos e axiomatização baseada em lógica difusa. A seguir, relacionamos alguns desses trabalhos, em ordem cronológica, onde buscamos enfatizar o tratamento que é dado a classes, propriedades, relações, instâncias e metaclasses para incorporar características da teoria de conjuntos e lógica difusa.

Parry (2004)

Silvia Calegari e Davide Ciucci (2006, p. 1) consideram que o artigo *A fuzzy ontology for medical document retrieval* de Parry (2004) foi o primeiro trabalho a utilizar ontologias difusas para organizar e recuperar informações em domínio de aplicação.

Parry (2004) usa ontologias difusas para a recuperação de documentos médicos. Considera que um mapeamento entre um termo específico e diversos documentos pode ter valor diferente, considerando-se diferentes interesses de diferentes usuários. Isto é, distintos termos podem ser associados por graus de relevância dentro do contexto da pesquisa delimitado pelo usuário, ou grupo de usuários.

Para expressar a característica desse mapeamento, adiciona um valor de grau de pertinência para cada termo que é atribuído para cada usuário, ou grupo de usuários de modo que os documentos recuperados a partir da ontologia possam refletir a adequação da informação recuperada. Os graus de relevância são atribuídos manualmente, de modo individual, ou coletivo, considerando-se a área profissional, o “status” e função. Os níveis de adequação são definidos como: “Oposto”, “Não

relacionado", "Levemente relacionado", "Moderadamente relacionado" e "Fortemente relacionado".

O mecanismo de recuperação de documentos apresenta os resultados, considerando a análise do corpus de documentos e o uso de mecanismo de relevância de documentos recuperados.

A despeito de seu pioneirismo, Calegari e Ciucci (2006) consideram que essa abordagem está mais próxima da teoria estatística do que da teoria difusa, possivelmente, em função do autor definir os graus de pertinência com base em levantamento de número de consultas realizadas envolvendo determinado termo, por exemplo.

Quan, Hui e Cao (2004)

Esses autores propõem uma abordagem denominada FOGA (*Fuzzy Ontology Generation frAmework*) para o desenvolvimento automático de ontologias difusas a partir de informação incerta.

Para tal, incorpora lógica difusa na *Análise Formal de Conceitos* para formar uma camada de conceitos difusos.

A partir dessa camada de conceitos difusos, constroi a hierarquia de conceitos e gera a ontologia difusa articulando esses conceitos e essa estrutura hierárquica onde eles se inserem.

Como estudo de caso, discute a aplicação da FOGA para gerar uma ontologia de *Web Semântica Erudita* a partir de um banco de dados de citações.

Pereira (2004) e Pereira, Ricarte e Gomide (2006)

Rachel Pereira, Ivan Ricarte e Fernando Gomide em artigo de 2006 (PEREIRA *et al.*, 2006), e em extensão a pesquisa desenvolvida em dissertação de Rachel Pereira na Unicamp (PEREIRA, 2004), propõem um modelo de representação do conhecimento e de recuperação de informação relevante baseado em uma ontologia relacional difusa.

Nesse trabalho a ontologia difusa é definida como um vocabulário de termos, associados entre si por uma relação difusa para representação em um domínio de conhecimento (PEREIRA *et al.* 2004, *apud* PEDRYCZ e GOMIDE, 1998).Cada classe da ontologia é relacionada a uma palavra por um grau de associação difusa.

Nesse trabalho, os autores apresentam como antecedentes pesquisas feitas para recuperação de informação a partir de *tesaurus* e redes conceituais difusas: Ogawa *et al.* (1991)[52], Horng *et al.* (2001)[53], Klir e Yuan (1995) e Takagi e Kawase (2001)[54].

Abulaish e Dey (2006)

A proposta de formalização de ontologias difusas por esses autores é altamente relevante, e nós já a apresentamos na Seção 6.2, quando tratamos da formalização de ontologias difusas. Entretanto, é interessante ainda acrescentar que o contexto de aplicação de suas ontologias difusas é o de incrementar a interoperabilidade entre ontologias distribuídas e que se sobrepõe, buscando níveis de "uniformidade" com que determinados conceitos são definidos em distintas ontologias.

De um lado, diferentes ontologias variam grandemente em termos do nível de detalhe de suas representações, assim como a natureza de sua especificação lógica subjacente. De outro lado, interoperabilidade entre diferentes ontologias torna-se essencial para obter ganhos a partir das ontologias existentes. Para tal, trabalham com um Descritor de conceito representado como uma relação difusa que incorpora o grau de valor de uma propriedade usando uma função de pertinência difusa.

Relações semânticas na ontologia como IS-A, HAS-PART, etc. também são associados a graus de força de associação. A forma de associação entre dois conceitos determina a "uniformidade" a qual esses dois conceitos têm sido definidos identicamente entre diferentes ontologias.

---

[52] OGAWA, Y., MORITA, T. e KABAYASHI, K. "A fuzzy document retrieval system using the keyword connection matrix and a learning method". Fuzzy Sets and Systems, 39, p 163-179, 1991.
[53] HORNG, Y., J, CHEN, S.M. e LEE, C.H., Automatically constructing multi-relationship fuzzy concept in fuzzy information retrieval systems", IEEE international Fuzzy Systems Conference, p. 606-609.
[54] TAKAGI, T. e KAWASE, K. A trial for data retrieval using conceptual fuzzy sets, IEEE Transactions on Fuzzy Systems, Vol. 9, No 4, p. 497-505, 2001.

Lam (2006)

Toby Lam (2006) propõe uma extensão chamada *Fuzzy Ontology Map* (FOM). FOM é uma matriz de conexão que coleta o valor de pertinência entre classes na ontologia, representando a informação incerta. Além disso, é definido um conjunto de algoritmos para inferir relacionamento difuso. Em sua versão original, a ontologia difusa é gerada pelo desenvolvedor da ontologia, sendo subjetiva. Uma versão futura incorporaria a capacidade de gerar a ontologia difusa de modo automático, tornando-a mais objetiva.

Calegari e Ciucci (2008, 2007a, 2007b, 2007c e 2006) Calegari e Sanchez (2007)

Silvia Calegari vem desenvolvendo trabalhos importantes com ontologias difusas em colaboração com Davide Ciucci (CALEGARI e CIUCCI, 2008, 2007a, 2007b, 2007c, e2006) e Elie Sanchez.(CALEGARI e SANCHEZ, 2007).

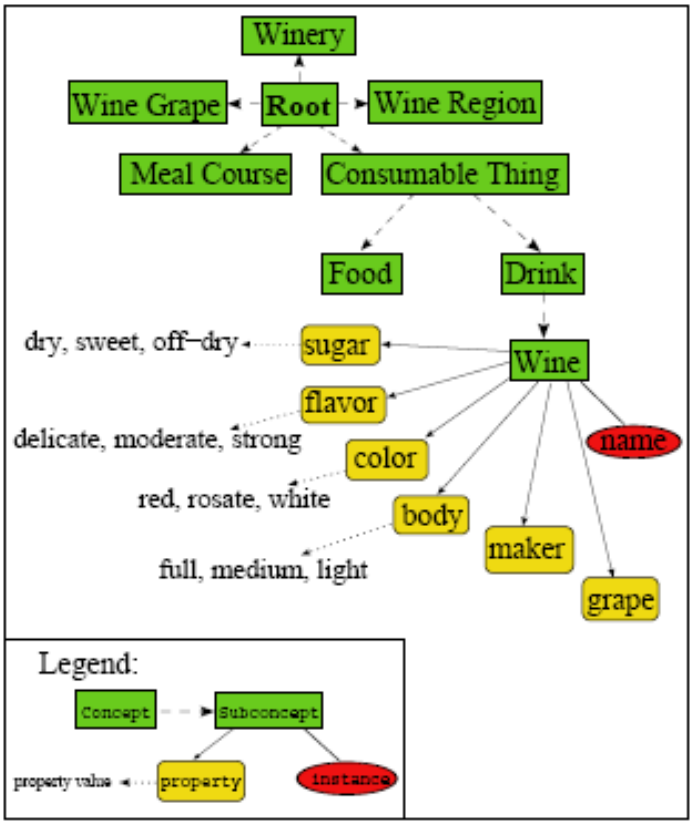

FIGURA 4.2 Ontologia de vinho (CALEGARI e CIUCCI, 2006, p. 5).

Suas ontologias são desenvolvidas no âmbito do projeto *KAON*, levado adiante pela Universidade de Karlsruhe (Alemanha), Universidade de Manchester (Inglaterra) e pelo *Research Center for Information* Technologies (FZI).

Yaguinuma (2007)

No Brasil, o uso de ontologias difusas para representar e recuperar informações também envolve pesquisas crescentes. Já destacamos os trabalhos de Pereira, Ricarte e Gomide (PEREIRA, 2004; PEREIRA, RICARTE e GOMIDE, 2006, 2005a, 2005b). Mais recentemente, um grupo de pesquisadores vinculados ao Departamento de Computação da Universidade Federal de São Carlos (UFSCar) vem publicando trabalhos aplicando ontologias difusas em distintas áreas, tais como para expansão semântica de consultas (YAGUINUMA, BIAJIZ E SANTOS, 2007b); mineração de dados semanticamente estendida através de ontologias difusas (ESCOVAR, YAGUINUMA, BIAJIZ, 2006); integração semântica de fontes de dados heterogêneos no domínio de Análise *Watershed* (FERRAZ, AFONSO, YAGUINUMA, BORGES E SANTOS, 2010), meta-ontologia para representação de informações imprecisas em ontologias (YAGUINUMA, SANTOS e BIAJIZ, 2007a) que permite modelar classes e relacionamentos difusos para serem herdados e/ou instanciados pelas ontologias específicas de domínio, de modo que estas sejam capazes de representar e realizar inferências sobre informações imprecisas.

Em sua dissertação Cristiane Yaguinuma (2007) já desenvolvia o sistema FOQuE para expansão de consulta semântica através de ontologias difusas. Este sistema permite recuperar resultados aproximados que satisfaçam aos requisitos do usuário, de acordo como parâmetros de expansão definidos por ele. As respostas adicionais recuperadas pelo sistema FOQuE são classificadas conforme o tipo de expansão realizada e a relevância para a consulta, melhorando, assim, a forma como os resultados são apresentados ao usuário.

Gu, Wang, Ling e Shi (2007)

Os autores apresentam a metodologia *Use-Case based Fuzzy Ontology Constructing* para construir uma ontologia difusa. Uma solução para representar

uma relação difusa também é fornecida. Um aspecto interessante desse trabalho é, justamente, a discussão da representação de uma relação difusa em OWL.

Segundo os autores, uma **relação difusa *R*** é um conjunto de triplas {<*x*,, $\mu_R(x,y)$>1 x ∈ X, y ∈ Y}. O grau de pertinência $\mu_R(x,y)$ é uma função de pertinência mapeada a partir do universo do discurso $X \times Y$ para o domínio dos números reais [0,1]. Para cada x ∈ X, y ∈ Y, $\mu_R(x,y)$ denota o grau de pertinência *R* entre *x* e *y*.

Considerando-se que uma ontologia difusa é uma extensão do domínio de uma ontologia, deve-se resolver o problema de representar o grau de pertinência em OWL (ou outra linguagem de representação).

Para tal, é necessário estender *owl:DatatypeProperty* de modo a permitir dois domínios ao mesmo tempo. Entretanto, isto é impraticável: OWL não alcança tal nível de expressividade. Uma solução viável é construir um conceito de Relação Difusa, que inclua dois *owl:ObjectProperty*: “domínio-1” e “domínio-2”. Os dois domínios denotam dois universos do discurso na Relação Difusa, e “temGrauDifuso” significa o grau de pertinência correspondente. Sempre que for necessário para um elemento da Relação Difusa, cria-se uma instância desse conceito e atribuem-se valores específicos para cada propriedade (Gu, Wang, Ling, Shim, 2007, p.592).

Ghorbel, Bahri e Bouazis (2009 e 2010)

O trabalho de Hanêne Ghorbel e colaboradores (GHORBEL *et al.*, 2009) é similar ao desenvolvido por Silvia Calegari e colaboradores descrito mais acima. Só que ao invés de utilizarem a plataforma *KAON*, eles fazem sua contribuição dentro do projeto *Protégé*[55] desenvolvido conjuntamente pela Universidade de Stanford (Estados Unidos), pelo *National Library of Medicine* e *The National Center for Biomedical Ontology*.

Dentro do *Protégé*, eles criaram uma ferramenta colaborativa semiautomática para a construção de modelos ontológicos difusos construídos como um *plug-in* à versão 3.3.1 do *Protégé*. A essa ferramenta eles deram o nome de Fuzzy *Protégé*.

---

[55] Vide: http://protege.stanford.edu/

A estratégia adotada por esses autores dá ênfase às definições de metaclasses para permitir a definição de funções de pertinência parametrizadas. Dá suporte a instanciação de conceitos e propriedades difusas e permite computação automática de graus de pertinência, e permite a realização de consultas baseadas em critérios difusos, nas ontologias difusas. Nesse último caso, realiza-se um pré-tratamento das instâncias dos conceitos difusos antes de processar a consulta.

A definição das metaclasses parte da consideração de funções associadas a conjuntos difusos, tais como: as funções trapezoidais, triangular, função L (*left-shoulder*), função R (*right-shoulder*). Para ilustração reproduzimos abaixo a função trapezoidal utilizada pelos autores para modelar a variável linguística “Jovem”.

$$\text{Pessoa\_Jovem} = \text{Pessoa} \cap \exists \text{temIdadeJovem}$$

O termo linguístico “Jovem” pode ser definido por uma função trapezoidal como mostrado na FIGURA 4.3, fórmulas matemáticas e gráfico.

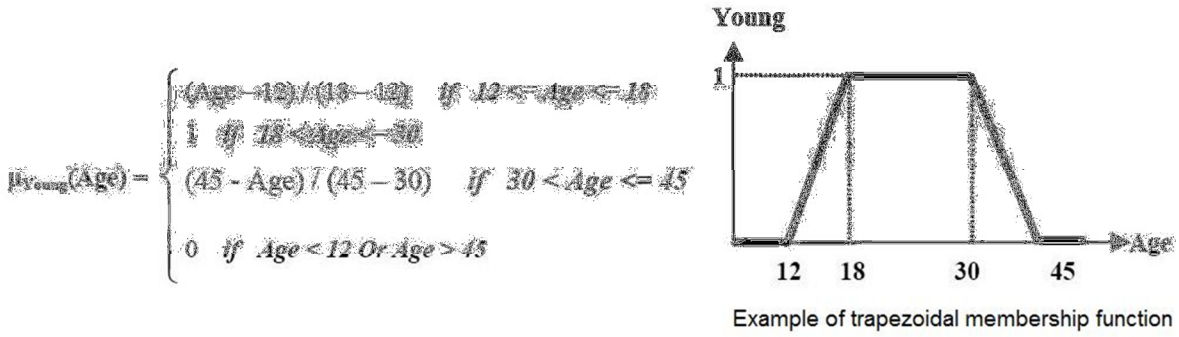

FIGURA 4.3 Exemplo para função de pertinência trapezoidal (GHORBEL *et al.*, 2009).

Na hierarquia dos componentes ontológicos estruturantes do projeto *Protégé*, os autores inserem metaclasses difusas que permitem a definição de funções difusas baseadas nessas curvas (trapézio, triângulo, função L (*left-shoulder*), função *R* (*right-shoulder*), conforme podemos ver na FIGURA 4.4.

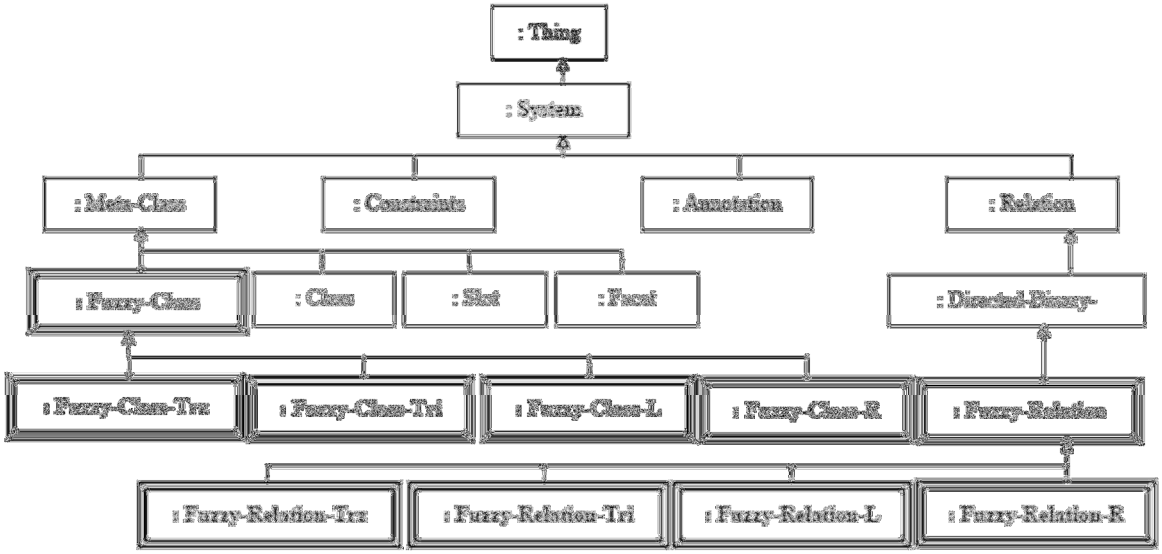

FIGURA 4.4 Hierarquia de metaclasses difusas (GHORBEL *et al.*, 2009, p. 2).

Para usar essas metaclasses temos que importa-las usando *plug-in* específico de Meta Data.

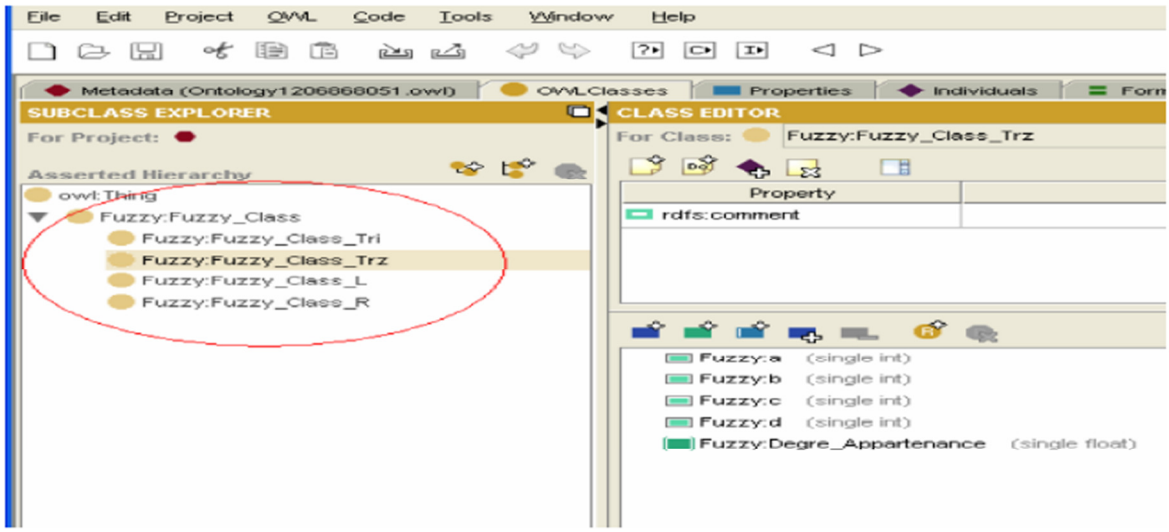

FIGURA 4.5 Metaclasses no *Fuzzy Protégé* (GHORBEL *et al.*, 2009, p. 2).

Mais recentemente, os autores publicaram outro artigo (GHORBEL *et al.*, 2010) que trata da construção de componentes ontológicos difusos no contexto da linguística e semântica de corpus, área emergente e com total afinidade com a Web Semântica.

4.4 Lógicas, Linguagens e Mecanismos de Inferência

Nesta seção, apresentaremos e discutiremos aspectos de algumas lógicas e linguagens lógicas que permitem a formalização, e axiomatização em ontologias e a

consequente implementação de mecanismos de inferência e raciocinadores automáticos.

4.4.1 Critérios para Avaliação de Lógicas e Linguagens

Lógicas e linguagens lógicas são desenvolvidas para fins específicos e com características, capacidades e limitações distintas. O conhecimento da conformação dessas lógicas e linguagens a determinados critérios propiciam-nos o conhecimento de garantias formais que nos serão exigidas conforme as necessidades de nosso domínio de aplicação e das soluções almejadas.

De um lado, diz-se que os requisitos de completude, juntamente com os requisitos de consistência e de decidibilidade constituem os requisitos fundamentais dos cálculos lógicos. De outro lado, a capacidade de expressividade é uma necessidade crescente dos sistemas formais, principalmente, em vista de aplicações em domínio de linguagem natural como o da Web Semântica.

Para compreender o alcance das lógicas, suas linguagens derivadas, e termos condições de avaliá-las melhor, apresentamos abaixo algumas considerações sobre critérios que sempre são levados em conta na apreciação e comparação dessas lógicas e linguagens.

Decidibilidade

Grosso modo, e para nossos fins, definimos decidibilidade como a capacidade da lógica, ou linguagem lógica derivar algoritmos que tenham fim. Isto é, dizer que uma linguagem é decidível é dizer que um programa implementado com base nessa linguagem não levará a uma condição de *loop* infinito.

Numa formulação mais rigorosa:

> "Uma frase ou fórmula bem-formada de uma teoria ou sistema formal é decidível se existe um algoritmo que permita determinar se a frase ou fórmula é um teorema do sistema; caso contrário, é indecidível. E uma teoria ou sistema formal é decidível se qualquer frase ou fórmula bem-formada do sistema for decidível."
> "O sistema da lógica proposicional clássica é decidível; mas, pelo Teorema da Indecidibilidade de Church, A lógica n-ádica de predicados é indecidível." (BRANQUINHO *et al.*, 2006, p. 228.)

Ainda em outras palavras, e em nosso contexto, o que a questão da decidibilidade nos impõe é o seguinte. Dado um conjunto de proposições envolvendo componentes conhecidos (classes, propriedades e instâncias) existirá algum algoritmo que permita saber, para qualquer proposição do conjunto e no fim de certo número finito de passos o valor de verdade atribuído à proposição?

Completude

A completude de uma lógica, ou linguagem lógica diz respeito à capacidade do sistema de derivar novas proposições, ou consequências válidas. Se um fato é consequência lógica de um conjunto de premissas, pode-se encontrar uma prova para este fato no sistema formal. O sentido desse termo fica mais bem esclarecido na seguinte definição:

> Intuitivamente, um sistema lógico é completo se tudo o que queremos derivar é derivável nele. Assim, uma formalização lógica é completa se todas as formas válidas de argumento são deriváveis no sistema; um sistema concebido para codificar o raciocínio matemático é completo se todas as verdades matemáticas podem ser nele derivadas, e assim por diante. Apesar de, explicada nesses termos, a noção parecer completamente informal, é possível defini-la com mais precisão. Um sistema de lógica é completo no sentido introduzido por Gödel se e somente se todas as fórmulas válidas bem-formadas são teoremas do sistema. Num sentido mais nítido, um sistema é completo se, para qualquer fórmula bem-formada *A*, ou *A* é um teorema, ou o sistema se tornaria inconsistente se *A* lhe fosse acrescentado como axioma. (BLACKBURN, 1997, p. 64)

Consistência

Uma lógica, ou linguagem lógica é consistente se todas as proposições que ela permite são simultaneamente verdadeiras. Por exemplo, não é permitido que *p* e ¬*p* (*não p*) sejam simultaneamente verdadeiras. Além disso, todas suas fórmulas-bem-formadas são demonstráveis (BLACKBURN, 1997, p. 73). Ou seja, o resultado de uma prova é consequência lógica das premissas.

Expressividade

A expressividade de uma lógica, ou linguagem lógica é a sua capacidade de expressar, isto é, sua capacidade de construir formalizações com significação. Uma expressão é uma formalização constituída por signos onde os signos não

meramente indicam, mas possuem efetivamente uma função significativa (HUSSERL, Investigações Lógicas, Investigação Primeira, Cap. I, § 1, *apud* MORA, Tomo II, p. 978).

As linguagens terão que se haver com a capacidade de expressar conceitos aparentemente simples como uma classe, ou relação de parentesco como "tio", expressar o significado de expressões facilmente compreensíveis para nós tais como "Sócrates é corajoso" ou, ainda, restrições de escalabilidade, etc.[56]

Entretanto, como indicaremos, tal tarefa não é tão simples do ponto de vista lógico e de sistemas formais. Também não é tão simples do ponto de vista ontológico. Um bom exemplo dessas dificuldades expressivas pode ser vislumbrada em detalhes no artigo recém-publicado *A Better Uncle For OWL* (KRÖTZSCH *et al.*, 2011), onde os autores mostram como a relação parental "tio" que não podia ser expressa em OWL 1, passa a ser expressa em OWL 2.

> Depois da publicação da Recomendação W3C de 2004 para a *Web Ontology Language OWL*[57], discussão do problema centrado sobre a "regra tio"
>
> $$\text{irmaoDe}(x,y) \wedge \text{parenteDe}(y,z) \rightarrow \text{tioDe}(x,y)$$
>
> que é fácil de estabelecer usando uma regra de linguagem simples como *Datalog*, mas não pode ser modelada no todo na versão da OWL de 2004. A partir da perspectiva do critério de projeto da OWL, uma dificuldade central é a que leva a raciocínios indecidíveis no resultado de uma linguagem combinada.
> Subsequentemente, um trabalho significativo foi desenvolvido, investigando a integração de lógicas descritivas (LDs), que formam a base para a OWL, e linguagens de regra (tipicamente *Datalog*). Conceitualmente, podem-se distinguir duas abordagens.
> De um lado, lógicas descritivas têm sido estendidas com características expressivas adicionais ao estilo de lógica descritiva, que tem tornado possível expressar certos tipos de regras. OWL 2, a revisão da recomendação W3C da OWL, de fato pode expressar a "regra tio" mencionada acima. Pela combinação de novos dispositivos da OWL 2 muitas regras com um "corpo em formato de árvore" pode ser expressa indiretamente. Decidibilidade é, não obstante, preservada. Muitas regras, entretanto, tais como
>
> $$\text{temPais}(x,y) \wedge \text{temPai}(x,z) \wedge \text{casado}(y,z) \rightarrow C(x)$$

---

[56] Vide artigo de Ian Horrocks. *OWL Rules, OK?.* Disponível em: <http://www.w3.org/2004/12/rules-ws/paper/42/>. Acessado em 01-02-2012.
[57] Vide: http://www.w3.org/TR/2004/REC-owl-features-20040210/.

> que define uma classe C de crianças cujos pais são casados, ainda não são expressáveis.(KRÖTZSCH *et al.*, 2011)

No nosso escopo, veremos que as abordagens clássicas sejam em lógica, ou linguagens lógicas, não possuem a capacidade de expressar modalidade, incerteza, imprecisão, ambiguidade, indeterminação, inexatidão, vaguidade; escalabilidade, ou não monotonicidade.

Balanceamento entre critérios

Alguns critérios de conformidade apresentam-nos como se fossem autoexcludentes, exigindo um balanceamento entre eles de modo a preservar um nível de satisfação com o resultado pretendido.

A história da lógica e da matemática ao longo do século XIX, e primeira metade do século XX, mostrou-nos como as aspirações ambiciosas dessas áreas foram repetidamente solapadas pela demonstração da impossibilidade de se constituir a partir delas sistemas formais rigorosamente completos, decidíveis, ou consistentes. Primeiro, houve o naufrágio do projeto logicista de Whitehead e Russel, inspirados nos esforços de Frege de reduzir a lógica à aritmética, depois, o constrangimento gerado por Gödel ao demonstrar a insubstância de fundamentação no seio da própria matemática, mais precisamente da lógica-matemática. Por fim, a demonstração independente, por parte de Alonzo Church e Alan Turing de que, nem mesmo no cálculo quantificacional elementar, não se pode elaborar nenhum procedimento de decisão.

Assim, do ponto de vista estrito do rigor lógico-matemático, não podemos contar com lógicas e linguagens que consigam, ao mesmo tempo, serem rigorosamente completas, consistentes e decidíveis.

Deveremos contentar-nos em termos uma lógica e uma linguagem que atenda satisfatoriamente às nossas necessidades específicas, mas sabendo que a ênfase dada a um critério, frequentemente, será construída em detrimento de outro critério.

Frequentemente, também teremos de abrir mão da capacidade de expressividade de uma linguagem em prol de sua decidibilidade, ou vice-versa. Veremos adiante,

por exemplo, como a extensão de ontologias formais com base em linguagens de marcação de regras afeta a decidibilidade lógica, impondo a adição de novas restrições para suplantar a possibilidade de um algoritmo levar a uma computação infinita.

4.4.2 Lógica Descritiva (LD) e suas Linguagens Usadas em Ontologias

A área de Lógica Descritiva começou sob o rótulo de sistemas terminológicos (*terminological systems*) para enfatizar que a linguagem de representação foi usada para estabelecer a terminologia básica adotada no domínio modelado. Mais tarde: linguagens de conceitos (*concept languages*), interesse movido em direção a propriedades de sistemas lógicos subjacentes, o termo lógicas descritivas (*Description Logics)* tornou-se popular (NARDI e BRACHMAN, 2007, p. 3).

Lógica Descritiva é a reconstrução lógica das chamadas linguagens de representação de conhecimento baseadas em *frames* (molduras), com o objetivo de fornecer uma semântica declarativa simples e no estilo de Tarski a fim de capturar o significado da maioria das características de uma representação estruturada do conhecimento (STRACCIA, 2006, p. 1). É um conjunto de fragmentos decidíveis de Lógica de Primeira Ordem e adequada para fazer raciocínio automatizado e vem despertando interesse crescente em vista de suas aplicações no contexto da Web Semântica.

Sistemas de Representação de Base de Conhecimento

Em Lógica Descritiva o conhecimento é divido em duas partes: TBox e ABox.

Uma TBox refere-se à construção e definição de conceitos e propriedades de um domínio (terminologias). É composto por axiomas de classes como, por exemplo, axiomas de equivalência, ou de inclusão.

A forma básica de declaração em uma Tbox é uma definição de conceito, que é realizado pela definição de um novo conceito em termos de outros conceitos previamente definidos. Por exemplo, uma mulher pode ser definida como uma pessoa feminina escrevendo-se esta declaração (NARDI e BRACHMAN, 2007, p. 14):

Mulher ≡ Pessoa ∪ Fêmea

Tais operações são diretamente relacionadas às formas e ao significado de declarações permitidas na Tbox. Nesse caso, tal declaração é usualmente interpretada como uma equivalência lógica que fornece condições, tanto necessária quanto suficiente, para classificar um indivíduo como mulher.

Uma ABox refere-se á especificação de indivíduos, através de asserções, de um ou mais conceitos. Essas asserções dizem respeito a fatos e tratam, por exemplo, de pertinência de classe, relação, relação de negação, igualdade, desigualdade.

Contem conhecimento extensional sobre o domínio de interesse, constituído por asserções sobre indivíduos, usualmente chamadas asserções de pertinência (NARDI e BRACHMAN, 2007, p. 19). Por exemplo:

Fêmea ∪ Pessoa (ANA)

Essa asserção estabelece que o indivíduo ANA é uma pessoa feminina. Dada a definição acima de mulher, pode-se derivar dessa asserção que ANA é uma instância do conceito Mulher.

Com tais recursos as lógicas descritivas permitem a execução de tarefas de classificação, como (HAASE *et al.*, 2006, p. 9):

- Checagem de instância: o indivíduo a está no conceito *C*?
- Subsunção de conceito: o conceito C é mais geral do que o *D*?
- Satisfabilidade de conceito: a definição do conceito *C* permite algumas instâncias desse conceito?
- Satisfatibilidade de base de conhecimento: o conhecimento combinado de TBox e ABox está livre de contradições?

No critério expressividade, a lógica descritiva sendo um fragmento de lógica de predicados de primeira ordem, não pode expressar, por exemplo, o seguinte:

- Expressões difusas: “frequentemente chove no verão”.
- Não-monotonicidade: “pássaros voam, pinguim é um pássaro, mas pinguim não voa”.
- Atitudes proposicionais: “Eva pensa que 2 não é um número primo” (é verdade que ela pensa isto, mas o que ela pensa não é verdade).
- Lógica modal:
  - possibilidade e necessidade: “é possível que chova hoje”

- modalidades epistêmicas: “Eva sabe que 2 é um número primo”
- lógica temporal: “Eu sempre estou com fome”
- lógica deôntica: “você deve fazer isto”.

4.4.3 OWL – Web Ontology Language

A Lógica Descritiva é a base da construção da linguagem OWL (*Web Ontology Language*), atual linguagem padrão para a representação em ontologias. A OWL é uma recomendação da W3C desde fevereiro de 2004.

OWL surge no contexto da Web Semântica (BERNERS-LEE *et al.*, 2001, SHADBOLD *et al.*, 2006) para permitir a representação de termos em vocabulários e seus inter-relacionamentos em uma ontologia. Ela foi criada a partir da linguagem para ontologias DAML+OIL, e buscou-se com a OWL mais facilidades para expressar significado e semântica em relação às linguagens XML, RDF e RDF-S. OWL vai além dessas linguagens em sua habilidade para representar conteúdo interpretável por máquina na Web e, permitindo a interoperabilidade semântica, torna a Web uma ferramenta capaz de tecer a almejada rede semântica.

Máquinas de busca (consultas) tradicionais retornam listas de recursos recuperados, oferecendo pouca ou nenhuma informação sobre as relações semânticas existentes entre eles. Consequentemente, o usuário tem de dispender uma quantidade substancial de tempo acessando-os, lendo-os e verificando a real relevância de seus recursos informativos. Nesse sentido, a Web Semântica irá solucionar vários problemas chaves das atuais arquiteturas de tecnologia da informação (DACONTA, 2003, *apud* ELLER, 2008, p. 2-21):

> - **Sobrecarga de informação:** na Web há uma quantidade imensa de informações não pertinentes que estão disponíveis e que são fornecidas pelos processos de busca.
> - **Informação indistinta, não classificada, não categorizada:** as ferramentas de busca enfrentam a dificuldade de executar pesquisas entre documentos que não estão diferenciados em termos de assunto, qualidade e relevância. Não são capazes de diferenciar uma informação comercial de uma educacional, ou informação entre idiomas, culturas e mídia.
> - **Integração de informações:** a variedade de fontes de informação distintas com diferenças sintáticas, semânticas e estruturais entre elas é muito grande, tornando o compartilhamento, a integração e a resolução de conflitos entre essas informações um problema de difícil solução.
> - **Heterogeneidade estrutural e semântica:** a heterogeneidade estrutural e semântica da informação na Web atualmente é imensa, e a maioria das

propostas de integração ainda adota soluções com alto índice de centralização, tornando o seu uso na Web inviável.

- **Conteúdo não estruturado:** as informações podem variar de não estruturadas, como imagens e vídeos, a semi-estruturadas, como arquivos de e-mail e páginas Web. Por serem criadas de forma autônoma, sem preocupação com regras de estruturação, catalogação e descrições de suas propriedades, essas informações são difíceis de serem abrangidas pelos mecanismos de pesquisa, ocasionando demora e ineficácia na sua localização. Outros problemas relacionados são: informações não localizadas devido às mudanças de URLs; recuperação de um número elevado de informações que, em sua maioria, não atendem às expectativas dos usuários; e recuperação de informações fora do contexto solicitado pelo usuário, devido a problemas de semântica e ambiguidade.

Para tratar esses problemas é necessário considerar questões relevantes como a utilização de metadados e ontologias, visando a busca de uma linguagem única, capaz de estruturar e representar conhecimento e regras. Nisso entra a OWL buscando organizar semanticamente a informação. Dentre outros recursos, faz uso da adição de mais vocabulário para descrever propriedades e classes; relações entre classes (por exemplo, disjunção), cardinalidade (por exemplo, “exatamente um”), igualdade, tipagem rica de propriedades, características de propriedades (por exemplo, simetria), e classes enumeradas. Uma ontologia OWL contém uma sequência de axiomas e fatos. Axiomas podem ser de vários tipos, por exemplo, axiomas envolvendo subclasses, ou classes equivalentes.

4.4.4 Extensões da OWL para Aprimorar Expressividade Atendendo a Critérios de Decidibilidade

A linguagem OWL é baseada em Lógica Descritiva que é uma família de lógicas projetada para ser tão expressiva quanto possível enquanto mantém sua capacidade de decidibilidade. A versão da OWL 1, e a OWL DL é baseada na lógica descritiva $\mathcal{SHOIN}^{(D)}$, e a versão OWL 2 é baseada na lógica descritiva $\mathcal{SROIQ}^{(D)}$. Tanto $\mathcal{SHIF}^{(D)}$, quanto $\mathcal{SHOIN}^{(D)}$, quanto $\mathcal{SROIQ}^{(D)}$, são versões da linguagem atributiva $\mathcal{AL}$.

Vamos esclarecer o significado desses nomes compostos pela agregação de rótulos. Abaixo apresentamos a TABELA 4.1 elaborada a partir do verbete *Description Logic* da Wikipedia[58].

[58] Vide: http://en.wikipedia.org/wiki/Description_logic

TABELA 4.1 Significados de rótulos usados na nomenclatura de linguagens baseadas em lógica descritiva.

Rótulo	**Significado**
(D)	Uso de propriedades de tipos de dados, valores de dados ou tipos de dados.
$\mathcal{AL}$	*Atributive language*. Linguagem básica que permite negação atômica (negação de nomes conceituais que não aparecem no lado esquerdo do axioma); interseção conceitual; restrições universais e quantificação existencial limitada.
$\mathcal{ALC}$	$\mathcal{ALC}$ é simplesmente $\mathcal{AL}$ com complemento de qualquer conceito permitido, não somente conceitos atômicos.
$\mathcal{F}$	Propriedades funcionais.
$\mathcal{H}$	Hierarquia de papeis (*roles*) (sub-propriedades – *rdfs:subPropertyOf*).
$\mathcal{I}$	Propriedades inversas
$\mathcal{N}$	Restrições de cardinalidade (*owl:cardinality*, *owl:maxCardinality*).
$\mathcal{O}$	Nominais (classes enumeradas de restrições a valores de objetos – *owl:oneOf*, *owl:hasValue*) Exemplo: João, Pedro, Carlos.
$\mathcal{Q}$	Restrições de cardinalidade qualificada (em *OWL 2,* restrições de cardinalidade mais completas que as permitidas à superclasse *owl:Thing*)
$\mathcal{R}$	Axiomas de inclusão de papeis (*roles*) de complexidade limitada: reflexividade e irreflexividade; disjunção de papeis.
$\mathcal{S}$	Uma abreviação para $\mathcal{ALC}$ com propriedades transitivas
$\mathcal{U}$	União de conceitos.

Com base nesta tabela fica mais fácil compreender a função das linguagens baseadas em lógica descritiva, como nas três linguagens seguintes.

$\mathcal{SHIF}^{(D)}$

Esta é uma linguagem $\mathcal{ALC}$ com papeis transitivos. Permite hierarquia de papeis, propriedades inversas e funcionais e, ainda, o uso de propriedades de tipos de dados, valores de dados, ou tipos de dados.

$\mathcal{SHOIN}^{(D)}$

É uma linguagem $\mathcal{ALC}$ com papeis transitivos; permite hierarquia de papeis, nominais, propriedades inversas, restrições de cardinalidade e, ainda, o uso de propriedades de tipos de dados, valores de dados, ou tipos de dados.

> $\mathcal{SHOIN}^{(D)}$ é uma lógica descritiva muito expressiva que fornece negação cheia, disjunção e uma forma restrita de forma universal de quantificação existencial. Fornece suporte a raciocínio com tipo de dados concretos, tais como cadeias de caracteres ou inteiros. Ao invés de tipo de dados concretos axiomatizados em lógica, $\mathcal{SHOIN}^{(D)}$ emprega uma abordagem onde as propriedades de tipos de dados concretos são encapsuladas nos assim chamados domínios concretos. Um domínio concreto é um par (Δ_D,Φ_D), onde Δ_D é a interpretação do domínio, eΦ_D é um conjunto de predicados de domínio concreto que vem com uma aridade *n* e uma interpretação pré-definida $d^D \subseteq \Delta_D^n$. Um domínio concreto admissível ***D*** é equipado com um procedimento de decisão para a satisfabilidade de conjunções finitas sobre predicados de domínio concreto. Checagem de satisfabilidade e domínios concretos admissíveis podem ser combinadas com raciocínio lógico para muitas lógicas descritivas. (HAASE e MOTIK, 2005, p. 1-2).

O editor de ontologias *Protégé* suporta $\mathcal{SHOIN}^{(D)}$.

$\mathcal{SROIQ}^{(D)}$

É uma linguagem $\mathcal{ALC}$ com papeis transitivos; permite axiomas de inclusão de papeis de complexidade limitada (reflexividade e irreflexividade; disjunções de papeis), propriedades inversas, restrições de cardinalidade qualificada e, ainda, o uso de propriedades de tipos de dados, valores de dados, ou tipos de dados.

A linguagem $\mathcal{SROIQ}$ representa uma extensão da linguagem $\mathcal{SHOIN}$, e foi especificada por Ian Horrocks, Oliver Kutz e Ulrike Sattler (HORROCKS *et al*., 2006), no artigo *The Even More Irresistible* $\mathcal{SROIQ}$.

> Grosseiramente falando, nós estendemos $\mathcal{SHOIN}$ com todos os meios que foram sugeridos para nós por desenvolvedores de ontologias como adições úteis à OWL-DL, as quais, adicionalmente, não afetam sua decidibilidade e praticabilidade. Nós consideramos axiomas de inclusão de funções complexas da forma $R \circ S < R$ *ou* $S \circ R < R$ para expressar a propagação de uma propriedade ao longo de outra, a qual tem se provado útil em terminologias médicas.
>
> Além disso, nós estendemos $\mathcal{SHOIN}$ com propriedades reflexivas, antisimétricas e irreflexiva; propriedades disjuntas e

> propriedade universal, e constructos ∃R.*Self*, permitido, por exemplo, a definição de conceitos tais como “narcisista”. Finalmente, nós consideramos asserções com propriedades negadas em ABox e restrições qualificadas de números. A lógica resultante é chamada de $\mathcal{SROIQ}$. (HORROCKS *et al*., 2006),

$\mathcal{SROIQ}$ deve ser uma base para as futuras extensões de OWL, e, com informamos acima, já tem sido adotada como a base lógica da OWL 1.1.

Além da especificação da $\mathcal{SROIQ}$, os autores Ian Horrocks, Oliver Kutz e Ulrike Sattler e outros colaboradores vem se notabilizando pela especificação e aprimoramentos e várias linguagens baseadas na lógica descritiva.

4.4.5 Extensões da OWL para Incorporar Lógica Difusa

Como vimos, uma das mais importantes características da lógica difusa é sua habilidade de realizar raciocínio aproximado, que envolve regras de inferência com premissas, consequências ou ambas delas contendo proposições difusas. No momento, já há linguagens para formalização e axiomatização de ontologias que foram estendidas para incorporar recursos de lógica difusa e há protótipos de raciocinadores difusos sendo desenvolvidos que tratar ontologias difusas.

Para estender a Lógica Descritiva para lidar com vaguidade, desde há alguns anos vêm sendo apresentadas propostas para integrar lógica difusa em lógica descritiva e em OWL. Possivelmente, os primeiros artigos a desenvolverem essa extensão foram *Generalizing term subsumption languages to fuzzy logic* (YEN, 1991), *A fuzzy description logic* (STRACCIA, 1998), *A description logic for vague knowledge* de 1998 (TRESP e MOLITOR, 1998, *apud* CARDOSO e LYTRAS, 2009) e *Reasoning within Fuzzy Description Logics* (STRACCIA, 2001). Eles foram seguidos pelos artigos *Fuzzy $\mathcal{ALC}$ with fuzzy concrete domains* (STRACCIA, 2005), *Towards a Fuzzy description logic for the semantic web* (STRACCIA, 2005), *Expressive Querying over Fuzzy DL-Lite Ontologies* (PAN *et al*., 2007) e *A Fuzzy Description Logic For The Semantic Web* (STRACCIA, 2006), dentre outros.

Mecanismos de inferência que lidam com as dificuldades impostas pela indeterminação permeadas pela linguagem natural em associação com as

ontologias difusas também vêm despertando crescente interesse. Vários esforços vêm sendo feitos para a criação, armazenamento, recuperação, manutenção e aplicação de ontologias utilizando teoria e tecnologia difusa. No desenvolvimento de lógicas e mecanismos de inferência difusos, podemos citar ainda os esforços de integração de lógica difusa com OWL e lógica descritiva (LD) em Stoilos *et al.* (2008, 2007a, 2005), Pan et al (2007 e 2008), Gu *et al.* (2007), e Wang *et al.* (2009), e Bobillo *et al.* (2011, 2009a, 2009b, 2008a, 2008b e 2008c), Bobillo e Straccia (2008) e Bobillo (2008).

O já citado trabalho de Umberto Straccia (STRACCIA, 2006) introduz a noção de Base de Conhecimento Difusa baseada na visão de uma ontologia para a Web Semântica onde conhecimento é expresso em uma ontologia baseada em Lógica Descritiva como uma tripla (*T,R,A*) onde *T*, *R* e *A* são respectivamente uma TBox, RBox e ABox . Então, usando uma ontologia difusa o conhecimento de um domínio é definido a fim de corresponder a uma base de conhecimento em Lógica Descritiva.

Os axiomas em uma base de conhecimento difusa *K* são agrupados em uma ABox difusa, uma TBox difusa, e uma RBox difusa, como segue:

> **RBox:** Uma função abstrata é um nome de função abstrata ou o inverso S^- de um nome de função abstrata *S* (nomes concretos de função não possuem inversa). Um RBox *R* consiste de um conjunto finito de axiomas de transitividade *trans(R)*, e axiomas de inclusão de função da forma $R \subseteq S$ e $T \subseteq U$, onde *R* e *S* são funções abstratas, e *T* e *U* são funções concretas. O fechamento reflexivo-transitivo do relacionamento de inclusão é denotado com $\subseteq^+$. Uma função que não tiver sub-funções transitivas é chamada de função simples.
>
> **TBox:** Uma TBox T consistem de um conjunto finito de axiomas de inclusão de conceitos $C \subseteq D$, onde *C* e *D* são conceitos. Por facilidade, nós usamos $C = D \in T$ no lugar de $C \subseteq D$, $D \subseteq C \in T$. Uma função abstrata simples *S* é chamada *funcional* se a interpretação da função *S* é sempre funcional. Uma função funcional *S* pode sempre ser obtida a partir de uma função abstrata por meio do axioma $T \subseteq (\leq 1\ S)$.Além disso, sempre que dizemos que uma função é funcional, dizemos que$T \subseteq (\leq 1\ S)$ está na TBox.
>
> **ABox:** Uma ABox consiste de um conjunto finito de conceito e axiomas de asserção de função e axiomas de igualdade/desigualdade individual*: C, (a, b):R, (a, c):T, a* $\approx$ *b e a* $\not\approx$ *b*, respectivamente.

> Uma base de conhecimento A $\mathcal{SHOIN}^{(D)}$ é composta por $K = (T, R, A)$ consiste de uma TBox T , uma RBox R, e uma ABox A. (STRACCIA, 2006, p. 3-4)

Essa definição é utilizada na especificação de abordagens difusas desenvolvidas por Silvia Calegari e colaboradores em *Fuzzy Ontology-Approach to improve Semantic Information Retrieval* (CALEGARI e SANCHEZ, 2007) e *Fuzzy Ontology, Fuzzy Description Logics and Fuzzy-OWL* (CALEGARI e CIUCCI, 2007).

> **Definição 1**. Uma Base de Conhecimento Difusa é um para definido como:
> $KB_F = (O_F, I)$
> onde O_F é uma Ontologia Difusa, e I é um conjunto de instâncias associadas com a ontologia difusa. Além disso, todo conceito $C \in \boldsymbol{C}$ é um conjunto difuso sobre o domínio de instâncias definidas como C: I → [0,1] (CALEGARI e CIUCCI, 2007).

Na TABELA 4.2 e na TABELA 4.3 podemos verificar como podem ser desenvolvidas expressões de classe, propriedades e respectivas ABox, TBox e RBox em ontologias difusas.

TABELA 4.2 Expressões de Classe e de Propriedade em Lógica Descritiva.

EXPRESSÕES DE CLASSE	
Nomes de classe	A, B
Conjunção	$C \cap D$
Disjunção	$C \cup D$
Negação	$\neg C$
Restrição existencial de propriedade	$\exists\ R.C$
Restrição universal de propriedade	$\forall\ R.C$
Próprio	$\exists S.\text{Self}$
Ao menos (no mínimo)	$\geq nS.C$
No máximo (Até)	$\leq nS.C$
Nominal	$\{\boldsymbol{a}\}$
PROPRIEDADES	
Nomes de propriedades	R, S, T
Propriedades simples	S, T
Propriedades inversas	R^-
Propriedade universal	U

TABELA 4.3 Axiomas de classe (TBox), papeis (RBox) e fatos (ABox).

TBox (axiomas de classe)	
Inclusão	$C \subseteq D$
Equivalência	$C \equiv D$
RBox (axiomas de papeis)	
Inclusão	$R_1 \subseteq R_2$
Axioma de inclusão de propriedade	$R_1^{(-)} \circ \ldots \circ R_n^{(-)} \subseteq R$
Transitividade	Tra(R)
Simetria	Sym(R)
Reflexividade	Ref(R)
Irreflexividade	Irr(S)
Disjunção	Dis(S,T)
ABox (fatos)	
Pertinência de classe	$C(a)$
Relação de propriedade	$R(a,b)$
Negativa de relação de propriedade	$\neg S(a,b)$
Igualdade	$a \approx b$
Desigualdade	$a \not\approx b$

Exemplos:

Homem(x) ∧ temCriança(x,y) → paiDe(x,y)
Homem ⊆∃ homemProprio **(Tbox)**
Homem ∘ temCriança ⊆ paiDe **(Rbox)**
Elefante(x) ∧ Rato(y) → maiorQue(x,y) **(produto conceitual)**
Homem(x) ∧ temIrmao(x,y) ∧ temCriança(y,z) → Tio(x)
casadoCom(x,y) ∧ ama(x,y) → Feliz(x)

4.4.6 Linguagem SWRL (*Semantic Web Rule Language*) de Definição de Regras

O grupo de trabalho da W3C para regras[59] foca principalmente em fornecer um formato intercambiável de regras (RIF - *Rule Interchance Format*) do que uma única linguagem para a Web Semântica. Os esforços desse grupo geraram a especificação da SPARQL.

Outra linguagem de especificação de regras é a RuleML. RuleML fornece um conjunto de linguagens de marcação para representar e intercambiar diferentes tipos de regras (STOILOS *et al.* 2006).

[59] Vide www.w3.org/2005/rules e Mochol *et al.* (2008).

Já a linguagem SWRL[60] de nosso interesse direto, é baseada em uma combinação das sub-linguagens OWL (nas versões OWL DL e OWL Lite) com a sub-linguagem RuleML (*Rule Markup Language*).

A descrição da SWRL submetida ao W3C[61] informa-nos que a SWRL não suporta OWL Full diretamente, pois não suporta classificação direta sobre classes ou propriedades. Isto é, constructos OWL Full tais como valores de classes, ou de propriedades não são suportados pela linguagem. Não podemos escrever uma regra que, por exemplo, deduza algum novo conhecimento baseado no fato que uma classe é uma subclasse direta de outra.

Deve ser notado que as garantias formais fornecidas por OWL e SWRL podem ser perdidas se essas extensões são usadas para deduzir novo conhecimento. Idealmente, tais recursos incorporados somente devem ser usados para consultas na ontologia. Na guia SWRL, por exemplo, essas bibliotecas incorporadas foram projetadas para serem usadas com a linguagem de consulta SQWRL[62].

Entretanto, de outro lado, implementações da SWRL, ativadas via *plug-ins* através de habilitação de guias em editores de ontologias, de tem duas extensões customizadas que suportam o uso de nomes de classe OWL, nomes de propriedades, nomes de indivíduos e tipos de definição de esquema XML (XSD) como argumentos embutidos. Com essas extensões, recursos incorporados podem ser definidos de modo a suportar operações OWL Full. Por exemplo, a guia SWRL fornece uma biblioteca que contém o **TBox** incorporado.

As regras são da forma de uma implicação entre um antecedente (corpo) e consequente (cabeça). O significado pretendido pode ser lido como: sempre que as condições especificadas no antecedente forem verdadeiras, então as condições especificadas no consequente deve também ser verdadeiras.

---

[60] Vide a submissão original de 21 de maio de 2004, http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/ e a última versão:http://www.w3.org/Submission/SWRL/. Ver também: http://protege.cim3.net/cgi-bin/wiki.pl?SWRLLanguageFAQ e o tutorial desenvolvido por Martin Kuba, 2012, em http://dior.ics.muni.cz/~makub/owl/.

[61] Vide: http://www.w3.org/Submission/SWRL/.

[62] ***S**emantic **Q**uery-Enhanced **W**eb **R**ule **L**anguage*; pronúncia-se em inglês *squirre.* Vide: http://protege.cim3.net/cgi-bin/wiki.pl?SQWRL

Um exemplo simples de regras SWRL e de sua aplicação é a da implicação da propriedade *Tio*.

$$temPais(?x_1,?x_2) \wedge temIrmao(?x_2,?x_3) \Rightarrow temTio(?x_1,?x_3)$$

Predicados SWRL

Em SWRL, os símbolos predicados podem incluir classes OWL, propriedades ou tipos de dados. Argumentos podem ser indivíduos OWL ou tipos de dados, ou variáveis referindo-se a eles. Todas as variáveis em SWRL são tratadas como universalmente quantificadas, com seu escopo limitado a uma dada regra. Deve-se observar que SWRL não suporta a negação de proposições atômicas, ou disjunção.

Mais detalhadamente, os predicados e, SWRL podem ser:

- Expressões de classe: expressões de classe arbitrárias, não somente classes nominadas.
- Expressões de propriedade: o único operador disponível em OWL 2 para criar expressões de propriedade é o inverso da propriedade objeto, entretanto o mesmo efeito pode ser obtido pela alteração dos argumentos da propriedade, então não é necessário usar expressões de propriedade em SWRL.
- Restrições em faixa de dados (contradomínio): especifica o tipo de valores de dados, como inteiro, data, união de alguns tipos de Esquema XML, tipos enumerados.
- *sameIndividual* e *differentIndividuals*: para especificar os mesmos indivíduos, ou diferentes indivíduos.
- Funções/recursos embutidos[63] no núcleo da SWRL– predicados especiais definidos na proposta da SWRL que pode manipular valores de dados, por exemplo para adicionar números que podem manipular valores, por exemplo para adicionar números.
- Funções/recursos embutidos customizados para a SWRL – pode-se definir seus próprios recursos embutidos usando código Java.

[63] Em inglês: *built-ins*.

O problema da indecidibilidade com a SWRL

Usando SWRL, podemos expressar mesmo o conceito *crianças de pais casados*. Entretanto, regras arbitrárias de SWRL poderão levar à indecidibilidade, em vista disso as assim chamadas regras LD-seguras (DL-safe) são implementadas em mecanismos.

Regras SWRL LD-Seguras[64] são um subconjunto restrito de regras SWRL. Estas regras possuem a propriedade desejável de decidibilidade. Decidibilidade é assegurada pela restrição de que regras devem operar somente com indivíduos conhecidos em uma ontologia OWL. Mais precisamente, todas variáveis em uma regra SWRL LD-Segura deve-se vinculara somente e indivíduos conhecidos em uma ontologia. Por razões complexas. A habilidade de vincular a indivíduos que não são conhecidos provoca indecidibilidade. Foi provado que restringindo regras para vincular somente indivíduos conhecidos assegura a decidibilidade (MOTIK, SATTLER e STUDER, 2004). Em outras palavras, LD-Seguras são regras aplicadas somente para nomear indivíduos, elas não se aplicam a indivíduos que não são nomeados, embora esses indivíduos sabidamente existam.

Pode não ser imediatamente óbvio porque variáveis em uma regra SWRL não deveriam se vincular a qualquer outra coisa que não indivíduos conhecidos. Entretanto, regras SWRL não são regras autônomas – elas são um tipo de axioma OWL e interagem com outros axiomas OWL em uma ontologia.

Considere por exemplo a seguinte regra[65]:

$$Veiculo(?v) \wedge Motor(?m) \wedge temMotor(?v, ?m) \rightarrow VeiculoMotorizado(?v)$$

que classifica um veículo como um veículo motorizado se ele tiver um motor. Claramente, esta regra irá classificar um indivíduo de classe *Veiculo* como um *VeiculoMotorizado* se ele tiver uma propriedade *temMotor* associada com um indivíduo de classe *Motor* como valor.

---

[64] Vide: http://protege.cim3.net/cgi-bin/wiki.pl?SWRLLanguageFAQ#nid9VC
[65] Vide http://protege.cim3.net/cgi-bin/wiki.pl?SWRLLanguageFAQ#nid9VF.

Assuma que nós definimos uma subclasse de *Veiculo* chamado *Carro* com a restrição associada (t*emMotoralgum Motor*) e definimos um único indivíduo dessa classe em nossa ontologia. Desde que estabelecemos que o carro tenha um motor, nós esperaríamos que o carro individual fosse classificado como um veículo motorizado. Entretanto, desde que não há motor específico declarado para a propriedade *temMotor* na ontologia (somente a afirmativa de que tem algum motor), uma implementação LD-Segura de um raciocinador SWRL não inferiria que o carro é *VeiculoMotorizado*. Fazendo assim, isso significaria que a variável *m* na regra seja limitada a algum indivíduo que não é explicitamente conhecido.

Há muitas outras situações nas quais variáveis em regras SWRL não podem ser vinculadas a indivíduos que não são conhecidos.

Claramente, LD-Segura restringe o poder expressivo da linguagem SWRL como um todo. LD-Segura pode produzir inferências incompletas – isto é, elas podem não gerar todas as deduções que são incorporadas por uma ontologia em particular. Entretanto, quaisquer deduções que são alcançadas são formalmente parecidas. É importante notar que LD-Segura é obtida pela restrição das inferências alcançadas por um raciocinador SWRL, não pela restrição da autoria das regras, elas mesmas, regras LD-seguras parecem exatamente como regras SWRL normais.

Entretanto, se recursos incorporados são usados vinculando seus argumentos, então regras podem tornar-se indecidíveis. Considere o seguinte exemplo[66]:

> *Motorista(?d) ∧ temIdade(?d, ?Idade) ∧ swrlb:add(?novaIdade, ?Idade, 1) → temIdade(?d, ?novaIdade)*

À primeira vista, esta regra parece incrementar a idade do motorista em uma unidade. Entretanto, a inferência da SWRL é monotônica de modo que ao invés de modificar a idade do motorista, essa regra gera um número infinito de idades para um motorista, cada idade acrescida de um em relação à idade prévia. Inferência com essa regra nunca irá terminar.

---

[66] Vide http://protege.cim3.net/cgi-bin/wiki.pl?SWRLLanguageFAQ#nidA3T.

4.5 Raciocinadores em Lógica Difusa

Um raciocinador é uma espécie de compilador de ontologias: uma ontologia bem projetada pode ser compilada para checar se o significado obtido é o pretendido. Além disso, um raciocinador pode ser usado como mecanismo de consultas em pequenas ontologias.

Raciocinadores típicos incluem checagem de consistência da ontologia (isto é, se existe uma interpretação de uma ontologia possível a partir da hierarquia de conceitos e de propriedades); checagem da consistência de conceito (se um conceito primitivo pode ser recuperado); checagem de subsunção (isto é, se uma classe é um subconjunto de uma outra classe), de equivalência (se uma classe é igual a outra); checagem de instanciação (isto é, se uma asserção é logicamente implicada por uma ontologia); checagem de satisfabilidade (verificação de cada conceito definido permitir instâncias desse conceito; em outra palavras, verificação se o conhecimento combinado de **TBox** e **ABox** estão livres de contradições) e verificação da possibilidade da base de conhecimento ontológica dar suporte a consultas individuais (HAASE *et al.*, 2006, p. 9).

Além disso, há inferência de informação que não é explicitamente contida dentro da ontologia revelando, por exemplo, dependências ocultas. Isto é, de posse dessa base de conhecimento, formalizada a partir de uma teoria lógica, temos um conhecimento imediato, ou factual que é obtido diretamente a partir da observação do domínio, mas temos também conhecimento derivado, ou seja, um conhecimento obtido através de inferência sobre o conhecimento imediato disponível. O conhecimento derivado é obtido mediante a aplicação de regras previamente definidas.

A possibilidade de utilizar um raciocinador é uma das principais vantagens de se usar um formalismo lógico tal como a OWL, em especial, a OWL-DL. As implementações de raciocinadores para serem integrados às ferramentas de desenvolvimento de ontologias formais, normalmente, seguem a interface especificada pelo Grupo de Implementação de Lógica Descritiva (*Description Logic Implementation Group* - DIG). Um raciocinador que segue a especificação proposta

pelo DIG é independente da ferramenta de edição de ontologias. Dentre outras especificações, é usado o protocolo HTTP para a comunicação entre o raciocinador e a ferramenta de desenvolvimento da ontologia.

Os raciocinadores mais referenciados são o *Pellet*, FACT, FACT++, Hoolet, DLP, o *KAON*2, o RACER. O *Pellet* já vem incorporado ao *Protégé*, o *KAON*2 faz parte do projeto *KAON*2 que envolve um ambiente completo para desenvolvimento e extração de ontologias. Os demais são plug-ins que podem ser incorporados aos editores de ontologias.

Dentre os raciocinadores OWL atuais que suportam SWRL, apenas *Pellet* e *KAON*2, implementam rigorosamente regras LD-Seguras. Entretanto, as implementações de guias (*tabs*) SWRL integradas às ferramentas de desenvolvimento de ontologias, via *plug-ins*, suportam de fato a LD-Segura. Por exemplo, a implementação atual de *Jess* que suporta tanto a SWRL, quanto a SQWRL, ignora a maior parte dos axiomas OWL quando realiza inferências, efetivamente vinculando variáveis a indivíduos conhecidos no contexto definido pela ontologia. Dessa forma, a *Jess* também suporta as regras LD-Seguras.

TABELA 4.4 Raciocinadores em LD clássica, lógicas e interface DIG suportados.

Raciocinador	LD Suportada	DIG Suportado
RACER, RACERPro	$\mathcal{SHIQ}^{(D)}$	*Sim*
Pellet	$\mathcal{SROIQ}^{(D)}$	*Sim*
Fact ++	$\mathcal{SROIQ}^{(D)}$	*Sim*
*KAON*2	$\mathcal{SHIQ}$	*Sim*
HERMIT	$\mathcal{SHIQ}$	*Não*

(BOBILLO *et al.*, 2008a, p. 46)

Alguns protótipos de raciocinadores vêm sendo desenvolvidos a partir da implementação de extensões difusas dessas linguagens.

FiRE

Um deles é o **FiRe -** ***Fuzzy Reasoning Engine***, baseado em linguagem difusa $\mathcal{SHIN}$ (f_{kp}-$\mathcal{SHIN}$) restrita à semântica de Zadeh. Este raciocinador suporta nominais, captura e raciocina com conhecimento difuso, permitindo o uso de alguns construtos DL, mas ainda não suporta restrições de cardinalidade e tipos de dados. Numa versão posterior, há a intenção de incorporar a expressividade da linguagem $\mathcal{SHOIQ}$. (STOILOS *et al.*, 2006).

A interface gráfica de sua plataforma possui três componentes: o painel de edição, o painel de serviços de inferência e painel de resultados.

DeLorean

Outro é o **DeLorean que** é o primeiro raciocinador que suporta DL $\mathcal{SROIQ}$ (BOBILLO *et al.*, 2008c). *A* $\mathcal{SROIQ}$ difusa estende a $\mathcal{SROIQ}$ para casos difusos adotando conceitos que denotam conjuntos difusos de indivíduos e funções que denotam relações binárias difusas. Axiomas também são estendidos para o caso difuso. Define operadores de igualdades, seus simétricos e negação. São muito úteis como linguagens ontológicas (BOBILLO, 2008, p. 1), e a DL $\mathcal{SROIQ}^{(D)}$é atualmente quase equivalente à OWL 1.1.

Reduz raciocínio em $\mathcal{SHOIN}$ difuso sob semântica de Zadeh a raciocínio em $\mathcal{SHOIN}$ rígido considerando algumas otimizações. Como consequência, permite reutilizar linguagens e recursos clássicos (editores, ferramentas, raciocinadores).

fuzzyDL

Um terceiro raciocinador é o **fuzzyDL** (BOBILLO e STRACCIA, 2008). Suporta tipos de dados concretos tais como reais, inteiros, cadeia de caracteres, e permite a definição de conceitos com representação explícita de funções de pertinência difusas. Permite modificadores tais como muito, mais ou menos, levemente que podem ser aplicados a conjuntos difusos para mudar a função de pertinência. Recursos de expressividade: conjuntos difusos explícitos; modificadores de conceitos; tipos de dados; defusificação e adequação de padrões. Como tal é um

raciocinador $\mathcal{SHIF}^{(D)}$. Suporta semântica de Zadeh e Lukasiewicz. A TABELA 4.5 associa os raciocinadores difusos com as LD difusas.

TABELA 4.5 Raciocinadores difusos e respectivas LD suportadas.

Raciocinador	***LD Suportada***
Fire	$\mathcal{SHIN}$
Delorean	$\mathcal{SROIQ}$
fuzzyDL	$\mathcal{SHIF}^{(D)}$

Bobillo e Straccia (2008, p.7) ainda comentam as características de outros raciocinadores difusos; GURDL, GERDS e YADLR.

5 Metodologia, Modelagem e Implementação de Classes e Regras de Classificação a partir de Estudo de Caso

5.1 Introdução ao Estudo de Caso

Para apresentar a metodologia, as ferramentas utilizadas e a forma de modelagem trataremos de um caso específico. Nosso estudo de caso constitui-se na avaliação dos dados de transações envolvendo serviços prestados por empresas, e informadas pelos tomadores como tendo sede fora de Belo Horizonte. O objetivo é a detecção de indícios relevantes de fraudes nestas transações.

Fraude envolvendo o local de incidência do ISSQN

Além dos casos comuns de fraudes com documentos fiscais (não emissão de nota fiscal, nota fiscal calçada, nota fiscal paralela, etc.), onde ocorre a total omissão da informação, ou declaração incorreta do valor do serviço, ou da atividade, da alíquota, de deduções, etc. há a fraude envolvendo o local de incidência do ISSQN. Nesse tipo de fraude há algumas especificidades como veremos a seguir.

Partindo-se da declaração apresentada pela empresa tomadora sediada em Belo Horizonte, há dois municípios envolvidos. O município da origem, onde, considerando a declaração, se pressupõe que se situa o estabelecimento prestador informado, e o do destino do serviço prestado, no caso, sempre Belo Horizonte. Nesse caso, uma questão se nos apresenta de imediato: qual o local de incidência do imposto?

A Lei Complementar à Constituição Federal n° 116 de 2003 (LC 116/2003) estabelece que o serviço considera-se prestado e o imposto devido no local do estabelecimento prestador ou, na falta do estabelecimento, no local do domicílio do prestador, exceto nas hipóteses previstas nos itens I a XXII do Art. 3° da LC 116/2003. As exceções previstas baseiam-se no exercício de determinadas atividades (Vide Anexo I).

Em vista desse dispositivo legal, a análise do local da incidência do imposto, leva em consideração o **local do estabelecimento prestador**, o **local de prestação do serviço**, e a **atividade vinculada ao serviço prestado**.

O estabelecimento prestador é definido pela LC 116/2003, da seguinte forma:

> Art. 4°. Considera-se estabelecimento prestador o local onde o contribuinte desenvolva a atividade de prestar serviços, de modo permanente ou temporário, e que configure unidade econômica ou profissional, sendo irrelevantes para caracterizá-lo as denominações de sede, filial, agência, posto de atendimento, sucursal, escritório de representação ou contato ou quaisquer outras que venham a ser utilizadas.

Tal definição traz dificuldades. É vago o entendimento do que seja permanente, ou temporário, ou unidade econômica, ou profissional. O dispositivo legal, por si só, não é suficiente para tratar de modo inequívoco, por exemplo, situações envolvendo o local de incidência do ISSQN:

> "De outra sorte, o aclamado princípio da territorialidade foi ainda incapaz de solucionar questões ensejadoras de conflitos de competência ativa em face da ocorrência de fatos como:
> - Pluralidade de estabelecimentos, localizados em municípios distintos, que participam da prestação do serviço;
> - Diversidade de etapas ou sequência de serviços executados em mais de um município;
> - Execução de serviço sem a participação de estabelecimento ou de pessoas jurídicas que dispensam estabelecimento;
> - Negócios jurídicos realizados mediante a utilização de novas tecnologias de informática e telecomunicações, que tornam dificultoso precisar o local das efetivas prestações, trazendo incertezas quanto a configuração do município onde se presta o serviço ou onde se localiza o estabelecimento."(FERNANDES e GOULART, 2007, p. 8)

A consideração da **atividade vinculada à prestação do serviço** também pode trazer dificuldades, em especial, se ela for constituída por distintas atividades tipificadas e tributadas distintamente. Algumas delas, inclusive, podendo não pertencer ao campo de incidência do ISSQN. Há casos emblemáticos de atividades que envolvem longas disputas por tipificação. Por exemplo, um trator para terraplenagem fornecido juntamente com um operador (motorista) é uma locação, ou um serviço de engenharia, ou, ainda, um serviço de fornecimento de mão-de-obra? Confecção de placas é serviço, ou industrialização?

No Direito, dizem que há zonas cinzentas, ou de penumbra na legislação que produzem disputas não somente entre o fisco e os contribuintes, mas também entre diferentes entes tributantes, como disputas entre município e Estado e entre municípios. Como se não bastassem as citadas zonas de penumbra abertas pela vaguidade dos termos legais no caso apontado, tanto o local do estabelecimento prestador, quanto a atividade vinculada ao serviço prestado são objetos de fraude.

O contribuinte pode optar por estabelecer-se em outro município em vista de uma alíquota menor, fiscalização inexistente, ou ineficiente, deduções de base de cálculo, ou outros benefícios, lícitos, ou ilícitos oferecidos pela municipalidade.

Se o estabelecimento é real e os benefícios lícitos, não há o que se objetar. Entretanto, há o domicílio fiscal simulado, em que o contribuinte cria ou transfere, mediante instrumento constitutivo, o domicílio para outro município, mas a transferência de fato não ocorre. A atividade, acompanhada de toda sua infraestrutura necessária, continua sendo exercida em Belo Horizonte.

Normalmente, para este tipo de fraude concorre o próprio contribuinte e os cedentes dos endereços. Muitas vezes, os contribuintes são orientados por escritórios de contabilidade. Indo além, constata-se ser relativamente comum que escritórios de contabilidade “domiciliem” essas empresas. Também tem sido comum o surgimento de um mercado paralelo de aluguel de endereços para domicílios fiscais simulados, gerando uma renda extra para as famílias residentes desses endereços. Há ainda casos de domicílios simulados em imóveis de parentes, ou mesmo em lotes vagos e em endereços inexistentes.

Outra situação possível é quando a própria administração municipal concorre para configurar o ilícito. Alguns dos mecanismos de incentivos a implantação de empreendimentos utilizados pelos municípios são isenções, doações e cessões de terrenos. Outros benefícios envolvem a redução de alíquotas e deduções na base de cálculo do ISSQN. Esses dois últimos instrumentos associados a isenções criam um cenário de verdadeira guerra fiscal entre os municípios.

Não é incomum os municípios cobrarem direta, ou indiretamente, explícita, ou dissimuladamente alíquotas inferiores à alíquota mínima de 2%, mediante conduta de seus agentes incorrendo em atos de improbidade administrativa ou crimes de responsabilidade. Tais municípios acabam por constituírem-se em paraísos fiscais municipais.

Havendo a suspeita de fraude, sua constatação é feita mediante realização de diligência no local, entrevistas, fotografias, e lavratura de boletim de ocorrência preenchido pela polícia do município. O domicílio simulado é descaracterizando através de processo administrativo, a empresa é incluída em relação pública de

empresas descaracterizadas disponibilizadas no sítio oficial da Secretaria de Finanças. A Secretaria Municipal de Finanças mantem em seu site uma relação de empresas cujos domicílios fiscais informados como sendo de outros municípios, não foram encontrados em diligências fiscais. Esta relação está disponível em: <http://www.pbh.gov.br/bhissdigital/download/Retencao.pdf>. Uma vez incluída nesta relação, os tomadores de serviços de Belo Horizonte ficam obrigados a reter o ISSQN dos serviços prestados por estas empresas.

5.2 Metodologia e Panorâmica do Sistema de Mineração de Dados com Suporte de Ontologias

A FIGURA 5.1 apresenta uma panorâmica do uso de duas ontologias agregando suas funcionalidades como suporte à mineração de dados. A primeira ontologia atua na etapa de pré-mineração de dados incorporando o conhecimento necessário para categorizar os atributos que serão usados como entrada no processo de mineração de dados.

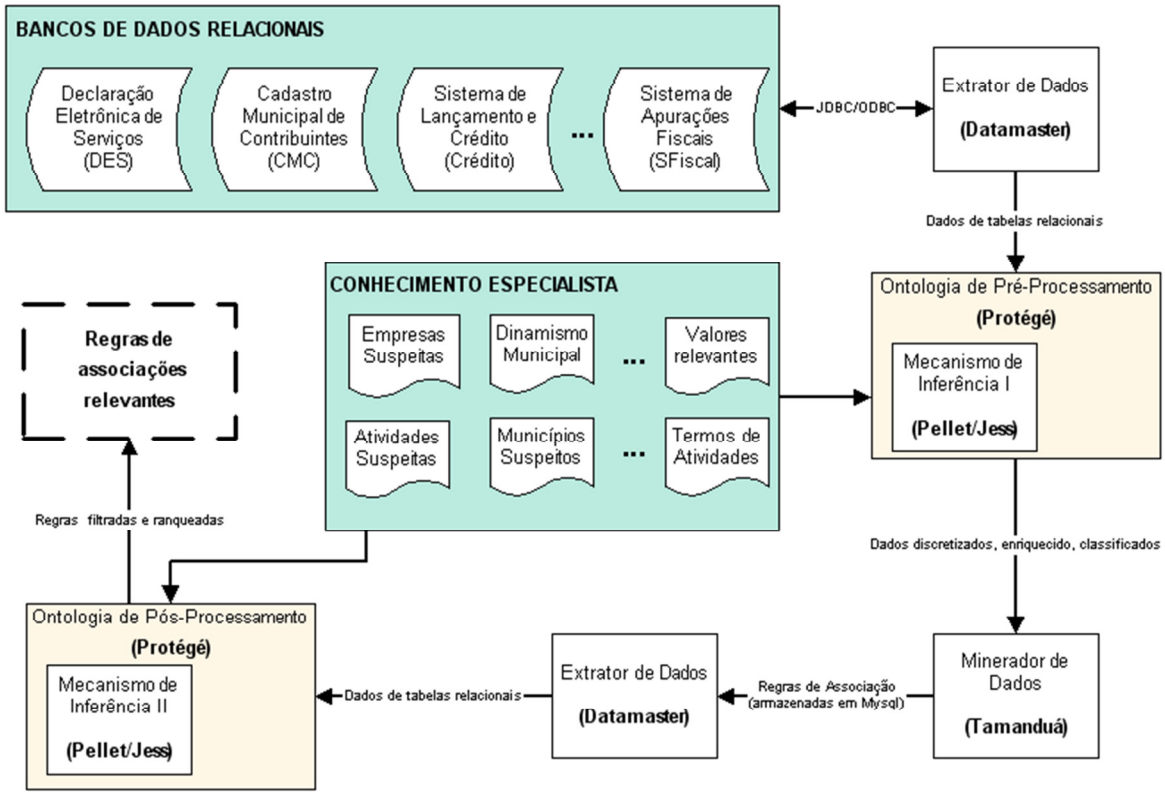

FIGURA 5.1 Panorâmica do uso de ontologias no suporte à mineração de dados.

A segunda ontologia atua na etapa de pós-mineração de dados, incorporando o conhecimento necessário para categorizar as regras geradas pela mineração.

O desenvolvimento do projeto envolve a construção dessas duas ontologias e respectivos mecanismos de inferência. Entretanto, o que distingue a solução desenvolvida é a modelagem, para fins de mineração, de atributos simples, tratados como indivisos, de modo a se obter atributo composto constituído mediante operador de agregação.

A modelagem dos atributos da mineração de dados, e da ontologia para classificá-los faz uso de noções, métodos e técnicas de sistemas difusos, e torna-se especialmente útil às organizações que lidam com conhecimentos cruciais de natureza vaga e subjetiva, típica de domínios sócio-humanos.

As ontologias explicitam e formalizam o conhecimento de natureza vaga e subjetiva gerando as categorias necessárias para a mineração de dados, e o atributo composto por agregação agiliza o processo de mineração de dados pela redução do número de atributos, e traz um nível de significação aos resultados capaz de orientar o seu processo de análise. Nesse último aspecto, esse atributo exerce a função de uma métrica subjetiva dos resultados obtidos.

5.2.1 Ontologia de Pré-mineração de Dados

A primeira ontologia atua na etapa de pré-mineração, com a preparação de dados a serem apresentados ao minerador. Na primeira ontologia, parte do conhecimento tem uma disposição mais estruturada e será capturado a partir de bancos de dados relacionais. As tabelas transferidas via conexão de dados, seja pelo esquema ODBC ou JDBC, integrarão diretamente a ontologia de pré-processamento. Nessa transferência, as tabelas dos bancos de dados serão traduzidas como classes, suas colunas como propriedades, e seus registros como instâncias de classe.

A ontologia de pré-mineração terá como domínio e escopo os serviços importados por empresas de Belo Horizonte, entendendo-se importação como sendo serviços prestados por empresas situadas em municípios distintos de Belo Horizonte, podendo tanto ser municípios brasileiros, quanto estrangeiros.

Em vista disso, incorporará classes, propriedades e instâncias relacionadas a transações realizadas entre essas empresas e declaradas pelos tomadores de serviços; termos associados a atividades; dados de municípios brasileiros; dados de empresas de fora de Belo Horizonte que prestam serviços (prestadores) a empresas de Belo Horizonte (tomadores dos serviços).

Mais especificamente, a primeira ontologia será usada para:

1. Classificar as empresas prestadoras de serviço por ramo de atividade, baseados na ocorrência de termos em suas razões sociais (Nome do Prestador).

   A única forma de extrair a informação sobre qual a possível natureza do serviço prestado é através da razão social da empresa prestadora de serviços, buscando-se por termos relacionados à determinadas atividades. Há inúmeras dificuldades nessa classificação, como a ocorrência de nomes poucos expressivos da atividade, uso de termos com erros de ortografia, termos estrangeiros, abreviaturas, prefixos, siglas, variações de sequência de termos, divisões distintas de termos, perda, ou inserção de termos, etc.
2. Classificar, categorizando para fins de mineração, os atributos rigorosos, bivalentes: *Unidade em BH*, *Multimunicipal*, *Com/Sem Retenção*, *Região Metropolitana*, *Colar Metropolitano* e *Paraíso Fiscal*.
3. Classificar, categorizando para fins de mineração, os atributos difusos, polivalentes: *Dinamismo Municipal*, *Consolidação e Tradição*, *Valor Total do Serviço*, *Percentual de ISS Retido* e *Suspeição de Transação*.

   A ferramenta de mineração de dados a ser usada, o *Tamanduá*, trabalha apenas com dados categorizados, não trabalhando com valores numéricos contínuos, tais como números reais. Em vista disso, são criadas categorias (classes) para intervalos de valores. Cada instância de valor deverá ser classificada.

   Nesse caso, pode-se atribuir um caráter difuso a esses valores, considerando-se a relevância deles. Isto é, ao invés de se atribuir nomes de valores aleatórios, ou baseados em critérios puramente estatísticos, pode-se atribuir significado aos intervalos, atribuindo-lhes rótulos linguísticos, agregando valor semântico à discretização exigida.

5.2.2 Ontologia de Pós-mineração de Dados

A ontologia de pós-mineração terá como domínio e escopo as regras obtidas no processo de mineração de dados. Ela será utilizada na classificação das regras de associação obtidas. Elas serão classificadas pelo mecanismo de inferência da ontologia com base no conhecimento especialista incorporado nesta ontologia. Mais especificamente, a segunda ontologia será usada para:

1. Classificar as regras resultantes da mineração de dados, conforme as métricas de interesse e parâmetros de sumarização.
2. Classificar as empresas, tanto tomadoras quanto prestadoras apresentam indícios mais significativos de fraude, ou sonegação do ISSQN.
3. Classificar os municípios fornecedores de serviços que apresentam indícios mais significativos de fraude, ou sonegação do ISSQN.
4. Classificar as instâncias das regras, isto é o conjunto de transações que apresentam indícios mais significativos de ocorrência de fraude, ou sonegação do ISSQN.

Cabe esclarecer que essa ontologia será instanciada com as regras resultantes da mineração de dados e suas respectivas métricas e parâmetros de sumarização. Estas regras da mineração de dados não devem ser confundidas com os axiomas e regras dos mecanismos de inferências da própria ontologia, como as regras de inferência da Seção 5.4 em SWRL. Por exemplo, uma regra resultante do processo de mineração de dados é: *Cidade-Uf → Suspeição*. A TABELA 5.1 apresenta instâncias da regra *Cidade-Uf→Suspeição*:

TABELA 5.1 Exemplos de instâncias da regra Cidade-Uf→Suspeição

Regra	**Métricas**			**Sumarizações**	
Cidade-Uf →Suspeição	**Convicção**	**Confiança**	**Suporte**	**Potencial ISS Sonegado (R$)**	**Total Transações**
Cuiaba-MT→9	1,4981	46,6667	0,1855	**55.988,64**	28
Saquarema-RJ →7	0,8875	26,3889	0,2476	**619.829,66**	57
Rio De Janeiro-RJ →7	0,7892	17,2185	0,1129	**348.039,63**	26
Nova Lima-MG →7	2,2365	70,7885	8,8532	**293.208,00**	2.038

A estas instâncias de regras estão vinculados os registros das transações cujos atributos foram efetivamente minerados. Para a instância da regra *Cidade-Uf→Suspeição*=”Cuiaba-MT→9”, temos alguns registros exibidos na TABELA 5.2.

TABELA 5.2 Exemplos de registros da instância de regra Cidade-Uf→Suspeição=”Cuiaba-MT→9”

Nome Tomador	Im	Nome Prestador	VlrTotalServ	VlrIssRetido	Suspeição
Nome_Tomador_A	1	Nome_Prestador_A	74.731,29	0,00	9
Nome_Tomador_B	2	Nome_Prestador_B	409.836,07	0,00	9
Nome_Tomador_A	1	Nome_Prestador_A	72.672,73	0,00	9
Nome_Tomador_A	1	Nome_Prestador_A	79.682,03	0,00	9
Nome_Tomador_A	1	Nome_Prestador_A	135.591,21	0,00	9
Nome_Tomador_A	1	Nome_Prestador_A	51.781,07	0,00	9
Nome_Tomador_A	1	Nome_Prestador_A	43.482,36	0,00	9
Nome_Tomador_A	1	Nome_Prestador_A	52.656,50	0,00	9
Nome_Tomador_A	1	Nome_Prestador_A	69.841,25	0,00	9
Nome_Tomador_A	1	Nome_Prestador_A	42.183,06	0,00	9
Nome_Tomador_A	1	Nome_Prestador_A	40.082,13	0,00	9
Nome_Tomador_B	2	Nome_Prestador_B	109.289,62	0,00	9
Nome_Tomador_C	3	Nome_Prestador_C	50.000,00	0,00	9
Nome_Tomador_B	2	Nome_Prestador_B	218.579,24	0,00	9
Nome_Tomador_B	2	Nome_Prestador_B	54.644,81	0,00	9
Nome_Tomador_B	2	Nome_Prestador_B	217.957,66	0,00	9
Nome_Tomador_B	2	Nome_Prestador_B	136.612,02	0,00	9
Nome_Tomador_B	2	Nome_Prestador_B	109.289,62	0,00	9
Nome_Tomador_B	2	Nome_Prestador_B	109.289,62	0,00	9

5.2.3 Construção de Atributo Composto via Operador de Agregação

A construção de um atributo composto, como o que indique o nível de suspeição da transação, envolve a consideração de atributos simples, tomados como indivisíveis que se agregam, via operador de agregação, para constituir um atributo que possa expressar o conjunto destes operadores simples.

Os atributos simples são ponderados de modo a se quantificar sua influência sobre o nível de suspeição. Nessa ponderação considera-se conhecimento especialista coletivo, de natureza vaga e subjetiva. A captura e preparação deste conhecimento não é nosso foco de interesse. Trata-se de área bem desenvolvida com métodos consolidados e amplamente usados pelas comunidades de inteligência. As

referências a esses métodos podem ser encontradas na **seção do Capítulo 7** que trata da Análise de Inteligência e conhecimento coletivo.

Em nossas simulações esses métodos não foram utilizados em vista da não existência de um corpo de especialistas que pudesse colaborar com a pesquisa. Usou-se a opinião de um único especialista, o próprio responsável pela pesquisa, que atribuiu pesos e quantificou opiniões sobre atributos simples. Por fim, também atribuiu peso ao atributo composto.

Atributos obtidos por agregação podem exercer duas funções. Primeiro, a de conduzir o processo de análise da mineração de dados. Essa capacidade de condução da análise é obtida por focar a relevância das regras a partir dos conjuntos de regras gerados. Isto é, as regras interessantes possuem este atributo como consequente. As regras onde estes atributos aparecerem sozinhos serão priorizadas para fins de análise. Desse modo, inicia-se o processo de análise da forma mais simples e objetiva possível, e apenas em uma fase posterior, caso seja possível e necessário, amplia-se o foco de análise para consequentes que misturem estes atributos compostos com atributos simples.

Segundo, a de constituir-se em métrica de avaliação do conjunto de regras gerado. Com vimos este atributo incorpora e expressa conhecimento especialista de natureza vaga e subjetiva, e constitui-se em métrica subjetiva valiosa para avaliar o grau de interesse das regras resultantes do processo de mineração de dados.

O uso desse atributo composto como elemento de incremento da performance mineração de dados, e como métrica de resultados parece-nos inédito, e dota a solução desenvolvida de caráter de originalidade.

5.3 Modelagem de Dados

Usaremos a abordagem para modelar componentes ontológicos difusos onde for identificada a ocorrência de vaguidade, e se pudermos associar à qualificação do conceito uma quantificação.

5.3.1 Dados originalmente disponíveis

Para apurarmos indícios de fraude, inicialmente, dispomos de: “Nome do Tomador de Serviços”, “Inscrição Municipal do Tomador”, “Nome do Prestador de Serviços”, “CNPJ do Prestador”, “Cidade do Prestador”, “Unidade Federativa do Prestador de Serviços”, “Valor Total do Serviço” e “Valor do ISS Retido”. De posse desses dados, temos o desafio de extrair informações que revelem a existência de indícios das fraudes comentadas acima, seja as envolvendo os valores, a não incidência, o local de prestação de serviço, o local do estabelecimento prestador, ou o domicílio fiscal simulado.

TABELA 5.3 Dados originalmente disponíveis

N°	Dado	Funções Específicas
1.	Nome Tomador do Serviço	Identifica Tomador não univocamente. Ajuda a identificar o setor de atividade do tomador em função de termos contidos no nome.
2.	Inscrição Municipal do Tomador	Identifica Tomador univocamente.
3.	Nome Prestador do Serviço	Identifica Prestador não univocamente. Ajuda a identificar o setor de atividade do prestador em função de termos contidos no nome.
4.	CNPJ do Prestador do Serviço	Identifica Prestador univocamente.
5.	Valor do Total do Serviço	Valor efetivo do serviço, sem qualquer tipo de deduções.
6.	Valor do ISSQN Retido	Valor efetivo do ISSQN retido pelo Tomador de Serviços.
7.	Cidade do Prestador	Identifica local de origem e, talvez, prestação do serviço. Em conjunto com UF faz a identificação unívoca de localidade.
8.	Unidade Federativa da Cidade	Identifica local de origem e, talvez, prestação do serviço. Em conjunto com Cidade faz a identificação unívoca de localidade.

Nesse estudo de caso, trabalhamos com dados extraídos da Declaração Eletrônica de Serviços (DES), no caso, declarada por empresas de Belo Horizonte, cadastradas no Cadastro Municipal de Contribuintes (CMC) que tomaram serviços de empresas sediadas fora de Belo Horizonte. Para o ano de 2010 foram identificadas 1.020.890 transações referentes a serviços importados, movimentando 6,83 bilhões de reais em serviços e propiciando uma arrecadação de ISSQN de aproximadamente 41,5 milhões de reais.

5.3.2 Enriquecimento de dados

Como vimos, nos dados que nos são ofertados, não há a informação do local de prestação do serviço e atividade vinculada à prestação do serviço. Além disso, não temos dados explícitos de possíveis deduções feitas na base de cálculo do ISSQN, da efetiva alíquota aplicada, qual a razão de possível não incidência, se há imunidade, isenção, venda conjugada, ou mero transporte. Diante da paupérie dos dados, buscamos meios de enriquecê-los das seguintes formas:

1. Empresa com ***Unidade em BH***: cruzando os dados de serviços importados com os dados do cadastro, adotando-se como critério de cruzamento o CNPJ Raiz, criou-se um campo informando se a empresa prestadora de serviços possui unidade em BH.
2. Empresa ***Multimunicipal***: criou-se um campo informando se uma empresa prestadora de serviços atua em mais de uma cidade (se ela possui atuação multimunicipal), com base na comparação de dados de transações de uma mesma empresa. É de se esperar que um tomador de serviços tome serviços de uma mesma unidade de um determinado prestador. Se ele declara a contratação de serviços de várias unidades de uma mesma empresa, este fato é indício de que a empresa prestadora possa estar realizando ajustes contábeis que, normalmente, são indícios de possibilidade de ocorrência de irregularidades tributárias não somente em vista do ISSQN, mas também de outros tributos. Por exemplo, uma empresa pode deixar de declarar que um serviço é obra para conseguir amenizar, ou evadir-se de obrigações trabalhistas e previdenciárias, tendo como efeito colateral, a alteração do local de incidência do ISSQN, em detrimento do município de Belo Horizonte.
3. Empresa ***Com/Sem Retenção***: informa se uma empresa prestadora de serviços teve serviços onde ocorreu a retenção, e outros em que não ocorreu a retenção, com base na comparação de dados de transações de uma mesma empresa. É um indício de irregularidade, pois é de se esperar que uma mesma empresa realize serviços com mesma atividade e, essa mesma atividade será sempre sujeita à retenção, ou não.
4. Município da ***Região Metropolitana***: informa se a cidade do prestador pertence à Região Metropolitana e, portanto, está mais sujeita à ocorrência de simulação de domicílio fiscal.

5. Município do ***Colar Metropolitano***: informa se a cidade do prestador pertence ao Colar Metropolitano e, portanto, está mais sujeita à ocorrência de simulação de domicílio fiscal.
6. Município ***Paraíso Fiscal***: informa se a cidade do prestador pertence à grupo de municípios suspeitos de serem paraísos fiscais, apurados com base na informação de municípios que adotam legislação com grandes benefícios fiscais (alíquotas reais reduzidas; e significativas deduções de base de cálculo), incentivando indiretamente a instalação de empresas virtuais, ou possuem alta incidência de domicílios fiscais descaracterizados.
7. Nível de ***Dinamismo Municipal***: através da informação da Cidade, criou-se um campo informando o nível de dinamismo da cidade. Adotamos a classificação de desenvolvimento municipal elaborada pelo estatístico Max Diniz Cruzeiro. Este levantamento está disponível em http://www.lenderbook.com/pesquisa/index.asp. Há outros levantamentos que poderiam ter sido utilizados, como o PIB municipal apurado pelo IBGE. Entretanto, a classificação desenvolvida por este estatístico leva em consideração simultânea de quarenta e duas variáveis que, potencialmente, possuem maiores condições de expressar o dinamismo municipal desejado. Além disso, as variáveis adotadas por este estatístico poderão ser utilizadas em estudos futuros para destacar determinados aspectos nos municípios.
8. Nível de ***Consolidação e Tradição***: criou-se um campo para anos de existência da empresa, através da Inscrição Municipal do Tomador dos Serviços agregamos a data de início de atividade da empresa tomadora de serviços existente no cadastro de contribuintes, com a intenção de ordenar as empresas de acordo com sua consolidação e tradição. Criou-se um campo para anos de existência da empresa. Através da Inscrição Municipal do Tomador dos Serviços agregamos a data de início de atividade destas empresas com a intenção de classificá-las de acordo com sua consolidação e tradição. A premissa é que quanto mais consolidada e tradicional for a empresa, menor a possibilidade dela envolver-se em irregularidades de forma sistêmica e contumaz. Também se baseia na constatação de que empresas criadas com a intenção de cometer crimes contra a ordem econômica e tributária possuem curta duração. Esses fatos, ainda associados à crença de que empresas novas, pelo pouco conhecimento do sistema tributário, ou pela

menor capacidade de cumprir todas as obrigações tributárias, são mais propensas a cometer irregularidades, leva-nos a atribuir níveis decrescentes de propensão a irregularidades com o aumento da idade da empresa.

Tal índice obviamente é apenas um indicativo a mais, a ser tratado com cautela. Há casos de empresas tradicionais e consolidadas envolvidas em irregularidades abrangentes, como os casos notórios das multinacionais Parmalat e Siemens. A real utilidade ou não desse indicativo, e de outros, também sempre será objeto de avaliação dentro das metodologias e modelagens adotada.

9. Relevância de ***Valor Total de Serviços***: quanto maior o Valor Total do Serviço, maior o imposto em potencial e mais relevante para uma auditoria.
10. Relevância de ***Percentual de ISS Retido***: o valor do ISS retido em cada transação, em si, não possui significado relevante. O que importa é o percentual do ISS que foi retido, pois, somente ele irá significar o grau de possibilidade de aumento da arrecadação. Se o percentual de ISS retido já representa a alíquota máxima, não há o que esperar de acréscimo de arrecadação. Se o valor indica que a alíquota está abaixo da máxima, há a possibilidade da ocorrência de uma tipificação simulada para abaixar o valor da alíquota, ou deduções indevidas. Se a alíquota é menor do que a mínima, pode estar ocorrendo deduções indevidas da base de cálculo, ou não incidência simulada. Portanto, quanto menor o percentual de ISS retido, para cada transação, maior a possibilidade de aumentar a arrecadação, em vista de possíveis irregularidades.

5.3.3 Discretização e fusificação

No processo de enriquecimento dos dados introduzimos informações qualitativas e quantitativas em forma de novos atributos. Alguns atributos possuem um domínio de valores discretos, outros um domínio de valores contínuos.

Para fins de mineração, estes atributos precisam ser categorizados. Para fins de modelagem, eles serão divididos entre atributos rigorosos, de um lado, e atributos vagos, de outro.

Os de natureza rigorosa serão modelados com graus de pertinência bivalentes, zero ou um. Isto é, determinada instância pertence ou não pertence à classe, ou conjunto rigoroso. Tais atributos são *Unidade em BH*, *Multimunicipal*, *Com/Sem Retenção, Região Metropolitana*, *Colar Metropolitano* e *Paraíso Fiscal*. Uma determinada empresa possui, ou não possui unidade em BH; atua, ou não atua em mais de um município. Uma determinada cidade pertence, ou não pertence à Região Metropolitana; pertence, ou não pertence ao Colar Metropolitano; é, ou não é paraíso fiscal.

TABELA 5.4 Atributos/Classes rigorosas

Atributos/Classes	Valor	Categoria	Grau de Pertinência
Unidade em BH **Multimunicipal** **Com/Sem Retenção**	0	0	0
Região Metropolitana **Colar Metropolitano** **Paraíso Fiscal**	1	1	1

Os de natureza vaga serão modelados difusamente com graus de pertinência polivalentes. Aos valores que eles possuem, e às categorias em que esses valores se incluem é necessário atribuir um grau de pertinência, no domínio de zero a um, à classe difusa representada por estes atributos. Tais atributos são *Dinamismo Municipal*, *Consolidação e Tradição*, *Valor Total do Serviço*, *Percentual de ISS Retido* e *Suspeição de Transação* são polivalentes. Seu domínio é constituído de vários valores reais, ou inteiros.

Os atributos rigorosos não possuem rótulo linguístico. A discretização assume valores 0, ou 1, correspondendo respectivamente às categorias 0, ou 1, conforme a instância do atributo em questão pertence, ou não à respectiva classe. Já nos atributos difusos, a discretização é realizada atribuindo-se categorias a valores, e a fusificação é realizada atribuindo-se grau de pertinência a esses mesmos valores.

O rótulo linguístico usado nos atributos difusos aproxima a tarefa de categorização à linguagem natural. É mais natural as pessoas atribuírem rótulos linguísticos às categorias, do que defini-las diretamente de acordo com valores numéricos que

expressam graus de pertinência. Metodologicamente falando, por exemplo, apresentamos um faixa de valores de um determinado atributo a especialistas, e solicitamos a eles que busquem definir quais subfaixas de valores corresponderiam aos rótulos “Irrelevante”, “Promissor” e “Relevante”. Estes rótulos indicam a capacidade da subfaixa de valores expressar sua “força” em relação ao atributo que está sendo avaliado.

TABELA 5.5 Atributo difuso para expressar nível de ***Dinamismo Municipal***

Rótulo Linguístico	Valor Mínimo	Valor Máximo	Categoria	Grau de Pertinência
Irrisório	1	2	1	1,0
Muito Irrelevante	3	3	2	0,9
Irrelevante	4	4	3	0,8
Pouco Promissor	5	5	4	0,7
Médio	6	6	5	0,6
Promissor	7	7	6	0,5
Relevante	8	8	7	0,4
Muito Relevante	9	9	8	0,1
Notável	10	10	9	0,0

Caso seja necessário um nível de granulação maior, a fim de se obter uma maior resolução da faixa de valores, pode-se atribuir esses rótulos em duas fases. Primeiro atribuindo-se com rótulos de menor granulação e, em um segundo momento, atribuindo-se subrótulos de maior granulação a cada um dos rótulos já tratados, como na TABELA 5.5. Começamos a definir as faixas de valores para “Irrelevante”, “Promissor” e “Relevante”. Uma vez estabelecida estas faixas de valores, dentro da subfaixa de valores de “Irrelevante” buscamos definir novas subfaixas para os rótulos “Irrisório”, “Muito Irrelevante” e “Irrelevante”. Do mesmo modo, definimos as subfaixas para “Promissor”, considerando os rótulos “Pouco Promissor”, “Médio”, “Promissor”; e para o rótulo “Relevante”, as subfaixas considerando os rótulos “Relevante”, “Muito Relevante” e “Notável”.

Já o grau de pertinência é atribuído considerando-se o rótulo linguístico do atributo simples, mas tomado em relação ao atributo composto que se está construindo. Dependendo do atributo simples, ele será tomado em proporção direta, ou indireta

aos valores. Por exemplo, o atributo simples *Dinamismo Municipal* influencia o atributo composto *Suspeição da Transação* de modo inverso. Isto é, quanto maior o dinamismo do município menos é possível que a transação oriunda deste município apresente irregularidades, pois é de se esperar que municípios dinâmicos como São Paulo, possuam empresas que prestem serviços a Belo Horizonte. De outro lado, transações oriundas de municípios com economia pouco dinâmica como Cabrobó em Pernambuco terão um nível de suspeição mais acentuado.

TABELA 5.6 Atributo difuso para expressar nível de ***Consolidação e Tradição da Empresa***

Rótulo Linguístico	Valor Mínimo	Valor Máximo	Categoria	Grau de Pertinência
Não Classificada	[112	112][67]	0	0,20
Recém Criada	(0	5]	1	1,00
Jovem	(5	10]	2	0,90
Madura	(10	20]	3	0,30
Consolidada	(20	50]	4	0,00
Tradicional	(50	∞)	5	0,00

TABELA 5.7 Atributo difuso para expressar relevância de ***Valor Total de Serviços***

Rótulo Linguístico	Rótulo linguístico com maior granulação	Valor Mínimo (R$)	Valor Máximo(R$)	Categoria	Grau de Pertinência
Irrelevante	Irrisório	0,00	15,55]	1	0,1
	Muito Irrelevante	(15,55	84,94]	2	0,2
	Irrelevante	(84,94	169,00]	3	0,3
Promissor	Pouco Promissor	(169,00	533,29]	4	0,5
	Médio	(533,29	2.299,86]	5	0,6
	Promissor	(2.299,86	7.345,33]	6	0,7
Relevante	Relevante	(7.345,33	17.593,91]	7	0,8
	Muito Relevante	(17.593,91	30.000,00]	8	0,9
	Notável	(30.000,00	∞)	9	1,0

[67] No caso dessa informação não estar disponível, o cálculo da idade da empresa considerou a data de início de atividades como sendo 1900.

Os atributos nível de *Consolidação e Tradição da Empresa* e *Percentual de ISS Retido* também influenciam o atributo *Suspeição da Transação* de modo inversamente proporcional, enquanto o atributo *Valor Total de Serviços* influencia de modo diretamente proporcional.

TABELA 5.8 Atributo difuso para expressar relevância de ***Percentual de ISS Retido***

Rótulo Linguístico	Valor Mínimo	Valor Máximo	Categoria	Interpretação plausível	Grau de Pertinência
Notável	[0	0]	1	Declaração de não incidência.	1,00
Relevante	(0,1	2]	2	Alíquota abaixo da mínima. Pode ter considerado alíquota errada, ou dedução de base de cálculo.	0,90
Médio	(2	5)	3	Declarou incidência, mas há a possibilidade de ter deduzido indevidamente, ou ter declarado alíquota menor.	0,50
Irrelevante	[5	5]	4	Alíquota máxima de BH. Considerando-se o Valor Total de Serviços como estando correto, não há potencial arrecadatório.	0,00

Os atributos nível de *Consolidação e Tradição da Empresa* e *Percentual de ISS Retido* também influenciam o atributo *Suspeição da Transação* de modo inversamente proporcional, enquanto o atributo *Valor Total de Serviços* influencia de modo diretamente proporcional.

Na TABELA 5.8 consideramos necessário introduzir a coluna “Interpretação plausível” para esclarecer o rótulo linguístico atribuído.

5.3.4 Atributo, ou Classe Difusa por Agregação

Além dessa consideração desses atributos, conjuntos, ou classes de caráter rigoroso e difuso, construiremos um atributo difuso adicional que irá expressar o nível de suspeição de cada transação, e será utilizada para ordenar as regras descobertas pelo minerador de dados. Essa variável terá o nome de nível de *Suspeição de Transação* e será constituída pela agregação de atributos simples,

anteriormente citados. A TABELA 5.9 apresenta os atributos que constituirão esse atributo composto.

TABELA 5.9 Atributos que constituirão o ***Nível de Suspeição de Transação***

Variáveis Rigorosas	Variáveis Difusas
1. Região Metropolitana 2. Colar Metropolitano 3. Paraíso Fiscal 4. Unidade em BH 5. Multimunicipal 6. Com/Sem retenção	7. Dinamismo Municipal 8. Consolidação e Tradição 9. Percentual de ISS Retido 10. Valor Total Serviço

em que, para *j* de 1 a 6, temos:

$$f\big(V(j)\big) = \begin{cases} 1 & \text{se } V(j) = [1,1] \\ 0 & \text{de modo contrário} \end{cases}$$

em que, para *j* de 7 a 10, temos:

$f\big(V(j)\big) = g(j)$, onde g(j) é o grau de pertinência da variável conforme tabelas acima (respectivamente, TABELA 5.5 à TABELA 5.8

Uniões e interseções de conjuntos difusos são agregadores imediatos. Normas triangulares e conormas são monotônicas, associativas, e satisfazem as condições de limite, eles fornecem uma ampla classe de operações de agregação associativa cujos elementos neutros são iguais a 1 e 0, respectivamente.

Nós não estamos, entretanto, restritos a eles como únicas alternativas disponíveis. Por exemplo, há o caso de agregadores formados por médias, onde além da monotonicidade e satisfação de condições limites, temos Idempotência e comutatividade (PEDRYCZ e GOMIDE, 2007, p. 121). De um modo geral, define-se uma função de agregação do seguinte modo[68]:

> "Formalmente, uma operação de agregação é uma função n-ária *g*:[0,1] →[0,1], satisfazendo os seguintes requerimentos:
>
> Monotonicidade $g(x_1, x_2,..., x_n) \geq g(y_1, y_2, ..., y_n)$ *se* $x_i > y_i$

[68]Para maiores detalhes sobre agregadores difusos, além de Pedrycz e Gomide (2007), vide ainda Dubois e Prade (1985 e 2004), Drewniak e Dudziak (2007) e Dudziak (2010).

> Condições limite $g(0,0,...,0) = 0$
>
> $g(1,1,....,1) = 1$
>
> Um elemento $e \in [0,1]$ é chamado de elemento neutro da operação de agregação g, e um elemento $I \in [0,1]$ é chamando assimilador (elemento de absorção) da operação de agregação g se para cada i= 1, 2,..., n, n>=2 e para todo $x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n \in [0,1]$ nós temos
>
> 1. $g(x_1, x_2,..., x_{i-1}, e, x_{i+1, ..., } x_n) = g(x_1, x_2,..., x_{i-1}, x_{i+1, ..., } x_n)$
> 2. $g(x_1, x_2,..., x_{i-1}, I, x_{i+1, ..., } x_n) = I$
>
> (PEDRYCZ e GOMIDE, 2007, p. 121)

Além de operadores de agregação básicos, ainda podemos desenvolver operadores para fins específicos, em tarefas de auxílio à tomada de decisão, como os apresentados em Merigó (2011), Ghosh, Meher e Shankar (2008), Takaci (2003) e Beliakov e Warren (2001). Selecionamos e apresentamos a seguir um operador que possui a vantagem de agregar distintos conjuntos difusos e ainda agregar as ponderações feitas por distintas pessoas, permitindo a consideração de opiniões distintas que, em conjunto, propiciam uma melhor expressão do conhecimento coletivo.

> **Operador MP-FUAO.** Seja Ψ o conjunto de números difusos. Um operador **MP-FUAO** (*multi person - fuzzy unified aggregation operator*) é um mapeamento MP-UAO: $\Psi^q \times \Psi^n \rightarrow \Psi$ que tem um vetor de pesos Z de dimensão q com $\sum_{k=1}^{q} z_k = 1$ e $z_k \in [0,1]$ e m vetores de pesos W de dimensão n com $\sum_{i=1}^{n} W_i^h = 1$ e $W_i^h \in [0,1]$, tais que:
>
> $$MPF\_UAO\left(\left(\tilde{a}_1^1, \ldots, \tilde{a}_1^q\right), \ldots, \left(\tilde{a}_n^1, \ldots, \tilde{a}_n^q\right)\right) = \sum_{h=1}^{m} \sum_{i=1}^{n} \tilde{C}_h \tilde{W}_i^h \tilde{a}_i$$
>
> onde $\tilde{C}_h$ é o grau de importância que cada conceito tem em uma agregação com $\tilde{C}_h \in [0,1]$ e $\sum_h^m \tilde{C}_h = 1$, $\tilde{a}_i = \sum_{k=1}^{q} z_k \tilde{a}_i^k$ e $\tilde{a}_i^k$ é o argumento variável fornecido por cada pessoa (ou especialista).
>
> (MERIGÒ, 2011)

Com este operador todos os atributos listados na TABELA 5.9, tanto os atributos rigorosos quanto os atributos difusos serão ponderados quanto a seu grau de influência no valor final da variável que expressa o nível de suspeição da transação. O operador acima permite que as ponderações atribuídas por várias pessoas sejam consideradas para se chegar a um valor que expresse o conhecimento de um grupo de especialistas.

Na TABELA 5.10 e TABELA 5.11 reunimos os graus de pertinência de cada atributo relacionado na TABELA 5.9. Consideramos como escopo local os graus de pertinência atribuídos aos rótulos linguísticos dos atributos 7 a 10, ou a pertinência, ou não atribuída aos atributos rigorosos de 1 a 6.

Na TABELA 5.11 consideramos como escopo global, o peso dado por um especialista à capacidade de cada atributo influenciar o atributo composto nível de *Suspeição da Transação*. Simulamos a utilização de opiniões de três especialistas (*i*: 1, 2 e 3). Inicialmente, solicitamos a eles que atribuam o nível de influência daquele atributo na faixa de valores de 1 a 0. Depois, normalizamos esses valores considerando que os pesos atribuídos por um determinado especialista deverão ter um somatório igual a 1.

TABELA 5.10 Atributos que constituirão o atributo composto *Suspeição de Transação* e respectivos pesos em escopo de relevância local.

Relevância em escopo local: *L(k)*										
Relevância local k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	0 ou 1	0 ou 1	0 ou 1	0 ou 1	0 ou 1	0 ou 1	*g(i)*	*g(i)*	*g(i)*	*g(i)*

TABELA 5.11 Atributos que constituirão o atributo composto Suspeição de Transação e respectivos pesos em escopo de relevância global, considerando-se as opiniões de três especialistas (*i*: 1, 2, 3).

Relevância em escopo global: *G(k)*										
Relevância global k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
i=1	**0,6**	**0,5**	**0,2**	**0,7**	**0,2**	**0,7**	**1,0**	**1,0**	**0,7**	**0,2**
i=1 normalizado	0,1034	0,0862	0,0345	0,1207	0,0345	0,1207	0,1724	0,1724	0,1207	0,0345
i=2	**0,4**	**0,3**	**0,1**	**0,8**	**0,4**	**0,4**	**0,8**	**0,9**	**0,6**	**0,1**
i=2 normalizado	0,0833	0,0625	0,0208	0,1667	0,0833	0,0833	0,1667	0,1875	0,1250	0,0208
i=3	**0,5**	**0,3**	**0,1**	**0,9**	**0,3**	**0,5**	**0,7**	**0,9**	**0,7**	**0,1**
i=3 normalizado	0,1000	0,0600	0,0200	0,1800	0,0600	0,1000	0,1400	0,1800	0,1400	0,0200

No processo de enriquecimento dos dados introduzimos informações qualitativas e quantitativas que irão colaborar para se classificar as transações em função de um nível de suspeição de transação. Os valores em si associados a estes atributos possuem um elevado grau de objetividade. Entretanto, a atribuição de um rótulo linguístico, os limites de suas faixas de valores, e os respectivos graus de pertinência de instâncias associadas a este rótulo envolvem a incorporação de um alto nível de vaguidade e subjetividade à modelagem desses atributos.

Ao se agregar esses atributos simples em um atributo composto, mais tratamento de subjetividade e vaguidade é incorporada à modelagem. Isso ocorre com a atribuição dos pesos com que esses atributos interferem globalmente na determinação do nível de suspeição da transação.

Assim, esse novo atributo Suspeição de Transação é intrinsecamente difuso, possuindo uma faixa de valores que também serão modelados atribuindo-se rótulos linguísticos e, para fins de mineração, uma categoria, conforme a TABELA 5.12.

TABELA 5.12 Atributo difuso para expressar nível de ***Suspeição de Transação***

Rótulo Linguístico	Rótulo linguístico com maior granulação	Valor Mínimo	Valor Máximo	Categoria
Irrelevante	Irrisório	[0,0000	0,3077]	1
	Irrelevante	(0,3077	0,3376]	3
Promissor	Pouco Promissor	(0,3376	0,3652]	4
	Médio	(0,3652	0,3893]	5
	Promissor	(0,3893	0,4285]	6
Relevante	Relevante	(0,4285	0,5661]	7
	Notável	(0,5661	1,0000]	9

5.4 Implementação de Classes no *Protégé* e Mecanismo de Inferência em SWRL

Em um primeiro momento, o que precisamos é construir a ontologia para incorporar esse conhecimento especialista e coletivo das classes e propriedades rigorosas e difusas. Com esse conhecimento incorporado na ontologia, podemos utilizá-la para checar a consistência da definição das classes, tornando-o compartilhável e divulgável. A construção da ontologia parte da definição da estrutura de classes, da definição de como classes se relacionam, quais são as propriedades das classes, os tipos dessas propriedades, as restrições e relacionamentos possíveis entre essas propriedades.

Na TABELA 5.13 relacionamos os dados originais e os dados adicionados destacando os que serão usados na mineração. Dos atributos que possuem categorias, os correspondentes aos itens 6, 7 e 8 já são classificados na origem, tendo seus valores originais zero, ou um, e não demandam a criação de classes na ontologia.

TABELA 5.13 Atributos resultantes da modelagem

N°	Campo	Campos a serem minerados	Categorias
1.	Nome do Tomador		
2.	Inscrição Municipal do Tomador	✓	
3.	Nome do Prestador		
4.	Cnpj do Prestador	✓	
5.	Cidade-Unidade Federativa	✓	
6.	**Unidade em BH**		0/1
7.	**Com/SemRetencao**		0/1
8.	**Multimunicipal**		0/1
9.	**Regiao Metropolitana**		0/1
10.	**Colar Metropolitano**		0/1
11.	**Paraíso Fiscal**		0/1
12.	**Dinamismo**		1/2/3/4/5/6/7/8/9
13.	**Consolidação e Tradição**		0/1/2/3/4/5/6
14.	**Valor Total dos Serviços**		1/2/3/4/5/6/7/8/9
15.	**Percentual do Valor do ISS Retido**		1/2/3/4
16.	Suspeição de transação	✓	1/2/3/4/5/6/7/8/9

Os correspondentes aos itens 9 a 15 demandam a criação de classes. Destes, dos itens 9 a 11 correspondem a classes rigorosas, e dos itens 12 a 15 a classes difusas.

O correspondente ao item 16 é uma classe difusa construída por agregação, como vimos na Seção 5.3.4.

As classes difusas possuem subclasses correspondentes aos seus rótulos linguísticos. Cada rótulo linguístico têm propriedades de valores mínimos e máximos que restringem seu domínio, assim como de categoria e grau de pertinência.

Em um segundo momento, precisamos instanciar a ontologia e ativarmos seus mecanismos de inferência para obtermos a classificação dessas instâncias.

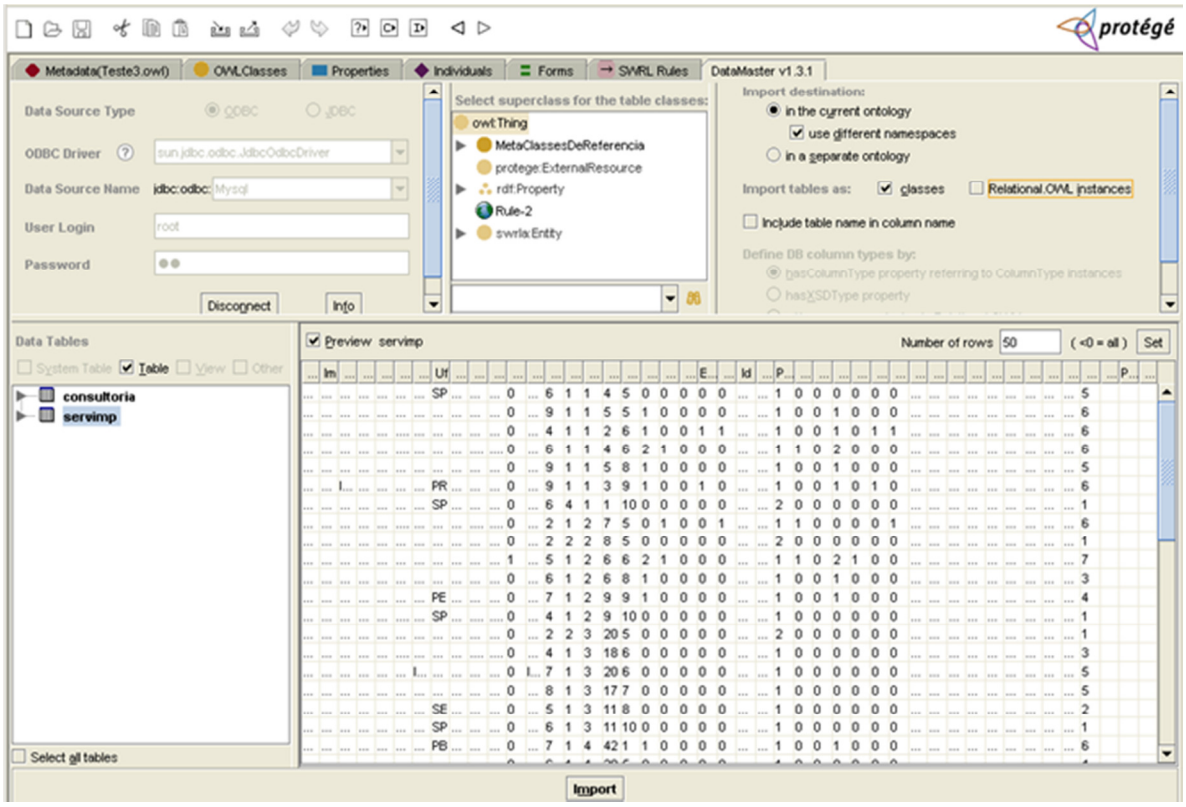

FIGURA 2.2 Tela do *Datamaster* para instanciação de ontologia com registros da tabela “servimp”, a partir de abertura de conexão ODBD com banco de dados *Mysql*.

A FIGURA 2.2 FIGURA 5.3 mostra as propriedades de uma instância inserida na ontologia através do *Datamaster*. Observe que “db:” corresponde à classe criada automaticamente pelo *Datamaster* dentro do *Protégé*.

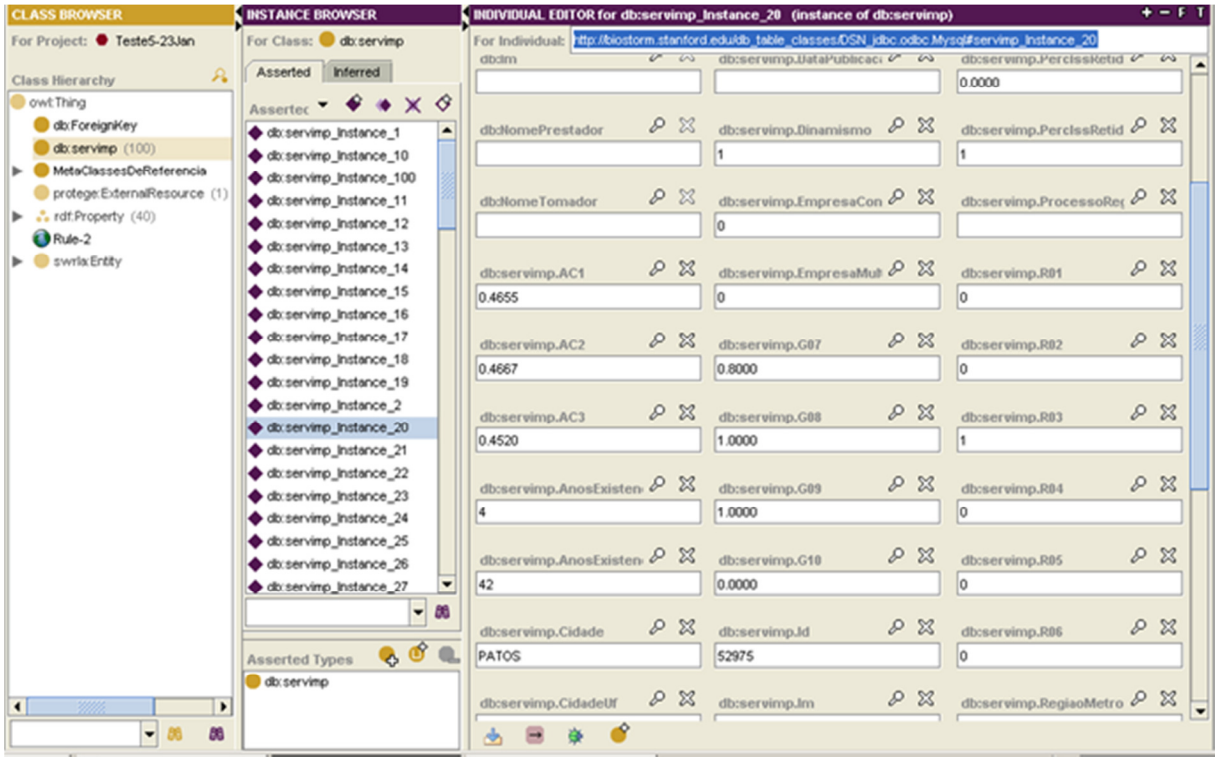

FIGURA 5.3 Instâncias de serviços importados migrados através do *Datamaster*.

Consideremos a seguinte tabela do banco de dados do Tamanduá. Essa tabela será instanciada na ontologia através do *Datamaster*. A tabela corresponderá a uma classe, suas colunas a propriedades da classe, e seus registros a instâncias.

```
CREATE TABLE  `tamandua_ds`.`table__servimp_servimp` (
  `NomeTomador` varchar(120) default NULL,
  `Im` char(10) default NULL,
  `NomePrestador` varchar(120) default NULL,
  `Cnpj` varchar(14) default NULL,
  `VlrTotalServ` char(18) default NULL,
  `VlrIssRetido` char(18) default NULL,
  `Cidade` varchar(32) default NULL,
  `Uf` char(2) default NULL,
  `UnidadeEmBh` char(1) default NULL,
  `CidadeUf` varchar(120) default NULL,
  `VlrTotalServDiscr` char(1) default NULL,
  `VlrIssRetidoDiscr` char(1) default NULL,
  `AnosExistenciaDiscr` char(1) default NULL,
  `AnosExistenciaInt` char(3) default NULL,
  `Dinamismo` char(2) default NULL,
  `DinamismoDiscr` char(1) default NULL,
```

```
`MunicipioParaisoFiscal` char(1) default NULL,
`RegiaoMetropolitana` char(1) default NULL,
`ColarMetropolitano` char(1) default NULL,
`EmpresaMultimunicipalDiscr` char(1) default NULL,
`EmpresaComSemRetencaoDiscr` char(1) default NULL,
`PercIssRetido` char(6) default NULL, --PercIssRetido (VlrIssRetido/VlrTotalServ)
`PercIssRetidoDiscr` char(1) default NULL,
`R01` char(1) default NULL, -- Grau de pertinência para Região Metropolitana
`R02` char(1) default NULL, -- Grau de pertinência para Colar Metropolitano
`R03` char(1) default NULL, -- Grau de pertinência para Município Paraíso Fiscal
`R04` char(1) default NULL, -- Grau de pertinência para Unidade em BH
`R05` char(1) default NULL, -- Grau de pertinência para Empresa multimunicipal
`R06` char(1) default NULL, -- Grau de pertinência para Com e sem retenção
`G07` char(6) default NULL, -- Grau de pertinência de VlrTotalServDiscr
`G08` char(6) default NULL, -- Grau de pertinência de PercIssRetido
`G09` char(6) default NULL, -- Grau de pertinência de Dinamismo
`G10` char(6) default NULL, -- Grau de pertinência de Consolidação e Tradição
`AC1` char(6) default NULL, -- Acumulador opinião especialista 1
`AC2` char(6) default NULL, -- Acumulador opinião especialista 2
`AC3` char(6) default NULL, -- Acumulador opinião especialista 3
`NivSuspeicaoTrans` char(6) default NULL, -- Nivel de Suspeicao da Transação considerando
                                          -- a opinião dos especialistas (AC1+AC2+AC3)
`NivSuspeicaoTransNorm` char(6) default NULL, -- Nível de Suspeição Normalizado
`NivSuspeicaoTransNormDiscr` char(1) default NULL, -- Suspeição Discretizado (categorizado)
`DataPublicacao` char(10) default NULL,
`ProcessoRegularizacao` char(15) default NULL,
) ENGINE=MyISAM DEFAULT CHARSET=latin1
```

As classes definidas na ontologia e suas propriedades podem ser utilizadas para classificar estas instâncias, checando a consistência dos dados em relação aos conceitos incorporados na ontologia. Já as regras em SWRL atribuem valores para os atributos das instâncias correspondentes às categorias e graus de pertinência dos atributos, da mesma instância, que serão agregados para compor o atributo composto. Também usaremos regras em SWRL para calcular o valor deste atributo composto e, para categorizar esse valor calculado. Também servirá para implementar o operador de agregação.Deste modo, valores relacionados direta, ou indiretamente (a serem agregados) a atributos a serem usados na mineração serão consistidos e categorizados.

5.4.1 Classes Rigorosas, Extensivas e Definidas por Enumeração

Na TABELA 5.4 trabalhamos com atributos, conceitos rigorosos, isto é uma determinada instância está totalmente dentro ou fora do conceito. Por exemplo, a

instância “Nova Lima” está na classe *Região Metropolitana*, mas não está na classe *Colar Metropolitano*, vide Anexo III.1 e Anexo III.2, respectivamente.

Em vista disso, associamos o grau de pertinência “1” da instância “Nova Lima” à classe *Região Metropolitana*, e grau de pertinência “0” da instância “Nova Lima” ao conceito *Colar Metropolitano*.

Classes dessa natureza podem ser definidas extensivamente, simplesmente, enumerando-se seus elementos constitutivos. Da mesma forma definimos a classe de municípios *Paraísos Fiscais* (Anexo III.3).

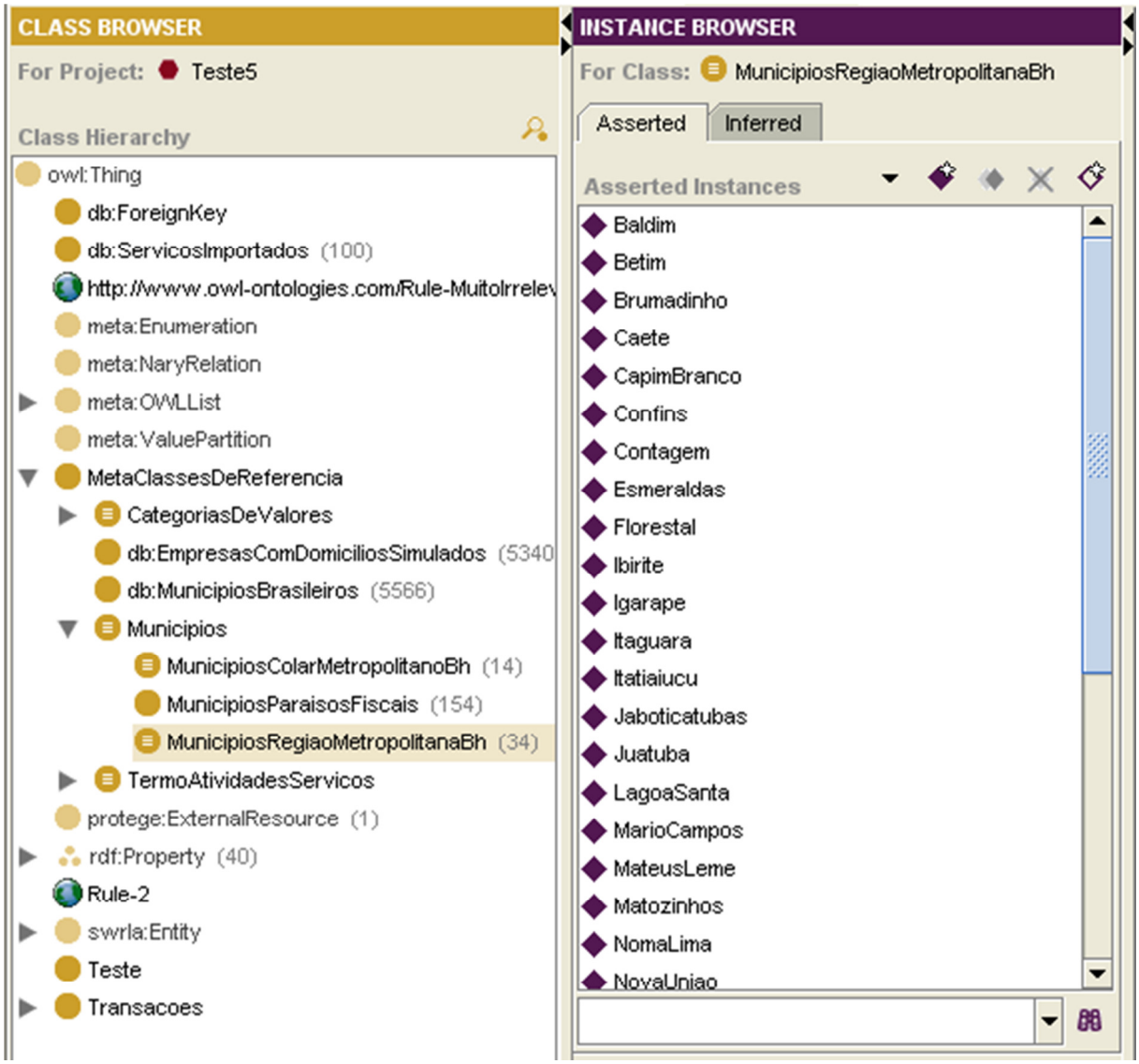

FIGURA 5.4 Classe extensiva, construída por enumeração no *Protégé* 3.4.6, que define os municípios da Região Metropolitana de Belo Horizonte.

Regras para classificação como município *Região Metropolitana*

De acordo com a TABELA 5.4 e Anexo III.1

1. db:ServImp(?p) ∧ db:Cidade(?p, ?a) ∧ MunicipiosRegiaoMetropolitana(?a) → db:RegiaoMetropolitana(?p, "1") ∧ db:R01(?p,"1.0") ∧ sqwrl:select(?p, ?a)[69]

Como exemplo do funcionamento desta regra SWRL, consideremos que uma determinada instância p retorna para a cláusula Cidade(?p, ?a), a="JUIZ DE FORA". Para a cláusula MunicipiosRegiaoMetropolitana(?a), agora sabendo que a="JUIZ DE FORA", temos o retorno "0". Logo, o consequente será falso e não retornará que a instância p refere-se a uma transação oriunda de uma cidade da Região Metropolitana. Logo, não haverá atribuição de valores para db:RegiaoMetropolitana e db:R01.

Caso tivéssemos como sendo um dos municípios relacionados no Anexo III.1, como a="BETIM", por exemplo, a regra retornaria que a instância p refere-se a uma transação oriunda de uma cidade da Região Metropolitana e atribuiria o valor "1" a db:RegiaoMetropolitana, e o valor "1.0" a atributo db:R01.

São similares as regras para avaliar se a instância p refere-se a uma transação oriunda de cidade pertencente ao Colar Metropolitano, ou é se oriunda de cidade Paraíso Fiscal.

Regras para classificação como município *Colar Metropolitano*

De acordo com a TABELA 5.4 e Anexo III.2

2. db:ServImp(?p) ∧ db:Cidade(?p, ?a) ∧ MunicipiosColarMetropolitano(?a) → db:ColarMetropolitano(?p,"1") ∧ db:R02(?p,"1.0") ∧ sqwrl:select(?p, ?a)

Regras para classificação como municípios *Paraíso Fiscal*

De acordo com a TABELA 5.4 e Anexo III.3

3. db:ServImp(?p) ∧ db:Cidade (?p, ?a) ∧ MunicipiosParaisosFiscais(?a) → db:ParaisoFiscal(?p, "1") ∧ db:R03(?p,"1.0") ∧ sqwrl:select(?p)

---

[69] O comando sqwrl:select não tem efeitos para fins de classificação. Colocamo-lo no consequente para fins de conferência da correção da regra, por ele gerar em tela a seleção realizada pelo antecedente.

As três regras a seguir diferem-se da estrutura das anteriores. pois, os atributos db:UnidadeEmBH, db:EmpresaMultimunicipal, db:ComSemRetencao já vem preenchidos do banco de dados. Esse preenchimento é realizado a partir de cruzamentos entre tabelas cadastrais. O conteúdos destes atributos serão utilizados para atribuir os valores aos graus de pertinência db:R04, db:R05, db:R06, respectivamente, de acordo com a TABELA 5.4.

4. db:ServImp(?p) ∧ db:UnidadeEmBh(?a) → db:R04(?p,?a) ∧ sqwrl:select(?p)

5. db:ServImp(?p) ∧ db:EmpresaMultimunicipal(?a) → db:R05(?p,?a) ∧ sqwrl:select(?p)

6. db:ServImp(?p) ∧ db:ComSemRetencao(?a) → db:R06(?p,?a) ∧ sqwrl:select(?p)

Nestas regras, os atributos db:R04, db:R05 e db:R06, correspondentes aos graus de pertinência, coincidentemente possuem os mesmos valores correspondentes às categorias de db:UnidadeEmBh, db:EmpresaMultimunicipal e db:ComSemRetencao. Estas regras permitem a atribuição destes graus de pertinência, considerando-se os valores já atribuídos às categorias.

A seguir apresentamos o desenvolvimento de classes difusas que irão utilizar as informações incorporadas até o momento na ontologia a fim de poder classificar transações, empresas e municípios em função de vagos conceitos de suspeição.

5.4.2 Classes Difusas

A modelagem apresentada nas TABELA 5.5 à TABELA 5.8 com atributos (conceitos, classes) difusos, e seus respectivos rótulos linguísticos acompanhados de seus respectivos graus de pertinência, adequam-se à ilustração da definição de ontologia difusa apresentada em Calegari e Ciucci (2008) apresentada na Seção 4.2.

> **Definição 1**. Uma ontologia difusa é uma ontologia estendida com valores difusos que são atribuídos mediante duas funções:
> g: (*Conceitos*∪*Instâncias*) × (*Propriedades*∪ Prop_val) → [0,1] e
> h: *Conceitos*∪*Instâncias* → [0, 1]

Nessa definição, a primeira função proposta, em vista do uso de rótulos linguísticos, adequa-se à nossa modelagem a fim de aprimorar o grau de compreensão e revelar o nível de relevância das regras a serem obtidas no processo de mineração de dados. Essa função é definida da seguinte forma:

$$g: (Conceitos \cup Instâncias) \times (Propriedades \cup Prop\_val) \rightarrow [0,1]$$

As variáveis, ou conceitos difusos que usaremos para ilustrar a aplicações dessa função serão: *Dinamismo*; *Consolidação e Tradição*, *Valor Total de Serviços* e *Percentual de ISS Retido*.

Por exemplo, consideremos na TABELA 5.5 os rótulos linguísticos associados ao atributo difuso *Dinamismo*. Consideremos, conforme essa definição, a associação entre o conceito *Dinamismo* e a instância "Belo Horizonte". Conforme a tabela do Anexo IV, Belo Horizonte está categorizada como uma cidade com *Dinamismo* **9**, tendo seu grau de pertinência ao conjunto difuso indicado pela variável linguística "Muito Relevante" que, por sua vez, possui o valor **0,9**. Usando a formalização acima, temos:

> *g(Dinamismo ∪ Belo Horizonte, temDinamismo ∪ temValorDoDinamismo) = g(Dinamismo de Belo Horizonte, Muito Relevante*) = **0,1**

A seguir apresentamos regras desenvolvidas em SWRL para atribuir as instâncias importadas às classes da ontologia.

Para cada atributo a ser agregado, ou a ser minerado é criado uma classe no Protégé. Para cada uma dessas classes são criadas subclasses no Protégé correspondentes aos rótulos linguísticos definidos para subfaixas de valores.

Para cada uma dessas subclasses definidas conforme o rotulo linguístico, associamos um atributo referente à categoria (temCategoria), ao grau de pertinência (temGrauPertinencia), um atributo referente ao valor inferior (temValorInferior) e um atributo referente ao valor superior (temValorSuperior).

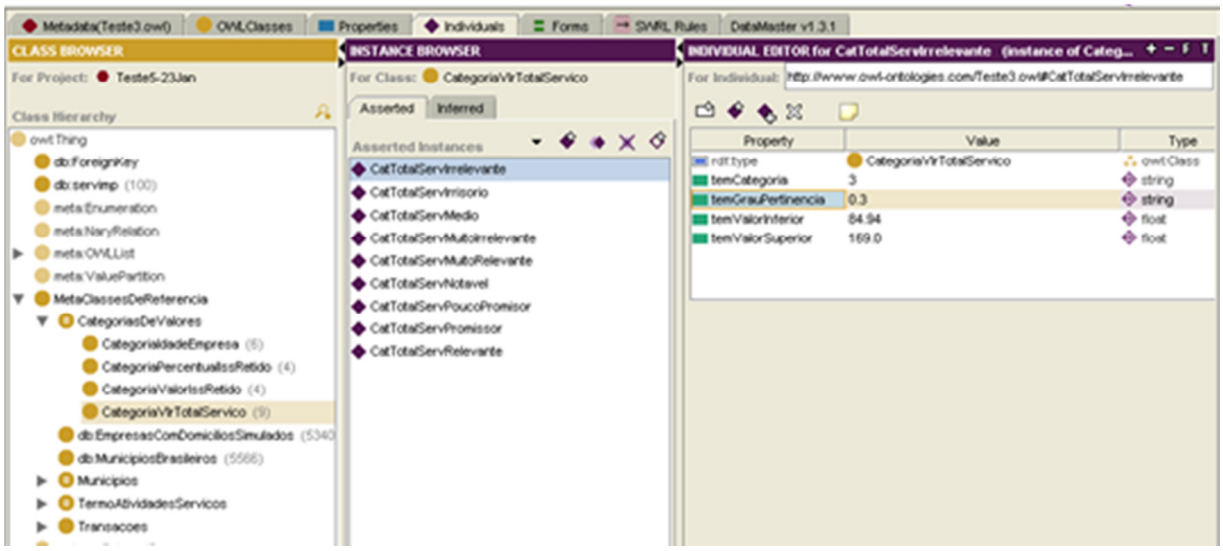

FIGURA 5.5 Classes usadas para classificar instâncias de *Valores Totais de Serviços.*

As regras para atribuir o grau de pertinência às instâncias verificam se o valor do atributo está entre os valores inferior e superior da faixa correspondente a cada um dos rótulos, e, caso positivo, retornam com o valor de pertinência e categoria.

Regras em SWRL também atribuirão valores para o atributo *Suspeição da Transação* através da implementação da função de agregação por regras SWRL.

Regras para Classificação como *Valor Total Dos Serviços*

De acordo com a TABELA 5.7.

Consideremos, por exemplo, que uma instância p possui o valor R$1200,00 para Valor Total dos Serviços (VlrTotalServ), isto é para a cláusula db:VlrTotalServ(?p.?a), teremos *a*=1200.00. Primeiro testaremos a pertinência deste valor para cada uma das subclasses da TABELA 5.7 (Irrissório, Muito Irrelevante,..., Notável). Este teste é feito recuperando-se os limites máximos e mínimos destas subclasses através da propriedades temValorInferior (variável *b*), e temValorSuperior (variável *c*). Testamos se a variável a é maior, ou igual ao valor mínimo *b* e menor que o valor máximo *c*. Para o antecedente que for verdadeiro, atribuiremos o grau de pertinência, e a categoria correspondente à subclasse que foi determinada.

Este raciocínio vigora para os quatro conjuntos de regras a seguir (regras de 7 a 34) referentes às atribuições de valores de categoria e grau de pertinência a *Valor Total*

dos Serviços, *Percentual do ISSQN Retido, Dinamismo*, e *Consolidação e Tradição do Tomador de Serviços*.

7. db:ServImp(?p) ∧ db:VlrTotalServ(?p, ?a) ∧ temCategoria(CatVlrTotalServIrrisorio, ?b) ∧ temGrauPertinencia(CatVlrTotalServIrrisorio, ?c) ∧ temValorInferior(CatVlrTotalServIrrisorio, ?d) ∧ temValorSuperior(CatVlrTotalServIrrisorio, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:VlrTotalServDiscr(?p, ?b") ∧ db:G07(?p, ?c) ∧ sqwrl:select(?p, ?a, ?b, ?c)

8. db:ServImp(?p) ∧ db:VlrTotalServ(?p, ?a) ∧ temCategoria(CatVlrTotalServMuitoIrrelevante, ?b) ∧ temGrauPertinencia(CatVlrTotalServMuitoIrrelevante, ?c) ∧ temValorInferior(CatVlrTotalServMuitoIrrelevante, ?d) ∧ temValorSuperior(CatVlrTotalServMuitoIrrelevante, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:VlrTotalServDiscr(?p, ?b) ∧ db:G07(?p, ?c) ∧ sqwrl:select(?p)

9. db:ServImp(?p) ∧ db:VlrTotalServ(?p, ?a) ∧ temCategoria(CatVlrTotalServIrrelevante, ?b) ∧ temGrauPertinencia(CatVlrTotalServIrrelevante, ?c) ∧ temValorInferior(CatVlrTotalServIrrelevante, ?d) ∧ temValorSuperior(CatVlrTotalServIrrelevante, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:VlrTotalServDiscr(?p, ?b) ∧ db:G07(?p, ?c) ∧ sqwrl:select(?p)

10. db:ServImp(?p) ∧ db:VlrTotalServ(?p, ?a) ∧ temCategoria(CatVlrTotalServPoucoPromisor, ?b) ∧ temGrauPertinencia(CatVlrTotalServPoucoPromisor, ?c) ∧ temValorInferior(CatVlrTotalServPoucoPromisor, ?d) ∧ temValorSuperior(CatVlrTotalServPoucoPromisor, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:VlrTotalServDiscr(?p, ?b) ∧ db:G07(?p, ?c) ∧ sqwrl:select(?p)

11. db:ServImp(?p) ∧ db:VlrTotalServ(?p, ?a) ∧ temCategoria(CatVlrTotalServMedio, ?b) ∧ temGrauPertinencia(CatVlrTotalServMedio, ?c) ∧ temValorInferior(CatVlrTotalServMedio, ?d) ∧ temValorSuperior(CatVlrTotalServMedio, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:VlrTotalServDiscr(?p, ?b) ∧ db:G07(?p, ?c) ∧ sqwrl:select(?p)

12. db:ServImp(?p) ∧ db:VlrTotalServ(?p, ?a) ∧ temCategoria(CatVlrTotalServPromissor, ?b) ∧ temGrauPertinencia(CatVlrTotalServPromissor, ?c) ∧ temValorInferior(CatVlrTotalServPromissor, ?d) ∧ temValorSuperior(CatVlrTotalServPromissor, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:VlrTotalServDiscr(?p, ?b) ∧ db:G07(?p, ?c) ∧ sqwrl:select(?p)

13. db:ServImp(?p) ∧ db:VlrTotalServ(?p, ?a) ∧ temCategoria(CatVlrTotalServRelevante, ?b) ∧ temGrauPertinencia(CatVlrTotalServRelevante, ?c) ∧ temValorInferior(CatVlrTotalServRelevante, ?d) ∧ temValorSuperior(CatVlrTotalServRelevante, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:VlrTotalServDiscr(?p, ?b) ∧ db:G07(?p, ?c) ∧ sqwrl:select(?p)

14. db:ServImp(?p) ∧ db:VlrTotalServ(?p, ?a) ∧ temCategoria(CatVlrTotalServMuitoRelevante, ?b) ∧ temGrauPertinencia(CatVlrTotalServMuitoRelevante, ?c) ∧ temValorInferior(CatVlrTotalServMuitoRelevante, ?d) ∧ temValorSuperior(CatVlrTotalServMuitoRelevante, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:VlrTotalServDiscr(?p, ?b) ∧ db:GrauTotalServ(?p, ?c) ∧ sqwrl:select(?p)

15. db:ServImp(?p) ∧ db:VlrTotalServ(?p, ?a) ∧ temCategoria(CatVlrTotalServNotavel, ?b) ∧ temGrauPertinencia(CatVlrTotalServNotavel, ?c) ∧ temValorInferior(CatVlrTotalServNotavel, ?d) ∧ temValorSuperior(CatVlrTotalServNotavel, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:VlrTotalServDiscr(?p, ?b) ∧ db:GrauTotalServ(?p, ?c") ∧ sqwrl:select(?p)

Regras para Classificação como *Percentual do Issqn Retido*

De acordo com a TABELA 5.8.

16. db:ServImp(?p) ∧ db: PercIssRetido (?p, ?a) ∧ temCategoria(CatPercIssRetNotavel, ?b) ∧ temGrauPertinencia(CatPercIssRetNotavel, ?c) ∧ temValorInferior(CatPercIssRetNotavel, ?d) ∧ temValorSuperior(CatPercIssRetNotavel, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:PercIssRetidoDiscr(?p, ?b) ∧ db:G08(?p,?c) ∧ sqwrl:select(?p)

17. db:ServImp(?p) ∧ db: PercIssRetido (?p, ?a) ∧ temCategoria(CatPercIssRetRelevante, ?b) ∧ temGrauPertinencia(CatPercIssRetRelevante, ?c) ∧ temValorInferior(CatPercIssRetRelevante, ?d) ∧ temValorSuperior(CatPercIssRetRelevante, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:PercIssRetidoDiscr(?p, ?b) ∧ db:G08(?p,?c) ∧ sqwrl:select(?p)

18. db:ServImp(?p) ∧ db: PercIssRetido(?p, ?a) ∧ temCategoria(CatPercIssRetMedio, ?b) ∧ temGrauPertinencia(CatPercIssRetMedio, ?c) ∧ temValorInferior(CatPercIssRetMedio, ?d) ∧ temValorSuperior(CatPercIssRetMedio, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:PercIssRetidoDiscr(?p, ?b) ∧ db:G08(?p,?c) ∧ sqwrl:select(?p)

19. db:ServImp(?p) ∧ db: PercIssRetido (?p, ?a) ∧ temCategoria(CatPercIssRetIrrelevante, ?b) ∧ temGrauPertinencia(CatPercIssRetIrrelevante, ?c) ∧ temValorInferior(CatPercIssRetIrrelevante, ?d) ∧ temValorSuperior(CatPercIssRetIrrelevante, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:PercIssRetidoDiscr(?p, ?b) ∧ db:G08(?p,?c) ∧ sqwrl:select(?p)

Regras para Classificação como *Dinamismo*

De acordo com a TABELA 5.5.

20. db:ServImp(?p) ∧ db:Dinamismo(?p,?a) ∧ temCategoria(CatDinamismoIrrisorio, ?b) ∧ temGrauPertinencia(CatDinamismoIrrisorio, ?c) ∧ temValorInferior(CatDinamismoIrrisorio, ?d) ∧ temValorSuperior(CatDinamismoIrrisorio, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:DinamismoDiscr(?p, ?b) ∧ db:G09(?p, ?c) ∧ sqwrl:select(?p, ?a, ?b, ?c, ?d)

21. db:ServImp(?p) ∧ db:Dinamismo(?p,?a) ∧ temCategoria(CatDinamismoMuitoIrrelevante, ?b) ∧ temGrauPertinencia(CatDinamismoMuitoIrrelevante, ?c) ∧ temValorInferior(CatDinamismoMuitoIrrelevante, ?d) ∧ temValorSuperior(CatDinamismoMuitoIrrelevante, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:DinamismoDiscr(?p, ?b) ∧ db:G09(?p, ?c) ∧ sqwrl:select(?p)

22. db:ServImp(?p) ∧ db:Dinamismo(?p,?a) ∧ temCategoria(CatDinamismoIrrelevante, ?b) temGrauPertinencia(CatDinamismoIrrelevante, ?c) ∧ temValorInferior(CatDinamismoIrrelevante, ?d) ∧ temValorSuperior(CatDinamismoIrrelevante, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:DinamismoDiscr(?p, ?b) ∧ db:G09(?p, ?c) ∧ sqwrl:select(?p)

23. db:ServImp(?p) ∧ db:Dinamismo(?p,?a) ∧ temCategoria(CatDinamismoPoucoPromissor, ?b) ∧ temGrauPertinencia(CatDinamismoPoucoPromissor, ?c) ∧ temValorInferior(CatDinamismoPoucoPromissor, ?d) ∧ temValorSuperior(CatDinamismoPoucoPromissor, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:DinamismoDiscr(?p, ?b) ∧ db:G09(?p, ?c) ∧ sqwrl:select(?p)

24. db:ServImp(?p) ∧ db:Dinamismo(?p,?a) ∧ temCategoria(CatDinamismoMedio, ?b) ∧ temGrauPertinencia(CatDinamismoMedio, ?c) ∧ temValorInferior(CatDinamismoMedio, ?d) ∧ temValorSuperior(CatDinamismoMedio, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:DinamismoDiscr(?p, ?b) ∧ db:G09(?p, ?c) ∧ sqwrl:select(?p)

25. db:ServImp(?p) ∧ db:Dinamismo(?p,?a) ∧ temCategoria(CatDinamismoPromissor, ?b) ∧ temGrauPertinencia(CatDinamismoPromissor, ?c) ∧ temValorInferior(CatDinamismoPromissor, ?d) ∧ temValorSuperior(CatDinamismoPromissor, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:DinamismoDiscr(?p, ?b) ∧ db:G09(?p, ?c) ∧ sqwrl:select(?p)

26. db:ServImp(?p) ∧ db:Dinamismo(?p,?a) ∧ temCategoria(CatDinamismoRelevante, ?b) ∧ temGrauPertinencia(CatDinamismoRelevante, ?c) ∧ temValorInferior(CatDinamismoRelevante, ?d) ∧ temValorSuperior(CatDinamismoRelevante, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:DinamismoDiscr(?p, ?b) ∧ db:G09(?p, ?c) ∧ sqwrl:select(?p)

27. db:ServImp(?p) ∧ db:Dinamismo(?p,?a) ∧ temCategoria(CatDinamismoMuitoRelevante, ?b) ∧ temGrauPertinencia(CatDinamismoMuitoRelevante, ?c) ∧ temValorInferior(CatDinamismoMuitoRelevante, ?d) ∧ temValorSuperior(CatDinamismoMuitoRelevante, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:DinamismoDiscr(?p, ?b) ∧ db:G09(?p, ?c) ∧ sqwrl:select(?p)

28. db:ServImp(?p) ∧ db:Dinamismo(?p,?a) ∧ temCategoria(CatDinamismoNotavel, ?b) ∧ temGrauPertinencia(CatDinamismoNotavel, ?c) ∧ temValorInferior(CatDinamismoNotavel, ?d) ∧ temValorSuperior(CatDinamismoNotavel, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:DinamismoDiscr(?p, ?b) ∧ db:G09(?p, ?c) ∧ sqwrl:select(?p)

Regras para classificação como nível de *Consolidação e Tradição do Tomador de Serviços*

De acordo com a TABELA 5.6.

29. db:ServImp(?p) ∧ db:AnosExistenciaInt(?p, ?a) ∧ temCategoria(CatIdadeEmpresaNaoClassificada, ?b) ∧ temGrauPertinencia(CatIdadeEmpresaNaoClassificada, ?c) ∧ temValorInferior(CatIdadeEmpresaNaoClassificada, ?d) ∧ temValorSuperior(CatIdadeEmpresaNaoClassificada, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:AnosExistenciaDiscr(?p, ?c) ∧ db:G10(?p, ?d) ∧ sqwrl:select(?p)

30. db:ServImp(?p) ∧ db:AnosExistenciaInt(?p, ?a) ∧ temCategoria(CatIdadeEmpresaRecemCriada, ?b) ∧ temGrauPertinencia(CatIdadeEmpresaRecemCriada, ?c) ∧ temValorInferior(CatIdadeEmpresaRecemCriada, ?d) ∧ temValorSuperior(CatIdadeEmpresaRecemCriada, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:AnosExistenciaDiscr(?p, ?c) ∧ db:G10(?p, ?d) ∧ sqwrl:select(?p)

31. db:ServImp(?p) ∧ db:AnosExistenciaInt(?p, ?a) ∧ temCategoria(CatIdadeEmpresaJovem, ?b) ∧ temGrauPertinencia(CatIdadeEmpresaJovem, ?c) ∧ temValorInferior(CatIdadeEmpresaJovem, ?d) ∧ temValorSuperior(CatIdadeEmpresaJovem, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:AnosExistenciaDiscr(?p, ?c) ∧ db:G10(?p, ?d) ∧ sqwrl:select(?p)

32. db:ServImp(?p) ∧ db:AnosExistenciaInt(?p, ?a) ∧ temCategoria(CatIdadeEmpresaMadura, ?b) ∧ temGrauPertinencia(CatIdadeEmpresaMadura, ?c) ∧ temValorInferior(CatIdadeEmpresaMadura, ?d) ∧ temValorSuperior(CatIdadeEmpresaMadura, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:AnosExistenciaDiscr(?p, ?c) ∧ db:G10(?p, ?d) ∧ sqwrl:select(?p)

33. db:ServImp(?p) ∧ db:AnosExistenciaInt(?p, ?a) ∧ temCategoria(CatIdadeEmpresaConsolidada, ?b) ∧ temGrauPertinencia(CatIdadeEmpresaConsolidada, ?c) ∧ temValorInferior(CatIdadeEmpresaConsolidada, ?d) ∧ temValorSuperior(CatIdadeEmpresaConsolidada, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧ swrlb:lessThan(?a, ?e) → db:AnosExistenciaDiscr(?p, ?d) ∧ db:G10(?p, ?e) ∧ sqwrl:select(?p)

34. db:ServImp(?p) ∧ db:AnosExistenciaInt(?p, ?a) ∧ temCategoria(CatIdadeEmpresaTradicional, ?b) ∧ temGrauPertinencia(CatIdadeEmpresaTradicional, ?c) ∧

temValorInferior(CatIdadeEmpresaTradicional, ?d) ∧
temValorSuperior(CatIdadeEmpresaTradicional, ?e) ∧ swrlb:greaterThanOrEqual(?a, ?d) ∧
swrlb:lessThan(?a, ?e) → db:AnosExistenciaDiscr(?p, ?d) ∧ db:G10(?p, ?e) ∧ sqwrl:select(?p)

Regras para cálculo do valor do atributo correspondente ao nível de *Suspeição da Transação*, mas para cada especialista considerado

De acordo com valores atribuídos de db:R01 até db:R10 conforme regras anteriores e pesos atribuídos por especialistas. De acordo também com a TABELA 5.11.

As três regras a seguir utilizam funções aritméticas de adição e multiplicação da SWRL para obter os resultados das funções de agregação, considerando-se os graus de pertinência atribuídos em escopo local (db:R01 a db:R10) e os pesos atribuídos por cada um dos três especialistas considerados.

Primeiro especialista, i=1.

35. db:ServImp(?p) ∧ db:R01(?p, ?k) ∧ db:R02(?p,?l) ∧ db:R03(?p,?m) ∧ db:R04(?p,?n) ∧ db:R05(?p, ?o) ∧ db:R06(?p,?q) ∧ db:R07(?p,?r) ∧ db:R08(?p, ?s) ∧ db:R09(?p, ?t) ∧ db:R10(?p, ?u) ∧ swrlb:multiply(?a, 0.1034, ?k) ∧ swrlb:multiply(?b, 0.0862, ?l) ∧ swrlb:multiply(?c, 0.0345, ?m) ∧ swrlb:multiply(?d, 0.1207, ?n) ∧ swrlb:multiply(?e, 0.0345, ?o) ∧ swrlb:multiply(?f, 0.1207, ?q) ∧ swrlb:multiply(?g, 0.1724, ?r) ∧ swrlb:multiply(?h, 0.1724, ?s) ∧ swrlb:multiply(?i, 0.1207, ?t) ∧ swrlb:multiply(?j, 0.0345, ?u) ∧ swrlb:add(?x,?a,?b,?c,?d,?e,?f,?g,?h,?i,?j) → db:AC1(?x)

Segundo especialista, i=2

36. db:ServImp(?p) ∧ db:R01(?p, ?k) ∧ db:R02(?p,?l) ∧ db:R03(?p,?m) ∧ db:R04(?p,?n) ∧ db:R05(?p, ?o) ∧ db:R06(?p,?q) ∧ db:R07(?p,?r) ∧ db:R08(?p, ?s) ∧ db:R09(?p, ?t) ∧ db:R10(?p, ?u) ∧ swrlb:multiply(?a, 0.0833, ?k) ∧ swrlb:multiply(?b, 0.0625, ?l) ∧ swrlb:multiply(?c, 0.0208, ?m) ∧ swrlb:multiply(?d, 0.1667, ?n) ∧ swrlb:multiply(?e, 0.0833, ?o) ∧ swrlb:multiply(?f, 0.0833, ?q) ∧ swrlb:multiply(?g, 0.1667, ?r) ∧ swrlb:multiply(?h, 0.1875, ?s) ∧ swrlb:multiply(?i, 0.1250, ?t) ∧ swrlb:multiply(?j, 0.0208, ?u) ∧ swrlb:add(?x,?a,?b,?c,?d,?e,?f,?g,?h,?i,?j) → db:AC2(?x)

Terceiro especialista, i=3

37. db:ServImp(?p) ∧ db:R01(?p, ?k) ∧ db:R02(?p,?l) ∧ db:R03(?p,?m) ∧ db:R04(?p,?n) ∧ db:R05(?p, ?o) ∧ db:R06(?p,?q) ∧ db:R07(?p,?r) ∧ db:R08(?p, ?s) ∧ db:R09(?p, ?t) ∧ db:R10(?p, ?u) ∧ swrlb:multiply(?a, 0.1000, ?k) ∧ swrlb:multiply(?b, 0.0600, ?l) ∧ swrlb:multiply(?c, 0.0200, ?m) ∧ swrlb:multiply(?d, 0.1800, ?n) ∧ swrlb:multiply(?e, 0.0600, ?o) ∧ swrlb:multiply(?f, 0.1000, ?q) ∧ swrlb:multiply(?g, 0.1400, ?r) ∧ swrlb:multiply(?h, 0.1800, ?s) ∧ swrlb:multiply(?i, 0.1400, ?t) ∧ swrlb:multiply(?j, 0.0200, ?u) ∧ swrlb:add(?x,?a,?b,?c,?d,?e,?f,?g,?h,?i,?j) → db:AC3(?x)

Regras para cálculo do nível de *Suspeição de Transação* total, considerando a média dos níveis calculados para cada especialista

38. db:ServImp(?p) ∧ db:AC1(?p, ?a) ∧ db:AC2(?p,?b) ∧ db:AC3(?p,?c) ∧ swrlb:add(?d,?a,?b,?c) ∧ swrlb:divide(?x,?d,3) → db: NivSuspeicaoTrans(?x)

Regras para classificação (categorização) do nível de *Suspeição de Transação*

De acordo com TABELA 5.12.

As regras a seguir funcionam de modo similar às regras 7 à 14, apenas com a diferença que, neste caso, não precisamos atribuir valores de grau de pertinência, mas apenas valores de categorias para que o atributo seja minerado.

39. db:ServImp(?p) ∧ NivSuspeicaoTrans(?p,?b) ∧ temCategoria(CatSuspeicaoIrrisorio, ?b) ∧ temValorInferior(CatSuspeicaoIrrisorio, ?c) ∧ temValorSuperior(CatSuspeicaoIrrisorio, ?d) ∧ swrlb:greaterThanOrEqual(?b, ?c) ∧ swrlb:lessThan(?b, ?d) → db:NivSuspeicaoTransDiscr(?p, “1”) ∧ sqwrl:select(?p, ?a, ?b, ?c, ?d)

40. db:ServImp(?p) ∧ NivSuspeicaoTrans(?p,?b) ∧ temCategoria(CatSuspeicaoIrrisorio,?b) ∧ temValorInferior(CatSuspeicaoIrrisorio, ?c) ∧ temValorSuperior(CatSuspeicaoIrrisorio, ?d) ∧ swrlb:greaterThanOrEqual(?b, ?c) ∧ swrlb:lessThan(?b, ?d) → db:NivSuspeicaoTransDiscr(?p, “3”) ∧ sqwrl:select(?p, ?a, ?b, ?c, ?d)

41. db:ServImp(?p) ∧ NivSuspeicaoTrans(?p,?b) ∧ temCategoria(CatSuspeicaoIrrelevante,?b) ∧ temValorInferior(CatSuspeicaoIrrelevante, ?c) ∧ temValorSuperior(CatSuspeicaoIrrelevante, ?d) ∧ swrlb:greaterThanOrEqual(?b, ?c) ∧ swrlb:lessThan(?b, ?d) → db:NivSuspeicaoTransDiscr(?p, “4”) ∧ sqwrl:select(?p, ?a, ?b, ?c, ?d)

42. db:ServImp(?p) ∧ NivSuspeicaoTrans(?p,?b) ∧ temCategoria(CatSuspeicaoPoucoPromissor?b) ∧ temValorInferior(CatSuspeicaoPoucoPromissor, ?c) ∧ temValorSuperior(CatSuspeicaoPoucoPromissor, ?d) ∧ swrlb:greaterThanOrEqual(?b, ?c) ∧ swrlb:lessThan(?b, ?d) → db:NivSuspeicaoTransDiscr(?p, “5”) ∧ sqwrl:select(?p, ?a, ?b, ?c, ?d)

43. db:ServImp(?p) ∧ NivSuspeicaoTrans(?p,?b) ∧ temCategoria(CatSuspeicaoMedio, ?b) ∧ temValorInferior(CatSuspeicaoMedio, ?c) ∧ temValorSuperior(CatSuspeicaoMedio, ?d) ∧ swrlb:greaterThanOrEqual(?b, ?c) ∧ swrlb:lessThan(?b, ?d) → db:NivSuspeicaoTransDiscr(?p, “6”) ∧ sqwrl:select(?p, ?a, ?b, ?c, ?d)

44. db:ServImp(?p) ∧ NivSuspeicaoTrans(?p,?b) ∧ temCategoria(CatSuspeicaoPromissor, ?b) ∧ temValorInferior(CatSuspeicaoPromissor, ?c) ∧ temValorSuperior(CatSuspeicaoPromissor, ?d) ∧ swrlb:greaterThanOrEqual(?b, ?c) ∧ swrlb:lessThan(?b, ?d) → db:NivSuspeicaoTransDiscr(?p, “7”) ∧ sqwrl:select(?p, ?a, ?b, ?c, ?d)

45. db:ServImp(?p) ∧ NivSuspeicaoTrans(?p,?b) ∧ temCategoria(CatSuspeicaoRelevante,?b) ∧ temValorInferior(CatSuspeicaoRelevante, ?c) ∧ temValorSuperior(CatSuspeicaoRelevante, ?d) ∧ swrlb:greaterThanOrEqual(?b, ?c) ∧ swrlb:lessThan(?b, ?d) → db:NivSuspeicaoTransDiscr(?p, “9”) ∧ sqwrl:select(?p, ?a, ?b, ?c, ?d)

46. db:ServImp(?p) ∧ NivSuspeicaoTrans(?p,?b) ∧ temCategoria(CatSuspeicaoNotavel,?b) ∧ temValorInferior(CatSuspeicaoNotavel, ?c) ∧ temValorSuperior(CatSuspeicaoNotavel, ?d) ∧ swrlb:greaterThanOrEqual(?b, ?c) ∧ swrlb:lessThan(?b, ?d) → db:NivSuspeicaoTransDiscr(?p, “9”) ∧ sqwrl:select(?p, ?a, ?b, ?c, ?d)

No Capítulo 6, Seção 6.3, após apresentarmos os resultados da mineração de dados, mostraremos critérios usados em regras SWRL para classificar os resultados da mineração instanciados na ontologia.

6 Testes, Simulações e Resultados

6.1 Introdução

A fim de verificar a efetiva utilidade e potencialidade da abordagem desenvolvida nesta pesquisa e da modelagem apresentada no capítulo anterior, foram realizadas simulações e testes com dados de serviços importados por empresas de Belo Horizonte.

Os dados de serviços importados apresentam características bem interessantes para nossos objetivos. De um lado, a escassez de informações, a precariedade dos dados, e a ampla dispersão estatística desses dados faz com que eles sejam um desafio à altura das pretensões da abordagem adotada. De outro, esses dados já vêm sendo laboriosamente tratados pela auditoria tributária há anos e, com isso, temos o conhecimento prévio de alguns resultados que necessariamente a abordagem pela mineração de dados também deveria resultar. Este conhecimento prévio de alguns resultados permitirá estabelecer alguns parâmetros de controle para validar a abordagem.

Vários atributos desses dados serão classificados a partir de uma ontologia que incorpora e formaliza conhecimento especialista. Na etapa prévia à mineração, este conhecimento formalizado na ontologia, permitirá classificar, categorizando, ou discretizando, os *Valor Total do Serviço*, *Percentual do ISSQN Retido*, *Idade da Empresa*, se a cidade do prestador é classificada como integrante da *Região Metropolitana*, ou *Colar Metropolitano*, se é município *Paraíso Fiscal*, seu nível de *Dinamismo*. Se o ISSQN de serviços de uma mesma empresa é retido em algumas ocasiões e outra não (*Com/Sem Retenção*). Se a empresa atua em mais de um município (*Multimunicipal*), se possui *Unidade em BH,* e, com base na agregação destes atributos, o nível de *Suspeição da Transação* de cada serviços prestado.

Além da discretização, necessária e indispensável para a mineração de dados, a ontologia permitirá classificar a atividade vinculada ao serviço prestado com base na ocorrência de termos no *Nome do Prestador*. Em nosso caso de estudo, dentre o universo de transações, selecionamos um subconjunto com base nas atividades prestadas de consultoria, gestão e representação.

TABELA 6.1 Algumas características dos dados de testes restrito a serviços prestados de consultoria, gestão e representação

Tipos de ocorrência	Número de Ocorrência	Valor Total de Serviços	Valor de ISS Retido
Transações	60.327	607.450.175,94	2.342.952,24
Transações com CNPJ em branco	6.824	57.945.963,02	577.392,71
Transações com CNPJ inválido (l<14)	260	2.788.805,19	10.478,99

TABELA 6.2 Características que levam a um alto nível de dispersão estatística dos dados de testes, restrito a serviços prestados de consultoria, gestão e representação

Tipos de ocorrência	Total de Ocorrências
Inscrições Municipais de Prestadores distintas	3.809
CNPJ de Tomadores distintos	5.732
Cidade + UF	787
Níveis de Suspeição	7

Na etapa posterior à mineração, este conhecimento especialista formalizado na ontologia, permitirá classificar as regras em categorias de relevância de acordo com as métricas geradas pelo *Tamanduá* e sumarizações de valores realizadas a partir dessas regras.

Além disso, a ontologia incorporará o conhecimento especialista de natureza vaga, através de classes, propriedades e instâncias difusas, representando rótulos linguísticos e seus respectivos graus de pertinência, conforme valores atribuídos nas TABELA 5.5 à TABELA 5.8

Este conhecimento especialista, de natureza vaga, será utilizado na abordagem difusa descrita na próxima seção.

A partir das simulações e testes realizados, teremos a avaliação da capacidade do minerador de detectar regras que revelem transações relevantes e também avaliaremos a capacidade da ontologia de instanciar dados oriundos de tabelas

relacionais, e de classificar os elementos a serem minerados e as regras produzidas pela mineração de dados, a partir do conhecimento incorporado e formalizado.

Além disso, avaliaremos os ganhos obtidos no processo global de recuperação de informação obtido pela abordagem difusa, usando conhecimento vago incorporado e formalizado em ontologias difusas, em contraste com a abordagem tradicional, clássica, que não utiliza o conhecimento de natureza vaga.

6.2 Simulações e Testes Realizados

6.2.1 Critérios Gerais de Avaliação de uma Mineração de Dados

A mineração de dados somente terá chance de se transformar em uma ferramenta efetivamente útil, a partir de uma necessária e clara definição de objetivos específicos. De modo contrário, teremos em mãos uma série de associações sem que tenhamos condições favoráveis de atribuir-lhes sentido e dotá-las de utilidade.

Para que possa se tornar interessante para as organizações ela deve apresentar resultados que sejam válidos, explicáveis e interpretáveis, com caráter de novidade e inesperabilidade, e, necessariamente, úteis (GONÇALVES, 2011).

Resultados válidos são aqueles relevantes, que possuem não somente expressividade estatística, mas também possuem a capacidade de alcançar os objetivos previamente estabelecidos. Em nosso caso, os resultados terão que propiciar apoio a decisões que darão início a ações fiscais com potencial de gerar incremento significativo, direto, ou indireto da arrecadação.

Resultados novos são aqueles que trazem um novo elemento para uma situação já conhecida. Por exemplo, sabemos que, de um modo geral, cidades do entorno de capitais estaduais e do Distrito Federal são usada para serem sedes de domicílios fiscais simulados, mas o aparecimento de uma cidade como, por exemplo, Paulista, no entorno do Recife, é uma novidade.

Resultados inesperados são aqueles que trazem informação surpreendente. Por exemplo, o aparecimento de cidades como Hortolândia-SP, Sarzedo-MG, ou

Alpinópolis-MG, dentre os resultados apresentados abaixo, ou as cidades de Pindoretama-CE e Mamanguape-PB, que surgiram em outros testes e simulações.

O mais importante, **resultados úteis** são aqueles que direcionam ações para alcançar os objetivos das organizações. Em nosso caso, as ações podem ser as de realização de diligências, e fiscalização de empresas prestadoras e tomadoras de serviços envolvendo transações suspeitas, para fins de descaracterização de domicílio fiscal dos prestadores, e obrigatoriedade de retenção por parte dos tomadores, e estabelecimento de convênios com prefeituras para troca de informações sobre domicílios fiscais simulados. Nesse último caso, por exemplo, a Prefeitura do Rio de Janeiro se beneficiaria com nossas informações sobre empresas com domicílios fiscais possivelmente simulados em Saquarema e Rio Bonito, que, por sua vez, beneficiaria a Prefeitura de Belo Horizonte com a informação sobre empresas com domicílios fiscais, possivelmente simulados, nas cidades do entorno de Belo Horizonte.

Não menos importante, pois são pré-requisitos para a utilidade, **resultados explicáveis e interpretáveis** são aqueles que podem ser entendidos e esclarecidos pelos analistas. Por exemplo, transações suspeitas oriundas de cidades pertencentes ao Colar e Região Metropolitana de Belo Horizonte, sugerem que casas de campo, sítios, ou imóveis de lazer de contribuintes e seus familiares sejam utilizados para simular domicílio fiscal de contribuintes efetivamente domiciliados em Belo Horizonte.

6.2.2 Tipos de Simulação e Teste

Para permitir a análise e discussão dos resultados que permitam avaliar a abordagem desenvolvida com uso de sistemas difusos realizaremos dois tipos de testes e simulações.

Primeiro, no que passaremos a chamar de **abordagem clássica**, buscaremos a associação de atributos originais dos dados e buscaremos avaliá-los e analisá-los com as medidas objetivas geradas pela mineração de dados, disponibilizadas pelo *Tamanduá*.

Segundo, no que já chamamos de **abordagem difusa**, buscaremos a associação de atributos originais agregados com a variável difusa de nível de *Suspeição da*

Transação, conforme modelada e definida na Seção 5.3.4. Essa variável difusa também será discretizada, conforme a TABELA 5.12: a ontologia classificará cada transação, atribuindo um nível discreto de suspeição.

Em uma ou outra abordagem, a mineração de dados terá que nos fornecer elementos que nos permitam tomar três tipos de decisão para revelar focos de fraude e sonegação considerando:

1. Municípios;
2. Prestadores e
3. Tomadores.

Uma vez identificados esses focos, o Fisco municipal poderá adotar ações fiscais que inibam a fraude e a sonegação.

Para fins de apresentação, discussão e avaliação dos resultados nessa tese, utilizaremos tão somente os dados referentes à primeira questão, fraude e sonegação focadas em municípios. O tratamento somente dessa questão já é capaz de produzir todos os elementos necessários às simulações, testes e análises necessárias. Desse modo, ainda atendemos às restrições legais, e as responsabilidades funcionais referentes ao sigilo dos contribuintes, pois mesmo disfarçando o nome, eventualmente, os contribuintes poderiam ser identificados por vias indiretas, através de outros dados.

Abordagem clássica

Neste tipo de simulação e teste, trabalharemos com atributos que permitam associar identificação do prestador, identificação do tomador, localidade de origem do serviço e valor total do serviço.

1. **Identificação do Prestador**: *Cnpj*
2. **Identificação do Tomador**: *Im*
3. **Localidade de origem do serviço**: *CidadeUf*
4. **Valor Total do Serviço Discretizado**: *VlrTotalServDiscr*

Este tipo de simulação levou a resultados apresentados nas TABELA 6.3, TABELA 6.5, TABELA 6.7, TABELA 6.8 e TABELA 6.9.

Abordagem difusa

Neste tipo de simulação e testes, a noção vaga de suspeição de uma transação conduzirá a mineração de dados e orientará toda a análise dos resultados. Como vimos, a noção de suspeição em uma transação está associada à possibilidade de ocorrência de domicílio fiscal simulado, à aplicação de alíquota incorreta, ao recolhimento a partir de unidade localizada fora de Belo Horizonte, a despeito do serviço ter sido prestado em unidade de Belo Horizonte, informação de atividade distinta da real, etc., que repercutem no valor e na territorialidade do tributo.

Esta variável difusa poderá estar presente no antecedente e no consequente de regras, e, ainda, servirá diretamente como uma medida subjetiva para avaliação da relevância das regras. Essa medida subjetiva, em conjunto com as medidas objetivas disponibilizadas pelo *Tamanduá* servirão de parâmetros para a ontologia classificar e ordenar as regras.

1. **Identificação do Prestador**: *Cnpj*
2. **Identificação do Tomador**: *Im*
3. **Localidade de origem do serviço**: *CidadeUf*
4. **Nível de Suspeição da Transação Discretizado**: *NivSupDiscr*

Com base nessa noção, as três questões foram redefinidas do seguinte modo:

1. Quais os principais municípios que originam transações suspeitas.
2. Quais são as principais empresas prestadoras de serviços que geraram transações suspeitas.
3. Quais as principais empresas tomadoras de serviços que geraram transações suspeitas.

Este tipo de simulação levou a resultados apresentados nas TABELA 6.4, TABELA 6.6, TABELA 6.8 e TABELA 6.9.

6.2.3 Tipos de Calibração na Mineração de Dados

O *Tamanduá* permite uma riqueza de testes e simulações com base na configuração de alguns parâmetros que permitem obter resultados com maior ou menor granularidade. Para isto, basta operar com calibrações em dois níveis. Primeiro,

realizando distintas restrições ao espaço amostral. Segundo, utilizando maiores recursos de máquinas que permitam a configuração de parâmetros estatísticos que manipulem uma maior gama de possibilidades combinatórias.

No primeiro tipo de calibração, o espaço amostral dos resultados apresentados foi o de todas as transações das empresas de consultoria, gestão e representação com níveis de suspeição maior ou igual a 6 (de acordo com a TABELA 5.12). Poderíamos obter novos resultados restringindo a base amostral, por exemplo, para transações oriundas somente de cidades da região e colar metropolitano, ou para transações com empresas que possuem unidade(s) em Belo Horizonte, ou para determinado estado, ou região, ou ainda, para determinados níveis de capacidade arrecadatória, etc. No *Tamanduá* as restrições ao espaço amostral podem ser configuradas como um dos parâmetros para a mineração de dados, como na FIGURA 6.1.

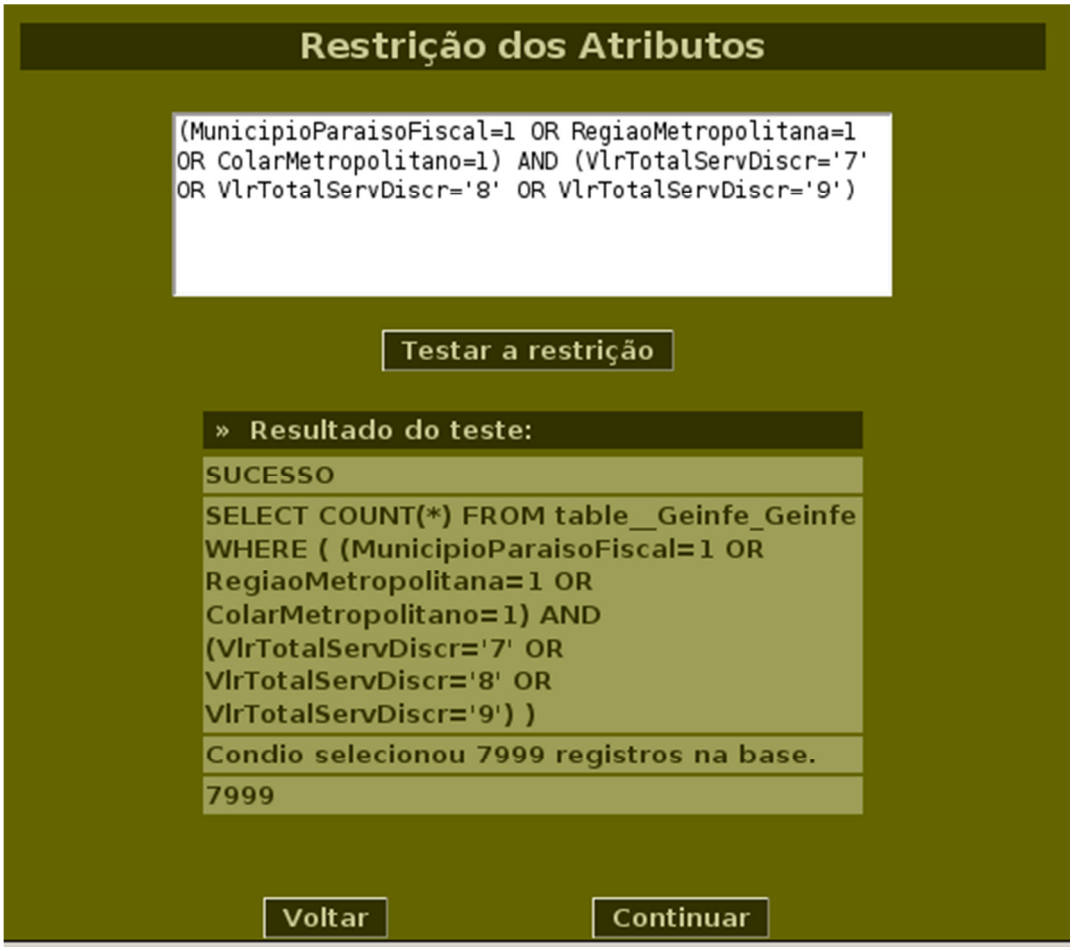

FIGURA 6.1 Calibração com base em restrição de espaço amostral.

No segundo tipo de calibração, os valores desses parâmetros deveriam ser ditados por experiências prévias com os resultados e com base em valores esperados, definidos a partir dessa experiência e conhecimento dos dados. Entretanto, em nosso caso, as definições dessas métricas, não corresponderam ao valor esperado. Primeiro, pelo desconhecimento do que se esperar. Segundo, pela imposição de dois fatores: a necessidade de gerar uma quantidade razoável de resultados, e a capacidade da máquina em processar as associações sem esgotar seus recursos.

Em nosso caso, para obtermos uma quantidade de resultados razoável, considerando-se a alta dispersão dos dados, tivemos que informar o suporte mais baixo possível, limitado tão somente à capacidade da máquina em processar. Valores de suporte e confiança inferiores a **0,1** produziram uma explosão combinatória que esgotavam os recursos da máquina utilizada e, consequentemente, não geravam resultados. Desse modo, obtivemos resultados satisfatórios adotando um suporte e confiança igual ou superior a **0,1** e **1,0**, respectivamente.

No *Tamanduá* suporte e confiança mínimas podem ser configuradas como é mostrado na FIGURA 6.2.

FIGURA 6.2 Calibração com base em suporte e confiança mínimos.

Poderíamos obter novos resultados utilizando uma máquina com mais recursos, ou explorando a capacidade de processamento paralelo do *Tamanduá* que permite distribuir a demanda fazendo o processamento em rede, usando os recursos de várias máquinas.

Com base nos parâmetros de suporte e confiança e definições de restrições, buscando-se similaridade na base amostral para fins de comparação, foram definidos os seguintes critérios para a mineração de dados nas duas abordagens.

Simulações e testes na abordagem clássica:

- Atributos: *CidadeUf*, *CnpjRaiz*, *Im* e *VlrTotalServDiscr.*
- Restrição: *MunicipioParaisoFiscais*=1 OU *RegiaoMetropolitana*=1 OU *ColarMetropolitano*=1 E (*VlrTotalServDiscr*>'7'), gerando 15.095 registros.
- Suporte: 0,1
- Confiança: 1,0
- Esquema de regra alvo: ***CidadeUf → VlrTotalServDiscr***

Simulações e testes na abordagem difusa:

- Atributos: *CidadeUf*, *CnpjRaiz*, *Im* e *NivSuspDiscr*
- Restrição: *NivSuspDiscr*>'6'), gerando 23.020 registros.
- Suporte: 0,1
- Confiança: 1,0
- Esquema de regra alvo: ***CidadeUf → NivSuspDiscr***

Aqui podemos observar que o atributo composto *Suspeição da Transação* permite a significativa redução do número de atributos usados na restrição de dados, e, potencialmente, de atributos, para a mineração de dados. Obtém-se, dessa forma, um ganho imediato em performance sem redução de significado. Isto é, vários atributos são transformados em um só, e como o conteúdo semântico dos atributos simples persiste no atributo composto através da operação de agregação, não há perda de informação relevante desses atributos. Além disso, em função do significado associado ao atributo composto, a análise de resultados também será facilitada. Como já foi dito anteriormente, esse atributo composto constitui-se em métrica subjetiva para avaliar os resultados da mineração de dados.

6.3 Resultados

Com base nos parâmetros de mineração de dados definidos acima, as TABELA 6.3 e TABELA 6.4 exibem as regras obtidas nos testes, considerando-se *CidadeUf* ocorrendo apenas no antecedente, com suas respectivas quantidades de regras.

TABELA 6.3 Resultados para *CidadeUf* no Antecedente na abordagem clássica

Nº	Antecedente	Consequente	Qtde
1.	CidadeUf	Im	176
2.	***CidadeUf***	***VlrTotalServDiscr***	***193***
3.	CidadeUf	CnpjRaiz	139
4.	CidadeUf	Im, VlrTotalServDiscr	118
5.	CidadeUf	Im, CnpjRaiz	90
6.	CidadeUf	VlrTotalServDiscr, CnpjRaiz	86
7.	CidadeUf	Im, VlrTotalServDiscr, CnpjRaiz	48
8.	**CidadeUf, Im**	**VlrTotalServDiscr**	**130**
9.	CidadeUf, Im	CnpjRaiz	95
10.	CidadeUf, Im	VlrTotalServDiscr, CnpjRaiz	50
11.	**CidadeUf, CnpjRaiz**	**VlrTotalServDiscr**	92
12.	CidadeUf, CnpjRaiz	Im	95
13.	CidadeUf, CnpjRaiz	VlrTotalServDiscr, Im	50
14.	CidadeUf, VlrTotalServDiscr	Im	130
15.	CidadeUf, VlrTotalServDiscr	CnpjRaiz	92
16.	CidadeUf, VlrTotalServDiscr	Im, CnpjRaiz	50
17.	CidadeUf, Im, VlrTotalServDiscr	CnpjRaiz	50
18.	CidadeUf, Im, CnpjRaiz	VlrTotalServDiscr	50
19.	CidadeUf, CnpjRaiz, VlrTotalServDiscr	Im	50
		Total de regras	**1784**

Na seção anterior destacamos que as regras de interesse imediato são as ***CidadeUf→VlrTotalServDiscr***, na abordagem clássica, e ***CidadeUf→NivSuspDiscr***, na abordagem difusa. Comparando-se a quantidade de regras deste tipo, podemos observar o ganho da abordagem difusa que obteve 112 regras, enquanto a abordagem clássica obteve 193 regras. Neste caso, obtivemos uma redução de 42% no número de regras.

A despeito deste ganho, o que mais nos interessa destacar é a objetividade e facilidade que o atributo composto *NivSuspDiscr* traz para a análise. Ele já nos traz de imediato a informação do nível de suspeição das transações que são instâncias destas regras. Através de um trabalho prévio de inteligência na modelagem e atribuição de pesos a atributos simples, de natureza vaga e subjetiva, que

constituem esse atributo composto, foi possível criar uma métrica subjetiva incorporada às próprias regras que definirão o sucesso da análise.

Este atributo que já nos traz a informação do nível de suspeição diz muito mais do que regras com a ocorrência de atributos como *VlrTotalServDiscr*. Estas regras demandarão um esforço de análise muito maior e, eventualmente, infrutífero.

Em um segundo momento, dependendo da disponibilidade de recursos humanos e de tempo, pode-se ampliar o escopo da análise e focar em regras, onde os atributos *VlrTotalServDiscr*, e *NivSuspDiscr* ocorrem em conjunto com outros atributos. Isto amplia o campo de análise, mas não necessariamente, pois muitas instâncias de regras podem ser redundantes em relação às analisadas no foco inicial, além da dificuldade em analisar as associações que podem até ser espúrias.

Retomaremos esta discussão nas subseções à frente: "Dificuldades mais específicas", "O desafio de se extrair significado pertinente da coocorrência de atributos" e "Problemas com as métricas objetivas".

TABELA 6.4 Resultados para *CidadeUf* no Antecedente na abordagem difusa

Nº	Antecedente	Consequente	Qtde
1.	CidadeUf	Im	110
2.	***CidadeUf***	***NivSuspDiscr***	***112***
3.	CidadeUf	CnpjRaiz	144
4.	CidadeUf	Im, NivSuspDiscr	94
5.	CidadeUf	Im, CnpjRaiz	75
6.	**CidadeUf**	**NivSuspDiscr, CnpjRaiz**	**135**
7.	CidadeUf	Im, NivSuspDiscr, CnpjRaiz	70
8.	CidadeUf, Im	NivSuspDiscr	125
9.	CidadeUf, Im	CnpjRaiz	101
10.	CidadeUf, Im	NivSuspDiscr, CnpjRaiz	93
11.	CidadeUf, CnpjRaiz	NivSuspDiscr	163
12.	CidadeUf, CnpjRaiz	Im	101
13.	**CidadeUf, CnpjRaiz**	**NivSuspDiscr, Im**	**93**
14.	CidadeUf, NivSuspDiscr	Im	122
15.	CidadeUf, NivSuspDiscr	CnpjRaiz	160
16.	CidadeUf, NivSuspDiscr	Im, CnpjRaiz	90
17.	CidadeUf, Im, NivSuspDiscr	CnpjRaiz	93
18.	CidadeUf, Im, CnpjRaiz	NivSuspDiscr	93
19.	CidadeUf, CnpjRaiz, NivSuspDiscr	Im	93
		Total de regras	**2067**

Nas TABELA 6.5 e TABELA 6.6 exibimos as instâncias das primeiras regras de interesse ***CidadeUf→VlrTotalServDiscr*** e ***CidadeUf→NivSuspDiscr***, em ordem de Potencial de ISSQN Sonegado. A fim de que se compreenda o relatório, esclarecemos o significado das colunas, consideremos as seguintes definições:

1. **Suporte:** percentual das transações que apresentam a "Cidade-Uf" que possuem o "Valor Total do Serviço"/"Suspeição da Transação" informado dentre todas as transações.
   Lembremos que o suporte expressa a relevância (significância) em uma associação. É o número de casos que contém *L* (*left*, esquerda, antecedente) e *R* (*right*, direita, consequente) dividido pelo número total de registros:

$$suporte\ (L \rightarrow R) = \frac{n^{\circ}\ transações\ suportadas{:}p(\ L \cup R)}{total\ de\ número\ de\ transações}$$

   Suporte mínimo: é o valor mínimo do suporte para que um conjunto de itens seja considerado. Ex: busca por suporte 50%: ocorrências de um determinado subconjunto em pelo menos 50% das amostras.

2. **Confiança:** percentual das transações que apresentam a "Cidade-Uf" que possuem o "Valor Total do Serviço"/"Suspeição da Transação" informado dentre as transações que apresentam a "Cidade-Uf".
   Lembremos que a confiança é a probabilidade condicional do consequente dado o antecedente. Número de registros que contém *L* e *R* dividido pelo número de registros que contém *L:*

$$confiança\ (L \rightarrow R) = \frac{suporte(L \cup R)}{suporte(L)} = \frac{p(L \cup R)}{p(L)}$$

   Confiança mínima é um limite para filtragem das associações descobertas pelo algoritmo.

3. **Convicção:** menor que 1 quer dizer que a regra complementar é mais forte.
   Lembremos que convicção mede o quanto a regra é mais forte em relação a sua regra complementar. Dada uma regra $L \rightarrow R$, convicção é a frequência com que *L* ocorre sem *R,* dividida pela frequência com que as duas ocorrem juntas:

$$convicção\ (L \rightarrow R) = \frac{p(L) \times p(\bar{R})}{p(L, \bar{R})}$$

   Do ponto de vista de análise, temos três casos:

   - Convicção=1: a regra e o seu complemento tem igual valor

- Convicção>1, a regra é mais forte que o seu complemento
- Convicção<1, a regra complementar é mais forte

Desta forma, a convicção é útil para verificar quando devemos avaliar a regra complementar.

Valor Potencial de Arrecadação: margem de arrecadação considerando-se a diferença entre o ISSQN calculado com base na alíquota de 2% e a alíquota utilizada. Trata-se apenas de um valor estimativo, tomando-se como base a alíquota de menor valor aplicável (2%). Na verdade, este valor pode ser menor, considerando-se que na análise da territorialidade o tributo não se destine a Belo Horizonte; ou pode ser maior, considerando-se a aplicabilidade de uma alíquota superior a 2%.

4. **Total de Transações**: número de transações consideradas no cálculo da regra.

TABELA 6.5 Métricas e sumarizações de municípios que originam transações com valor total de serviços relevantes na abordagem clássica

Cidade-Uf	Vlr. Total Serviços	Convicção	Confiança	Suporte	Potencial ISS Sonegado (R$)	Total Transações
Rio De Janeiro-RJ	9	1,0876	26,5363	1,8880	**1.142.709,79**	285
Saquarema-RJ	9	1,4156	43,5582	0,4704	**702.941,01**	71
Sao Paulo-SP	9	1,0835	26,2584	3,3521	**539.142,05**	506
Brasilia-DF	9	1,1181	28,5388	0,8281	**245.252,29**	125
Araxa-MG	9	0,9655	17,2413	0,0992	**220.261,69**	15
São Paulo-SP	7	0,8473	52,7763	6,7373	**202.573,11**	1017
Uberlandia-MG	9	1,1612	31,1897	0,6426	**180.365,95**	97
Curitiba-PR	9	1,0925	26,8656	0,3577	**165.257,40**	54
Sao Paulo-SP	8	1,0133	20,9652	2,6764	**148.090,95**	404
Santana Parnaiba-SP	9	1,0927	26,8786	0,6161	**147.260,86**	93
Nova Lima-MG	9	1,0267	22,1748	1,3779	**144.928,94**	208
Salvador-BA	9	1,0249	22,0430	0,2716	**142.004,41**	41
Rolante-RS	9	0,0000	100,0000	0,1126	**117.790,56**	17
Rio De Janeiro-RJ	7	0,8544	53,1657	3,7827	**103.385,19**	571
Nova Lima-MG	7	0,8852	54,7974	3,4051	**85.082,92**	514
Rio De Janeiro-RJ	8	1,0048	20,2979	1,4442	**79.725,52**	218
Paulista-PE	9	1,1123	28,1690	0,1325	**69.774,56**	20
Sabara-MG	9	1,0282	22,2891	0,2451	**68.263,96**	37
Porto Alegre-RS	9	1,1985	33,3333	0,2915	**67.522,60**	44
Goiania-GO	9	1,1508	30,5699	0,3909	**65.313,99**	59
Rio Acima-MG	9	0,8968	10,9070	0,6293	**62.488,04**	95

Observe que pode haver uma mesma cidade com dois, ou três níveis de *Valor Total de Serviços*.

Convicção, *Confiança* e *Suporte* são métricas objetivas. *Potencial ISS Sonegado*, *Total Transações* são sumarizações que também expressam a relevância de resultados. Por fim, *Suspeição* é uma métrica subjetiva produzida pela abordagem difusa.

TABELA 6.6 Métricas e sumarizações de municípios que originam transações suspeitas na abordagem difusa

Cidade-UF	Suspeição	Convicção	Confiança	Suporte	Potencial ISS Sonegado (R$)	Total Transações
Cuiaba-MT	9	1,4981	46,6667	0,1855	**55.988,64**	28
Saquarema-RJ	7	0,8875	26,3889	0,2476	**619.829,66**	57
Rio De Janeiro-RJ	7	0,7892	17,2185	0,1129	**348.039,63**	26
Nova Lima-MG	7	2,2365	70,7885	8,8532	**293.208,00**	2.038
Curitiba-PR	6	31,5496	98,9011	0,3910	**169.589,46**	90
Uberlandia-MG	6	0,6041	42,6087	1,0643	**149.598,92**	245
Contagem-MG	6	1,3117	73,5697	11,4509	**135.119,20**	2.636
Saquarema-RJ	6	1,3138	73,6111	0,6907	**113.880,39**	159
Rio Acima-MG	7	1,0346	36,8534	4,4570	**88.449,61**	1.026
Sabara-MG	6	1,3868	75,0000	1,0425	**87.832,96**	240
Sao Paulo-SP	6	0,4721	26,5595	3,0148	**84.259,71**	694
Paulista-PE	6	16,6415	97,9167	0,2042	**74.339,70**	47
Uberlandia-MG	7	1,5333	57,3913	1,4335	**71.313,36**	330
Rio Bonito-RJ	6	3,9555	91,2351	0,9948	**71.098,53**	229
Rio De Janeiro-RJ	6	2,0136	82,7815	0,5430	**68.756,10**	125
Barueri-SP	6	5,2187	93,3566	1,1599	**68.240,02**	267
Santana Parnaiba-SP	6	2,9036	88,0597	0,2563	**58.799,79**	59
Santarem-PA	6	0,0000	100,0000	0,1651	**55.423,45**	38
Nova Lima-MG	6	0,48977	29,2115	3,6533	**53.192,76**	841
Porto Velho-RO	6	6,8184	94,9153	0,2433	**49.483,49**	56
Contagem-MG	7	0,8880	26,4304	4,1138	**42.453,62**	947
Brasilia-DF	6	4,0737	91,4894	0,1868	**41.489,17**	43

Observe que pode haver uma mesma cidade com dois níveis de *Suspeição* distintos.

Estes atributos são instanciados na ontologia de pós-processamento dos resultados, usamos uma regra SWRL para selecionar a regra como sendo interessante. Um dos critérios de seleção é considerar interessante as regras com *Convicção* maior do que 1, com *Potencial de ISS Sonegado* superior a 25 mil, *Potencial de ISS Sonegado per capita* (dividido pelo número de transações) superior a 1 mil. O primeiro critério indica a força objetiva da regra, o segundo critério indica o potencial de arrecadação, o terceiro critério indica um menor custo-benefício de fiscalização.

Uma vez selecionadas, as regras podem ser ranqueadas por nível de *Suspeição de Transação*.

6.3.1 Discussão dos Resultados

Dificuldades usuais que revelam a abrangência e complexidade do problema

Nossas dificuldades com a mineração de dados foram as usualmente apresentadas em distintos relatos de fracassos em projetos desse tipo (por exemplo, GONÇALVES, 2004 e 2002). As dificuldades usuais são:

1. Dificuldades com dados:
   1.1. Dados ruins, errados, incompletos, insuficientes;
   1.2. Ocorrência de dados e associações relevantes, mas com baixa frequência, invisíveis à mineração de dados e
   1.3. Dados mal utilizados.
2. Dificuldades com regras geradas:
   2.1. Excesso de regras;
   2.2. Regras óbvias;
   2.3. Regras redundantes;
   2.4. Regras contraditórias;
   2.5. Regras sem sentido e
   2.6. Geração de falsos positivos,
3. Dificuldades com a mineração:
   3.1. Configuração dos parâmetros de suporte e confiança;
   3.2. Definição dos atributos a serem minerados: excesso de atributos, atributos pouco, ou nada significativos e
   3.3. Uso de diferentes atributos já intimamente correlacionados;
4. Dificuldades em operacionalizar e inserir a mineração de dados na organização:
   4.1. Limitações de utilidade;
   4.2. Limitações de autonomia e independência: necessidade do conhecimento do negócio e entendimento dos dados para a análise dos resultados;
   4.3. Limitações de facilidade de uso no trabalho de rotina e
   4.4. Limitações de eficácia, de desempenho e produtividade.

Conforme podemos observar nas dificuldades relacionadas, não há problemas com as ferramentas computacionais adotadas. Normalmente, as ferramentas portam-se excepcionalmente bem, atuando em seus propósitos com extrema eficiência. Em nosso caso, a ferramenta *Tamanduá* respondeu com presteza gerando resultados em segundos, a partir dos elementos que lhe eram fornecidos. O *Tamanduá* disponibiliza algoritmos clássicos de mineração de dados finamente implementados de forma paralelizável e em plataforma de *software* livre. Sua visualização de resultados era restrita, mas suficientemente expressiva.

Isso obviamente evidencia que as dificuldades envolvendo mineração de dados não originam-se a partir de um problema computacional, mas diante de um problema informacional amplo e de maior complexidade, envolvendo não somente aspectos técnicos, mas também humanos. Em problemas dessa natureza a vocação da Ciência da Informação em lidar com questões inter, multi e transdisciplinares tem plena oportunidade de se sobressair.

Dificuldades mais específicas

Um minerador de dados pode atuar em dois sentidos. O de identificar e mensurar padrões e, quando se tem algum atributo em escala de tempo, o de identificar e mensurar tendências, ou alterações de comportamento.

Ele oferece-nos atributos que ocorrem simultaneamente, métricas da força dessa associação entre os atributos e, uma vez que se disponha de um atributo temporal, as alterações dessas coocorrências ao longo do tempo. O sucesso da mineração de dados dependerá da capacidade de dotarmos essas coocorrências de significado, usar as métricas disponíveis para selecionar e ordenar regras e usar as variações no tempo para detectar tendências, ou alterações de comportamento.

Em nosso caso de estudo, nosso sucesso é ameaçado, **primeiro, por não dispormos desse atributo temporal** que permitiria detectar e analisar alterações de padrões ao longo do tempo. Segundo, teremos **dificuldades em atribuir sentido às coocorrências dos atributos trabalhados na abordagem clássica**: como atribuir sentido, em nosso caso, como associar fraude e sonegação à mera coocorrência de *CidadeUf*, *CNPJ do Prestador*, *Inscrição Municipal do Prestador*, e *Valor Total do Serviço*? Terceiro, há as **limitações do alcance das métricas**

objetivas em sua capacidade de destacar a relevância das regras resultantes do processo de mineração de dados e, portanto, são referências limitadas para as tarefas de seleção e ordenação.

Os problemas com a falta de um atributo temporal e, consequentemente, o desconhecimento de variações de padrões, são evidentes. Passemos então, à discussão mais detida dos outros problemas detectados, e como eles nos afetam e indicam a via da abordagem difusa como solução.

O desafio de se extrair significado pertinente da coocorrência de atributos

Mineração de dados parece ser talhada para *marketing*, vendas e detecção de fraudes. O exemplo clássico é a cesta de supermercado, onde a associação de produtos via mineração de dados levam à conclusão de quais produtos tendem a ser comprados conjuntamente e, em vista disso, podem orientar desde campanhas de *marketing* até a disposição dos produtos nas prateleiras. Por exemplo, pode analisar dados de vendas de varejo para descobrir produtos aparentemente não relacionados que frequentemente são comprados em conjunto, como fraldas e cervejas[70]. Questões básicas como as seguintes podem ser facilmente respondidas:

Descrição: “Quem comprou este produto também comprou...”

Predição: “Quem quer comprar este produto, também comprará...”

Na cesta de supermercado temos a associação de atributos de natureza similar, associam-se produtos vendidos. Nesse caso, fica fácil atribuir até de modo prévio a significação de coocorrências, seja lá entre que atributos forem. Entretanto, essa situação de facilidade de pré-significação das regras não é o comum em situações complexas como as trabalhadas na área de auditoria tributária. Muito diferente é associar atributos de natureza diversa como empresas, cidades e valores de serviços e buscar uma significação para isso.

---

[70]Alusão ao famoso *case* de mineração de dados do *Wal-Mart* que descobriu que, às sextas-feiras, o volume de vendas de cerveja crescia praticamente na mesma proporção que o de fraldas. Ao colocar os dois produtos próximos, obteve-se um significativo aumento de vendas. A análise dos dados levou à interpretação de que pais com crianças pequenas deixavam de sair de casa para cuidar dos filhos, e as cervejas passavam a serem tomadas em casa, em especial, em dias de jogos como nas quintas-feiras e finais de semana.

TABELA 6.7 Resultados para Cidade-UF=”CAMPINAS-SP”

Métricas/Sumarizações	1 ° resultado	2 ° Resultado	3º Resultado
Relevância Vlr. Total Serviços	Notável	Relevante	Muito Relevante
Convicção	1,0601	0,9407	0,9756
Confiança	24,6269	57,4627	17,9104
Suporte	0,2186	0,5101	0,1590
Potencial ISS Sonegado (R$)	37.656,54	17.740,35	11.565,80
Total de Transações	33	77	24

O significado óbvio é que são significativas as frequências de ocorrências, entretanto, não nos trazem informações consistentes de indícios de fraude, ou sonegação. Qual o sentido, como interpretar sob o ponto de vista de quem busca indícios de fraude e sonegação, uma regra que associa, por exemplo, as transações originadas do município de Campinas, Estado de São Paulo?

A mineração destaca este resultado na abordagem clássica, mas, em verdade, este não se apresenta como um resultado relevante. Campinas é um pólo tecnológico e cidade com alto dinamismo, de modo que é natural que exerça um papel relevante na importação de serviços para empresas sediadas em Belo Horizonte.

Consideremos agora cidades que são destacadas tanto pela abordagem clássica, quanto pela abordagem difusa. Consideremos Rio Bonito, no Estado do Rio de Janeiro, TABELA 6.8 e Paulista, no Estado de Pernambuco, TABELA 6.9.

TABELA 6.8 Resultados para Cidade-UF=”RIO BONITO-RJ”

Métricas/Sumarizações	Abordagem Clássica	Abordagem Difusa
Relevância de Vlr. Total de Serviços	Relevante	---
Nível de Suspeição	---	6-Promissor
Convicção	1,4449	3,9555
Confiança	72,3077	91,2351
Suporte	0,6227	0,9948
Potencial ISS Sonegado (R$)	19.392,58	71.098,53
Total de Transações	94	229

No resultado da TABELA 6.8, a mineração de dados destaca uma cidade do entorno da cidade do Rio de Janeiro que nos é conhecida, justamente pela expressividade de quantidade de ocorrências e valores associados.

No resultado da TABELA 6.9, a mineração de dados destaca uma cidade que nunca foi objeto de atenção por parte da auditoria tributária. Pesquisas adicionais esclarecem-nos que se trata de uma cidade da região metropolitana do Recife. Entretanto, na abordagem clássica, tanto numa cidade como em outra, não saberíamos, a não ser após a avaliação individualizada dos registros de regras e pesquisas adicionais, se tais valores seriam indícios de fraude e sonegação, ou não.

TABELA 6.9 Resultados para Cidade-UF="PAULISTA-PE"

Métricas/Sumarizações	Abordagem Clássica	Abordagem Difusa
Relevância de *Valor Total de Serviços*	Notável	---
Nível de *Suspeição*	---	6-Promissor
Convicção	1,112342	16,641529
Confiança	28,169014	97,916667
Suporte	0,132494	0,204170
Potencial ISS Sonegado (R$)	69.774,56	74.339,70
Total de Transações	20	47
Total Valor de Serviços	3.488.727,80	3.716.984,76

Já na abordagem difusa, o atributo de *Nível de Suspeição* é capaz de agregar significação direta. Este atributo foi ontologicamente trabalhado, atuando como uma propriedade definida como feixe, a partir da copresença de sub-atributos apresentados na TABELA 5.10 (*Região Metropolitana*; *Colar Metropolitano*; *Município Paraíso Fiscal*; *Unidade em BH*; *Empresa multimunicipal*; *Com/Sem retenção*; *Dinamismo Municipal*; *Consolidação e Tradição*; *Percentual de ISS Retido*; e *Valor Total de Serviço*).

A abordagem clássica, de um lado, destaca ocorrências não relevantes e ainda, de outro lado, inibe a emergência de ocorrências relevantes. O critério único de frequência e suas métricas objetivas associadas apresentam resultados e direcionam a análise a ocorrências e critérios que, em verdade, afastam-nos de nosso objetivo. Como exemplos, podemos citar a omissão das cidades de SARZEDO-MG, PAULINEA-SP, MATOZINHOS-MG, MARABA-PA, PINHAIS-PR, HORTOLANDIA-SP, SÃO JOAQUIM DE BICAS-MG E MARAVILHAS-MG. Também podemos citar que mesmo quando a abordagem clássica detecta a ocorrência da

regra, o critério de sumarização “Total de Transações” é minimizado, como se pode constatar através das TABELA 6.8 e TABELA 6.9.

É paradoxal, teremos que lidar com métricas objetivas que afastam-nos de nosso objetivo de análise.

Problemas com as métricas objetivas

Como vimos nas Seção 3.4.2 e Seção 3.4.4, e na **FIGURA 3.5**, o processo de mineração de dados vincula-se à tarefa de descrição de uma base de conhecimento empírico. Entretanto, o que se espera da mineração de dados não é um mero retrato da realidade. Não desejamos a representação de algo que somos capazes de ver e reconhecer de imediato. Desejamos uma revelação da realidade não óbvia.

O minerador de dados não deverá retratar o visível, mas tornar visível algo que não está ao alcance de nossa percepção imediata, condicionada e entorpecida por vários fatores. Ele deverá ser capaz de trazer algo novo, revelar-nos associações obliteradas, em especial, pelo excessivo volume de dados, mas, também, por nossas predisposições cognitivas.

Remontando aos primórdios da criação de parâmetros de análise estatística, o minerador de dados exerce a função de um gramaticógrafo da ciência, como imaginado por Karl Pearson (1857–1936). Pearson foi o responsável pela divulgação, estabelecimento e desenvolvimento dos parâmetros analíticos da regressão e correlação estatísticas inventadas por seu mestre Sir Francis Galton (1822-1911). Em 1895, Pearson editou o livro “A Gramática da Ciência” (PEARSON, 1905) onde advogava que a ciência somente pode ser baseada em um claro conhecimento dos fatos, através da apreciação de sua sequência e correlação (em seus termos: “significância relativa”, PEARSON, 1905, p. 6).

Entretanto, já em seus primórdios ficou evidente, **primeiro, que a correlação falha em capturar dependência**, isto é, pode haver variáveis que apresentam uma forte dependência estatística, mas apresentam correlação fraca, ou nula. **Segundo, a correlação não esclarece a natureza da relação de determinação entre as variáveis correlacionadas**. Como consequência, ela por si só, não nos revela os elementos para explicar, compreender, ou interpretar a relação de determinação.

A primeira limitação da correlação foge ao escopo dessa tese e, acreditamos, seus efeitos não serão significativos para nossos objetivos. Já o reconhecimento da segunda limitação, é a origem de grandes dificuldades em mineração de dados, e valoriza nossa solução proposta, a abordagem difusa.

As TABELA 6.5, TABELA 6.6, TABELA 6.8 e TABELA 6.9 apresentam as métricas de convicção, confiança e suporte para as *Cidades-Uf* consideradas, associadas ao atributo de *Valor Total do Serviço*, para a abordagem clássica, e/ou ao atributo nível *de Suspeição*, para a abordagem difusa.

O processo de avaliação de regras se inicia com a avaliação de suporte e confiança, seguida de uma avaliação do incremento provido pelas regras de maior confiança e finalmente verificando a sua convicção, que pode indicar a necessidade de avaliar outras regras. Em nosso caso, um valor baixo para a métrica convicção com o atributo *Valor Total do Serviço*, ou nível de *Suspeição* no consequente, indica-nos que a associação é responsável por grande (a quase totalidade; ou a totalidade, caso a convicção seja zero) parte do número das transações referentes à *Cidade-Uf* informada. Dessa forma, e em princípio, o índice de convicção pode servir para estabelecer prioridades de fiscalização, considerando que uma vez solucionada o problema representado por tal nível de *Suspeição*, o problema com tal "Cidade-UF" estará solucionado. Entretanto, essas métricas consideradas em separado, dizem-nos sobre a força das associações, ou seus complementos, mas nada nos dizem sobre o nível de sonegação, ou fraude associado aos registros das regras.

Em resumo, a relevância não está necessariamente relacionada à frequência de ocorrência e suas métricas derivadas. As métricas objetivas, em verdade tornam-se valiosas após o uso prévio de uma métrica subjetiva, como a representada pelo atributo de nível de *Suspeição* da transação.

Conhece-te a Ti Mesmo – Necessidade do conhecimento do domínio

Uma grande limitação das regras geradas pelos mineradores de dados, a de que não nos revela os elementos para explicar, compreender, ou interpretar a relação de determinação.

Esta limitação reforça o caráter oracular dos mineradores em dois aspectos. Primeiro, eles revelam-nos uma realidade oculta, mas muitas vezes em linguagem,

em associações que precisam ser decifradas. Segundo, como todo oráculo, justamente para podermos decifrar seus propósitos, ele exige de nós um amplo conhecimento do domínio com o qual estamos lidando: **"conhece-te a ti mesmo"** advertia o Oráculo de Delfos em seus pórticos. A mensagem do oráculo revelada em tom metafórico, velada, só alcança um significado, é absorvida e dotada de utilidade por aquele que conhece a si mesmo. Do contrário, é mensagem jogada ao vento.

É nessa apregoada tarefa do "conhece-te a ti mesmo" que as ontologias se inserem e vem a nós para auxiliar-nos com as dificuldades vinculadas ao necessário conhecimento do negócio. Pois, as ontologias são capazes de auxiliar-nos de forma decisiva na operacionalização e inserção da mineração de dados em uma organização. Isto se faz através de seus recursos de incorporação e formalização do conhecimento do negócio e do entendimento dos dados para a análise dos resultados, sobrepujando as limitações de autonomia e independência, as limitações de utilidade, facilitando o uso da mineração de dados no trabalho de rotina e aprimorando a eficácia, o desempenho e produtividade.

Sua importância e capacidade torna-se ampliada em aplicações em domínios sócio-humanos pela capacidade de incorporar conhecimento vago em ontologias difusas. Conhecimento vago é comum nesses domínios e, em caso de serviços de inteligência, é quase todo seu material de trabalho.

Incorporação e formalização de conhecimento vago

Utilizamos variáveis difusas para indicar suspeição de município de origem, do valor da transação, do ISS retido, da empresa prestadora, e da empresa tomadora para compor o índice de nível de *Suspeição de Transação*. Este índice de nível de suspeição foi utilizado com sucesso com três finalidades. Primeiro, compor como atributo as regras, de modo a possibilitar a resposta objetiva às questões propostas, sem maiores dificuldades de interpretação. Segundo, servir como métrica subjetiva, auxiliar às métricas objetivas e de sumarização, para a seleção e ordenação das regras detectados pelo minerador. Terceiro, servir como métrica subjetiva para a

seleção e ordenação dos registros individualizados de cada instância de regra[71] selecionados.

Os resultados alcançados usando essa abordagem podem ser apreciados na TABELA 6.6, TABELA 6.8 e TABELA 6.9 também apresentam essa comparação para casos específicos.

Expressividade, decidibilidade e classificação

Além disso, o uso das ontologias deve ser avaliado em sua capacidade de expressar, decidir e classificar os dados a serem preparados (discretizados e selecionados) para a mineração, e a classificação dos dados (esquemas de regras, registros de regras e respectivas métricas) fornecidos pela mineração. Essa classificação e seleção utiliza os recursos internos da ontologia desenvolvida no *Protégé*, e do mecanismo de inferência estendido pelas regras *SWRL* propiciadas pelo *Jess*.

Tanto a ontologia desenvolvida, quanto o mecanismo de regras apresentado na Seção 5.4 funcionaram a contento, não havendo dificuldades com as classificações.

Instanciação, migração de dados

A ontologia deverá permitir a instanciação a partir de Sistema de Gerenciamento de Bancos de Dados (SGBD) relacional. Uma vez instanciados, os dados serão classificados e essa classificação deverá ser exportada para uso no minerador de dados, ou uso em editores, ou planilhas.

O *Datamaster* (Anexo I.2) permite a conexão com o banco de dados relacional através de conexões ODBC, ou JDBC. Testamos as conexões ODBC com *Mysql* e *Postgresql*. Os dispositivos de conexão via *Mysql* vêm embutidos no *Datamaster* e a conexão foi obtida sem problemas. Já a tentativa de conexão via *Postgresql* apresentou dificuldades, exigiria muito mais esforços de ajustes e configurações e, em vista disso, foi abandonada.

---

[71] Esclarecendo, chamamos de regra “CidadeUf → NivSuspDiscr”. Uma instância desta regra: “RIO BONITO-RJ→6”. Registros desta instância são todos os registros da tabela ServImp, com todos seus atributos, que possuem o atributo CidadeUf=”RIO BONITO-RJ” e o atributo “NivSuspeicaoTransNormDiscr`=6.

A real dificuldade encontrada foi a de instanciar dados em larga escala com o *Datamaster*. A importação de grande volume de dados leva a problemas de escalabilidade e desempenho. Para tentar gerenciar este problema pode-se ajustar o tamanho de pilha da máquina virtual *Java* disponível para o *Protégé*.

Nos testes realizados, alcançamos sucesso na instanciação de transações envolvendo os dados de serviços prestados e importados por Belo Horizonte, abaixo de sete mil registros. Acima desse número, o *software* aparentemente travava, não gerava resultados em tempo razoável e era abortado. Funcionando, a instanciação é lenta, e, caso seja realmente em larga escala, torna-se inviável de ser realizada em somente um passo.

Tais problemas ainda são um entrave para a plena adoção dessas ferramentas de importação e do uso de ontologias para classificar e fazer inferências com grandes volumes de dados. Auer e Ives (2007) apresentam uma solução alternativa e híbrida, integrando os dois esquemas (ontologias e bancos de dados) pela codificação da informação implícita das ontologias usando um conjunto de regras de inferência completas para *SHOIN*. Essas regras de inferência podem ser traduzidas em consultas sobre uma instância de SGBD relacional, e os resultados das consultas (representando inferências) podem ser adicionados de volta a este banco de dados. Subsequentemente, aplicações de banco de dados podem fazer uso direto dessa inferência, conhecimento previamente implícito, por exemplo, na anotação de banco de dados biomédicos. Segundo eles, comparando com raciocinadores de lógica descritiva nativa, OWLDB fornece significativamente maior escalabilidade e capacidades de consulta, sem sacrificar performance com respeito à inferência.

A solução que nós adotamos para gerar resultados em tempo satisfatório foi construir consultas e procedimentos em SQL similares ao mecanismo de inferência ontológico, e gerar a classificação desejada diretamente no banco de dados.

6.3.2 Avaliação dos Resultados

A Prefeitura Municipal de Belo Horizonte publica em seu sítio na Web, a relação de empresas que devem sofrer retenção obrigatória de ISSQN na fonte em vista do prestador de serviços não ter sido localizado no endereço informado por ele mesmo,

configurando-se caso de domicílio fiscal simulado. Esta relação pode ser acessada no seguinte endereço: http://www.pbh.gov.br/bhissdigital/download/Retencao.pdf.

Este trabalho de descaracterização de domicílio fiscal simulado vem sendo realizado sistemática e continuamente desde 2003, quando apurou-se que este fato era responsável por uma significativa evasão de receitas do Tesouro Municipal.

Este levantamento permite-nos avaliar a validade dos resultados da mineração através da comparação entre o número de transações oriundas de domicílio fiscal simulado e o número de transações detectadas na mineração de dados, conforme a TABELA 6.10. O número de transações oriundas de domicílio fiscal simulado foi obtido cruzando os registros de transações referentes a serviços importados, usados na mineração de dados, com a relação oficial de empresas descaracterizadas agrupando-se por cidade. O número de transações detectadas na mineração de dados foi obtido agrupando-se os resultados da mineração por cidade.

TABELA 6.10 Validação dos resultados, índice de detecção de transações envolvendo domicílios fiscais simulados

CidadeUf	N° de transações oriundas de domicílio fiscal simulado	N° de transações detectadas na mineração de dados	Percentual de acerto
RIO ACIMA-MG	1.954	1.953	99,95
LAGOA SANTA-MG	163	162	99,39
ESMERALDAS-MG	158	158	100,00
BETIM-MG	102	102	100,00
PRUDENTE DE MORAIS-MG	78	63	80,77
PARA DE MINAS-MG	75	70	93,33
NOVA LIMA-MG	74	74	100,00
MARAVILHAS-MG	72	32	44,44
OURO BRANCO-MG	72	70	97,22
JEQUITIBA-MG	63	59	93,65
JUATUBA-MG	42	42	100,00
JABOTICATUBAS-MG	41	37	90,24
OLIVEIRA-MG	40	40	100,00
VESPASIANO-MG	38	38	100,00
ALFREDO VASCONCELOS-MG	34	0	0,00
PIRACEMA-MG	32	6	18,75
CAETE-MG	30	30	100,00
CONTAGEM-MG	26	23	88,46
ITABIRITO-MG	26	26	100,00

TABELA 6.11 Validação dos resultados, índice de detecção de transações envolvendo domicílios fiscais simulados (continuação...)

CidadeUf	N° de transações oriundas de domicílio fiscal simulado	N° de transações detectadas na mineração de dados	Correção (%)
FLORESTAL-MG	21	21	100,00
SANTA LUZIA-MG	21	21	100,00
BALDIM-MG	19	19	100,00
ITAGUARA-MG	18	18	100,00
SAO JOAQUIM DE BICAS-MG	17	17	100,00
BRUMADINHO-MG	15	15	100,00
TAQUARACU DE MINAS-MG	14	14	100,00
PEDRO LEOPOLDO-MG	12	12	100,00
IGARAPE-MG	10	10	100,00
ITAUNA-MG	10	10	100,00
BELO VALE-MG	9	9	100,00
OLIVEIRA FORTES-MG	6	6	100,00
CAPIM BRANCO-MG	3	3	100,00
CARMOPOLIS DE MINAS-MG	2	2	100,00
IBIRITE-MG	1	1	100,00
MATOZINHOS-MG	1	1	100,00
RIO CASCA-MG	**1**	**0**	**0,00**
TOTAL	**3.300**	**3.164**	**95,88**

Esta apuração envolve a realização de diligência de reconhecimento nesses endereços e, portanto, limitam-se a cidades relativamente próximas a Belo Horizonte, num raio não superior a 200 Km. Para esses casos, **obtivemos uma taxa de sucesso global de 95,88%**. As taxas de insucesso mais significativas, como as de MARAVILHAS, ALFREDO VASCONCELOS, e PIRACEMA, estão associadas à baixa relevância dos valores transacionados: o minerador de dados não descreveu essas regras em vista de seus baixos valores de serviços associados.

Além dessa distância geográfica, podemos dispor dos casos de SAQUAREMA-RJ, RIO BONITO-RJ, SANTANA DE PARNAÍBA-SP e BARUERI-SP, notoriamente conhecidos, e que foram destacados na mineração de dados. Nesse último caso, o de Barueri, na análise o nível de suspeição é amenizado considerando-se a constatação de real núcleo de dinamismo econômico no município representado pelo centro empresarial de Alfaville.

Com base nesses resultados consideramos que a mineração de dados pela abordagem difusa foi altamente bem sucedida em confirmar a ocorrência de casos já conhecidos. Os novos casos apontados merecem a atenção da auditoria tributária, mediante análises mais aprofundadas e ações fiscais de apuração, a fim de se confirmar a relevância das ocorrências.

7 Conclusão

A compreensão do fenômeno da vaguidade e de seus efeitos de incerteza, imprecisão, inexatidão, ou indeterminação é importante para a Ciência da Informação. Essa compreensão é necessária para avaliarmos as soluções disponíveis e desenvolvermos soluções próprias na representação, recuperação, validação e análise da informação sujeita às influências deste fenômeno.

Também a compreensão das peculiaridades da fundamentação do conhecimento em domínios sócio-humanos, e o reconhecimento de que estas peculiaridades exigem a adoção de abordagens específicas e distintas das de domínios de ciências naturais levam à prospecção de abordagens mais adequadas para se tratar a informação nos domínios sócio-humanos.

O uso de metodologias e técnicas da área de sistemas difusos já se mostram adequadas ao tratamento do fenômeno da vaguidade e das peculiaridades dos domínios sócio-humanos. Além disto, a área de sistemas difusos é ativa e em constante evolução, o que a torna bastante promissora para o suporte de desenvolvimento de aprimoramentos e novas soluções.

Mineração de dados exige um pleno conhecimento do domínio de aplicação. A incorporação e formalização de conhecimento especialista através de ontologias difusas é capaz não somente de dar suporte e aprimorar os projetos de mineração de dados, mas também é elemento vital para sua própria viabilidade e sucesso. Onde abordagens “clássicas” falharam, há chances desta abordagem ser bem sucedida.

O uso da ontologia disciplina a modelagem do conhecimento e agrega ao sistema de mineração de dados suas vantagens usuais de explicitar o conhecimento, validá-lo do ponto de vista da consistência, completude e expressividade lógica, além de torná-lo compartilhável, reutilizável, e interoperável.

O uso de lógica e sistemas difusos para modelar e apresentar o conhecimento em linguagem natural dá-nos ganhos em nossa capacidade de analisar, explicar, compreender e interpretar os fatos do domínio de conhecimento empírico evidenciados nas regras resultantes.

Apresentamos o estado da arte em ontologias difusas, com distintas formalizações de estruturas ontológicas, e ainda relacionamos os esforços no sentido de se desenvolverem raciocinadores difusos em lógica descritiva. Esta área também revelou-se ativa, em constante evolução e promissora para dar suporte a soluções em domínios sócio-humanos.

Considerando-se e descontando-se as predisposições individuais e subjetivas na criação de variáveis difusas que possam expressar o conhecimento de natureza vaga, podemos criar mediante funções difusas de agregação, métricas subjetivas que, atuando em conjunto com as métricas objetivas e dados de sumarização, permitem selecionar e ordenar adequadamente as regras de interesse. Desse modo, questões que tradicionalmente exigiam uma extensa e custosa análise, agora podem ser respondidas de modo mais direto e rápido a partir dos dados, modelando-se e construindo-se variáveis difusas e métricas subjetivas adequadas à solução de cada problema.

A abordagem por computação suave, raciocínio aproximado e, em especial, a modelagem baseada em lógica difusa adotada é capaz de apresentar uma solução alternativa em casos de precariedade de dados: pode-se associar conhecimento especialista na criação de um atributo a partir de aspectos copresenciais que percorra todo o processo de mineração de dados, participando das regras, e atuando com medida subjetiva.

Os resultados apresentados sustentam fortemente que o uso de ontologias difusas no suporte à mineração de dados possui aplicações potenciais em auditoria tributária. É dado suporte para a auditoria identificar indícios de fraude e sonegação nos serviços importados por empresas de Belo Horizonte e ainda extrapolar tal aplicação para outras situações de auditoria em amplos contextos. Essa capacidade permitirá utilizar melhor os recursos humanos em auditoria e minimizar a fraude e sonegação, aumentar a eficiência e produtividade, e apresentar resultados mais apurados.

Pesquisas Futuras

Ao longo da tese chegamos a pesquisar, apresentar, comentar e esboçar alguns aspectos que não tiveram a oportunidade de serem plenamente desenvolvidos como os raciocinadores *DeLorean* e *fuzzyDL* por não estarem plenamente desenvolvidos e não puderem ser utilizados. Uma continuidade natural de nosso trabalho será aguardar sua liberação e incorporá-los em nossos testes, simulações e avaliações.

Além disto, podemos citar como uma continuidade natural a modelagem por hipercubos nebulosos que leva à questão dos espaços conceituais; a relevância do uso de ontologias no contexto da governança eletrônica no âmbito dos documentos e declarações fiscais; a acoplagem de técnicas de análise de inteligência e captura de conhecimento coletivo à abordagem desenvolvida; e a prospecção e identificação de tipos de determinações de diversas naturezas e relevantes para o processo de análise de inteligência.

Espaços Conceituais

No que diz respeito à modelagem por hipercubos, imaginamos utilizá-la na modelagem de conceitos mais complexos e com atributos difusos, tais como os conceitos de obra, construção civil e serviços de engenharia, assim como na modelagem de atributos associados ao conceito de dinamismo municipal.

A modelagem via hipercubos é apenas uma via aberta que hoje pode ser traçada a partir dos polítopos de Coxeter às tesselações de Voronoi e que obras como as de Peter Gärdenfors e Dominic Widdow[72] permitem-nos conjecturar a real possibilidade de articularmos conceitos em espaços conceituais dotados de um arcabouço lógico e matemático sólido, para representar a realidade[73]. Tal via é típica da Ciência da Informação por requerer a junção de vários saberes de profissionais "alfabéticos" e "numéricos", em especial, linguistas, filósofos e lógicos dando suporte a matemáticos geômetras, algebristas e analíticos, para o desenvolvimento da necessária axiomatização da área.

---

[72] Peter Gärdenfors: *Conceptual Spaces – The Geometry of Thought* (GÄRDENFORS, 2000). Dominic Widdows: *Geometry and Meaning* de 2004 (WIDDOWS, 2004).

[73] Vide ainda: em Aisbett e Gibbon (2001), em Rickard (2006), e Rickard, Aisbett e Gibbon (2007), Rickard e Yager (2007) e Gärdenfors (2005).

Ontologias, Governo Eletrônico, declarações e documentos fiscais eletrônicos

Recentemente, o W3C Brasil (*World Wide Web Consortium*[74], sucursal brasileira[75]) criou um grupo de trabalho (W3C Brasil, 2011) dedicado ao tema de modelagem conceitual e interoperabilidade semântica em governo eletrônico (GT Ontologias[76]). Isso reflete a crescente importância que vem sendo dada à capacitação de interoperabilidade dos inúmeros sistemas desenvolvidos pelas várias esferas de governo.

A área tributária é a área que mais demanda tecnologia da informação, e crescentes necessidades de serviços Web. Seja para o contribuinte apresentar declarações e emitir documentos, seja para solicitar informações referentes à suas situações perante o Fisco, essas demandas vão surgido em crescente com a introdução em nível nacional da Nota Fiscal Eletrônica (NF-e), do Sistema Público de Escrituração Digital (SPED), Escrituração Fiscal Digital (EFD), e da Declaração Anual do Simples Nacional (DASN). Além desses há inúmeros outros documentos e declarações, aos quais se adicionam outros exigidos em níveis estaduais e municipais. No caso da Prefeitura Municipal de Belo Horizonte, podemos citar a Declaração Eletrônica de Serviços (DES), DES-IF (Declaração Eletrônica de Serviços de Instituições Financeiras) e a Nota Fiscal de Serviços Eletrônica (NFS-e).

Como vimos, ontologias permitem o compartilhamento de estruturas conceituais, informações, facilitam a comunicação e interoperabilidade entre sistemas. Desse modo, considerando-se o contexto do domínio, a introdução e consolidação do uso de ontologias no âmbito da administração pública dará suporte ao desenvolvimento e consolidação dos sistemas que vêm sendo demandados e constituem as camadas de comunicação entre o Governo e o cidadão. Tal via pode ser adotada entre o Fisco e os contribuintes no desenvolvimento e implantação das declarações e documentos fiscais eletrônicos e sua auditoria.

No esteio, voltando ao foco da tese, e no âmbito da auditoria tributária da PBH, haverá casos de estudo que muito poderão se beneficiar dessa abordagem. Há grandes subdomínios a serem explorados, por serem notoriamente campeados por

---

[74] Vide: http://www.w3.org/
[75] Vide: http://www.w3c.br/Home/WebHome
[76] Vide: http://www.w3c.br/GT/GrupoOntologias

sonegação e fraudes das mais variadas formas. Dentre eles podemos citar os serviços de engenharia e construção civil; serviços de propaganda e publicidade, sociedades de profissionais liberais, e serviços associados, ou prestados por instituições financeiras. Ontologias desenvolvidas nesses subdomínios seriam de interesse de todos os municípios brasileiros.

Análise de Inteligência

Considerando-se a área de Análise de Inteligência, escopo principal de aplicação de nossa solução, destacamos a possibilidade de se aplicar nesta área tanto ontologias, quanto a solução conjugada descrita. Tal proposição, em verdade, já se encontra além da mera possibilidade, pois, comunidades de inteligência já vêm adotando ontologias no auxílio de suas tarefas.

Uma ilustração do uso de ontologias pelas comunidades de inteligência pode ser vislumbrada nos trabalhos apresentados nos encontros científicos denominados *Ontologies for the Intelligence Community (OIC)*. A relação e acesso direto aos artigos publicados, nos encontros realizados de 2007 a 2009, estão disponíveis através dos sítios: <http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-555/>; <http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-440/> e <http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-299/>.

Além desse fórum há trabalhos de ontologias e mineração de dados em blogs e redes sociais publicados nos encontros da *Semantic Technology For Intelligence, Defense, And Security (STIDS)* realizados em 2010, 2011 e já com chamadas de trabalhos para outubro de 2012. Para visualizar publicação de encontros passados e a chamada para o evento a se realizar, acesse <http://stids.c4i.gmu.edu/index.php>. Há também o livro *Ontologies and Semantic Technologies for Intelligence*[77] de 2010 que reúne uma série de trabalhos publicados nesses encontros.

Além desses desenvolvimentos, há de se verificar a adoção de métodos de análise coletiva de informação, considerando intersubjetividade e consenso tais como os já citados Métodos Delphi e Método de Hipóteses Concorrentes (ver Seção 5.4.5

[77] OBRST, L., JANSSEN, T., CEUSTERS, W. (Eds.). Ontologies and Semantic Technologies for Intelligence. *Frontiers in Artificial Intelligence and Applications*. Volume 213, January 2010, p. 236.

dessa tese) oriundos da área de Análise de Inteligência. Tais métodos buscam capturar e gerenciar o conhecimento coletivo e amenizar as predisposições cognitivas na determinação desse conhecimento. Em todos os métodos há vantagens e desvantagens em sua adoção que devem ser bem avaliadas pelo analista de inteligência. Voltemos à figura que nos é mais cara em todo este trabalho, fazendo a vinculação dela com as técnicas de Análise Estruturada de Inteligência.

A despeito de não se originar do meio, pudemos avaliar que o uso do método GUHA – *General Unary Hypotheses Automaton* (HAJÉK, HOLEÑA e RAUCH, 2010) na construção e teste de hipóteses em mineração de dados, assim como da ferramenta *4ft-Miner* inspirado no método GUHA (SVÁTEK e RAUCH, 2005) podem também ser promissores. O método GUHA vem sendo desenvolvido e aprimorado por Pétr Hajék, um dos grandes axiomatizadores da lógica difusa, desde 1976. Do ponto de vista da mineração de dados, o método GUHA foi um dos primeiros métodos a serem desenvolvidos.

Na área de sistemas difusos é comum a preocupação com a captura e gerenciamento do conhecimento coletivo através da opinião de especialista, conforme podemos comprovar na relação abaixo.

- TANINO, T. On group decision making under fuzzy preferences*. In:* J. Kacprzyk, M. Fedrizzi (Eds.).*Multiperson Decision Making Using Fuzzy Sets and Possibility Theory.* Kilowatt Academic Publishers, Dordrecht, 1990, p. 172-185.
  - AMBIGUOS, Ishikawa, M.SHIGA, T. e TOMIZAWA, R. Tactic e H. Mileage. The max-min Delpi method and fuzzy Delphi method via fuzzy integration. *Fuzzy Sets and Systems*, Vol. 55, 1993, p. 241-253.
- LAN, Jibin; HE, Liping e WANG, Zhongxing. A New Method for Fuzzy Group Decision Making Based on α-Level Cut and Similarity.*Fuzzy Systems and Knowledge Discovery (FSKD),* 2005, p. 503-513.
- BARDOSSY, A.; Duckstein, L e BOGARDI. Combination of fuzzy numbers representing expert opinions. *Fuzzy Sets and Systems*, Vol. 57, 1993, p. 173-181.
- LEE, H. S. Optimal consensus of fuzzy opinions under group decision making environment. *Fuzzy Sets and Systems*, Vol. 132, 2002, p. 303-315.
- FEDRIZZI, M. e KACPRZYK, J. On measuring consensus in the setting of fuzzy preference relations. *In:* J. Kacprayk e M. Roubens, Eds. *Non-conventional preference Relations in Decision Making*. Springer, Berlin, 1988, p. 129-141.

- HSU, H. M. e CHEN, C. T. Aggregation of fuzzy opinions under group decision making. *Fuzzy Sets and Systems*, Vol. 79, 1996, p. 279-285.
- KACPRZYK, J. e FEDERATION, M. A soft measure of consensus in the setting of partial (fuzzy) preferences. *Eur. J. OR,* Vol. 34, 1988, p. 315-325.
- KACPRZYK, J.; FEDERATION, M. e , H. Group decision making and consensus under fuzzy preferences and fuzzy majority. *Fuzzy Sets and Systems*, Vol. 49, 1992, p. 21-31.
- NURMI, H. Approaches to collective decision making with fuzzy preference relations. *Fuzzy Sets and Systems*, Vol. 6, 1981, p. 249-259.
- WILLIAMS, Jon e STEELE, Nigel. Difference, distance and similarity as a basis for fuzzy decision support based on prototypical decision classes. *Fuzzy Sets and Systems*, Vol. 131, 2002, p. 35-46.

Uma abordagem mais abrangente na avaliação de limitações, condicionantes e métodos que se utilizam de juízos humanos para medições é feita na área de psicologia. Uma descrição sucinta dessas abordagens é feita por Douglas Hubbard (2010) no Capítulo 12 de seu livro *How to Measure Anything – Finding the Value of "Intangibles in Business,* com destaque para os métodos Lens[78] e Rasch[79].

Prospecção e identificação de determinações de diversas naturezas, em especial, de determinações causais

As ontologias incorporam naturalmente relações taxonômicas de caráter hierárquico. Essas relações possibilitam a construção da estrutura da ontologia. Entretanto, uma grande área de interesse e aplicabilidade é a consideração de outros tipos de relações[80] como elementos estruturantes, vinculantes e manifestadores de aspectos semânticos dos conceitos incorporados às ontologias, e que são úteis em uma análise mais refinada dos resultados do processo de descrição da base de conhecimento empírico (no caso, via mineração de dados).

---

[78] BRUNSWIK, Egon. Representative Design and Probabilistic Theory in a Functional Psychology. *Psychological Review*, 62, 1955, p. 193-217.

[79] RASCH, G. On General Laws and the Meaning of Measurement in Psychology, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, University of California Press, 1980, p. 321-334.

[80] Vide, por exemplo:
WINSTON, M. E.; CHAFFIN, R.; HERRMANN, D. *A taxonomy of part-whole relations*. 1987. Disponível em: http://www.cogsci.rpi.edu/csjarchive/1987v11/i04/p0417p0444/main.pdf. Acesso em 20-12-2006.
BITTNER, Thomas; DONNELLY, Maureen e SMITH, Barry. Individuals, universals, collections: On the foundational relations of ontology. *In*: *Proceedings of the International Conference on Formal Ontology in Information Systems*, FOIS04, 2004.

Os mineradores de dados ao descreverem a realidade através de associações com base em critérios de frequência, descobrem uma série de associações que podem ser classificadas como distintos tipos de relações. Essas relações são determinações que podem ser mereológicas (todo/parte), de identidade, tipo *Cambridge*[81], micro-macro[82] e causais (KIM, 1974), semânticas e de naturezas diversas.

Haverá determinações parte-todo que apresentarão alto nível de correlação, possuindo confiança de 100%, ou próxima a 100%[83]. No nosso estudo de caso, o melhor exemplo dessa ocorrência é a determinação que associa cidade à unidade federativa[84]. Uma determinação desse tipo, relacionada à identificação de localidade é irrelevante, não tendo nenhum valor para a análise e deve ser tratada com uma modelagem de dados adequada.

Já outras determinações parte-todo como a que puder associar sócios, responsáveis e contadores (as partes) numa rede social (o todo) indicando a possível existência de associação econômica seria altamente relevante. No nosso caso, pudemos associar matriz e filiais de uma mesma empresa através do CNPJ raiz. Uma base de conhecimento especialista que distinga as associações parte/todo de interesse será de relevante utilidade.

Outro tipo de distinção de determinação relevante seria a do tipo Cambridge:

> "Uma 'mudança Cambridge' é dita ocorrer a um objeto se há um predicado verdadeiro dele em um tempo, mas falso em um tempo mais tarde. (de acordo com Geach, isto era o critério de

---

[81] Falsas relações causais associadas a mudanças, tal como a associação entre o toque de sirenes em Manchester anunciando o fim da jornada de trabalho, com o pôr do sol em Londres. A introdução desse termo é atribuída a Peter Geach. GEACH, P.T. *Mental Acts, their Content and their Objects*, London: Routledge and Kegan Paul. *God's relation to the World* as repr. in Geach, *Logic Matters*, Oxford: Blackwell, (1957), 1972, 318-27.

[82] Na verdade, o grande debate envolve a indeterminação micro, considerando-se o domínio caracterizado pela indeterminação e as repercussões da mecânica quântica, e a determinação macro, caracterizado pelo "domínio natural", naturalmente perceptível por nós. Vide o debate:

- GLYMOUR, Bruce e SABATÉS, Marcelo. Micro-Level Indeterminism and Macro-Level Determinism. *The Proceedings of the Twentieth World Congress of Philosophy*, Vol. 10, 2001, p. 11-18.
- SPERRY, R. W. Macro- versus Micro-Determinism. *Philosophy of Science,* Vol. 53, N° 2, 1986, p. 265-270.

[83] Numa relação com 5565 municípios foram detectados 276 nomes idênticos.

[84] Daí termos conjugado os campos de "Cidade" e "Uf" em um único campo "CidadeUf" e ter submetido à mineração de dados somente o campo conjugado.

> ‘mudança’ advogado por filósofos ilustres de Cambridge tais como Russell e McTaggart). Assim todas as mudanças reais são mudanças Cambridge – ao menos aquelas que são representáveis por predicados – mas o contrário claramente não é verdadeiro”. Deixe-nos emprestar o termo “Cambridge” com uma leve modificação: nós dizemos “mudança Cambridge” ou “evento Cambridge” onde Geach diria “mera mudança Cambridge”, etc.

Kim (1974, p. 22-32) toma emprestada essa definição de “mera mudança Cambridge” explicitada por Peter Geach para estendê-la e designar o evento Cambridge, ou mudança Cambridge para aludir a determinações do tipo:

> o evento a “morte de Sócrates” determina o evento “Xantipa tornar-se viúva”

O que está em questão é o questionamento da natureza da determinação. Kim argumenta, assim com Geach que aqui não temos uma relação causal. O evento “a morte de Sócrates” não é causa do evento “Xantipa tornar-se viúva”. A despeito de serem eventos distintos e imediatos, não ocorrem no mesmo espaço e não são simétricos. O que ocorre é que o evento “Xantipa tornar-se viúva” é parasitário do evento “morte de Sócrates”.

Da mesma forma, em nosso contexto, a omissão de pagamento do contribuinte não é causa dele ficar inadimplente. Trata-se também de uma determinação do tipo Cambridge. Nem a morte de Sócrates deve ser confundida com a causa da viuvez de Xantipa, e nem a omissão do pagamento do tributo com a causa da inadimplência do contribuinte. Ignorarmos isso, leva-nos a abstermo-nos da busca das causas reais e, consequentemente, comprometermos fatalmente nossa capacidade de explicar, interpretar e compreender a realidade.

Se quisermos realmente explicar a causa da inadimplência do contribuinte, teríamos que experimentar recorrer, por exemplo, a fatores macro-econômicos, a mudanças de legislação, ou a contingências (desastres, acidentes, roubos, etc.). Também podem ser decorrentes da livre adoção de práticas fraudulentas de nível sistêmico, seja pela mera adoção de planejamento tributário equivocado e ilícito, após a contratação de determinada empresa, ou profissional consultor (contador, advogado, preposto, etc.), ou à livre associação com o crime organizado.

Dentre os diversos tipos de determinações (relações) que podem ser de interesse, destacam-se as relações causais. As relações causais possuem a capacidade de estabelecer um nexo, uma cadeia, uma rede ou campo causal que suportariam efetivamente as tarefas de explicação útil e, em especial, da predição de ocorrências. Essas capacidades de explicar e prever são as capacidades, por exemplo, que permitiriam a uma auditoria de detecção de fraudes a reprimir e prevenir os mecanismos de realização dessas fraudes.

Em uma tarefa de análise de inteligência, a análise causal, seja considerando a causa como um componente natural, ou como uma ilusão útil para a compreensão da natureza, é um elemento chave para as tarefas de explicação e interpretação (compreensão) que, por sua vez, dão oportunidade à tarefa de predição, a antecipação do que pode ocorrer.

No caso da tarefa de predição, ela é inerente à atividade de análise de inteligência. É tal discernimento que se identifica com a inteligência fiscal, a capacidade de prever, de antecipar e não meramente reagir. Enquanto a auditoria tributária, ou qualquer serviço que se pretenda serviço de inteligência se contentar, ou se especializar em tratar relações do tipo Cambridge, em verdade, jamais alcançarão o patamar de um real serviço de inteligência.

Hoje a auditoria tributária encontra-se no patamar de buscar descrições, e nessas descrições atém-se quase exclusivamente a determinações do tipo Cambridge. Tal tarefa é importante, em vista de detectar comportamentos que levaram à omissão, ou à sonegação fraudulenta do imposto. Entretanto, o sucesso na tarefa de explicação é apenas um dos aspectos da auditoria. Sua real eficiência baseia-se na tarefa de explicação, interpretação e, a partir daí, em sua capacidade de realizar a tarefa de predição.

Paralisar nossa análise em determinações do tipo Cambridge, permite que detectemos a inadimplência e, consequentemente, tomemos atitudes no sentido de saná-la, pela repressão (aplicação de penalidades) e cobrança. Entretanto, não elevamos a análise ao nível da análise de inteligência que permitirá que tomemos a frente de todo o processo e possamos passar a adotar uma atitude de predição.

Essa situação fica evidente considerando-se a simples lei empírica capturada da observação dos fatos, amplamente reconhecida pelos órgãos fiscalizadores e procuradorias em todo o mundo:

> “quanto mais tempo se leva para se constituir e cobrar o crédito tributário, menor a possibilidade de recuperá-lo”

Há muito é sabido que quanto maior o tempo para a cobrança, maior a chance das empresas desaparecerem, assim como seus sócios e responsáveis, ou, por várias vias, lícitas, ou ilícitas, incrementa-se a blindagem financeira dos sócios.

A dívida ativa na União encontra-se na faixa de 2 trilhões de reais (D’ANGELO, 2012). No município de Belo Horizonte encontra-se na faixa de 4 bilhões de reais, em 2008 (FURBINO, 2008). A maior parte desses recursos é moeda podre, incapacitada de se traduzir em real riqueza pública. Outra parte costuma se volatilizar em anistias e descontos que premiam a fraude, a sonegação, e promovem a injustiça fiscal, incentivam a recorrência ao débito, e confirmam a tese de que é bom negócio dever para o Estado.

O desenvolvimento da capacidade da auditoria fiscal via a adoção de técnicas de análise estrutura de inteligência, privilegiando as tarefas de predição, reduziriam o volume de recursos a serem inscritos em Dívida Ativa. O serviço de inteligência sabendo de antemão da abertura de licitações de obras, realizando o monitoramento de empresas de planejamento tributário, supervisionando a movimentação de sócios e de seus prepostos, a constituição informal de consórcios e outros tipos de agrupamentos econômicos, etc. a antecipação da própria ocorrência do fato gerador do tributo elevará a capacidade de arrecadação.

Referências

ABULAISH, Muhammad e DEY, Lipika. Interoperability among Distributed Overlapping Ontologies – A Fuzzy Ontology Framework. *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web, 2006, 7 p.*

AGRAWAL, R. e SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. *In: 20th International Conference on Very Large Data Bases*, 1994, p. 487–499.

AISBETT, Janet e GIBBON, Greg. A general formulation of conceptual spaces as a meso level representation. *Artificial Intelligence*, Vol, 133, 2001, p. 18-232.

ANSCOMBE, G. Elizabeth M. *Intencion*. Ediciones Paidós, 1991 (1957).

ASCHOFF, Felíx-Robinson; SCHMALHOFER, Franz e van ELST, Ludger. Knowledge Mediation: A Procedure for the Cooperative Constrution of Domain Ontologies. *In:Proceedings of the ECAI-2004*, Workshop on Agent-mediated Knowledge Management, AMKM, 2004, p. 29-38.

AUER, Sören e IVES, Zachary G. *Integrating Ontologies and Relational Data.* Department of Computer & Information Science, Technical Reports (CIS). University of Pennsylvania, 2007, 13 p. Disponível em: http://repository.upenn.edu/cgi/viewcontent.cgi?article=1753&context=cis_reports. Acesso em 12-12-2011.

BAESENS, B.; VIAENE, S. e VANTHIENEN, J. (2000). *Post-Processing of Association Rules*. Research Report 0020, 2000.

BAX, Marcello Peixoto e COELHO, Eduardo de Mattos Pinto Coelho. Compromissos Ontológicos e Pragmáticos em Ontologias Informacionais: Convergências e Divergências. *Revista Datagramazero*, aceito para publicação, junho de 2012.

BAYARDO Jr., Roberto e AGRAWAL, Rakesh. Mining the Most Interesting Rules. *Fifth ACM SIGKDD Int. Conf On Knowledge Discovery and Data Mining*, 1999, p. 145-154.

BELIAKOV, Gleb and WARREN, Jim.Appropriate choice of aggregation operators in fuzzy decision support systems*, IEEE transactions on fuzzy systems*, Vol. 9, N° 6, 2001, p. 773-784.

BELOHLAVEK, Radim; KLIR, George J.; LEWIS, Harold W. e WAY,. Eileen C. Concepts and fuzzy sets: Misundertandings, misconceptions, and oversights. *International Journal of Approximate Reasoning*, Vol. 51, 2009, p. 23-34.

BEN-GAL, Irad. Outlier Detection. *In:* Maimon, Oded e Rokach, Lior (Eds *The Data Mining and Knwoledge Discovery Handbook*, Springer, 2005, p. 131-148.

BENEDIKTSSON, Jon Atli e SWAIN, Philip H. Consensus Theoretic Classification Methods. IEEE, *Transactions on Systems, Man & Cybernetics*, Vol. 22 N° 4, 1992, p. 688-704

BENTA, Kuderna-Iulian; RARAU, Anca e CREMENE, Marcel. Ontology Based Affective Context Representation. Technical University of Cluj-Napoca, Romania, EATIS, 2007.

BERNERS-LEE, Tim; HENDLER, James e LASSILA, Ora. The Semantic Web – A new form of Web contente that is meaningful to computers will unleash a revolution of new possibilities. Scientific American. Maio, 2001. Disponível em <http://www.scientificamerican.com/article.cfm?id=the-semantic-web>. Acesso em 01-02-2012.

BITTNER, Thomas; SMITH, Barry e DONNELLY, Maureen. Systems of Granular Partitions. 2010. Disponível em: http://ontology.buffalo.edu/smith/articles/BittnerSmithDonnelly.pdf. Acesso em 01-10-2010.

BITTNER, Thomas; DONNELLY, Maureen e SMITH, Barry. Endurants and perdurants in directly depicting ontologies. *AI Communications*, v. 17, Issue 4, October, 2004, p. 247-258.

BITTNER, Thomas e SMITH, Barry. A Theory of Granular Partitions. Foundations of Geographic Information Science, . M. Duckham, M. F. Goodchild and M. F. Worboys, eds., London: Taylor & Francis Books, 2003a, 117-151. Disponível em: http://ontology.buffalo.edu/smith/articles/partitions.pdf. Acesso em 01-10-2010.

BITTNER, Thomas e SMITH, Barry. Granular Spatio-Temporal Ontologies. AAAI Symposium: Foundations and Applications of Spatio-Temporal Reasoning FASTR, 2003b, p. 12-17. Disponível em: http://ontology.buffalo.edu/smith/articles/granular_ontologies.pdf. Acesso em 01-10-2010.

BITTNER, Thomas e SMITH, Barry. Vague Reference and Approximating Judgments.Spatial Cognition and Computation, 3: 2, 2003c, p. 137–156. Disponível em: http://ontology.buffalo.edu/smith/articles/vraj.pdf. Acesso em 01-10-2010.

BITTNER, Thomas e SMITH, Barry. A taxonomy of partitions. 2001a. Disponível em <http://ontology.buffalo.edu/smith/articles/Bittner-Smith-cosit01.pdf>. Acesso em 01-10-2010.

BITTNER, Thomas e SMITH, Barry. Granular Partitions and Vagueness. *In:* Christopher Welty and Barry Smith (eds.), Formal Ontology and Information Systems, New York: ACM Press, 2001b, p. 309–321. Disponível em: http://ontology.buffalo.edu/smith/articles/vagueness.pdf. Acesso em 01-10-2010.

BLACK, Max. Vagueness: An Exercise in Logic & Analysis, *Philosophy of Science*, 4 (1937), 427-455 – In: *Margins of Precision: Essays in Logic and Language*, 1970.

BLACKBURN, Simon. *Dicionário Oxford de Filosofia*, Jorge Zahar Editor, 1997, 437 p.

BOBILLO, Fernando e STRACCIA, Umberto. Fuzzy ontology representation using OWL 2. *International Journal of Approximate Reasoning*, 52, 2011, p. 1073-1094.

BOBILLO, Fernando; DELGADO, Miguel; GÓMEZ-ROMERO, Juan e STRACCIA, Umberto. Fuzzy description logics under Gödel semantics. *International Journal of Approximate Reasoning*, 52, 2009, p. 494-514.

BOBILLO, Fernando; DELGADO, Miguel e GÓMEZ-ROMERO, Juan da Costa. Crisp Representations and Reasoning for Fuzzy Ontologies. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 17, N°. 4, 2009, p. 501-530.

BOBILLO, Fernando. *Managing Vagueness in Ontologies*. PhD Dissertation, University of Granada, Spain, 2008, 298 p.

BOBILLO, Fernando; DELGADO, Miguel e GÓMEZ-ROMERO, Juan da Costa. A Crisp Representation for Fuzzy *SHOIN* with Fuzzy Nominals and General Concept Inclusions. P.C.G. et al. (Eds.), *URSW* 2005-2007, LNAI 5327, 2008a, p. 174-188.

BOBILLO, Fernando; DELGADO, Miguel e GÓMEZ-ROMERO, Juan da Costa. Optimizing the Crisp Representation of the Fuzzy Description Logic *SROIQ*. *In:* P.C.G. et al. (Eds.), *URSW 2005-2007*, LNAI 5327, 2008b, p. 189-206,.

BOBILLO, F.; DELGADO, Miguel e GÓMEZ-ROMERO, J. DeLorean: A Reasoner for Fuzzy OWL 1.1.. *In:Proceedings of the 4th International Workshop on Uncertainty Reasoning for the Semantic Web* (URSW 2008). Karlsruhe, Alemanha, 2008c.

BOBILLO, F. e STRACCIA, U. fuzzyDL: An expressive fuzzy description logic reasoner *In:IEEE International Conference on Fuzzy Systems*, 2008, 923-930 .

BRADLEY, Jeremy. Fuzzy Logic as a Theory of Vagueness: 15 Conceptual Questions. *In:Views on Fuzzy Sets and Systems from Different Perspectives Studies in Fuzziness and Soft Computing*, Volume 243, 2009a, p. 207-228.

BRADLEY, Jeremy. Fuzzy Logic as a Theory of Vagueness: 15 Conceptual Questions. Versão resumida do artigo publicado em livro, 2009b. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=ECD413C1E08BC15FF0B1D369E1EE3614?doi=10.1.1.145.649&rep=rep1&type=pdf>. Acesso em 01-10-2010.

BRANQUINHO, João; DESIDÉRIO, Murcho; GOMES, Nelson Gonçalves. *Enciclopédia de termos lógico-filosóficos*. Martins Fontes, 2006.

BRÄSCHER, Marisa. A Ambigüidade na Recuperação da Informação. *Datagrama Zero*, Revista de Ciência da Informação, v. 3, n° 1, fev., 2002.

BRUHA, Ivan e FAMILI, A. Posprocessing in Machine Learning and Data Mining. *ACM SIGKDD Explorations Newsletter - Special issue on Scalable data mining algorithms,*Volume 2 Issue 2, Dec. 2000.

BURGES, Christopher J. C. Geometric Methods for Feature Extraction and Dimensional Reduction. *In:* Maimon, Oded e Rokach, Lior (Eds).*The Data Mining and Knwoledge Discovery Handbook*, Springer, 2005, p. 59-92.

BURGESS, John Alexander. In defence of an indeterminist theory of vagueness. *Monist* , Volume 81 Número 2, 1998, p. 233-253.

BURGESS, John Alexander. Vague Identity: Evans Misrepresented, *Analysis* 49, 1989, p. 112-119.

CALEGARI, Silvia e CIUCCI, Davide. Towards A Fuzzy Ontology Definition and a Fuzzy Extension of an Ontology Editor. *In: Enterprise Information Systems – Lecture Notes In Business Information Processing*, Volume 3, Part. 3, 2008, p. 147-158.

CALEGARI, Silvia e CIUCCI, Davide. Fuzzy Ontology, Fuzzy Description Logics And Fuzzy-Owl. *In:* Masulli, F.; Mitra, S. e Pasi, G. (Eds.): Wilf 2007, Lnai 45768, 2007, p. 118-126.

CALEGARI, Silvia e CIUCCI, Davide. Fuzzy Ontology, Fuzzy Description Logics And Fuzzy-Owl. *In: Proceedings Of Wilf 2007*. Volume 4578 Of Lncs., 2007, In Printing.

CALEGARI, Silvia e CIUCCI, Davide. Fuzzy Ontology And Fuzzy-Owl In The Kaon Project. *In: Fuzzy Ieee 2007. Ieee* International Conference On Fuzzy Systems, 2007, no prelo.

CALEGARI, Silvia e CIUCCI, Davide. Integrating Fuzzy Logic In Ontologies. ICEIS, 2, 2006, p. 66-73.

CALEGARI, Silvia e SANCHEZ, E. A Fuzzy Ontology-Approach to improve Semantic Information Retrieval. *In*: BOBILLO, F., DA COSTA, P.C.G., D'AMATO, C., FANIZZI, N., Fung, F., LUKASIEWICZ, T.; MARTIN, T.; NICKLES, M.; PENG, Y., POOL, M., SMRZ, P.; VOJTAS, P., (eds.): *Proceedings of the Third ISWC Workshop on Uncertainty Reasoning for the Semantic Web - URSW'07.* Volume 327 of CEUR Workshop Proceedings., CEUR-WS.org, 2007. Disponível Em <Http://Sunsite.Informatik.Rwth-Aachen.De/Publications/Ceur-Ws/Vol-327/Pos_Paper3.Pdf>. Acesso Em 01-10-2010.

CAO, Longbing. Domain-Driven Data Mining: Challenges and Prospects. *Transaction on Knowledge And Data Engineering*, V. 22, n° 6, June, 2010, p. 755-769.

CARVALHO, Veronica Oliveira de. *Generalização de regras de associação utilizando conhecimento de domínio e avaliação do conhecimento generalizado*. Universidade de São Paulo, São Carlos, 2007.

CHAWLA, Sanjay; DAVIS, Joseph e PANDEY, Gaurav. On Local Pruning of Association Rules Using Directed Hypergraphs. *Proceedings of the 20th International Conference on Data Engineering* (ICDE'04), 2004.

CHIBENI, Silvio Seno. What is ontic vagueness?. 3rd Principia *International Symposium*, 8-11, September, 2003. Disponível em <http://www.unicamp.br/~chibeni/public/whatisonticvagueness.pdf>, acesso em 1/10/2010.

CHIZI, Barak e MAIMON, Oded. Dimension Reduction and Feature Selection. *In:* Maimon, Oded e Rokach, Lior (Eds).*The Data Mining and Knwoledge Discovery Handbook*, Springer, 2005, p. 93-112.

COELHO, Eduardo de Mattos Pinto Coelho; BAX, Marcello Peixoto e MEIRA JÚNIOR. Wagner. Suporte de ontologias aplicadas à mineração de dados por regras de associação. *Proceedings of Joint IV Seminar On Ontology Research in Brazil and VI International Workshop on Metamodels, Ontologies, Semantic Technologies*, Gramado, Brazil, September 12-14, 2011, p. 171-176.Disponível em <http://ceur-ws.org/Vol-776/ontobras-most2011_paper20.pdf>, acesso em 28-12-2011.

COELHO, Eduardo de Mattos Pinto Coelho; BAX, Marcello Peixoto e MEIRA JÚNIOR. Wagner. As várias naturezas dos compromissos em ontologias informacionais. *3° Ontobrás – Seminário de Pesquisa Em Ontologia No Brasil*, Florianópolis/SC, 30 e 31 de agosto de 2010.

COELHO, Eduardo de Mattos Pinto Coelho; BAX, Marcello Peixoto e MEIRA JÚNIOR. Wagner. "Ontologias nebulosas no suporte a mineração de dados". *International Conference on Information Systems and Technology Management, 7th CONTECSI. São Paulo, 2010.*

COELHO, Eduardo de Mattos Pinto Coelho; BAX, Marcello Peixoto e MEIRA JÚNIOR. Wagner. Ontologias nebulosas no suporte a mineração de dados. *X Encontro Nacional da Associação Nacional de Pesquisa em Ciência da Informação* (Enancib X). João Pessoa, Paraíba. Disponível em: <http://dci2.ccsa.ufpb.br:8080/jspui/handle/123456789/585>, acesso em 12-12-2009.

CORCHO, O.; FERNÁNDEZ-LÓPEZ, M.; GÓMEZ-PÉREZ, A. e LÓPEZ-CIMA, A. Building legal ontologies with methontology and webode. *In:Law and The Semantic Web*, 2003, p. 142-157.

CORCHO, Oscar; LÓPEZ-FERNÁNDEZ, Mariano e PÉREZ, Asunción Gómez. *OntoWeb* - Technical Roadmap v1.0. IST Project IST-2000-29243, Ontoweb Consortioum, 2001. Disponível em: <http://www.sti-innsbruck.at/fileadmin/documents/deliverables/Ontoweb/D1.1.1.pdf>. Acesso em 01-02-1010.

D'ÂNGELO, ANA. Brasileiros devem R$ 2 trilhões ao Fisco. Correio Braziliense, 15/04/2012. Disponível em: <https://conteudoclippingmp.planejamento.gov.br/cadastros/noticias/2012/4/15/receita-aperta-fiscalizacao-e-divida-chega-a-r-2-trilhoes>. Acesso em 01-02-2012.

DCC/UFMG, *Manual do Tamanduá*. Versão 1.1 , 23 de Agosto de 2006. Departamento de Ciência da Computação. 2006. Disponível na Web em <http://tamandua.speed.dcc.ufmg.br>, acesso em setembro-2006.

DOMINGUES, Ivan. *Epistemologia das Ciências Humanas. Tomo I: Positivismo e Hermenêutica – Durkheim e Weber*. Edições Loyola, 2004, 671 p.

DOMINGUES, Marcos Aurélio; REZENDE, Solange Oliveira. Using Taxonomies to Facilitate the Analysis of Association Rules. *Proceedings of ECML/PKDD'05 The Second International Workshop on Knowledge Discovery and Ontologies* (KDO-2005), Porto, Portugal, 2005, p. 59-66.

DREWNIAK, Józef; DUDZIAK, Urszula. Preservation of properties of fuzzy relations during aggregation processes. *Kybernetika*, Vol. 43, N° 2, 2007, p. 115-132.

DUBOIS, Diddier e PRADE, Henri. On the use of aggregation operations in information fusion processes. *Fuzzy Sets and Systems*, Vol. 142, 2004, p. 143-161.

DUBOIS, Didier e PRADE, Henri. Fuzzy Sets – A Convenient Fiction for Modeling Vagueness and Possibility. *IEEE Transactions On Fuzzy Systems*, Volume 2, Número 1, Fevereiro, 1994.

DUBOIS, Diddier e PRADE, Henri. A review of fuzzy set aggregation connectives. *Information Sciences*, Vol. 36, 1985, p. 85-121.

DUDZIAK, Urszula. Weak and graded properties of fuzzy relations in the context of aggregation process. *Fuzzy Sets and Systems*, Vol. 161, 2, 2010, p. 216-233.

DUMMETT, Michel. Bivalence and Vagueness, *Theoria*, 61, 1995, p. 201-216, 1995.

DUMMETT, Michel. Wang's paradox, *Synthese*, 30, 1975, p. 301-324.

EDGINGTON, Dorothy. Vagueness by degrees. *In:* KEEFE, Rosanna e SMITH, Peter (eds.). *Vagueness: A Reader,* Capítulo 17, 1997, p. 317. p. 294-316,

ÉGRÉ, Paul. *Vagueness, Ambiguity and Perceptual Bistability*. 2009. Disponível em <http://esslli2009.labri.fr/documents/vic09-paul-egre.pdf>. Acesso em 01-10-2010.

ELLER, Markus Pereira. *Anotações Semânticas de Fontes de Dados Heterogêneas* - Um Estudo de Caso com a Ferramenta *Smore*. Dissertação de Mestrado, Departamento de Informática e Estatística, Universidade Federal de Santa Catarina, 2006.

ENOCH, David. Epistemicism and nihilism about vagueness: what's the difference? *Philosophical Studies*,Volume, 133, 2007, p. 285-311.

ESCOVAR, Eduardo L. G.; YAGUINUMA, Cristiane A. e BIAJIZ, Mauro. Using Fuzzy Ontologies to Extend Semantically Similar Data Mining. XXI Simpósio Brasileiro de Banco de Dados, 2006, p.16-30.Disponível Em<Http://Www.Lbd.Dcc.Ufmg.Br:8080/Colecoes/Sbbd/2006/002.Pdf>. Acesso Em 01-10-2010.

EVANS, Garreth. Can there be vague objects?, Analysis 38, 1978, p. 208. Reimpresso *In:* KEEFE, Rosanna e SMITH, Peter. *Vagueness: A Reader,* Capítulo 17, 1997, p. 317.

FAYYAD, U.; UTHURUSAMY, R. Data Mining And Knowledge Discovery In Databases. *Communications Of The Acm*, New York, V. 39, N. 11, p. 26, Nov. 1996.

FERMÜLLER, Christian G. Revisiting Gile's Game: Reconciling Fuzzy Logic and Supervaluation.*In:* Games: Unifying Logic, Language, and Philosophy Logic, Epistemology, and the Unity of Science, 2009, Volume 15, III, p. 209-227.

FERMÜLLER, Christian G. e KOSIK, Robert. Combining supervaluation and degree based reasoning under vagueness. *In:* HERMANN, Miki e VORONKOV, Andrei (Eds.). *Logic for Programming, Artificial Intelligence, and Reasoning. Proceedings of 13th International Converence*, LPAR 2006, p. 212-226. Disponível em <http://www.logic.at/people/chrisf/superluk-corr.pdf>. Acesso em 01-10-2010.

FERNANDES, Eugênio Eustáquio Veloso e GOULART, Maurício Carlos de Paula. A Guerra Fiscal em Matéria de ISS e a Exigência de Tributo e Imposição de Obrigações Acessórias a Contribuintes Não Estabelecidos no Território do Município. Secretaria Municipal de Finanças de Belo Horizonte, 2007. Disponível em <http://www.sinfisco.com.br/files/artigos/guerrafiscal.pdf>. Acesso em 01-02-2012.

FERRAZ, V. R. T.; AFONSO, G. ; YAGUINUMA, C.; BORGES, S.; SANTOS, M. T. P. Fuzzy Ontology-based Semantic Integration of Heterogeneous Data Sources in the Domain of Watershed Analysis. *In: 2º Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais (WCAMA/CSBC 2010).* Porto Alegre: SBC, 2010, p. 565-574.

FINE, Kit. Vagueness, Truth and Logic, *Synthese* 30, 1975, p. 265-300.

FODOR, Jerry e LEPORE, Ernest. What Cannot be Valuated Cannot be Valuated, and it Cannot be Supervaluated Either. *Journal of Philosophy*, 93, 1996, p. 516-535.

FULLÈR, Robert. What is fuzzy logic and fuzzy ontology?. *KnowMobile National Workshop*, October 30, Helsinki, Finland,2008. Disponível em <http://users.abo.fi/rfuller/otaniemi-2.pdf>. Acesso em 01-10-2010.

FURBINO, Zulmira. Prefeitura de BH terceiriza a dívida ativa. Estado de Minas, 11/07/2008. Disponível em: <http://wwo.uai.com.br/UAI/html/sessao_4/2008/07/11/em_noticia_interna,id_sessao=4&id_noticia=71150/em_noticia_interna.shtml>. Acesso em 01-02-2012.

GÄRDENFORS, Peter. Conceptual Spaces – The Geometry of Thought. A Bradford Book, MIT, 2000, 307 p.

GÄRDENFORS, Peter. Chapter 37: Concept Learning and Nonmonotonic Reasoning. *In*: COHEN, Henri e LEFEBVRE, Claire (Eds.) *Handbook Of Categorization In Cognitive Science*. Elsevier, 2005, p. 824-843.

GHORBEL, Hanêne; BAHRI, Afef e BOUAZIS, Rafik. *Fuzzy Protégé for Fuzzy Ontology Models*. 2009 Disponível em: <http://protege.stanford.edu/conference/2009/abstracts/S10P2Ghorbel.pdf>. Acesso em 01-02-2012.

GHORBEL, Hanêne; BAHRI, Afef e BOUAZIS, Rafik. Construction des composants ontologiques flous à partir de corpus de données sémantiques floues. *Actes du XXVIII°Congrés INFORSID*, Marseille, mai, 2010, p 361-376.

GHOSH, Ashih; MEHER, Saroj K. e SHANKAR, B. Uma. A novel fuzzy classifier based on product aggregation operator. *Patter Recognition*, Vol. 41, 2008, p. 961-971.

GOGUEN, J. A. The Logic Of Inexact Concepts. *Synthese*, 19, 1968-69, p. 325-373.

GONÇALVES, Eduardo Corrêa. *Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas*. Universidade Federal Fluminense, Niterói, 2005.

GONÇALVES, Eduardo Corrêa. *Data mining com a ferramenta WEKA*. III Fórum de Software Livre de Duque de Caxias (III FSLDC). Escola Nacional de Ciências Estatísticas (IBGE/ENCE). 2011.

GONÇALVES, Lóren P. F. Um estudo sobre a confiabilidade de ferramentas de mineração de dados. *Revista do CCEI - Centro de Ciências da Economia e Informática (Rev. CCEI - URCAMP)*, v.8, n.14, agosto, 2004, p. 37-47.

GONÇALVES, Lóren P. F. e Freitas. H. Ferramentas de mineração de dados: algo confiável? Porto Alegre; RS: *Anais do XXXVII CLADEA*. Outubro de 2002, anais em CD-ROM.

GOTTWALD, Siegfried, Many-Valued Logic.*The Stanford Encyclopedia of Philosophy*, 2000. http://plato.stanford.edu/entries/logic-manyvalued/. Acesso em 01-10-2010.

GRANGER, G.-G. *A ciência e as ciências*. São Paulo, EDUSP, 1994, 85 p.

GRZYMALA-BUSSE, Jerzy e GRZYMALA-BUSE, Witold J. Handling Missing Attribute Values. *In: The Data Mining and Knwoledge Discovery Handbook*, Springer, 2005, p. 37-58.

GU, Hun-Mao; WANG, Xun; LING, Yun e SHI, Jin-Qin. Building a Fuzzy Ontology of Edutainment Using OWL. *In:* SHI, Y. et al. (Eds.): *ICCS* 2007, Parte III, LNCS 4489, 2007, p. 591-594.

HAACK, Susan. Filosofia das lógicas.Editora Unesp, (1978), 1998.

HAACK, Susan. Deviant Logic, Fuzzy Logic: Beyond The Formalism. The University Of Chicago Press,1994.

HAASE, Peter; HITZLER, Pascal e KRÖTZSCH, Marcus. *Practical Reasoning with OWL and DL-Safe Rules*. ESWC, 2006. Disponível em: <http://korrekt.org/talks/2006/Kroetzsch_Practical-Reasoning-OWL-DL-safe-Rules_ESWC2006.pdf>. Acesso em 01-02-2012.

HAASE, Peter e MOTIK, Boris. A Mapping System for The Integration of OWL-DL Ontologies.*Proceedings of the First International Workshop on Interoperability of Heterogeneous Information Systems (IHS 2005*), 2005. Disponível em: <http://www.cs.ox.ac.uk/boris.motik/pubs/hm05mapping.pdf>. Acesso em 01-02-2012.

HÁJEK, Petr, HOLEŇA, Martin, RAUCH, Jan. “The GUHA method and its meaning for data mining”. *Journal of Computer and System Scienc*e, 2010, 76, 1, p. 34-48.

HÁJEK, Petr. On Vagueness, Truth Value and Fuzy Logics. *Studia Logica*, 91, 2009, p. 367-382.

HEMPEL, C. G. Vagueness and Logic -.*Philosophy of Science,* 6, 1939, p. 163-180.

HORNG Y.-J; CHEN, S.-M e LEE, C.-H. Automatically constructing multirelationship fuzzy concept in fuzzy information retrieval systems. *In: IEEE International Fuzzy Systems Conference,* 2001, p. 606-609.

HORROCKS, Ian; KUTZ, Oliver e SATTLER, Ulrike. The Even More Irresistible SROIQ. 2006.Disponível em <http://www.cs.ox.ac.uk/people/ian.horrocks/Publications/download/2006/HoKS06a.pdf> Acesso em 01-02-2012.

HOUAISS, Antônio. Dicionário Eletrônico Houaiss da língua portuguesa. Versão monousuário 2.0.1, Outubro de 2007. Editora Objetiva Ltda.

HUBBARD, Douglas W.*How to Measure Anything – Finding the Value of "Intangibles in Business.* Second Edition, John Wiley & Sons, 2010. p. 304 p.

HÜLLERMEIER, Eyke. Fuzzy methods in machine learning and data mining: Status and prospects. *Fuzzy Sets and Systems*, 156, 2005, p. 387-406.

HUSSERL, Edmund (1900-1). *Investigaciones logicas*. MORENTE, Manuel G. e GAOS,. Jose (trad.). Revista de Occidente, Madrid, 1976, 777 p.

HYDE, Dominic. Sorites Paradox. 2005. *In:*E. Zalta (ed.), *The Stanford encyclopedia of* philosophy 2ª ed., revised and expanded. (1ª. Ed. 1997) Disponível em: <http://plato.stanford.edu/entries/sorites-paradox/>. Acessado em 01-02-2010.

INSTITUTO BRASILEIRO DE PLANEJAMENTO TRIBUTÁRIO (IBPT) . (Coord.) AMARAL, Gilberto Luiz do; OLENIKE, João Eloi e VIGGIANO, Letícia Mary Fernandes do Amaral. Estudo Sobre o Verdadeiro Custo da Tributação Brasileira. 2008, 19 p.

KEIM, Daniel. Information Visualization and Visual Data Mining. *IEEE Trans. On Visualization and Computer Graphics*, Vol. 7, N° 1, January/March, 2002, p. 100-107.

KHATIB, Sam Al. *The Semantics of Vagueness: Supertruth, Subtruth, and The Cooper Principle.* Simon Fraser University, Summer, 2008.

KIM, Jaegwon. (1974). Noncausal Connections *In:*KIM;J. (1993). *Supervenience and Mind*. Cambridge, Cambridge University Press, 1993.

KLIR, G. J. e YUAN, B. *Fuzzy Sets And Fuzzy Logic*, Prentice Hall: Upper Saddle River, EUA, 1995.

KOHAVI, Ron. *Data Mining and Visualization*. National Academy of Engineering, 2000.

KORPIPÂÂ, Panu. Visualizing constraint-based temporal association rules. *Artificial Intelligence for Engineering Design, Analysis and Manufactoring*, 2001, p. 401-410.

KOSKO, Bart. The Probability Monopoly. *IEEE Transactions On Fuzzy Systems*, Vol. 2, N° 1, February, 1994, p. 32-33.

KOSKO, Bart. Fuzziness Vs. Probability. *Int. J. General Systems,* Vol. 17, 1990, p. 211-240.

KOZAKI, Kouji; SUNAGAWA, Eiichi; KITAMURA, Yoshinoby e MIZOGUCHI, Riichiro. A Framework for Cooperative Ontology Construction Based on Dependency Management of Modules. *International Workshop on Emergent Semantics and Ontology Evolution*, ESOE, Busan-Korea, November, 2007, p. 33-44

LAKOFF, George Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts, *Journal of Philosophical Logic* 2, 1973, p. 458-508.

LAM, Toby H. W. Fuzzy Ontology Map – A Fuzzy Extension of the Hard-Constraint Ontology. *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06),* 2006.

LEWIS, D. K. Vague Identity: Evans Misunderstood, *Analysis*, 48, 1988, p. 128-130.

LI, Yanhui; XU, Baowen; Lu, Jianjiang e KANG, Dazhou. *Reasoning with Fuzzy Ontologies*.2006. Disponível em < http://ceur-ws.org/Vol-210/paper19.pdf >. Acesso em 01-02-2010.

LIU, B.; HSU, W.; CHEN, S. e Ma, Y. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, Vol 15, N° 5, Setembro, 2000, p. 47-55.

LIU, B.; HSU, W.; WANG, K. e CHEN, S. Visually Aided Exploration of Interesting Association Rules. *Proc. Pacific-Asia Conf.l Knowledge Discovery and Data Mining (PAKDD),* 1999, p. 380-389.

MACHINA, Kenton.Truth, Belief and Vagueness, *Journal of Philosophical Logic* 5, 1976, p. 47-78. Reimpresso *In:* KEEFE, Rosanna e SMITH, Peter. *Vagueness: A Reader,* Capítulo 11, 1997, p. 174-203.

MAHESH, K. *Ontology Development For Machine Translation: Ideology And Methodology.* New Mexico State University, Computing Research Laboratory Mccs, 1996, p. 292.

MAIMON, Oded e ROKACH, Lior. Introduction To Knowledge Discovery in Databases. Chapter 1, *In: The Data Mining and Knwoledge Discovery Handbook*, Springer, 2005, p. 1-17,

MALETIC, Jonathan I e MARCUS, Andrian. Data Cleansing. *In:* Maimon, Oded e Rokach, Lior (Eds). *The Data Mining and Knwoledge Discovery Handbook*, Springer, 2005, p. 21-36.

MANSINGH, Gunjan, KWEKU-MUATA Osei-Bryson, REICHGELT, Han. Using Ontologies to facilitate post-processing of association rules by domain experts. *Journal Information Sciences*, Volume 181, Issue 3, February, 2011, p. 419-493.

MARINICA, Claudia e GUILLET, Fabrice. Knowledge-Based Interactive Postmining of Association Rules Using Ontologies, I*EEE Transactions on Knowledge and Data Engineering*, vol. 22, n. 6, Feb. 2010, p. 784-797.

MARINICA, Claudia; GUILLET, Fabrice; Briand, Henri. Post-Processing Of Discovered Association Rules Using Ontologies, *IEEE International Conference On Data Mining Workshops*, 2009, p.126-133.

MARINICA, Claudia; GUILLET, Fabrice e BRIAND, Henri. Post-Processing of Discovered Association Rules Using Ontologies.*The Second International Workshop on Domain Driven Data Mining (DDDM 2008), IEEE International Conference on Data Mining (ICDM 2008),* Nantes, França, 2008. Disponível em: <http://datamining.it.uts.edu.au/dddm/dddm08/slides/DDDM08-7.pdf >. Acesso em 01-02-2011.

MEHLBERG, Henryk. Truth and Vagueness, *The Reach of Science*, 1958. *In:* KEEFE, Rosanna e SMITH, Peter. *Vagueness: A Reader,* Capítulo 6, 1997, p. 85-88.

MELANDA, Edson Augusto. *Pós-processamento de Regras de Associação*. Tese, USP-São Carlos, outubro, 2004.

MERIGÓ, José M. A Unified Model for Fuzzy Aggregation Operators and its Application in Group Decision Making.*Advances in Intelligent Systems Research*, Vol. 1, N° 1, July 2011, p. 965-972.

MINSKY, Marvin. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster. 2006, 400 p. Disponível em: < http://web.media.mit.edu/~minsky/eb6.html>

MITRA, Sushmita e PAL, Sankar K. Fuzzy sets in pattern recognition and machine intelligence. *Fuzzy Sets And Systems*, 156, 2005, p. 381-386.

MITRA, Sushmita. Data Mining in Soft Computing Framework: A Survey. *IEEE Transactions on Neural Networks*, Vol. 13, N° 1, January, 2002, p. 3-14.

MOCHOL, Malgorzata; CREGAN, Anne e VRANDECIC, Denny. Exploring OWL and rules: a simple teaching case. *Int. J. Teaching and Case Studies*, Vol. 2008, p. 299-318. Disponível em: <http://www.aifb.kit.edu/images/6/6f/2008_1933_Mochol_Exploring_OWL_a_1.pdf>. Acesso em 01-02-2012.

MORA, J. F.Compromisso Ontológico.*Dicionário de Filosofia*, Loyola, 2ª edição, 2004, p. 512.

MORA, J. Ferrater.Paradoxo. *Dicionário de Filosofia*, Edições Loyola, 2ª edição, 2004a, p. 2200-2206.

MORA, J. Ferrater.Polivalente.*Dicionário de Filosofia*, Edições Loyola, 2ª edição, 2004b, p. 2313-2316.

MORA, José Ferrater. Probabilidade. *Dicionário de Filosofia*, Tomo III (K-P). Edições Loyola, 2004, p. 2373-2376.

MORA, J. Ferrater.Vaguidade. *Dicionário de Filosofia*, Edições Loyola, 2ª edição, 2004c, p. 2963-2965.

MOTIK, Boris; SATTLER, Ulrike e STUDER, Rudi. *Query Answering for OWL-DL with Rules*. 2004. Disponível em <http://www.cs.ox.ac.uk/people/boris.motik/pubs/mss04dl-safe.pdf>. Acesso em 01-02-2012.

MULLIGAN, Kevin; SIMONS, Peter e SMITH, Barry. What's Wrong With Contemporary Philosophy? *Topoi,* Volume 25, Números 1-2, 2006, p. 63-67.

MURCHO, Desidério. Verofuncional. *In:* BRANQUINHO, João; DESIDÉRIO, Murcho; GOMES, Nelson Gonçalves. *Enciclopédia de termos lógico-filosóficos*. Martins Fontes, 2006, p. 802.

NARDI, D. e BRACHMAN, R. J. An Introdution to Description Logics.Edit: Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi e Peter F. Patel-Schneider. *The Description Logic Handbook. Theory, Implementation and Applications.* Second Edition. Cambridge. 2003, 2007, p. 1-44.

NEVES, João Manuel Poças Marques das. *Ambiente de Pós-processamento para regras de associação*. Dissertação, Fac. de Economia, Univ. do Porto, Out. de 2002.

NOVÁK, Vilém. Are fuzzy sets a reasonable tool for modeling vague phenomena? *Fuzzy Sets and Systems*, N° 156, 2005, p. 341-348.

OGAWA, Y.; MORITA, T. e KOBAYASHI, K. A Fuzzy Document Retrieval System Using The Keyword Connection Matrix And A Learning Method. *Fuzzy Sets And Systems.* V. 39, 1991, p. 163-179.

PAN, Jeff Z.; STAMOU, Giorgos; STOILOS, Giorgo; TAYLOR, Stuart e THOMAS, Edward. Scalable Querying Services over Fuzzy Ontologies. *In:*Proc. of the 17th International World Wide Web Conference (WWW2008).ACM, Abril 21-25, Beijing, China, 2008, p. 575-584. Disponível em <http://www.abdn.ac.uk/~csc280/PSSTT08.pdf>. Acesso em 01-10-2010.

PAN, Jeff Z.; STAMOU, Giorgos; STOILOS, Giorgo; e THOMAS, Edward. Expressive Querying over Fuzzy DL-Lite Ontologies. *In: Proc. of 2007 International Workshop on Description Logics (DL2007).* 2007. Disponível em < http://www.csd.abdn.ac.uk/~jpan/pub/>. Acesso em 01-10-2010.

PARRY, David. A fuzzy ontology for medical document retrieval. *The Australasian Workshop on DataMining and Web Intelligence* (DMWI2004), Dunedin. Conferences in Research and Practice in Information Technology, Vol. 32, 2004, p. 121-126.

PEARSON, Karl. The Grammar of Science. London, Adomand, Charles Black. Second Edition, January, 1905. Disponível em: <http://archive.org/download/grammarofscience00pearuoft/grammarofscience00pearuoft.pdf>. Acesso em 01-02-2012.

PEDRYCZ, W. e GOMIDE, F. *An Introduction To Fuzzy Sets: Analysis And Design*. Mit Press: Cambridge, Eua, 1998.

PEIRCE, Charles Sander. Vague. In: *Dictionary of Philosophy and Psychology*, (ed.) BLADWIN, J. M. New York: Macmillan, 1902, p. 748.

PEREIRA, R.; RICARTE, I. L. M. ; GOMIDE, F. A. C. Fuzzy Relational Ontological Models in Information Search Systems. *In:* Elie Sanchez. (Org.). *Fuzzy Logic and the Semantic Web.* Amsterdam: Elsevier, 2006, p. 395-412.

PEREIRA, R.; RICARTE, I. L. M. ; GOMIDE, F. A. C. . Relational ontology in information retrieval systems. *In: Eleventh International Fuzzy Systems Association World Congress*, 2005, Beijing. *Fuzzy Logic, Soft Computing and Computational Intelligence.* New York: Springer, 2005a. v. I. p. 509-514.

PEREIRA, Rachel; RICARTE, Ivan e GOMIDE, Fernando. Ontologia relacional fuzzy em sistemas de recuperação de informação. *In: Anais do XXV Congresso da Sociedade Brasileira de Computação*, p. 672 a 681, julho, 2005b.

PEREIRA, Rachel. *Modelo ontológico relacional fuzzy em sistemas de recuperação de informação textual*. Dissertação de Mestrado, Unicamp, novembro de 2004, 99 p.

PETRATUS, Panagiotis. Information Retrieval Systems: A Perspective on Human Computer Interaction. *Issues in Informing Science and Information Technology*, Volume 3, 2006.

POPPER, Karl. Two Kinds of Definition. *In:The Open Society and Its Enemies*. 1945.

PRINZ, Jesse. Vagueness, Language, and Ontology. 1998. Disponível em <http://ejap.louisiana.edu/EJAP/1998/prinz98.html>. Acesso em 01-10-2010.

PYLE, Dorian. *Data Preparation For Data Mining*. Morgan Kaufmann, 1999, 537 p.

QUAN, Thanh Tho; Hui, Siu Cheung e CAO, Tru Hoang. FOGA: A Fuzzy Ontology Generation Framework for Scholarly Semantic Web. *In: Knowledge Discovery and Ontologies (KDO-2004*). Workshop em ECML;PKDD. 2004.

RAFFMAN, Diana. Vagueness and Context-Relativity, *Philosophical Studies* 81, 1996, p. 175-92.

RESTALL, Greg. Chapter 5: Vagueness and bivalence. *In:* Logic – An Introduction. McGill-Quenn's University Press, Montreal & Kingston, 2004, 225 p.

RICKARD, John e YAGER, Ronald R. Hypercube Graph Representations and Fuzzy Measures of Graph Properties. *IEEE Trans. On Fuzzy Systems,* Vol. 15, N° 6, December, 2007, p. 1278-1293.

RICKARD, John T.; AISBETT, Janet e GIBBON, Greg.Reformulation of the theory of conceptual spaces. *Information Sciences*, Vol. 177, N° 21, 2007, p. 4539-4565.

RICKARD, John. A concept geometry for conceptual spaces. *Fuzzy Optim. Decis Making,* Vol. 5, 2006, p. 311-329.

RORTY, Richard. How many grains make a heap?. London Review of Books, Vol. 27, N° 2, January, 2005. Disponível em http://www.lrb.co.uk/v27/n02/richard-rorty/how-many-grains-make-a-heap. Acessado em 01-01-2010.

RUSSELL, Bertrand. Vagueness, *Australasian Journal of Philosophy and Psychology*, 1, 1923, p. 84-92

RUSSELL, Bertrand. On Denoting. *Mind,* v. 14, p. 479-493, 1905.

SAHAR, Sigal. Interestingness Measures – On Determining What Is Interesting. *In: The Data Mining and Knwoledge Discovery Handbook*, Springer, 2005, p. 1057-1068.

SALMON, Wesley C (1980). Probabilistic Causality *In:*SOSA, E. e TOOLEY, M. (Eds.). Causation, Oxford, Oxford University Press, p. 137-153, 1997.

SANFORD, David H. Competing semantics of vagueness: Many values versus super-truth. *Synthese*, V. 33, N. 2-4, 1976, p. 195-210.

SANTOS, Daniel Parente Lemos. Lima-Marques, Mamede (Orientador). Um arcabouço teórico para autoria de documentos visando atenuar o surgimento do fenômeno da ambigüidade. Unb, Fac. de Economia, Administração, Contabilidade e Ciência da Informação e Documentação. Dissertação, Brasília, 2006.

SAUERLAND, Uli. Vagueness in Language: The Case Against Fuzzy Logic Revisited. 2009. Disponível em < http://semanticsarchive.net/Archive/DQxYTUwY/Sauerland.pdf>. Acesso em 01-10-2010.

SEARLE, John R. Intencionalidade. Martins Fontes, 2002, 390 p.

SEIFERT, Jeffrey W. *Data Mining: An Overveiw*. CRS Report for Congress, RL 31798, 2004.

SHADBOLT, Nigel; HALL, Wendy e BERNERS-LEE, Tim. The Semantic Web Revisited. *IEEE Inteligent Systems*, May/June, 2006, p 96-101. Disponível em: <http://eprints.ecs.soton.ac.uk/12614/1/Semantic_Web_Revisted.pdf> . Acesso em 01-02-2012.

SHETH, Amit; RAMAKRISHNAN, Cartic e THOMAS, Christopher. Semantics for the Semantic Web: The Implicit, the Formal and the Powerful. Int. Journal on Semantic Web & Information Systems, Volume 1, 2005, p. 1-18.

SILBERCHATZ, A.; TUSHILIN, A. What makes patterns interesting in kwnowledge discovery systems. *IEEE Trans. Knowledge & Data Engineering*. Vol. 8, N° 6, 1996.

SILVA, Célio Fernando de Souza, *Serviço de Inteligência* Fiscal - Acompanhamento de Contribuintes e Responsáveis Tributários do ISSQN. Documento de Circulação Interna, SMF-PBH, Belo Horizonte, Janeiro de 2006a.

SILVA, Lúcio Buzon da. *Ambigüidades da língua portuguesa: recorte classificatório para a elaboração de um modelo ontológico*. Unb, Fac. de Economia, Administração, Contabilidade e Ciência da Informação e Documentação. Dissertação, Brasília, 2006b.

SIMOVICI, Dan A. e DJERABA, Chabane. *Mathematical Tools For Data Mining*. Springer, 2010, 616 p.

SLANEY, JOHN. A logic for vagueness. Disponível em http://users.cecs.anu.edu.au/~jks/logic_for_vagueness.pdf.Artigo não publicado adaptado do relatório TR-ARP-15-1988. Australian National University, 1988, 32 p.

SMITH, Barry e BROGAARD, Berit. Quantum Mereotopology. Preprint,2001. Disponível em http://ontology.buffalo.edu/smith/articles/qm.pdf, acessado em jan, 2010.

SMITH, Nicholas J. J. Fuzzy Logic and Higher-Order Vagueness, 2010. Disponível em < http://www-personal.usyd.edu.au/~njjsmith/papers/SmithFuzLogHOVag.pdf>. Acesso em 01-10-2010.

SMITH, Richard L. *Statistics of Extremes, with applications in environment, insurance and finance*. 2003. Disponível em: <http://www.stat.unc.edu/postscript/rs/semstatrls.ps>. Acesso em 01-02-2010.

SOARES, Pedro. Sorites. *In*: BRANQUINHO, João; DESIDÉRIO, Murcho; GOMES, Nelson Gonçalves. Enciclopédia de termos lógico-filosóficos. Martins Fontes, 2006, p. 713-721.

SOARES, Pedro. Vagueza. *In:* BRANQUINHO, João; DESIDÉRIO, Murcho; GOMES, Nelson Gonçalves. *Enciclopédia de termos lógico-filosóficos*. Martins Fontes, 2006, p. 787.

SOERGEL, D. The rise of ontologies or the reinvention of classification. *Journal of the American Society for Information Science*. October, 50, 12, 1999 p. 1119-1120.

SORENSEN, Roy. Vagueness. *In:*E. Zalta (ed.), *The Stanford encyclopedia of philosophy*, 2006. Disponível em <http://plato.stanford.edu/entries/vagueness/>, acessado em 01-02-2010.

SORENSEN, Roy. *A brief history of the paradox* - Philosophy and the labyrinths of the mind.Oxford. 2003, 394 p.

SORENSEN, Roy A. Vagueness, Blurriness, and Measurement. *Synthese* 75, 1988, p. 45-82.

SOUZA, Viviane Dal Molin de e CARVALHO, Deborah Ribeiro. Avaliação das Regras de Associação Descobertas Sob a Perspectiva do Usuário em Relação de Medidas Objetivas. *Gestão* - Revista Científica De Administração e Sistemas de Informação, Volume 8, número 8, jan/jun 2007.

STOILOS, Giorgos; STAMOU, Giorgos; PAN, Jeff Z.; SIMOU, Nick e TZOUVARAS, Vassilis. *Reasoning with the Fuzzy Description Logic f-SHIN*: Theory, Practice and Applications. da COSTA, P.C.G. *et al.* (Eds.): *URSW 2005-2007*, 2008, p. 262-281.

STOILOS, Giorgos; STAMOU, Giorgos; TZOUVARAS, Vassilis e HORROCKS, Ian. Reasoning with Very Expressive Fuzzy Description Logics. *Journal of Artificial Intelligence Research*, Vol. 30, 2007, p. 273-320.

STOILOS, Giorgos; SIMOU, Nikos; STAMOU, Giorgos e KOLLIAS, Stefano. Uncertainty and the Semantic Web. *IEEE Intelligent Systems*, 2006, p. 84-87.

STOILOS, Giorgos, STAMOU, Giorgos, TZOUVARAS, Vassillis; PAN, Jeff Z., e HORROCKS, Ian. Fuzzy OWL: Uncertainty and the Semantic Web. *In: International Workshop of OWL: Experiences and Directions*, Galway, Ireland, 2005.

STRACCIA, Umberto. A Fuzzy Descriptio Logic for the Semantic Web. *Proc. in Capturing Intelligence*: Fuzzy logic and the semantic Web, Elie Sanchez ed., Elsevier, 2006. Disponível em <http://www.win.tue.nl/~aserebre/ks/Lit/Straccia2006.pdf>. Acesso em 01-02-2012.

STRACCIA, U. Towards a Fuzzy Description Logic for the Semantic Web Preliminary Report. *In: ESWC 2005,* Vol. 3532 of LNCS, Springer-Verlag. 2005, p. 167-181.

STRACCIA, Umberto. Fuzzy ALC with Fuzzy Concrete Domains. *In*: *Proceedings of the International Workshop on Description Logics (DL-05), 2005*.

STRACCIA, Umberto. Reasoning within fuzzy description logics. *Journal of Artificial Intelligence Research*, 14, 2001, p. 137–166.

STRACCIA, Umberto. A Fuzzy Description Logic. *In*: *Proceedings of AAAI-98, 15th National Conf. on Artificial Intelligence*, Madison, Wisconsin, 1998, pp. 594-599.

SVÀTEK, Vojtěch e RAUCH, Jan Rauch. *Ontology-Enhanced Association Mining*. EWMF/KDO, 2005, p. 163-179. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=5D7667B217F26F84FBF9D613FF45E75F?doi=10.1.1.66.7810&rep=rep1&type=pdf

TAKACI, A. General aggregation Operators Acting On Fuzzy Number Induced By Ordinary Aggregation Operators. *Novi Sad. J. Math.*, Vol. 33, N° 2, 2003, p. 76-76.

TAKAGI, T. e KAWASE, K. A Trial For Data Retrieval Using Conceptual Fuzzy Sets. *IEEE Transactions On Fuzzy Systems*, Vol.9, N° 4, 2001, p. 497-505.

TAKAHASHI, A. e BEDREGAL, B.R.C. T-as, T-Coas, Complementos e Implicações Intervalares. *Tendências em Matemática Aplicada Computacional*, Vol. 7, N°1, 2006, p.139-148. http://www.sbmac.org.br/tema/seletas/docs/v7/15-Ta-Be.pdf

TAN, P. N; KUMAR, V; SRIVASTAVA J. *Selecting the Right Interestingness Measure for Association Patterns*. Canadá, 2002. Disponível em: <http://www.dbis.informatik.hu-berlin.de/dbisold/lehre/WS0405/KDD/paper/TKS02.pdf>, Acesso em 01-02-2011.

THEARLING, Kurt; BECKER, Barry; DeCoste, Dennis; MAWBY, Bill; PILOTE, Michel e SOMMERFIELD. Visualizing Data Mining Models. *In:* FAYYAD, Usama; GRINSTEIN, Georges e WIERSE, Andreas (Eds.), *Information Visualization in Data Mining and Knowledge Discovery.* Morgan Kaufman, 2001, p. 205-222.

TRESP, Cristopher e MOLITOR. Ralf. A description logic for vague knowledge. *In: Proceedings ECAI-1998*, p. 361-365. J. Wiley & Sons. apud CARDOSO, Jorge e LYTRAS, Miltiadis D. *Semantic Web Engineering in the Knowledge Society*.p. 331

TSENG, Ming-Cheng; LIN, Wen-Yang e JENG-Rong. Incremental maintenance of generalized association rules under taxonomy evolution. *Journal of Information Science*, XX (X), 2007, p. 1-22.

TYE, Michael. Vague Objects. *In:* Mind, Volume 99, Número 396, 1990, p. 535-557.

UNGER, Peter. There Are No Ordinary Things, *Synthèse* 41, 1979, p. 117-54.

VAN KERKHOVE, Bart.Between Semantics and Pragmatics: *In*: Search of a Logic for Vagueness, *Internal Survey Report* , 2002.

VAN KERKHOVE, Bart & Vanackere, Guido Vagueness-Adaptive Logic: A Pragmatical Approach to Sorites Paradoxes, *Studia Logica*, 68.p. 1-30, 2001.

VARZI, Achille C. Supervaluationism and Its Logics. *Mind*, 116, 2007, p. 633-676.

VARZI, Achille & Collins, John. Unsharpenable Vagueness, *Philosophical Topics* 28, p. 1-10, 2002. Disponível em http://www.columbia.edu/%7Eav72/papers/PhilTopics_2001.pdf. Acesso em 01-10-2010.

VARZI, A. Vagueness, Logic, and Ontology. *The Dialogue*, n° *1*, 2001a, p. 135-54. Disponível em: <http://www.columbia.edu/~av72/papers/Dialogue_2001.pdf>. Acesso em 01-10-2010

VARZI, Achille. Vagueness in Geography. *Philosophy and Geography*, Volume 4, 2001b, p. 49-65.

VERMA, Roop Rekha. Vagueness & the Principle of Excluded Middle *Mind*, Volume 79, Número 313, 1970, p. 67-77.

W3C Brasil. Dados abertos governamentais. 2011. Disponível em <http://www.w3c.br/pub/Materiais/PublicacoesW3C/dados-abertos-governamentais.pdf>. Acesso em 01-12-2012.

WANG, Hailong; MA, Zongmin; YAN, Li e CHENG, Jingwei. Chapter II: A Review of Fuzzy Models for the Semantic Web.*In*: Zongmin, M. e Wang, H. (Eds.). The Semantic Web for Knowledge and Data Mangement. Northeastern University, China, 2009, p. 23-37.

WEISS, Gary M. Mining with Rare Cases. *In:* Maimon, Oded e Rokach, Lior (Eds *The Data Mining and Knwoledge Discovery Handbook*, Springer, 2005, p. 765-776.

WILLIAMSON, Timothy. *Vagueness*. London: Routledge, 1994.

WRIGHT, Crispin. Rosenkranz on Quandary, Vagueness and Intuitionism, *Mind*, 112, 2003, p. 465-74.

WU, Chin-Ang; LIN, Wen-Yang e WU, Chuan-Chun. Facilitating Active Multidimensional Association Mining with User Preference Ontology. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2009, p. 437-440.

YAGUINUMA, C. A.; SANTOS, M. T. P.; BIAJIZ, M. Meta-ontologia Difusa para Representação de Informações Imprecisas em Ontologias. *In:* II Workshop on Ontologies and Metamodeling in Software and Data Engineering (WOMSDE), 2007, João Pessoa (PB). p. 57-67. Disponível em <http://www2.dc.ufscar.br/~cristiane_yaguinuma/publications/womsde2007.pdf> Acesso em 01-10-2010.

YAGUINUMA, C. A.; BIAJIZ, M.; SANTOS, M. T. P. Sistema FOQuE para Expansão Semântica de Consultas Baseada em Ontologias Difusas. *In:* XXII SBBD, 2007, João Pessoa (PB). p. 208-222. Disponível em <http://www2.dc.ufscar.br/~cristiane_yaguinuma/publications/FOQuE_SBBD_2007.pdf>. Acesso em 01-10-2010.

YANG, Li. Pruning and Visualizing Generalized Association Rules In Parallel Coordinates. *Transaction on Knowledge And Data Engineering*, Vol. 17, n° 1, January, 2005, p. 60-70.

YANG, Ying; WEBB, Geoffrey I e WU, Xindong. Discretization Methods. *In:* Maimon, Oded e Rokach, Lior (Eds). *The Data Mining and Knwoledge Discovery Handbook*, Springer, 2005, p. 113-130.

YAO, Y. Y.; ZHONG, N. An Analysis of Quantitative Measures Associated with Rules. Pacific-Asia Conference on Knowledege Discovery and Database. 1999.

YEN, John. Generalizing term subsumption languages to fuzzy logic. *In: Proceedings IJCAI*, 1991, p. 472-477.

ZADEH, Lotfi A. Toward a perception-based theory of probabilistic resoning with imprecise probabilities. *In: Journal of Statistical Planning and Inference*, 105, 2002, p. 233-264.

ZADEH, Lotfi A. Fuzzy Logic, Neural Networks, and Soft Computing. Communications of The ACM. Volume 37, Número 3, 1994a, p. 77-84.

ZADEH, Lotfi A. Soft Computing and Fuzzy Logic. *IEEE Software*, Nov, 1994.

ZADEH, L.A. A Theory of Approximate Reasoning.*In:*D. Mitchie J.E. Hayes and L.I. Mikulich, editors, *Machine Intelligence 9*. Wiley Masson, New York, 1979.

ZADEH, Lofti A. Fuzzy logic and approximate reasoning, *Synthese*, 30, 1975, p. 407–428.

ZADEH, L. A.Fuzzy Sets.*Information Control*, Volume 8, 1965, p. 338-353.

Anexo I - Ferramentas

Neste anexo relacionamos e comentamos alguns aspectos de ferramentas usadas no desenvolvimento do projeto.

As ferramentas foram escolhidas em vista da adoção de padrões internacionalmente recomentados e adotados. São também de livre acesso e uso. Além disso, considerou-se a disponibilidade de documentação satisfatória e a existência de comunidades de usuários ativos. Essas comunidades propiciam o desenvolvimento e a disponibilidade de significativas publicações relatando uso dessas ferramentas, além de um suporte informal e espontâneo constituído por listas de discussões, e pesquisas relacionadas.

Essas ferramentas e metodologias guiam a construção da modelagem conceitual possibilitando a detecção automática de inconsistências da modelagem e permitindo seu aprimoramento.

I.1 Editor de Ontologias e Plataforma de Integração de Ferramentas de Desenvolvimento

O *software* usado na modelagem conceitual e constituição das bases de conhecimento é o *Protégé*, versão 3.4.6.

O *Protégé* é um *software* livre, aberto e gratuito desenvolvido pela Universidade de Stanford. Há vários *plug-ins* do *Protégé* desenvolvidos independentemente, também livres e gratuitos, que podem ser facilmente integrados a essa ferramenta e que agregam facilidades e funcionalidades. Por exemplo, alguns dos *plugins* mais interessantes é o *DataMaster* que permite a integração com banco de dados e o *Jess* que permite o desenvolvimento de mecanismos de inferência baseados em regras, como veremos adiante.

Como editor, o *Protégé* permite introduzir um vocabulário de conceitos, estruturar e hierarquizar atributos de conceitos (através da estrutura de classes, propriedades e instâncias); relacionamentos entre conceitos; características lógicas de relacionamentos; **restrições** de domínio e contradomínio(faixa), restrições com quantificadores (existencial e universal), propriedades de relações (simetria, transitividade, reflexividade), cardinalidade de relações, etc. Além disso, o *Protégé*

permite o desenvolvimento de ontologias em equipe, de modo cooperativo, com recursos de gerenciamento do trabalho coletivo.

A versão 3.4.6 permite a integração direta da modelagem conceitual com bancos de dados através de drivers ODBC e JDBC.

O *Protégé*, além de ser uma ferramenta de desenvolvimento e de classificação, pode ser usada como um meio de consulta para usuários comuns que queiram tirar dúvidas a respeito do significado e dos relacionamentos das várias entidades modeladas.

O uso do *Protégé* é amplamente documentado em vários tutoriais e artigos relacionados podem ser obtidos em: http://protege.stanford.edu/doc/users.html, http://owl.cs.manchester.ac.uk/tutorials/protegeowltutorial/ e http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html.

Já existe a versão 4.0.2 do *Protégé*, mas ele não substitui a versão 3.4.6, pois, trabalha com a versão OWL 2.0 que dá ênfase ao tratamento de metaclasses. A versão 3.4.6 dá ênfase à integração com os mecanismos de inferência dos raciocinadores, e à OWL-DL que enfatiza a descrição dos conceitos em lógica descritiva. Essa integração com os raciocinadores e a OWL-DL é mais interessante para nossos objetivos do que o uso de metaclasses e expansões expressivas propiciadas pela OWL 2.0.

I.2 Instanciador de Ontologias a Partir de SGBD Relacional

Instâncias podem ser adicionadas a ontologias através de formulários, ou métodos semiautomáticos, ou automáticos a partir de texto, planilhas, imagens, páginas *web* (HTML, XML), ou bancos de dados relacionais. Essa capacidade é crucial para a ampliação do uso de ontologias.

Além de variadas técnicas descritas em artigos, há algumas opções permitem essa integração tais como *SpreadSheetMaster, XMLMaster*, *DataGenie, Ontobase,* e *DataMaster*.

O *DataMaster* é um *plug-in* para o *Protégé* que dá suporte à importação de estrutura de esquemas e dados de bancos de dados relacionais. Lida tanto com ontologias

baseadas em OWL quanto em frames. O usuário pode selecionar o tipo de conexão ao banco de dados (ODBC ou JDBC), o nome da fonte de dados, o nome do usuário e a senha para acessar o banco de dados.

Por essas funcionalidades e sua integração à versão 3.4 do *Protégé,* o *DataMaster* foi mais uma das razões para a escolha do *Protégé* 3.4 como ferramenta de desenvolvimento. Para maiores detalhes veja: http://protegewiki.stanford.edu/wiki/*DataMaster*.

A FIGURA I.1 apresenta um exemplo de tela do *Datamaster* mostrando a escolha de uma conexão e tabela a ser transferida para uma ontologia.

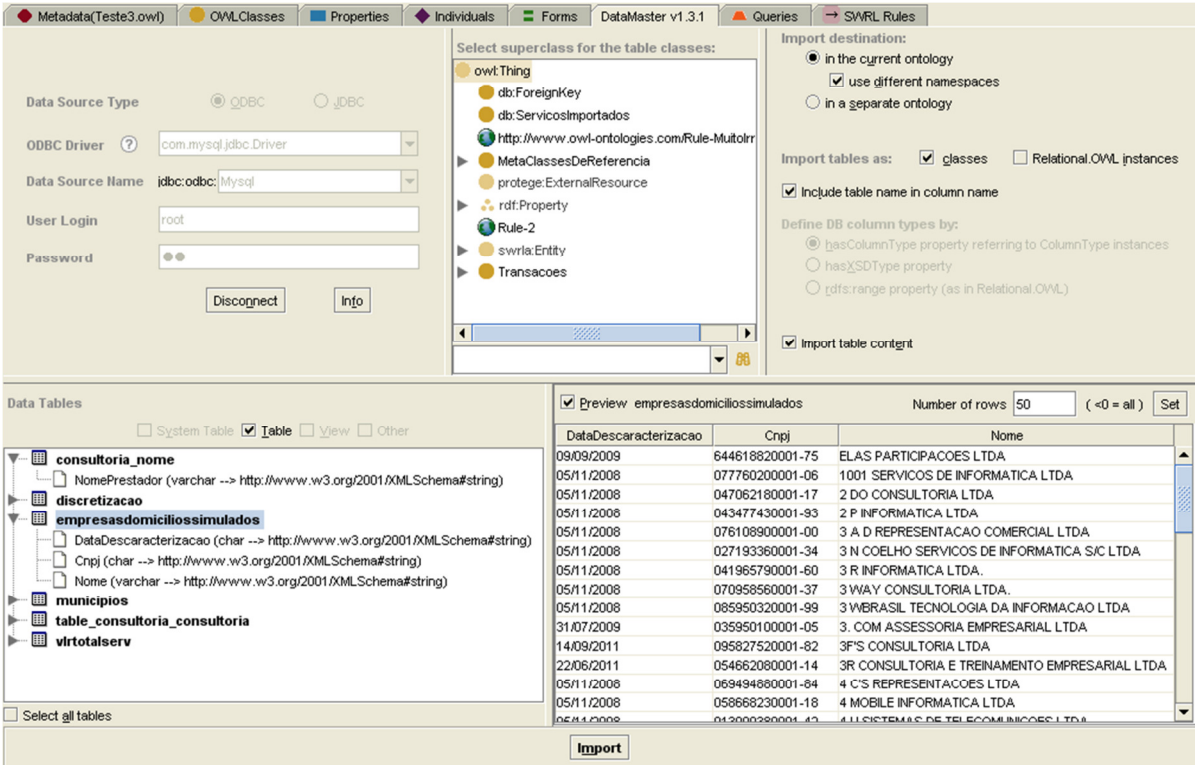

FIGURA I.1 Exemplo de tela do *plug-in DataMaster*.

No *DataMaster a*s tabelas são importadas como classes, as colunas como propriedades e o conteúdo como instâncias. Na FIGURA I.2 apresentamos a tela do *DataMaster* e suas opções de configuração.

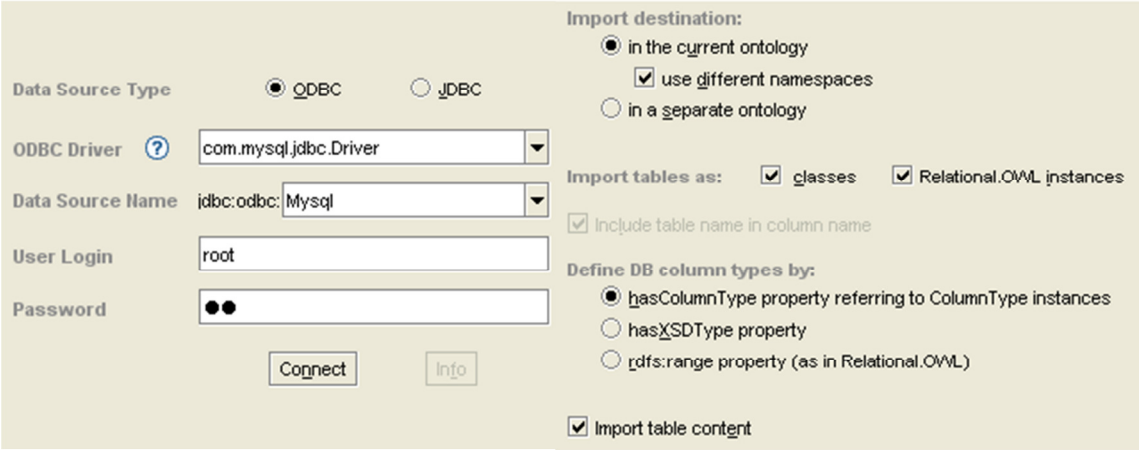

FIGURA I.2 Detalhamento de configuração de importação de esquemas e conteúdo de banco de dados relacional.

I.3 Raciocinador

Os raciocinadores e a OWL-DL permitem uma ampla verificação automática da consistência da modelagem conceitual e, ainda, a geração de regras e mecanismos de inferências a partir de regras desenvolvidas manualmente, ou inferidas a partir da modelagem conceitual desenvolvida.

O raciocinador a ser usado é o *Pellet* 1.5.2 que já vem no *Protégé*. Do ponto de vista histórico, é o primeiro raciocinador a suportar inteiramente OWL DL, incluindo nominais e tipos de dados. Na versão corrente, suporta $\mathcal{SROIQ}^{(D)}$e, consequentemente, OWL 2, com exceção de tipo de dados n-ários. É implementado em Java, tem múltiplas interfaces para acessá-lo, inclusive sua própria API, e está disponível sem ônus (BOBILLO, 2008, p. 45).

O *Pellet* permite integração com aplicativos desenvolvidos a partir de padrões distintos. Ele implementa o padrão OWL-DL e também o conjunto de regras SWRL LD-seguras. Para aplicativos em código aberto, *Pellet* pode ser usado sob os termos da licença AGPL version 3.Para maiores detalhes vide: http://www.mindswap.org/2003/pellet/ e http://clarkparsia.com/pellet.

I.4 Extensor da Inferência - Suporte de Implementação de Regras

O *plug-in* SWR*L* para o *Protégé* dá suporte ao uso de regras. No *Protégé* a guia SWRL requer o mecanismo de regra *Jess Rule Engine* para executar regras SWRL, ou consultas SQWRL.

Jess pode ser licenciada para uso comercial e está disponível sem custo para uso acadêmico e governamental. Para maiores detalhes vide: www.jessrules.com.

Desenvolveremos regras declarativas nas linguagens SWRL e SQWRL, seja para a classificação de instâncias carregadas pelo *DataMaster*, seja para os processos de fusificação e desfusificação, e para o mecanismo de inferência global.

I.5 Minerador de Dados

O *Tamanduá* é uma plataforma de serviços de mineração de dados e vem sendo usada para apoio à gestão e decisão governamentais. Esse minerador foi desenvolvido pelo laboratório *e-Speed* (Laboratório de Análise de Modelagem de Desempenho de Sistemas de Computação) do Departamento de Ciência da Computação da UFMG (Universidade Federal de Minas Gerais). Sua eficácia se estabelece em vista de sua facilidade de uso, da interoperabilidade com sistemas existentes e a sua escalabilidade – a arquitetura modular permite fácil replicação e rápida adaptação aos cenários de uso.

Do ponto de vista operacional, o *Tamanduá* permite as seguintes parametrizações e opções de visualização:

1. Definição de suporte e confiança mínimos
2. Seleção dos atributos a serem minerados
3. Definição de restrição (ões) (minerar apenas determinados valores das categorias definidas)
4. Seleção dos Antecedentes (dentre os campos selecionados para a mineração)
5. Seleção dos Consequentes (dentre os campos selecionados para a mineração)
6. Seleção de regras

7. Seleção de escala (métricas objetivas). Obtêm-se esquemas de regras que podem ser avaliadas conforme as seguintes métricas:
    a. *Lift*
    b. *Leverage*
    c. Convicção
    d. Confiança
    e. Suporte
8. Regras: a combinação dos Antecedentes e dos Consequentes que obtiveram os suportes e confianças mínimos
9. Filtros: pode-se filtrar Antecedentes e Consequentes.
10. Cada ponto do gráfico corresponde a um, ou mais esquemas de regra que acionado leva ao seu detalhamento. No detalhamento, temos:
    - Relata % dos registros em que antecedente tiveram determinado conseqüente.
    - Relata % de ocorrência em relação ao total dos dados.
    - Relata comentários:
        - Se apresenta a confiança esperada, acima, ou abaixo.
        - Qual a convicção.
    - Apresenta a opção de visualizar os registros (transações) de cada regra.
        - Guia que mostra transações e opção de exportar dados em formato csv.
        - Guia de Sumário que relata somatório de atributos numéricos.
11. A partir do detalhamento, podemos recuperar os registros de regras do esquema acionado que podem ser exportados em arquivo “.csv”

Anexo II – Local de Incidência do ISSQN

Lei Complementar Nº 116, de 31 de Julho de 2003**:** Dispõe sobre o Imposto Sobre Serviços de Qualquer Natureza, de competência dos Municípios e do Distrito Federal, e dá outras providências.

Art. 3º O serviço considera-se prestado e o imposto devido no local do estabelecimento prestador ou, na falta do estabelecimento, no local do domicílio do prestador, exceto nas hipóteses previstas nos incisos I a XXII, quando o imposto será devido no local:
I – do estabelecimento do tomador ou intermediário do serviço ou, na falta de estabelecimento, onde ele estiver domiciliado, na hipótese do § 1º do art. 1º desta Lei Complementar;
II – da instalação dos andaimes, palcos, coberturas e outras estruturas, no caso dos serviços descritos no subitem 3.05 da lista anexa;
III – da execução da obra, no caso dos serviços descritos no subitem 7.02 e 7.19 da lista anexa;
IV – da demolição, no caso dos serviços descritos no subitem 7.04 da lista anexa;
V – das edificações em geral, estradas, pontes, portos e congêneres, no caso dos serviços descritos no subitem 7.05 da lista anexa;
VI – da execução da varrição, coleta, remoção, incineração, tratamento, reciclagem, separação e destinação final de lixo, rejeitos e outros resíduos quaisquer, no caso dos serviços descritos no subitem 7.09 da lista anexa;
VII – da execução da limpeza, manutenção e conservação de vias e logradouros públicos, imóveis, chaminés, piscinas, parques, jardins e congêneres, no caso dos serviços descritos no subitem 7.10 da lista anexa;
VIII – da execução da decoração e jardinagem, do corte e poda de árvores, no caso dos serviços descritos no subitem 7.11 da lista anexa;
IX – do controle e tratamento do efluente de qualquer natureza e de agentes físicos, químicos e biológicos, no caso dos serviços descritos no subitem 7.12 da lista anexa;
X – (VETADO)
XI – (VETADO)
XII – do florestamento, reflorestamento, semeadura, adubação e congêneres, no caso dos serviços descritos no subitem 7.16 da lista anexa;
XIII – da execução dos serviços de escoramento, contenção de encostas e congêneres, no caso dos serviços descritos no subitem 7.17 da lista anexa;
XIV – da limpeza e dragagem, no caso dos serviços descritos no subitem 7.18 da lista anexa;
XV – onde o bem estiver guardado ou estacionado, no caso dos serviços descritos no subitem 11.01 da lista anexa;
XVI – dos bens ou do domicílio das pessoas vigiados, segurados ou monitorados, no caso dos serviços descritos no subitem 11.02 da lista anexa;
XVII – do armazenamento, depósito, carga, descarga, arrumação e guarda do bem, no caso dos serviços descritos no subitem 11.04 da lista anexa;
XVIII – da execução dos serviços de diversão, lazer, entretenimento e congêneres, no caso dos serviços descritos nos subitens do item 12, exceto o 12.13, da lista anexa;
XIX – do Município onde está sendo executado o transporte, no caso dos serviços descritos pelo subitem 16.01 da lista anexa;
XX – do estabelecimento do tomador da mão-de-obra ou, na falta de estabelecimento, onde ele estiver domiciliado, no caso dos serviços descritos pelo subitem 17.05 da lista anexa;
XXI – da feira, exposição, congresso ou congênere a que se referir o planejamento, organização e administração, no caso dos serviços descritos pelo subitem 17.10 da lista anexa;
XXII – do porto, aeroporto, ferroporto, terminal rodoviário, ferroviário ou metroviário, no caso dos serviços descritos pelo item 20 da lista anexa.
§ 1º No caso dos serviços a que se refere o subitem 3.04 da lista anexa, considera-se ocorrido o fato gerador e devido o imposto em cada Município em cujo território haja extensão de ferrovia, rodovia, postes, cabos, dutos e condutos de qualquer natureza, objetos de locação, sublocação, arrendamento, direito de passagem ou permissão de uso, compartilhado ou não.
§ 2º No caso dos serviços a que se refere o subitem 22.01 da lista anexa, considera-se ocorrido o fato gerador e devido o imposto em cada Município em cujo território haja extensão de rodovia explorada.
§ 3º Considera-se ocorrido o fato gerador do imposto no local do estabelecimento prestador nos serviços executados em águas marítimas, excetuados os serviços descritos no subitem 20.01

Anexo III – Classes Enumeradas

III.1 Municípios da Região Metropolitana

N°	Município	N°	Município
1.	BALDIM	2.	MATOZINHOS
3.	BETIM	4.	NOVA LIMA
5.	BRUMADINHO	6.	NOVA UNIAO
7.	CAETE	8.	PEDRO LEOPOLDO
9.	CAPIM BRANCO	10.	RAPOSOS
11.	CONFINS	12.	RIBEIRAO DAS NEVES
13.	CONTAGEM	14.	RIO ACIMA
15.	ESMERALDAS	16.	RIO MANSO
17.	FLORESTAL	18.	SABARA
19.	IBIRITE	20.	SANTA LUZIA
21.	IGARAPE	22.	SÃO JOAQUIM DE BICAS
23.	ITAGUARA	24.	SÃO JOSE DA LAPA
25.	ITATIAIUCU	26.	SARZEDO
27.	JABOTICATUBAS	28.	TAGUARA
29.	JUATUBA	30.	TAQUARACU DE MINAS
31.	LAGOA SANTA	32.	VESPASIANO
33.	MARIO CAMPOS		
34.	MATEUS LEME		

III.2 Municípios do Colar Metropolitano

N°	Município
1.	BARAO DE COCAIS
2.	BELO VALE
3.	BONFIM
4.	FORTUNA DE MINAS
5.	FUNILANDIA
6.	INHAUMA
7.	ITABIRITO
8.	ITABIRA
9.	MOEDA
10.	PARA DE MINAS
11.	PRUDENTE DE MORAIS
12.	SANTA BARBARA
13.	SAO JOSE DE VARGINHA
14.	SETE LAGOAS

III.3. Municípios com Benefícios Fiscais e Abrigos de Empresas Virtuais (Paraísos Fiscais Municipais)[85]

N°	Município	UF	N°	Município	UF
1.	CRUZEIRO DO SUL	AC	2.	CUIABA	MT
3.	RIO BRANCO	AC	4.	RONDONOPOLIS	MT
5.	MACAPA	AP	6.	SINOP	MT
7.	BARREIRAS	BA	8.	VARZEA GRANDE	MT
9.	CAMACARI	BA	10.	ANANINDEUA	PA
11.	FEIRA DE SANTANA	BA	12.	MARABA	PA
13.	ILHEUS	BA	14.	PARAUAPEBAS	PA
15.	ITABUNA	BA	16.	SANTAREM	PA
17.	VITORIA DA CONQUISTA	BA	18.	CAMPINA GRANDE	PB
19.	CAUCAIA	CE	20.	JOAO PESSOA	PB
21.	JUAZEIRO DO NORTE	CE	22.	MAMANGUAPE	PB
23.	SOBRAL	CE	24.	PATOS	PB
25.	CACHOEIRO DE ITAPEMIRIM	ES	26.	CARUARU	PE
27.	CARIACICA	ES	28.	IPOJUCA	PE
29.	SERRA	ES	30.	JABOATAO DOS GUARARAPES	PE
31.	VILA VELHA	ES	32.	OLINDA	PE
33.	ANAPOLIS	GO	34.	PAULISTA	PE
35.	APARECIDA DE GOIANIA	GO	36.	PETROLINA	PE
37.	GOIANIA	GO	38.	TERESINA	PI
39.	ITUMBIARA	GO	40.	CASCAVEL	PR
41.	JATAI	GO	42.	FOZ DE IGUACU	PR
43.	RIO VERDE	GO	44.	LONDRINA	PR
45.	URUACU	GO	46.	MARINGA	PR
47.	IMPERATRIZ	MA	48.	PARANAGUA	PR
49.	BARBACENA	MG	50.	PONTA GROSSA	PR
51.	DIVINOPOLIS	MG	52.	SAO JOSE DOS PINHAIS	PR
53.	GOVERNADOR VALADARES	MG	54.	BELFORD ROXO	RJ
55.	IPATINGA	MG	56.	CAMPOS DOS GOYTACAZES	RJ
57.	JUIZ DE FORA	MG	58.	DUQUE DE CAXIAS	RJ
59.	MONTES CLAROS	MG	60.	ITAGUAI	RJ
61.	POCOS DE CALDA	MG	62.	MACAE	RJ
63.	PONTE NOVA	MG	64.	MAGE	RJ
65.	POUSO ALEGRE	MG	66.	NILOPOLIS	RJ
67.	RIBEIRAO DAS NEVES	MG	68.	NITEROI	RJ
69.	SETE LAGOAS	MG	70.	NOVA FRIBURGO	RJ
71.	TEOFILO OTONI	MG	72.	NOVA IGUACU	RJ
73.	UBERABA	MG	74.	PETROPOLIS	RJ
75.	UBERLANDIA	MG	76.	RIO BONITO	RJ
77.	VARGINHA	MG	78.	SAO GONCALO	RJ
79.	CAMPO GRANDE	MS	80.	SAO JOAO DE MERITI	RJ
81.	CORUMBA	MS	82.	SAQUAREMA	RJ
83.	VOLTA REDONDA	RJ	84.	ARACAJU	SE
85.	MOSSORO	RN	86.	AMERICANA	SP
87.	PORTO VELHO	RO	88.	ARARAQUARA	SP
89.	BOA VISTA	RR	90.	BARUERI	SP

[85] Essa relação foi construída com base em levantamentos em sítios na Web que ofereciam serviços de constituição de empresas reais e virtuais e indicavam esses municípios como possuindo condições favoráveis para o menor recolhimento de tributos, em vista de vários fatores. Em alguns casos mais notórios, como os casos de Saquarema e Rio Bonito, no Rio de Janeiro, chegamos a verificar a legislação tributária e confirmar a existência de amplas facilidades e benefícios. Em todo o caso, essa lista é apenas ilustrativa para fins das simulações realizadas nessa tese.

Continuação ...

N°	Município	UF	N°	Município	UF
91.	CAMAQUA	RS	92.	BAURU	SP
93.	CANOAS	RS	94.	CAMPINAS	SP
95.	CAXIAS DO SUL	RS	96.	CARAPICUIBA	SP
97.	GRAVATAI	RS	98.	CUBATAO	SP
99.	NOVO HAMBURGO	RS	100.	DIADEMA	SP
101.	PASSO FUNDO	RS	102.	EMBU	SP
103.	PELOTAS	RS	104.	FRANCA	SP
105.	PORTO ALEGRE	RS	106.	GUARUJA	SP
107.	RIO GRANDE	RS	108.	GUARULHOS	SP
109.	SANTA MARIA	RS	110.	HORTOLANDIA	SP
111.	URUGUAIANA	RS	112.	INDAIATUBA	SP
113.	VIAMAO	RS	114.	ITAPEVI	SP
115.	BLUMENAU	SC	116.	ITAQUAQUECETUBA	SP
117.	BRUSQUE	SC	118.	JACAREI	SP
119.	CACADOR	SC	120.	JUNDIAI	SP
121.	CHAPECO	SC	122.	LIMEIRA	SP
123.	CRICIUMA	SC	124.	MARILIA	SP
125.	CURITIBANOS	SC	126.	MAUA	SP
127.	ITAJAI	SC	128.	MOGI DAS CRUZES	SP
129.	JARAGUA DO SUL	SC	130.	OSASCO	SP
131.	JOACABA	SC	132.	PIRACICABA	SP
133.	JOINVILLE	SC	134.	POA	SP
135.	LAGES	SC	136.	PRAIA GRANDE	SP
137.	MAFRA	SC	138.	PRESIDENTE PRUDENTE	SP
139.	NAVEGANTES	SC	140.	RIBEIRAO PRETO	SP
141.	RIO DO SUL	SC	142.	RIO CLARO	SP
143.	RIO NEGRINHO	SC	144.	SANTO ANDRE	SP
145.	SAO BENTO DO SUL	SC	146.	SANTOS	SP
147.	SAO FRANCISCO DO SUL	SC	148.	SAO BERNARDO DO CAMPO	SP
149.	SAO JOSE	SC	150.	SAO CARLOS	SP
151.	SAO MIGUEL DO OESTE	SC	152.	SAO JOSE DO RIO PRETO	SP
153.	TUBARAO	SC	154.	SAO JOSE DOS CAMPOS	SP

Anexo IV – Índice de Dinamismo do Município

Ranq.	Município	U.F	Cluster	Taxa de Desenvolvimento
1º	SAO PAULO	SP	10	92,83
2º	RIO DE JANEIRO	RJ	9	52,28
3º	BELO HORIZONTE	MG	9	37,89
4º	SAO CAETANO DO SUL	SP	8	36,81
5º	PORTO ALEGRE	RS	9	36,40
6º	DIADEMA	SP	8	36,23
7º	CURITIBA	PR	9	36,00
8º	TABOAO DA SERRA	SP	8	35,92
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
5563º	GUAJARA	AM	3	28,47
5564º	ENVIRA	AM	3	28,01