UM MODELO TEMPORAL-RELACIONAL PARA CLASSIFICAÇÃO DE DOCUMENTOS

FERNANDO HENRIQUE DE JESUS MOURÃO

UM MODELO TEMPORAL-RELACIONAL PARA CLASSIFICAÇÃO DE DOCUMENTOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Wagner Meira Júnior

Belo Horizonte

Novembro de 2009

Mourão, Fernando Henrique de Jesus
D1234p Um Modelo Temporal-Relacional para Classificação de Documentos / Fernando Henrique de Jesus Mourão. — Belo Horizonte, 2009
xxiv, 92 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Wagner Meira Júnior

1. Classificação. 2. Redes Complexas. 3. Análise Temporal. I. Título.

CDU 519.6*82.10

[Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufmg`.

Agradecimentos

Se me perguntassem o que eu desejaria ser,
Certamente responderia:

Nada para muitos,
Uma boa pessoa para poucos,
Alguém especial para raras pessoas,
E para mim ...
Apenas alguém que jamais queira mudar seu passado,
E que sempre almeje cada segundo de seu futuro.

Aos muitos para os quais nada sou,
Meu obrigado por não se incomodarem comigo.
Aos poucos que me vêem como um ser bom,
Meu muito obrigado pelas energias positivas oferecidas e recebidas.
Às raras pessoas que me acham especial,
Minha eterna gratidão pela demonstração diária de carinho
E a Deus, minha fé incondicional
Por cada segundo de vida e sentimento de orgulho experimentado.
Não pelo que sou, pois somos vários,
Mas pelo que faço para mudar o que sou
Para cada uma das pessoas que são boas e especiais para mim!

Fernando H. J. Mourão

"Veni, Vidi, Vici"
(Julius Caesar)

Resumo

Classificação Automática de Documentos (CAD) representa atualmente um dos mais relevantes e desafiadores problemas de pesquisa em Recuperação de Informação. Apesar do grande número de técnicas existentes, poucas levam em consideração características da linguagem humana. Como discutido em trabalhos recentes [Montejo-Raez et al., 2008; Chen, 1995], entender e considerar tais características pode beneficiar CAD. Dessa forma, neste trabalho propomos uma representação para documentos textuais, através de uma rede de termos, baseada fundamentalmente em conceitos lingüísticos, em particular conceitos associados a relacionamentos entre termos. Usando a representação proposta, também apresentamos um algoritmo relacional para CAD que explora tais relacionamentos. Uma avaliação experimental deste algoritmo mostrou que é possível alcançar resultados comparáveis ao SVM em quatro coleções reais. Além disso, sua simplicidade, eficiência de execução e inexistência de um complexo ajuste de parâmetros, são características que tornam nosso algoritmo uma interessante alternativa ao SVM. Uma análise detalhada também mostrou que existem várias dimensões nas quais este algoritmo relacional pode ser melhorado.

Dada a sua relevância, particular atenção pode ser dada à dimensão temporal. De fato, evoluções naturais ocorrem a todo momento modificando definições e observações previamente realizadas sobre a rede de termos. Como apontado por estudos recentes [Alonso et al., 2007; Mourão et al., 2008], considerar o tempo pode ser muito útil na área de Recuperação de Informação. A fim de incorporar a dimensão temporal em nosso algoritmo, atribuímos a cada relacionamento de nossa rede informação sobre o momento de sua construção. Avaliando simples versões temporais do algoritmo proposto, observamos que considerar a evolução temporal permitiu melhorar o desempenho do nosso classificador relacional, por prover informações mais precisas sobre o comportamento de cada termo. Uma avaliação preliminar sobre outras dimensões de análise, tais como escassez de informação e uso de atributos dos relacionamentos, também mostrou que a consideração de tais dimensões pode prover melhorias ao algoritmo proposto. Além disso, dada a generalidade das propriedades lingüísticas utilizadas

como base neste trabalho, acreditamos que nossa proposta pode ser efetiva em vários domínios de aplicação de CAD.

Palavras-chave: Recuperação de Informação, Mineração de Texto, Classificação e Agrupamento, Modelagem de Redes Complexas, Análise Temporal.

Abstract

Automatic Document Classification (ADC) is one of the most relevant and challenging research problems in Information Retrieval. Despite the large number of ADC techniques already proposed, few of them take into consideration characteristics of the human language. As discussed in recent studies [Montejo-Raez et al., 2008; Chen, 1995], understanding and considering such characteristics may benefit ADC. Therefore, in this work we propose a new network-based representation for textual documents that is based on fundamental concepts of Linguistic, in particular those associated with relationships between terms. Using the proposed model, we also introduce a relational algorithm for ADC which exploits such relationships. Experimental evaluation of this algorithm shows that it achieves results that are comparable to SVM in four real datasets. In addition, its simplicity, execution efficiency and a simple parameter tuning are characteristics that make our algorithm an interesting alternative to SVM. A deeper analysis also shows that there are several dimensions in which relational algorithms may be enhanced.

Due to its relevance, particular attention is given to the temporal dimension. In fact, changes occur spontaneously at every moment affecting settings and observations made previously on the term network. Considering this evolving behavior may be very useful in the area of Information Retrieval [Alonso et al., 2007]. In order to incorporate the temporal dimension to our algorithm, we attach to every relationship of our network information about the moment of its construction. The evaluation of simple temporal versions of the proposed algorithm showed that considering the temporal evolution has improved the performance of our relational classifier, by providing more accurate information about the behavior of each term. A preliminary assessment of other dimensions of analysis, such as information scarcity and the use of attributes of relationships, also showed that more elaborated techniques to address such dimensions may benefit the proposed algorithm. Further, considering the generality of the linguistic concepts incorporated in this work, we believe that our proposal may be equally successful in various ADC application domains.

Keywords: Information Retrieval, Text Mining, Classification and Clustering, Complex Network Modeling, Temporal Analysis.

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Sumário

Capítulo 1

Introdução

Nas duas últimas décadas, a humanidade foi capaz de gerar uma quantidade de dados sem precedentes em sua história. De fato, estudos diversos [Gantz et al., 2008] estimam que o volume de dados digitais correspondente a este período supera ao produzido em todo o restante de nossa história. Este aumento da quantidade e disponibilidade de dados deve-se, em grande parte, ao surgimento da INTERNET, permitindo a consolidação de um grande repositório amplamente acessível. Como afirmado em Mendelson [2002], *"O grande problema educacional da atualidade é ensinar às pessoas a ignorar a informação irrelevante, a recusar-se a saber de coisas, antes que fiquem sufocadas. Muita informação é tão ruim quanto nenhuma."*. Dessa forma, organizar e encontrar os recursos informacionais apropriados para satisfazer as necessidades dos usuários passou a figurar como um dos problemas mais estudados e desafiadores em Ciência da Computação.

Embora estudos apontem a explosão da quantidade de imagens e vídeos como grande precursora do aumento de dados digitais nos próximos anos [Gantz et al., 2007], a quantidade de dados textuais ainda configura-se como uma parcela representativa dos dados existentes. Assim, a organização e recuperação deste tipo de dados é uma área de crescente interesse. Neste contexto, emerge a área de Recuperação de Informação (RI), que estuda o armazenamento, organização e sobretudo recuperação de dados e metadados relacionados a documentos. Dentre as diversas tarefas estudadas em RI, podemos destacar a tarefa de Classificação Automática de Documentos (CAD), dada sua relevância no processo de organização e recuperação de dados textuais. CAD é definida como a tarefa de inferir a categoria semântica à qual um documento pertence, dado um conjunto discreto e finito de categorias conhecidas. Dentre as várias aplicações desta tarefa podemos citar a construção de filtros de *spam*[1] e documentos,

[1]DEFINIR

diretórios de tópicos e bibliotecas digitais, bem como a identificação de estilos de escrita e auxílio à navegação e pesquisa na Web, dentre outras. Métodos tradicionais de CAD usualmente seguem uma estratégia supervisionada por usar informações de um conjunto de documentos pré-classificados. Existem, na literatura, diversas abordagens para CAD, tais como modelos Associativos [Liu et al., 1998], Vizinhos mais Próximos [Salton & McGill, 1986], Bayesianos [Salton & McGill, 1986], *Support Vector Machines (SVM)* [Joachims, 2006], dentre outros.

Embora haja um grande e crescente número de propostas para CAD, poucas consideram propriedades lingüísticas para a classificação de documentos. Entende-se por propriedades lingüísticas regras sintáticas e semânticas que definem a forma como um documento foi gerado. Por exemplo, alguns termos podem apresentar diferentes capacidades discriminativas quanto ao contexto de discurso no qual ocorrem (i.e., categoria semântica da comunicação). Uma das propriedades lingüísticas mais importantes refere-se à forma de utilização dos termos na comunicação. Os padrões de uso dos termos na linguagem não são completamente aleatórios e independentes. Ou seja, termos se relacionam obedecendo certas regras a fim de estabelecer a semântica da comunicação. Textos são organizados como sentenças, que são compostas por palavras que interagem entre si. Logo, considerar tais relacionamentos pode ser importante para uma modelagem apropriada dos distintos contextos de discurso. Mais ainda, determinados termos tendem a apresentar importantes relacionamentos que compõem comportamentos relacionais distintos em cada contexto de discurso. Definimos como comportamento relacional de um termo o conjunto de termos (i.e., vocabulário) com os quais este termo se relaciona em cada contexto de discurso. Por exemplo, os relacionamentos entre o termo *DNA* e alguns termos do contexto de criminologia (tais como *perícia*, *investigação* e *crime*) compõem um comportamento relacional do termo *DNA* neste contexto. Ou seja, antes a simplesmente considerar os termos isoladamente, ou mesmo co-ocorrências mais freqüentes entre termos, podemos considerar comportamentos relacionais de cada termo, de forma a identificar o vocabulário com o qual o termo se relaciona.

A relevância do uso de relacionamentos para a classificação é, inclusive, mostrada por diversos estudos na literatura [Macskassy & Provost, 2004, 2003]. Modelos que consideram a rede de relacionamentos entre os objetos (i.e., modelos relacionais) muitas vezes apresentam resultados significativamente melhores que os de modelos que a ignoram. Apesar dessa relevância, estudos que consideram propriedades lingüísticas em CAD, usualmente, ignoram os relacionamentos entre termos. Tais estudos objetivam apenas utilizar medidas estatísticas das palavras (e.g., freqüência de uso), informações sintáticas dos documentos (e.g., se uma palavra é um substantivo ou um verbo), ou

mesmo a semântica, baseado em uma simples análise gramatical [Montejo-Raez et al., 2008; Chen, 1995].

Outra importante observação sobre a comunicação é que tais relacionamentos podem variar em intensidade ao longo do tempo como conseqüência, por exemplo, da evolução da linguagem ou do surgimento e desaparecimento de algumas áreas de pesquisa. Redes de relacionamentos são, em geral, altamente dinâmicas, crescendo e modificando-se rapidamente ao longo do tempo. De fato, evoluções naturais ocorrem a todo momento modificando definições e observações previamente realizadas. Como argumenta Alonso et al. [2007] *"o tempo é uma importante dimensão da informação em qualquer cenário e que pode ser muito útil na área de Recuperação de Informação."*. Assim, modelos de classificação relacionais baseados em todo o histórico da rede podem apresentar um desempenho deteriorado, uma vez que importantes informações sobre mudanças comportamentais da rede são perdidas [Mourão et al., 2009]. Neste contexto, um grande desafio consiste em selecionar a granulação temporal apropriada dos dados a ser considerada para a modelagem. Por exemplo, em uma rede de co-autoria em artigos científicos, podemos estar interessados em classificar os pesquisadores de acordo com a área de interesse. Neste cenário, é importante considerar que autores publicam em períodos diferentes, sobre tópicos de pesquisa distintos devido a vários fatores, tais como novos interesses, novos contatos, surgimento de novas áreas etc. Em Koren [2009], é inclusive mostrado que *"abordar a dinâmica temporal dos dados pode ter um impacto sobre o desempenho mais significativo que projetar algoritmos de aprendizado mais complexos para recomendação"*.

Este trabalho objetiva definir uma família de classificadores relacionais para CAD, baseada na análise de relações entre termos presentes em documentos, que seja robusta à mudanças naturais inerentes às coleções de documentos. Para tanto, realizamos uma ampla discussão teórica que visa mostrar como propriedades lingüísticas dos documentos podem beneficiar a tarefa de classificação. Tais propriedades, inclusive, ancoram a utilização de informações contidas nos relacionamentos entre termos para a CAD.

A fim de explorar tais relacionamentos, definimos uma rede de termos para uma coleção de documentos. Nesta rede, cada termo distinto presente em documentos da coleção representa um nodo, e há um relacionamento entre dois termos se eles co-ocorrerem em pelo menos um documento da coleção. Tal representação visa tornar a identificação dos contextos de discurso em que os termos ocorrem mais flexível e dependente dos relacionamentos estabelecidos entre os termos. Como termos, por imposições próprias da linguagem, ocorrem em diversos contextos de discurso, essas características são de grande importância para o surgimento de modelos relacionais de classificação mais robustos. Além disso, nosso modelo de representação relacional

permite-nos explorar importantes propriedades encontradas em redes na classificação, tal como a propriedade de **Homofilia** [Mcpherson et al., 2001], a qual estabelece que indivíduos "similares" [2] se relacionam mais freqüentemente na rede. Em nossa rede, isso representa dizer que termos usados no mesmo contexto de discurso estão relacionados mais freqüentemente na rede. Essa é uma propriedade importante e necessária para o bom desempenho de vários algoritmos relacionais de classificação.

A família de classificadores relacionais apresentada baseia-se na análise de vizinhança, sobre a rede de termos definida, para realizar a classificação, valendo-se de uma simples e intuitiva máxima freqüentemente utilizada: *"me digas com quem andas que direi quem és"*. Ou seja, definimos inicialmente para cada termo, presente em um documento a ser classificado, sua classe baseado nos termos aos quais ele se relaciona na rede. Posteriormente, determinamos a classe do documento através de um processo de votação ponderada das classes estimadas para cada termo do documento. Esses algoritmos exploram diferentes funções de ponderação e estratégias de análise de vizinhança. Apesar de sua simplicidade, mostramos que os algoritmos propostos são capazes de superar métodos simples de CAD baseados em *BAG of words* [3], alcançando resultados comparáveis ao SVM, em quatro coleções de documentos reais. Além disso, mesmo sendo transdutivo (i.e., nossa estratégia 'projeta' uma parte da rede contendo apenas os termos presentes em cada documento de teste), nosso algoritmo pode apresentar um tempo de classificação até 60% menor que o tempo alcançado pelo SVM, em virtude do simples esquema de votação.

A fim de incorporar a dimensão temporal em nossos algoritmos, atribuímos a cada relacionamento de nossa rede a informação de quando a relação foi construída. Este momento corresponde para CAD ao momento no qual o documento, em que os termos ocorrem, foi publicado. Abordar tal dimensão temporal em CAD representa uma forma de obter informações mais precisas sobre o comportamento de cada termo. Podemos considerar isso, inclusive, como um aprimoramento da máxima mencionada para: *"me digas com quem andas,* **quando***, que direi quem és"*. Dessa forma, separamos os relacionamentos em momentos unitários distintos, estendendo nossa rede de termos para um multigrafo temporal apto a capturar informações de diferentes momentos. A partir deste multigrafo, avaliamos diversas estratégias de considerar essa informação temporal em CAD. Utilizando uma estratégia de ponderação temporal dos relacionamentos, conseguimos verificar a hipótese que considerar a evolução temporal pode melhorar o desempenho do nosso algoritmo relacional. Além disso, avaliações preliminares sobre outras dimensões de análise, tais como escassez de informação e uso de atributos dos

[2] Baseado em alguma função ou propriedade de similaridade pré-estabelecida.

[3] Métodos baseados na simples ocorrência de atributos assumidamente independentes.

relacionamentos, também mostrou que a consideração de tais dimensões pode prover melhorias ao algoritmo proposto.

Por fim, é importante ressaltar o caráter interdisciplinar deste trabalho. Valendo-se de uma argumentação teórica baseada na **Lingüística**, justificamos a modelagem através de **Redes Complexas** para uma importante tarefa de **Recuperação de Informação**. Além disso, dada a generalidade das propriedades lingüísticas utilizadas como base neste trabalho, acreditamos que nossa proposta pode ser efetiva em vários domínios de aplicação de CAD.

1.1 Contribuições

Podemos sumarizar as principais contribuições deste trabalho como segue:

1. Ampla discussão teórica sobre propriedades lingüísticas dos termos e sua utilidade para CAD;

2. Proposta de um modelo de representação de documentos textuais baseado nos relacionamentos entre termos;

3. Proposta de uma família de algoritmos simples, intuitivos, eficientes e eficazes, baseado na análise de vizinhança, para CAD;

4. Incorporação do aspecto temporal nos algoritmos relacionais propostos para CAD;

5. Validação dos conceitos e algoritmos propostos em coleções reais.

1.2 Organização da Dissertação

Este trabalho possui mais 7 capítulos, organizados como segue. O capítulo 2 apresenta os principais conceitos envolvidos, bem como a formalização do problema abordado e a definição precisa da rede de termos utilizada. O capítulo 3 sumariza os principais trabalhos relacionados. A fim de facilitar o entendimento, tais trabalhos são divididos entre as principais áreas de interseção deste trabalho. O capítulo 4 apresenta uma discussão teórica baseada em conceitos lingüísticos, visando demonstrar a utilidade e aplicabilidade destes conceitos para a CAD. Posteriormente, no capítulo 5 descrevemos a família de algoritmos relacionais de classificação propostos, bem como avaliamos estes algoritmos em quatro bases reais. Em seguida, o capítulo 6 apresenta uma extensão dos

algoritmos relacionais propostos, a fim de tratar a evolução temporal dos relacionamentos presentes na rede de termos. No capítulo 7, discutimos uma proposta de extensão do nosso algoritmo relacional visando abordar o problema de escassez de informação. Avaliamos também formas simples de estender os algoritmos propostos de forma a incorporar atributos relevantes associados a cada relacionamento da rede de termos. E, finalmente, no capítulo 8 sumarizamos os aspectos positivos e limitações inerentes ao modelo representacional e algoritmos relacionais propostos, bem como apresentamos as principais conclusões do trabalho e direções de pesquisas proeminentes.

Capítulo 2

Conceitos Básicos

A definição de conceitos tais como predição, classificação, modelos de classificação, classificadores e a correlação entre tais conceitos é freqüentemente confusa e não consensual na literatura. Dessa forma, definiremos neste capítulo, de maneira mais precisa, cada um dos conceitos adotados no presente trabalho, e assumidos por grande parte dos trabalhos relacionados. Além disso, apresentamos importantes conceitos relacionados aos modelos de classificação relacionais, foco do nosso trabalho, uma definição formal do problema de classificação automática de documentos abordado, bem como uma descrição precisa da rede de termos definida.

2.1 Conceituação

Podemos definir a predição como a tarefa de estimar valores futuros de uma dada variável de interesse mediante um conhecimento prévio sobre o comportamento da mesma. Conseqüentemente, um modelo de predição é definido como um processo pelo qual um conjunto de regras, premissas e representações são criadas ou escolhidas para realizar a tarefa de predição. Sob o prisma científico, podemos dividir a predição em duas formas distintas: a predição sobre dados contínuos e a predição sobre dados categóricos.

A predição sobre dados contínuos consiste em definir uma relação matemática entre duas variáveis de valores contínuos, ou discretos sobre um intervalo infinito ou arbitrariamente grande. Modelos de regressão são os mais comuns para esses cenários. Tais modelos são vistos como ferramentas estatísticas que quantificam a relação entre variáveis dependentes e independentes. Podemos, assim, definir um modelo de regressão como um modelo de predição para dados contínuos. Já a predição de dados categóricos é usualmente referenciada como um problema de classificação. Ou seja, é

vista como o problema de assinalar classes a objetos de determinado domínio a partir do conhecimento de determinadas características destes objetos, dado um conjunto finito de classes distintas. Dessa forma, um modelo de classificação é tido aqui como um modelo de predição para dados categóricos.

Diversos modelos de classificação são propostos e avaliados na literatura. Dentre os mais importantes e populares podemos destacar as árvores de decisão, modelos estatísticos, modelos vetoriais e modelos baseados em reconhecimento de padrões, dentre outros [Sebastiani, 2002]. Tais modelos são, inclusive, aplicados aos diversos variantes do problema de classificação existentes, tais como classificação binária, *single-label* (i.e., um objeto pode estar associado a apenas uma classe dentre várias) e *multi-label* (i.e., um objeto pode estar associado a várias classes) [Tsoumakas & Katakis, 2007]. Cabe salientar ainda que, independentemente do problema abordado, um modelo de classificação consiste para nós em um conjunto de premissas básicas, que impõe direta ou indiretamente regras e representações específicas sobre dados, a serem seguidas pelo processo de classificação. Assim, os algoritmos de classificação, ou classificadores, passam a ser vistos como instâncias específicas de implementação, com diferenças procedimentais, de modelos. Por exemplo, os classificadores ID3 e C4.5 [Quinlan, 2003] são instâncias de implementação do modelo de árvores de decisão, ou seja, embora apresentem diferenças operacionais adotam as mesmas premissas sobre os dados de entrada.

Dada a generalidade da tarefa de classificação, ela atualmente é aplicada e estudada em diversos domínios distintos. A classificação de usuários em perfis, bem como de amostras celulares em tipos ou mesmo de documentos em classes semânticas, ilustram a grande aplicabilidade desta tarefa. Estudos focados em cada um desses domínios de aplicações objetivam identificar propriedades inatas ao domínio de estudo, que permitam melhorar a tarefa de classificação. Cada domínio possui propriedades intrínsecas distintas que determinam o sucesso de aplicação de modelos de classificação diferentes. Neste trabalho, objetivamos estudar a classificação automática de documentos (CAD), *single-label*, identificando e avaliando propriedades deste domínio de forma a justificar e melhorar modelos de classificação para documentos.

Os modelos de CAD tradicionais baseados em atributos (i.e., *attribute-based*), em geral, modelam seus dados como uma coleção de amostras de dados independentes e identicamente distribuídos, e apenas as informações contidas em tais atributos são consideradas. Tais modelos são comumente referenciados como *BAG of Words*. Entretanto, grande parte dos dados reais são relacionais, onde diferentes amostras estão relacionadas entre si. Por exemplo, para a classificação de artigos científicos podemos definir relacionamentos entre os documentos através das citações entre eles. Além

disso, estudos de diversos domínios [Macskassy & Provost, 2003; Sen & Getoor, 2007; Chakrabarti et al., 1998] constataram uma alta qualidade semântica presente em tais relacionamentos, facilitando consideravelmente a tarefa de classificação. A classificação de páginas Web [Chakrabarti et al., 1998], patentes [Couto et al., 2006] e mesmo de proteínas [Vazquez et al., 2003], figuram entre os cenários em que relacionamentos entre os objetos podem beneficiar a classificação. Baseado nisso, diversos dos modelos tradicionais foram modificados ou estendidos de forma a considerar tais relações, constituindo assim os modelos relacionais de classificação. Assim, mais especificamente, estamos interessados em investigar o problema de CAD considerando para isso modelos relacionais.

2.2 Modelos Relacionais de Classificação

De forma a permitir uma discussão mais ampla do modelo relacional para CAD apresentado, bem como dos principais trabalhos referentes a modelos relacionais, algumas terminologias precisam ser definidas. Apresentamos aqui os principais conceitos referentes a tais modelos e referenciados nos capítulos seguintes.

Um conjunto significativo de técnicas de classificação em dados relacionais pode ser visto como centrada em nodos, uma vez que elas focam em um simples nodo por vez [Macskassy & Provost, 2007], analisando sua vizinhança. Tais métodos são divididos na literatura em dois grupos distintos:

- **modelos relacionais:** tentam responder a seguinte questão: dado um vértice V_i e seus vizinhos em uma rede G de relacionamentos entre objetos de um domínio, como a classe de V_i pode ser estimada? Por exemplo, classificadores relacionais podem combinar informações de atributos locais de um vértice com informações de atributos de seus vizinhos para inferir classes [Sen & Getoor, 2007]. Tais modelos consideram para análise apenas os vizinhos de cada vértice V_i cujas classes são conhecidas, estabelecendo que a classe de um objeto em G depende apenas dos atributos de seus vizinhos com classes conhecidas. Como a análise de vizinhos em vários níveis (i.e., separados por várias arestas) pode ser um processo muito caro, modelos relacionais, em geral, garantem a viabilidade do processo adotando uma premissa Markoviana:

  $$P(X_i \mid G) = P(X_i \mid N_i)$$

  onde N_i é um conjunto de vizinhos imediatos do vértice V_i tal que $P(X_i \mid N_i)$ é independente de $G - N_i$ (i.e., $P(X_i \mid N_i) = P(X_i \mid G)$).

- **modelos coletivos:** são mais complexos e tentam tratar o problema de classificação quando a classificação de um vértice depende da classificação dos seus vizinhos e vice-versa. Ou seja, neste caso há também uma inter-dependência entre os objetos com classes desconhecidas. Dados modelados dessa forma são definidos como *networked-data*, pois assume-se que durante o processo de aprendizado vértices com classes desconhecidas estão conectados não somente a vértices com classes conhecidas, mas também a vértices com classes desconhecidas.

A exibição de modelos de classificação sobre dados relacionais através desta organização é útil por fornecer uma maneira de descrever abordagens distintas de forma a destacar as semelhanças e diferenças entre elas, tornando mais fácil a comparação entre as técnicas e um levantamento mais preciso dos trabalhos relacionados. Neste trabalho, antes a considerar modelos coletivos, analisamos modelos relacionais que assumem a premissa Markoviana.

2.3 Definição do Problema

Uma vez apresentados os principais conceitos referentes ao problema de classificação e a modelos relacionais, nosso próximo passo consiste em definir precisamente o problema abordado neste trabalho. O problema de classificação aqui abordado pode ser descrito mais formalmente como segue.

Definição 2.3.1 *Seja $G = (V; E; X; Y)$ um grafo onde V é o conjunto de todos os vértices existentes; E o conjunto de todas as arestas conhecidas; X o conjunto de atributos associados aos vértices; Y o conjunto de atributos associados às arestas. Cada vértice V_i está relacionado a um atributo único X_i, e cada aresta E_j pode estar relacionada a um subconjunto Y^j de Y. Seja também X^k um subconjunto de X cujos valores sejam conhecidos e, de forma contrária, X^u o subconjunto de atributos de X cujo valor é desconhecido, onde $X^k \cup X^u = X$ e $X^k \cap X^u = \emptyset$. Considere também que E corresponde apenas ao conjunto de arestas cujos atributos Y^j sejam conhecidos no grafo, ou seja, todos os valores de Y^j são conhecidos para toda aresta $E_j \in E$. Dessa forma, o problema de classificação é definido como o processo de inferir os valores de $X_i \in X^u$, ou definir a distribuição de probabilidade sobre estes valores, utilizando para isso todas as informações conhecidas de G.*

Para a tarefa de CAD, podemos definir o grafo G de diversas maneiras. Por exemplo, comumente na literatura G é definido como uma rede de documentos. Cada documento de uma coleção representa um nodo $V_i \in V$, e cada aresta $E_i \in E$ representa

uma citação entre quaisquer documentos V_i e V_k. A fim de inferir a classe de documentos não rotulados da coleção, utiliza-se apenas a porção de documentos da coleção cujas classes são conhecidas. Dessa forma, tem-se, na verdade, dois conjuntos distintos de documentos que são derivados em dois grafos $G_{treino} = (V_{treino}; E_{treino}; X_{treino}; Y_{treino})$ e $G_{teste} = (V_{teste}; E_{teste}; X_{teste}; Y_{teste})$. G_{treino} possui todos os valores dos atributos X_{treino} de seus vértices corretos e conhecidos. Ou seja, $X^k = X_{treino}$ e $X^u = \emptyset$. Já G_{teste} possui todos os valores de atributos X_{teste} desconhecidos e necessitam ser estimados. Em ambos grafos X representa o conjunto de classes dos documentos, e Y um conjunto de informações relevantes dos relacionamentos. Assim, utiliza-se as informações presentes no grafo G_{treino} para inferir os valores do grafo G_{teste}. É importante salientar que os grafos G_{treino} e G_{teste}, neste caso, são distintos, não compartilhando nenhuma informação. Em particular, eles apresentam estruturas topológicas distintas, tornando assim a tarefa de classificação mais difícil, dado que é necessário decidir o que deve ser considerado.

2.4 Rede de Termos

Como discutido acima, o grafo G pode ser definido de diversas formas. No exemplo citado, definimos V como um conjunto de documentos e E como relacionamentos entre tais documentos. Por utilizar-se de documentos, consideramos que a definição apresentada possui uma **granulação de documentos**. Diversos trabalhos utilizam definições similares a esta para classificar documentos relacionais (i.e., que possuem algum tipo de relação entre si) [Chakrabarti et al., 1998; Couto et al., 2006]. Diferentemente, a definição de rede adotada apresenta uma **granulação de termos**, por definir V como um conjunto de termos e E como relações entre os termos. A fim de explorar as informações dos relacionamentos definidos pela linguagem, construímos uma rede na qual os relacionamentos são definidos entre termos que co-ocorrem em um mesmo documento.

Mais formalmente, seja D um conjunto de documentos de treino, e X o conjunto de classes distintas presentes em D, em que cada documento $D_i \in D$ está associado a uma única classe $X_j \in X$. Considere também $T_i = \{t_1, t_2, \cdots, t_k\}$ o conjunto de termos distintos que ocorrem em um documento de treino D_i, e T^D o conjunto de todos os termos distintos observados em D. Dessa forma, definimos um grafo não direcionado $G = (V, E, X, Y)$, tal que V representa o conjunto de vértices de G, E o conjunto de arestas, X o conjunto de classes dos documentos e Y atributos relacionados às arestas. Cada termo presente em T^D corresponde a um vértice em V, ou seja, $|V| = \left|T^D\right|$, e haverá uma aresta entre dois termos distintos se eles co-ocorrerem em pelo menos

um dos documentos de D. Logo, cada documento $D_i \in D$ representa uma *clique* [1], definindo um conjunto de relacionamentos entre um conjunto específico de termos, uma vez que todos os termos de T_i se relacionarão. Além disso, para cada aresta $E_{t_l - t_y} \in E$, definimos dois atributos. O primeiro consiste na classe $X_i \in C$ na qual os termos t_l e t_y co-ocorrem mais freqüentemente em D. Denominaremos a classe X_i como a classe dominante da aresta $E_{t_l - t_y}$. O segundo atributo representa a *"intensidade"* do relacionamento entre t_l e t_y. Tal intensidade objetiva quantificar a relevância, para a tarefa de classificação, de cada relacionamento presente na rede.

Diversas formas de mensurar a intensidade podem ser propostas [Bernstein et al., 2003]. Adotamos aqui o conceito de **Predominância** (*Pred.*) [Rocha et al., 2008], associada à classe dominante, que representa a porcentagem de documentos em que $E_{t_l - t_y}$ foi observada na classe dominante X_i, tal como mostrado na fórmula 2.1.

$$Pred(E_{t_l - t_y}, X_i) = \frac{df(E_{t_l - t_y}, X_i)}{\sum_{k=1}^{M} df(E_{t_l - t_y}, X_k)} \tag{2.1}$$

onde *df* representa a função *document frequency*, que contabiliza o número de documentos distintos em que cada aresta $E_{t_l - t_y}$ é observada em uma classe X_j. Já a variável M representa o número total de classes presentes em X (i.e., $M = |X|$).

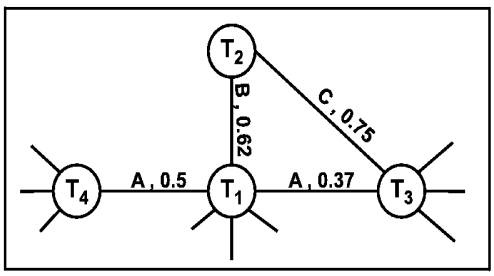


Figura 2.1. Exemplo de Rede de Termos

A figura 2.1 apresenta um exemplo de rede de termos. Analisando a figura e a definição apresentada, um ponto a observar é que os atributos X são na verdade associados aos documentos. Ao contrário dos documentos, cada termo está relacionado a diversos valores de X. Ao invés de associar cada termo à classe dominante, tal como feito para os relacionamentos, optamos por definir este atributo dinamicamente, a partir de uma análise da vizinhança de cada termo. Por exemplo, podemos associar ao termo T_1, da figura 2.1, a classe A, por ocorrer na maioria de seus relacionamentos. Dessa forma, projeções distintas das vizinhanças de cada nodo provêem valores distintos

[1] Uma clique corresponde a um subgrafo tal que para todo par de nodo existente, há uma aresta que os conecta.

para X, tornando a identificação do contexto de discurso mais flexível e dependente do subconjunto de relacionamentos analisados. Uma projeção é definida aqui como uma indução de um subgrafo G', a partir do grafo G, baseada em algum critério bem definido de seleção de um subconjunto de nodos $X' \in X$, e todos os relacionamentos estabelecidos entre os nodos de X'. Cada uma dessas projeções define o que chamamos de Comportamento Relacional dos termos. Um comportamento relacional consiste no conjunto de termos (i.e., vocabulário) com os quais cada termo se relaciona nos distintos contextos de discurso. A análise do comportamento relacional dos termos é promissora para CAD dado que contextos de discurso distintos definem comportamentos relacionais diferentes para os termos. Com isso, fenômenos naturais inerentes à linguagem, tal como a utilização de um mesmo termo em distintos contextos de discurso, passam a ser mais facilmente capturados, provendo formas mais robustas de se modelar domínios de CAD.

Cabe ressaltar que definições similares de rede de relacionamento entre termos existem [Cancho & Sole, 2001a,b]. Entretanto, diferentemente de propostas anteriores, não limitamos a distância de ocorrência entre termos em um documento para definir os relacionamentos. A eliminação desta restrição deve-se a duas razões principais. Primeiro, tal restrição impõe uma filtragem significativa sobre os relacionamentos entre termos, reduzindo a quantidade de informação disponível para a tarefa de classificação. Este problema não existe para os trabalhos prévios visto que não objetivam utilizar a rede de termos para CAD. A segunda razão é que mesmo relacionamentos entre termos distantes em um documento podem prover informações relevantes para a classificação. Por exemplo, a co-ocorrência entre os termos *feromônio* e *otimização*, mesmo que em sentenças separadas, pode ser uma evidência útil para a classificação de um documento sobre *otimização de colônia de formigas*.

2.5 Sumário

Neste capítulo apresentamos os principais conceitos utilizados em nosso trabalho. Iniciamos nossa discussão definindo a classificação como uma predição para dados categóricos. Ou seja, classificação é vista como o problema de assinalar classes a objetos de determinado domínio a partir do conhecimento de determinadas características destes objetos, dado um conjunto finito de classes distintas. Posteriormente, definimos especificamente como objetivo de estudo o problema de classificação automática de documentos (CAD), *single-label*, identificando e avaliando propriedades deste domínio de forma a justificar e melhorar modelos de classificação para documentos. Dentre os

vários modelos existentes para CAD, nosso foco está na análise de modelos relacionais dos dados, visto que grande parte dos dados reais são naturalmente relacionais.

Por focar em modelos relacionais, apresentamos e discutimos alguns conceitos associados a tais modelos. Os algoritmos para CAD propostos são baseados em modelos relacionais que assumem uma premissa Markoviana sobre os dados. Em seguida, o problema de CAD sobre dados relacionais foi também formalmente definido neste capítulo. Além disso, apresentamos algumas definições de grafos para o contexto de CAD, diferenciando grafos que apresentam uma **granulação de documentos**, por modelar relacionamentos entre documentos, de grafos que apresentam **granulação de termos**, que consideram relacionamentos entre termos. Finalmente, apresentamos nossa definição de grafo com granulação de termos, a ser utilizada por nossas propostas de algoritmos, e discutimos algumas de suas propriedades.

Capítulo 3

Trabalhos Relacionados

Uma das principais características do nosso trabalho é a interdisciplinaridade dos conceitos e propostas discutidos. Trabalhos da área de Redes Complexas, bem como de Lingüística são amplamente abordados em conjunto com trabalhos próprios da área de Recuperação de Informação, que estudam especificamente CAD. Dessa forma, a fim de melhor distinguir e apresentar os trabalhos referenciados, os agruparemos em cada uma destas áreas. Com isso, além de permitir uma melhor identificação das principais abordagens de cada área, conseguimos identificar interseções e diferenças entre tais abordagens.

3.1 Modelos Relacionais

Neste trabalho definimos documentos como um conjunto de dados relacionais, antes a um simples conjunto de termos isolados. Tal visão é comum não somente a documentos mas a diversos outros objetos de estudo em distintos domínios. Essa emergente tendência para modelagem de dados reais através de uma perspectiva relacional, bem como a disponibilidade de computadores potentes para processar esses dados, tornou redes complexas uma área de estudo de recorrente interesse [Newman, 2003]. Devido à sua capacidade em modelar uma grande variedade de aplicações, redes complexas têm sido aplicadas em campos tais como economia, esporte, medicina, entre outros. Conforme declarado por Wilson [1998], *"O maior desafio, hoje, não só em biologia celular e ecologia, mas de toda a ciência, é a correta e completa descrição dos sistemas complexos."*. Portanto, a modelagem de dados através de redes complexas se encontra em ampla expansão e podemos dividir os esforços de estudo na literatura em dois grupos principais: a modelagem descritiva e a modelagem preditiva.

A modelagem descritiva visa propiciar uma representação adequada para

os estudos de análise e entendimento da rede global, de um domínio de interesse [Dorogovtsev & Mendes, 2002; Archdeacon, 1996]. Estudos que buscam analisar a rede global focam em encontrar formas de sumarizar e explicar comportamentos que ocorrem na rede [Albert et al., 1999; Albert & Barabasi, 2002]. O intuito destes trabalhos é identificar fenômenos bem definidos, intrigantes e algumas vezes comuns a diversas redes. Além disso, esses estudos procuram definir modelos matemáticos, tais como modelos de Grafos Randômicos [Erdos & Renyi, 1960], Livres de Escala [Barabasi & Bonabeau, 2003] e Mundos Pequenos [Watts, 1999], que permitam entender e explicar tais fenômenos.

A modelagem preditiva, por sua vez, visa o estudo da tarefa de predição através das redes complexas, bem como tarefas de sistema de recomendação, campanhas mercadológicas, dentre outras, propondo modelos de predição, ou classificação, para diversas aplicações [Du et al., 2007; Said et al., 2008; Calderón-Benavides et al., 2004]. Tais estudos focam, principalmente, em investigar três premissas básicas: (1) o comportamento individual tende a ser consistente ao longo do tempo; (2) o comportamento de um grupo pode explicar comportamentos individuais; e (3) indivíduos similares tendem a se relacionar mais freqüentemente e se comportar de maneira semelhante. Essas premissas são, sobretudo, avaliadas quanto ao seu poder preditivo em diversos domínios distintos, sendo inclusive base para diversos modelos de classificação.

Grande parte dos modelos de classificação relacionais existentes são baseados na premissa (3), conhecida como homofilia [Mcpherson et al., 2001], ou auto-correlacionamento [Neville & Jensen, 2007]. Esses modelos são considerados como importantes linhas de base, pois a homofilia, em geral, é preponderante principalmente em redes sociais e de artefatos que possuem intervenção humana. O método relacional de vizinhança é o mais simples exemplo de classificador baseado neste modelo [Macskassy & Provost, 2004]. Para cada nodo, avalia-se sua vizinhança a fim de estimar a classe na qual a maior parte dos relacionamentos são observados ou que os nodos vizinhos pertençam. Este método adota a premissa Markoviana, discutida anteriormente, para avaliar as informações contidas na vizinhança. Outro método que adota esta premissa é o método relacional probabilístico de vizinhança [Macskassy & Provost, 2004], no qual, diferentemente do anterior, que gera apenas um valor como saída, gera uma distribuição de probabilidade para cada valor possível de saída. Outros trabalhos [Macskassy & Provost, 2003; Provost et al., 2003], investigaram, sobre diversos domínios, o ganho ao utilizar informações de vizinhança dos nodos. Simples modelos relacionais baseados em homofilia foram, então, propostos, alcançando resultados comparáveis a modelos mais complexos.

Para tratar os desafios da aprendizagem relacional, modelos complexos de classifi-

cação foram propostos recentemente. Por exemplo, modelos probabilísticos relacionais (MPRs) [Schmidt, 2000] são modelos baseado em grafos direcionados que estendem redes Bayesianas. Outro exemplo são as árvores de probabilidade relacionais (APRs), um método de aprendizado baseado em árvores de decisão que codificam a distribuição de probabilidade para as classes [Neville et al., 2003]. Taskar et al. [2004], por outro lado, usaram redes Markovianas relacionais (RMRs), baseadas em campos randômicos condicionais para seqüência de dados, para modelar dependência entre páginas Web, a fim de prever o tipo de cada página. Em Bernstein et al. [2003] é proposto um modelo vetorial relacional, em analogia ao modelo vetorial utilizado na área de Recuperação de Informação. O modelo proposto, basicamente, abstrai a estrutura da rede representando entidades por vetores adjacentes. Ou seja, cada entidade é representada através de informações, ponderadas ou não, de ligação (e.g., presença de *link*) entre a entidade e todas as demais entidades da rede. Em Macskassy & Provost [2004] também foi proposto um método no qual busca-se aprender como diferentes configurações das classes dos vizinhos afetam a classe de uma entidade a ser classificada. Já em Sen & Getoor [2007], os autores realizam um estudo comparativo entre vários modelos de classificação coletiva propostos na literatura. Outro estudo que investiga a classificação coletiva é o realizado em Neville & Jensen [2007]. Neste trabalho é proposto um modelo de classificação que corresponde à junção de modelos Bayesianos com Markovianos, definindo os chamados modelos de redes de dependências (MRDs). Dentre as aplicações de maior interesse atualmente para tais modelos, podemos destacar a tarefa de predição de relacionamentos em redes (i.e., *link prediction*). Em Taskar et al. [2003]; Bilgic et al. [2007], inclusive, são propostos modelos relacionais probabilísticos capazes de obter bons desempenhos para essa tarefa.

Para tratar a CAD, antes a modelos mais elaborados, adotamos os simples modelos relacionais que assumem homofilia na rede, e empregamos a premissa Markoviana para análise. A escolha por tais modelos, deve-se à usual eficiência de execução, comprovada eficácia em diversos domínios, e, sobretudo, à forte intuitividade por trás dos conceitos utilizados. Assim como árvores de decisão, por exemplo, o modelo adotado é facilmente interpretável por humanos.

3.2 Classificação de Documentos

Considerando agora os esforços em CAD, centrados na área de Recuperação de Informação, percebemos uma ampla diversidade de propostas para este problema atualmente. Tais propostas variam desde simples modelos, tais como modelos vetoriais, no qual

algoritmos como KNN [Salton & McGill, 1986] e Rocchio [Salton & McGill, 1986] se baseiam, a modelos mais complexos como *Support Vector Machines* (SVM) [Joachims, 2006] e Redes Neuronais [Li & Park, 2009]. Embora diferentes, a respeito de conceitos e complexidades algorítmicas, a maioria destes modelos tradicionais de RI adotam a mesma premissa: as amostras são independentes e identicamente distribuídas [Getoor, 2002]. Entretanto, grande parte dos dados reais são relacionais, em que diferentes entidades estão relacionadas entre si. Por exemplo, documentos podem ser vistos como um conjunto de termos que interagem entre si, antes a um simples *BAG of words*. Podemos também, considerar relacionamentos entre documentos como conseqüência de citações entre eles.

Baseado nessa observação, alguns modelos para CAD foram propostos, explorando o conhecimento sobre os relacionamentos entre entidades. Assim, modelos de classificação que consideram a rede (i.e., modelos relacionais) muitas vezes apresentam resultados significativamente melhores que os de modelos que a ignoram [Macskassy & Provost, 2004]. Entretanto, a maioria dos modelos relacionais para CAD consideram apenas citações explícitas entre os documentos, tais como citações entre artigos ou *links* entre hipertextos. Em Chakrabarti et al. [1998], por exemplo, os autores propuseram novas formas nas quais a informação latente em hiperlinks pode ser explorada em um classificador que utiliza informações de vizinhança. Simples modelos relacionais foram também propostos e testados para a classificação de documentos conectados tais como patentes [Chakrabarti et al., 1998] e páginas WEB [Couto et al., 2006]. O domínio de artigos científicos foi também analisado em Taskar et al. [2001], onde os autores propuseram uma classe de modelos para domínios relacionais que capturam dependências probabilísticas entre instâncias relacionadas.

Alguns trabalhos recentes [Schenker et al., 2003, 2004], objetivam modelar documentos através de grafos. Tais estudos mantêm a estrutura inerente aos documentos originais modelando cada documento como um grafo distinto, ao invés de um vetor. Em Markov & Last [2005], foi também proposta uma nova abordagem híbrida para a classificação de documentos WEB, combinando ambas representações, grafos e vetores. Baseados nesta representação, grande parte dos estudos estendem algoritmos de classificação tradicionais (e.g., algoritmo KNN), definindo medidas de distância entre grafos para comparar documentos distintos. Ou seja, o foco de tais estudos consiste em usar a informação estrutural contida em cada documento para melhorar sua classificação. Assim, esses métodos estão sempre limitados à granulação de documentos de uma coleção, uma vez que consideram informações relacionadas a cada documento. Com isso, importantes interações definidas pela linguagem na construção da comunicação humana não são modeladas por tais métodos. Diferentemente, neste trabalho

objetivamos usar a informação contida nas interações entre termos distintos de um domínio. Assim, criamos uma grande rede com todos os termos que ocorrem em cada documento de uma coleção e, através da identificação de co-ocorrências entre termos nos documentos, inferimos a classe na qual a co-ocorrência de um subconjunto específico de termos é mais provável de ocorrer. Como discutida no capítulo 2, esta rede apresenta a granulação de termos, por usar informações relacionadas aos termos.

Classificadores associativos [Veloso et al., 2006; Liu et al., 1998] parecem similares no sentido de também explorarem relacionamentos entre termos, mas apresentam grandes diferenças conceituais. A principal diferença refere-se ao modo como os relacionamentos são considerados no processo de classificação. Enquanto modelos associativos focam na identificação e uso de co-ocorrências isoladas mais relevantes (i.e., usualmente as mais freqüentes), modelos relacionais focam nos comportamentos relacionais de cada termo, compostos por todas as co-ocorrências do termo com todos os demais termos do domínio. Como dito, um comportamento relacional está associado a uma vizinhança de relacionamentos para cada termo, permitindo identificar a classe em que o termo ocorre com base no vocabulário associado a tal vizinhança. A definição de uma vizinhança também permite realizar projeções sobre a rede, a fim de considerar apenas um subconjunto de relacionamentos de interesse. Dessa forma, classes distintas podem ser preditas para um mesmo termo de acordo com a projeção realizada sobre sua vizinhança. Além disso, todos os relacionamentos podem ser representados em modelos relacionais, enquanto em modelos associativos, usualmente, mantêm-se apenas relacionamentos que ocorrem com uma freqüência acima de um limiar mínimo. Outro aspecto importante relacionado a modelos relacionais é que relacionamentos indiretos entre indivíduos podem ser facilmente definidos, extrapolando a premissa Markoviana, descrita no capítulo 2. Ou seja, além de examinar os vizinhos de cada termo, algoritmos que também consideram os vizinhos destes vizinhos, e assim por diante, podem ser definidos.

3.3 Estudos Lingüísticos

Como nosso modelo relacional para CAD é baseado em algumas observações lingüísticas sobre a comunicação humana, outro conjunto de métodos que merecem nossa atenção são aqueles baseados em estudos lingüísticos [Montejo-Raez et al., 2008]. Tais métodos objetivam expandir os conjuntos de dados com informações novas, chamadas *features* lingüísticas, a serem utilizadas por classificadores tradicionais em adição aos termos dos documentos (i.e., *features* tradicionais), comumente utilizadas. As *features* lingüís-

ticas usadas para treinar os classificadores são, por exemplo, as chamadas informações *POS-tags* tais como categoria sintática de uma palavra, como substantivo ou verbo, estrutura das frases, sentido das palavras, dentre outras propriedades gramaticais [Chen, 1995]. Em Aizawa [2001], por exemplo, palavras compostas, tais como *categorização texto*, foram definidas em adição a palavras individuais, por considerar que palavras compostas sejam *features* melhores que palavras isoladas. Os bons resultados alcançados mostram a utilidade deste tipo de informação. O uso de *POS-tags* foi também estudado em Moschitti & Basili [2004]. Usando *POS-tags* como *features* lingüísticas, os autores compararam os resultados alcançados pelo SVM e Rocchio com as versões tradicionais, usando apenas termos como features. A adição de *features* lingüísticas, neste caso, não promoveu ganhos significativos sobre o uso de termos apenas.

Diferentemente, em Gee & Cook [2005], os autores codificaram frases como grafos, definindo as arestas através de diferentes combinações de dimensões lingüísticas, tais como ordem das palavras, elementos sintáticos e elementos semânticos. Usando tais grafos, subgrafos freqüentes são identificados e usados para a classificação de documentos. Embora esta técnica tenha superado o desempenho alcançado por métodos *BAG of words*, sua aplicabilidade é bastante restrita dado o alto custo computacional associado. Há também estudos que usam as estruturas das frases como informações adicionais para classificadores [Furnkranz et al., 1998], ou combinam informações providas pela análise lingüística com métodos estatísticos de classificação [Bigi et al., 2001; Bingham et al., 2003]. Diferentemente, em Furnkranz [1998], é analisado o efeito de se utilizar N-gramas (i.e., seqüências de palavras de tamanho N) para a CAD. Esse estudo mostrou que seqüências de tamanho 2 e 3 são mais úteis mas, seqüências maiores reduzem a qualidade da classificação.

Outro modelo de classificação que considera seqüências de termos são os Modelos Estatísticos de Seqüência (MES) [Denoyer et al., 2001]. Tais modelos são usados em uma variedade de tarefas na área de processamento de linguagem natural. Dentre as aplicações comumente auxiliadas por esses modelos podemos citar a captura de dependência entre palavras, e a realização de inferências sobre seqüências. Em geral, os MES se baseiam na observação de que muitos documentos têm uma estrutura seqüencial que pode ser explorada em RI. Alguns documentos têm uma estrutura genérica. Este é o caso, por exemplo, de periódicos ou textos de conferências, que são compostos por título, resumo, introdução etc. Em documentos compostos de partes distintas, a distribuição de alguns termos relevantes pode ser diferente para cada parte, e este tipo de informação é capturada para desempenhar tarefas como CAD. Em Mittendorf & Schäuble [1994], um dos primeiros trabalhos em MES, foi proposto um modelo probabilístico baseado em *Hidden Markov Models* para a recuperação e

classificação de documentos. Embora esses modelos consigam apresentar resultados superiores a modelos Bayesianos simples, requerem, em geral, um alto custo computacional. Como se baseiam na análise da ordenação e organização dos termos presentes nos documentos, dado o número de termos existentes em uma linguagem, a quantidade de possíveis seqüências a serem avaliadas pode ser muito grande. Certamente pode-se reduzir este custo selecionando termos relevantes. Entretanto, este problema consiste em um tipo de seleção de *features*, que objetiva identificar quais termos são os mais importantes para um dado domínio. Com isso, a qualidade da classificação torna-se muito dependente da qualidade da seleção de *features* realizada.

É importante salientar que, diferentemente dos trabalhos lingüísticos mencionados acima, não estamos interessados em definir novos tipos de informações, baseado em aspectos sintáticos e semânticos dos textos. Tampouco objetivamos explorar a ordem de ocorrência dos termos nos documentos. Neste trabalho, focamos na informação provida pelos termos e pelos relacionamentos entre eles. Usamos os conceitos definidos pelo campo da Lingüística com o intuito de justificar a proposta de modelo relacional apresentada para CAD. Assim, acreditamos ser este o primeiro trabalho que define um algoritmo relacional de classificação sobre uma rede de relacionamento entre termos.

3.4 Análise Temporal

Embora a dimensão temporal seja reconhecidamente importante em diversos domínios, sua utilização é negligenciada por basicamente todos os algoritmos tradicionais de classificação de textos. Tanto os estudos de Redes Complexas, quanto os de Lingüística e da própria Recuperação de Informação, em geral, desconsideram a evolução temporal da comunicação.

Avaliando inicialmente a área de Redes Complexas, grande parte dos trabalhos, como dito, adotam a premissa de que o comportamento individual na rede tende a ser consistente ao longo do tempo. Assim, esses trabalhos focam na análise de uma "foto" estática da rede. Entretanto, a evolução temporal acarreta transformações naturais que podem modificar o comportamento observado da rede, invalidando parcial ou completamente os modelos estáticos construídos. Alguns trabalhos recentes analisam tal evolução sobre modelos descritivos de construção das redes, identificando tendências comportamentais ao longo do tempo [Leskovec et al., 2008; Kossinets et al., 2008; Crandall et al., 2008; Sharan & Neville, 2007]. Em Guo et al. [2007], os autores propuseram um modelo de evolução das redes ao longo do tempo, a fim de entender os mecanismos que determinam o surgimento de relacionamentos em redes complexas.

Outros trabalhos, como os discutidos em Hopcroft et al. [2003, 2004] buscam entender como é o processo de criação e extinção de comunidades em redes sociais. Considerando modelos preditivos, especialmente aplicados à CAD, não encontramos nenhum trabalho que considere a evolução temporal das redes definidas nos diversos estudos discutidos na seção 3.1.

Quanto à área de RI, grande parte dos esforços para entender o impacto da evolução temporal em CAD são divididos em duas grandes áreas de estudo: Classificação Adaptativa de Documentos e Tendência de Tópicos (*Concept Drift*). Classificação Adaptativa de Documentos [Cohen & Singer, 1999] engloba um conjunto de técnicas que visam contornar os problemas relacionados aos aspectos temporais, melhorando a efetividade e a precisão dos classificadores. Essas melhorias são alcançadas através de uma adaptação incremental e eficiente dos classificadores [Liu & Lu, 2002], e trazem pelo menos três grandes desafios. O primeiro é a definição de um conceito de contexto (i.e., partição de dados semanticamente significante), e como ele pode ser explorado para se obter melhores modelos de classificação. O segundo desafio consiste em criar modelos de forma incremental, e o terceiro está relacionado à eficiência computacional dos classificadores gerados, comumente elevado neste caso. A área de Tendência de Tópicos (*Concept Drift*) [Tsymbal, 2004; Widmer & Kubat, 1996] considera que tanto conceitos utilizados nos documentos quanto os interesses dos usuários se modificam ao longo do tempo, e que essas mudanças podem tornar inconsistentes modelos de classificação construídos com dados antigos. A partir dessa premissa, a área de Tendência de Tópicos visa identificar essas mudanças, e manter o modelo de classificação consistente com os conceitos atuais. *Concept Drift*, usualmente, assume que essa tendência é um fenômeno global sobre as coleções de documentos. Ou seja, assume-se que todos os conceitos de um domínios "evoluem" da mesma forma e com a mesma "intensidade". Entretanto, como exposto em Koren [2009], há graduais mudanças isoladas e não sincronizadas que não são capturadas por técnicas que adotam essa tendência global dos dados. Em documentos, subconjuntos de termos podem apresentar tendências de conceitos distintas, em momentos distintos.

Com o intuito de entender melhor a evolução temporal sobre coleções de documentos e seu impacto sobre a CAD, alguns estudos recentes analisaram mais a fundo tal aspecto [Mourão et al., 2008; Rocha et al., 2008; Salles et al., 2009]. Em Mourão et al. [2008], os autores identificam e caracterizaram três efeitos temporais das coleções de documentos que podem afetar o desempenho dos classificadores. Em Rocha et al. [2008], é apresentado o conceito de **contextos temporais**, que correspondem a porções de documentos que minimizam o impacto destes efeitos no desempenho dos classificadores. Além disso, os autores propuseram um algoritmo genérico para encontrar tais contex-

tos no cenário de CAD. O algoritmo proposto, *Chronos*, consiste em identificar o mais longo período no qual certas *features* permanecem estáveis em um conjunto de treino. Este período, assim, define um subconjunto de documentos de treino nos quais os termos ocorrem. Aplicando uma heurística gulosa simples, baseada nestes contextos, para classificar os documentos, os autores alcançaram ganhos significativos sobre a versão atemporal do algoritmo guloso. Em Salles et al. [2009], os autores avaliam formas de ponderação temporal dos dados, baseado em sua distância temporal para um ponto de referência, em alguns classificadores tradicionais. Os resultados obtidos demostram que a dimensão temporal é capaz de melhorar modelos tradicionais simples baseados em *BAG of words*.

Por fim, avaliamos os esforços sobre a análise temporal em estudos da Lingüística. Tarefas em Lingüística Computacional (CL) normalmente focam apenas no conteúdo de um documento, dando pouca atenção ao contexto em que foi ele produzido. Em Liebscher [2004], os autores avaliam a mudança lexical ao longo das décadas em coleções de publicações acadêmicas, mostrando que as mudanças podem ser bastante acentuadas durante um período, relativamente, curto de tempo. Há também alguns trabalhos em CL [Luo & Zincir-Heywood, 2004] que utilizam o termo *"Temporal"* para referenciar a ordenação cronológica dos termos dentro de um documento. Tais trabalhos são similares aos Modelos Estatísticos de Seqüência, anteriormente discutidos, e não abordam a evolução temporal das coleções.

Dessa forma, no presente trabalho propomos um modelo simples de classificação relacional, que também assume homofilia. Entretanto, diferentemente dos outros, consideramos também a inclusão da dimensão temporal na análise da rede.

3.5 Sumário

Discutimos neste capítulo os principais esforços existentes na literatura referentes à tarefa de CAD, considerando para isso aspectos lingüísticos, relacionais e temporais dos documentos. Dada a variedade de áreas de interseção apresentada por nosso trabalho, organizamos essa discussão nas principais áreas de estudo abordadas. Com isso, além de permitir uma melhor identificação das principais abordagens de cada área, conseguimos identificar interseções e diferenças entre tais abordagens.

Iniciando nossa discussão pelos estudos sobre dados relacionais, argumentamos inicialmente que a crescente disponibilidade deste tipo de dados tornou a área de Redes Complexas uma das mais estudadas atualmente. Tanto modelos descritivos dos dados, que visam propiciar um entendimento da rede global, quanto modelos predi-

tivos, que objetivam desempenhar a tarefa de predição através das redes complexas, são amplamente estudados. Neste contexto, modelos de classificação estão dentre os de maior interesse, dada sua aplicabilidade. Grande parte dos modelos de classificação relacionais existentes são baseados em uma premissa conhecida como homofilia. Assim, diversos trabalhos definem simples modelos relacionais que assumem que indivíduos similares se relacionam mais freqüentemente na rede. A simplicidade, fácil intuição e eficiência alcançada por esses modelos simples são usualmente decisivas para sua utilização, antes a modelos mais elaborados, em diversos estudos. Devido a tais caraterísticas, também adotamos modelos relacionais simples, baseados em homofilia.

Abordando agora trabalhos em CAD, centrados na área de Recuperação de Informação, percebemos uma ampla diversidade de propostas, que variam desde simples modelos, tais como modelos vetoriais, a modelos mais complexos, tais como SVM. Embora diferentes, a respeito de conceitos e complexidades algorítmicas, a maioria destes modelos de RI adotam a mesma premissa: as amostras são independentes e identicamente distribuídas. Entretanto, grande parte dos dados reais são relacionais, em que diferentes entidades estão relacionadas entre si. Apesar de existirem estudos em RI que consideram tais relacionamentos, eles estão normalmente limitados à granulação de documentos, avaliando apenas relações entre documentos distintos. Diferentemente, estamos interessados em utilizar informações sobre relacionamentos entre termos.

Quanto aos trabalhos conduzidos sob o prisma da Lingüística, em geral, eles objetivam definir novos tipos de informações a serem utilizadas por classificadores tradicionais. Informações tais como categoria sintática de uma palavra, como substantivo ou verbo, estrutura das frases, sentido das palavras, são extraídas do texto e incorporadas às entradas dos classificadores. Outros estudos procuram explorar a estrutura organizacional dos documentos, representando-os através de grafos, ou seqüências de palavras encontradas nos textos. Nosso trabalho difere destes mencionados por não propor a utilização de nenhum tipo de informação adicional. Simplesmente usamos os conceitos definidos pelo campo da Lingüística com o intuito de justificar a proposta de modelo relacional apresentada.

Finalmente, realizamos um levantamento sobre trabalhos que incorporam a dimensão temporal na tarefa de CAD. Nossa busca revelou que tanto os estudos de Redes Complexas, quanto os de Lingüística e da própria Recuperação de Informação, em geral, desconsideram a evolução temporal da comunicação. Alguns esforços nessa direção existem mas, em geral, formas robustas e eficientes para este problema ainda carecem ser propostas. Dessa forma, no presente trabalho apresentamos um simples modelo de classificação relacional, que considera a dimensão temporal na análise da rede.

Capítulo 4

Uma Perspectiva Lingüística

O entendimento sobre a linguagem humana pode beneficiar diversas tarefas em Recuperação de Informação [Montejo-Raez et al., 2008]. Por exemplo, sabe-se que um subconjunto de termos é usado em grande parte dos contextos de falas de uma linguagem. Essa informação pode ser importante para ignorar tais termos, ou torná-los menos relevantes, no processo de classificação. Embora este tipo de entendimento usualmente beneficie tarefas básicas, a maioria dos esforços atuais o ignora, focando no desenvolvimento de técnicas cada vez mais complexas e menos intuitivas. Neste capítulo abordamos algumas das propriedades lingüísticas que podem beneficiar a CAD. Mais especificamente, discutimos as vantagens de considerar os relacionamentos entre termos, comparado à utilização de termos individualmente, baseando nossa discussão em algumas observações sobre a linguagem humana.

Iniciamos nossa discussão descrevendo as coleções de documentos utilizadas para avaliar os conceitos e hipóteses abordados neste trabalho. Posteriormente, apresentamos algumas características intrigantes da linguagem, relacionadas aos termos, que explicam a eficiente cognição humana na comunicação. Mais especificamente, discutimos detalhadamente o impacto da simples utilização de termos em CAD. Posteriormente, discutimos os benefícios de se considerar os relacionamentos entre termos em CAD. Por fim, mostramos que a rede de termos utilizada para modelar os relacionamentos entre termos apresenta uma importante propriedade necessária para o bom desempenho de algoritmos relacionais.

4.1 Coleções de Documentos

Todas as análises realizadas no presente trabalho foram realizadas sobre quatro coleções de documentos reais, com características distintas. A primeira coleção (ACM) contém

Coleção	Número de Vértices	Número de Arestas	Densidade
ACM	56.449	8.602.858	152,40
MD	268.576	108.596.246	404,34
AG	251.642	59.523.131	236,54
NT	104.439	150.440.634	1440,46

Tabela 4.1. Informações Básicas das Redes

termos que ocorrem em aproximadamente 25.000 artigos de Ciência da Computação presentes na biblioteca digital da ACM, publicados anualmente entre 1980 e 2001. Em termos de classes, utilizamos apenas as 11 classes do primeiro nível do esquema da ACM. A segunda coleção (MD) compreende termos de artigos de Medicina disponibilizados pela biblioteca digital da MedLine. Essa coleção apresenta 861.454 artigos publicados anualmente de 1970 a 1985, divididos em 7 classes distintas. Nosso terceiro conjunto de documentos (AG) consiste de artigos presentes na *AG news corpus*, publicados diariamente. Os artigos foram coletados de 2000 fontes de notícias distintas, através de um coletor acadêmico denominado *ComeToMyHead* [Corso et al., 2005]. Tal conjunto de artigos possui 835.795 artigos publicados em 573 dias distintos entre 17/08/2004 a 20/02/2008, e organizados em 11 classes de notícias distintas. Por fim, a quarta coleção (NT) é composta por artigos da *Nature*[1], que são semanalmente publicados e organizados em 5 classes distintas. Nesta coleção há 7.964 artigos publicados entre 01/01/2005 a 21/12/2006. Todas as coleções passaram por um pre-processamento a fim de remover palavras *"stopwords"* dos documentos. É importante salientar também que cada documento de todas as coleções descritas é atribuído a apenas uma classe.

As coleções ACM, MD e AG são compostas por títulos e resumo dos artigos, quando disponíveis. Já a coleção NT é composta por documentos completos. A tabela 4.1 apresenta as principais características das redes de informação, construídas tal como descrito na seção 2.1, correspondentes a cada uma de nossas coleções. Podemos notar que as redes construídas são grandes, possuindo um grande número de arestas, e uma alta densidade (i.e., alto número de arestas por vértice). Em particular, a rede correspondente à coleção NT possui a mais alta densidade, devido a esta coleção ser composta por documentos completos. Embora os tamanhos das redes sejam, em geral, grandes, isso não representa um problema para nossos algoritmos de classificação, como veremos no capítulo 5. Como usamos pequenas projeções desta rede a cada momento, o processo de classificação não se torna computacionalmente caro.

[1] *Nature* corresponde a um dos mais antigos periódicos científicos destinado a uma ampla variedade de áreas de estudo.

4.2 Análise de Termos

Nesta seção discutimos algumas propriedades relacionadas a forma de utilização dos termos na comunicação humana. Primeiramente, analisamos algumas propriedades lingüísticas intrigantes relacionadas a forma como a comunicação é definida através dos termos. Em seguida, contrastamos as funções lingüísticas dos dois principais grupos de termos existentes na linguagem. Por fim, avaliamos o impacto de se utilizar um grupo específico de termos, os chamados termos especializados, em CAD.

4.2.1 Características da Linguagem

O ponto inicial de nossa análise são os diversos esforços a fim de entender a linguagem humana [Hauser et al., 2002]. Estudos diversos, tais como aqueles elaborados por Zipf [1949], permitiram grandes avanços, em termos práticos, nesta direção. Por exemplo, foi observado que o tempo de reação para identificarmos se um termo pertence ou não ao nosso vocabulário é menor que 100 milisegundos [Ke, 2007]. Ainda não está claro como somos capazes de realizar esta tarefa tão rapidamente, considerando que o vocabulário de uma linguagem possui, em geral, milhares de termos distintos. Trabalhos mais recentes [Ke, 2007; Cancho & Sole, 2001b] demonstraram que a linguagem humana emprega mais freqüentemente apenas um restrito conjunto destes termos, permitindo uma cognição mais eficiente e uma assimilação mais rápida da comunicação.

Outros estudos, modelando textos como redes, encontraram características intrigantes. Em Cancho & Sole [2001a], os autores definiram uma rede de informação, similar à definição apresentada na seção 2.4. Isto é, cada termo do texto representa um vértice, e haverá uma relação entre dois termos se eles co-ocorrerem em uma mesma sentença e estiverem separados por no máximo X outros termos. Usualmente o valor de X corresponde a 1 ou 2 termos. Uma análise sobre esta rede detectou tanto propriedades de *Mundos Pequenos* quanto de redes *Livres de Escala*, em relação à freqüência de ocorrência dos termos. Tais observações permitem-nos entender melhor a alta eficiência associada à cognição humana. A propriedade de *Mundos Pequenos* indica que poucos passos são necessários para conectar quaisquer termos da rede. Além disso, a propriedade de *Livres de Escala* demonstra que um número restrito de palavras são usadas para compor a maior parte da comunicação.

A fim de verificar essas propriedades em nossas coleções, construímos as redes de informações, tal como definido na seção 2.4, correspondentes a cada coleção. Analisando as redes criadas, identificamos as mesmas características, tal como mostrado na tabela 4.2 e nos gráficos da figura 4.1. A propriedade de redes *Livres de Escala* pode ser

observada através do comportamento modelado por leis de potência [Newman, 2005], como mostrado nos gráficos da figura 4.1. Já a propriedade de *Mundos Pequenos* é reconhecida por apresentar um pequeno caminho mínimo médio $\overline{K}$, similar ao observado em redes randômicas com mesmo número de nodos, aliado a um alto coeficiente de clusterização (CC) [2], muito maior que o das respectivas redes randômicas. Como podemos observar na tabela 4.2, os valores de $\overline{K}$ são muito similares em ambas versões da rede, enquanto os valores de CC para as redes reais são maiores que os encontrados nas versões randômicas por um fator superior a 150 para ACM, MD e AG e a 30 para a NT.

Coleção	Métrica			
	CC		$\overline{K}$	
	Real	Randômica	Real	Randômica
ACM	0.909	0.0053	2.328	2.116
MD	0.913	0.0030	2.211	1.867
AG	0.910	0.0019	2.346	2.019
NT	0.869	0.0276	1.972	1.451

Tabela 4.2. Métricas de Rede

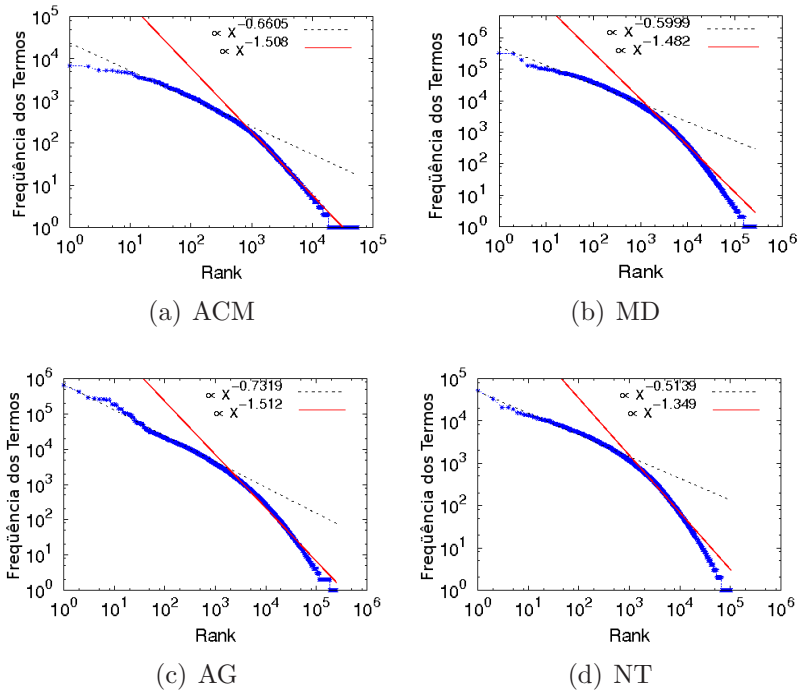
Apesar da intuição associada a tais observações, ainda não é claro seu impacto sobre CAD. Neste contexto, uma importante constatação, obtida através das propriedades de redes *Livres de Escala* observadas, é a existência de um limitado e distinto conjunto de termos, que se relacionam caracterizando cada contexto de fala na comunicação. Assim, consideramos que os documentos de cada classe definem um contexto de fala distinto. Tal consideração reforça nossa intuição que modelos nos quais termos são considerados isoladamente são mais fracos que aqueles que estamos propondo.

Em Cancho & Sole [2001b], os autores identificaram que, diferentemente da lei de Zipf [3], existem dois regimes distintos da freqüência de ocorrência dos termos por rank. Este comportamento sugere que, baseado na freqüência de ocorrência, podemos dividir todos os termos usados na comunicação geral em dois grupos: básicos e especializados. Por compor o *kernel* da linguagem[4], os termos básicos são também chamados termos *kernel*. Esses termos são definidos como os termos mais freqüentemente utilizados na comunicação e são modelados por uma lei de potência. Os termos especializados, como o nome sugere, são usados em contextos de fala específicos (i.e., classes) e são caracterizados por uma freqüência de ocorrência mais baixa, sendo definidos por uma lei de potência com decaimento maior que o observado para os termos kernel. Assim, a

[2] Esta medida quantifica a tendência dos nodos em uma rede se agruparem.

[3] A lei de Zipf estabelece que a freqüência de ocorrência do i-ésimo evento é inversamente proporcional ao seu rank [Adamic, 2000].

[4] conjunto de termos mais freqüentemente utilizados na maior parte da comunicação realizada por uma comunidade de indivíduos.



(a) ACM (b) MD

(c) AG (d) NT

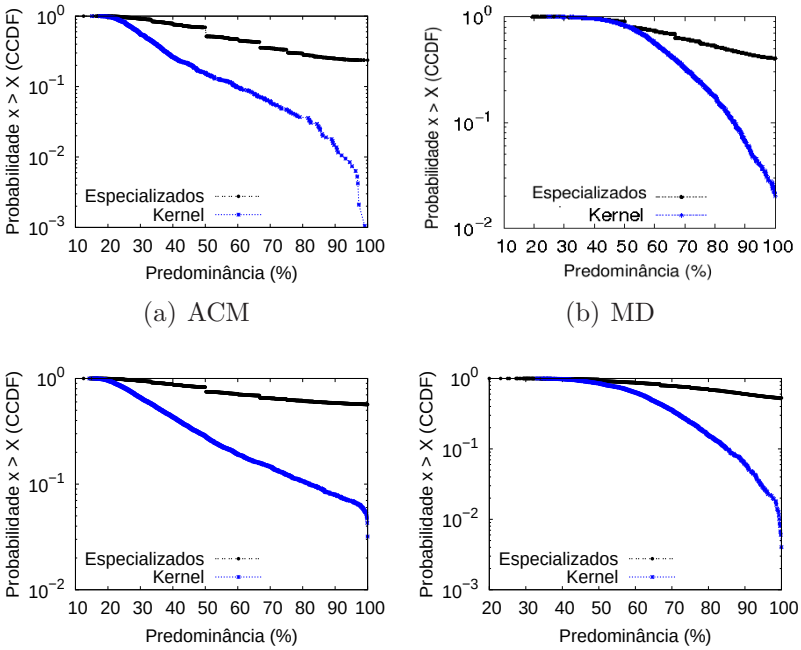
Figura 4.1. Freqüência de Ocorrência dos Termos por *Rank*

delimitação entre esses dois grupos ocorre no ponto de interseção entre as duas leis de potência. Analisando novamente os gráficos da figura 4.1, podemos observar claramente os dois regimes mencionados em todas as coleções analisadas. Note que a lei de potência com menor expoente modela apenas os termos de menor rank, que correspondem aos termos mais freqüentes nas coleções. A segunda lei de potência, que modela a cauda da distribuição, engloba um número maior de termos distintos que ocorrem menos freqüentemente.

4.2.2 Termos Kernel vs. Termos Especializados

Com base na discussão da seção anterior, podemos observar que a principal diferença entre os dois grupos de termos consiste na função lingüística desempenhada por cada um. O kernel de uma linguagem define a linguagem básica de comunicação usada pelos membros de uma comunidade. Grande parte da comunicação pode ser realizada utilizando apenas este restrito conjunto de termos. Entretanto, tais termos são, usualmente, empregados em um amplo número de contextos de fala distintos, estando associados a vários contextos simultaneamente e sendo, conseqüentemente, menos discriminativos para a tarefa de classificação. Demonstramos isso calculando a *Predominância* de cada termo, tal como discutido na seção 2.4. A *Predominância*, neste caso, mede o grau de exclusividade de um termo em relação às classes nas quais ele ocorre. A ocorrência

de um termo em uma classe é definida como o número de documentos distintos da classe nos quais o termo aparece. A figura 4.2 mostra a *Complementary Cumulative Distribution Function* ($CCDF$) para todos os termos kernel em nossas bases de dados. Observamos que a probabilidade desses termos possuírem uma *Predominância* alta é muita baixa, confirmando nossa hipótese de baixo poder discriminativo dos termos kernel.



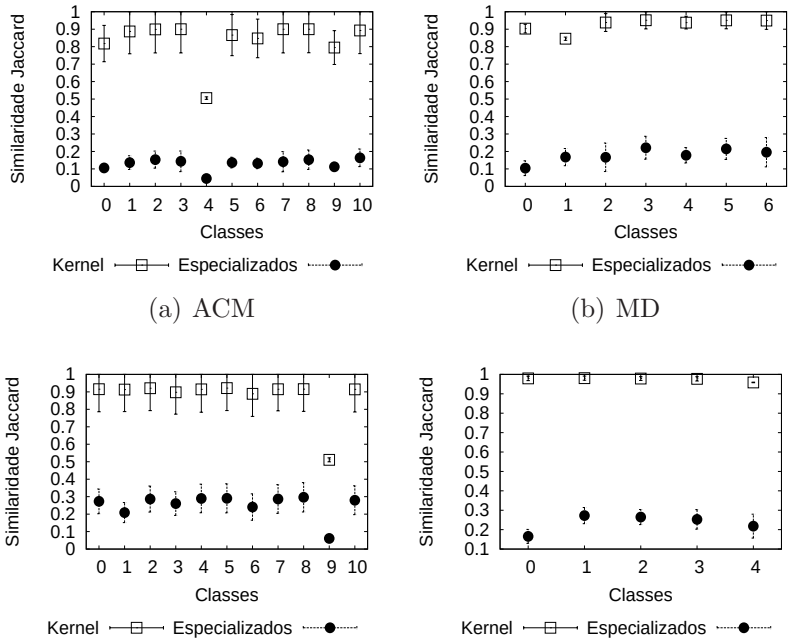
(a) ACM (b) MD

(c) AG (d) NT

Figura 4.2. Distribuição de Predominância

Os termos especializados, por outro lado, apresentam a função lingüística de definirem contextos de fala mais específicos. É intuitivo que haja uma grande diversidade de contextos distintos, assim como um grande número de termos especializados distintos. Nossa hipótese é que termos especializados estão associados a poucos contextos distintos, representando um conjunto de termos mais efetivo para diferenciar os contextos e, conseqüentemente, melhor classificar documentos. Avaliamos tal hipótese definindo dois tipos de vocabulários para cada classe: um usando apenas os termos kernel e outro usando apenas os termos especializados. Em seguida, calculamos a similaridade média entre os vocabulários de todas as classes, considerando o vocabulário de termos kernel e especializados separadamente. Para o cálculo da similaridade, usamos o coeficiente de *Jaccard*. Os resultados são mostrados na figura 4.3. Assim, podemos confirmar que termos kernel são amplamente utilizados, apresentando similaridades altas entre todas as classes. Em contrapartida, as similaridades entre os vocabulários de termos especializados são significativamente menores, mostrando que as classes possuem uma

baixa interseção com relação aos termos especializados. Um ponto interessante a mencionar é que quando usamos o *Cosseno* como medida de similaridade, a distinção entre termos kernel e especializados se torna menos proeminente. Usando Jaccard a maioria das classes possuem uma similaridade média elevada (acima de 80%), enquanto para o Cosseno esta similaridade entre classes reduz sensivelmente (aproximadamente 50%), mostrando que a freqüência de ocorrência de cada termo representa uma informação, de fato, útil para a diferenciação entre classes.



(a) ACM (b) MD

(c) AG (d) NT

Figura 4.3. Similaridade *Jaccard*

4.2.3 Uso de Termos Especializados em CAD

Definindo CAD como a tarefa de determinar o contexto de fala no qual um documento pertence, com base na discussão anterior, o uso de termos especializados demonstra ser uma estratégia promissora. Entretanto, o uso desses termos somente pode resultar em dois grandes desafios para CAD, descritos a seguir.

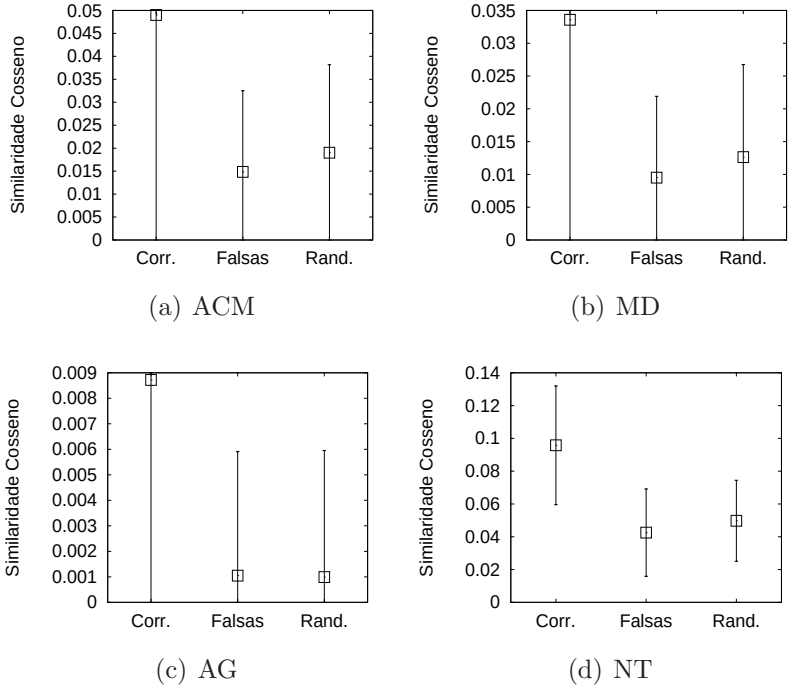
Reuso de Termos: Embora termos especializados estejam relacionados a poucos contextos de fala, seu uso, em geral, não está restrito a um único contexto. Com isso, identificar corretamente o contexto de fala de aparição dos termos, e conseqüentemente de cada documento a ser classificado, se torna mais difícil, degenerando a qualidade das classificações realizadas. O reuso de termos é observado devido a dois fatores principais. Primeiramente, existem algumas limitações impostas pela própria linguagem, em particular fenômenos tais como polissemia, em que um mesmo termo

apresenta diversos significados. O segundo fator consiste no mesmo conceito ser aplicado a diferentes contextos. Um exemplo seria o uso de conceitos genéticos na área de máquinas de aprendizado. Assim, o reuso de termos é um fator inerente à comunicação e à evolução da linguagem e podemos facilmente identificar sua ocorrência. Considerando novamente a figura 4.3, podemos observar que embora a similaridade média entre as classes, considerando o vocabulário de termos especializados, é baixa, ela é diferente de zero. Ou seja, mesmo para vocabulários especializados, contextos de fala distintos compartilham um subconjunto de termos. O reuso também explica o fato da *Predominância* dos termos especializados ficar abaixo do máximo possível na figura 4.2 (i.e., ficar abaixo de 1).

Super-especialização de Classes: Dado que a freqüência de ocorrência dos termos especializados é baixa, vocabulários definidos apenas por tais termos geram uma representação restritiva e *ad-hoc* para as classes. Este comportamento dificulta a tarefa de classificação visto que se torna elevada a probabilidade de um novo documento, a ser classificado, apresentar uma similaridade baixa com todas as classes. Isso ocorre devido aos termos presentes neste novo documento terem grandes chances de não estarem dentre os termos, em geral de baixa freqüência, utilizados para descrever as classes. Podemos verificar essa hipótese através do seguinte experimento. Dividimos nossas coleções de documentos em um conjunto de treino (90%) e um conjunto de teste (10%). Posteriormente, definimos um vocabulário especializado para cada classe no conjunto de treino e calculamos a similaridade entre cada documento de teste e todos os vocabulários construídos no treino. Como a freqüência de ocorrência dos termos mostrou-se relevante para diferenciar as classes, adotamos a similaridade Cosseno para este experimento.

A figura 4.4 mostra a similaridade média para todos os documentos de teste, considerando a classe correta de cada documento (*Corr.*), bem como a similaridade média dos documentos com todas as outras classes (*Falsas*). Podemos observar que a similaridade da classe correta não difere muito da similaridade das outras classes. Isso mostra que os termos especializados presentes no conjunto de teste diferem daqueles encontrados no treino, mesmo quando consideramos o mesmo contexto de fala. Assim, esses termos perdem sua capacidade de identificar um contexto de fala no qual ocorrem. Para mostrar isso, contrastamos os valores de similaridades encontrados para os documentos de teste com valores encontrados para documentos gerados randomicamente com o mesmo número de termos. Ou seja, geramos um conjunto de documentos equivalente ao de teste, no qual cada documento possui termos randomicamente selecionados, tornando o documento desprovido de significado e, conseqüentemente, não possuindo um contexto de fala. As similaridades para esses documentos (*Rand.*) com os vocabu-

lários do treino são mostradas na figura 4.4, e não diferem muito da similaridade para as classes corretas. Ou seja, documentos desprovidos de contexto de fala apresentam similaridades semelhantes a documentos reais quando usamos apenas termos especializados para descrever as classes. Dessa forma, o uso de termos especializados somente não gera uma descrição ampla dos contextos de fala.



(a) ACM (b) MD

(c) AG (d) NT

Figura 4.4. Similaridade Cosseno

Assim, duas questões fundamentais relacionadas à tarefa de classificação emergem: Como podemos identificar o contexto de fala de um termo em um documento específico? Como podemos definir de maneira mais ampla esse contexto? As respostas para tais perguntas podem ser encontradas na análise dos relacionamentos entre os termos. Como em diferentes contextos de fala subconjuntos distintos de termos, sobretudo especializados, são utilizados, a análise do comportamento relacional de cada termo provê informações importantes sobre o contexto de aparição dos termos, definindo uma modelagem mais adequada destes contextos.

4.3 Análise de Relacionamentos

Nesta seção abordamos como os relacionamentos que consideramos podem auxiliar a tarefa de classificação. Termos interagem de diversas formas na linguagem humana. Textos são organizados como sentenças, que são compostas por palavras que interagem entre si. De fato, alguns conjuntos de termos possuem uma maior probabilidade

de co-ocorrerem [Cancho & Sole, 2001a], e este tipo de relacionamento contém informações importantes não capturadas quando consideramos os termos individualmente. Um fato que demostra essa importância é que se um texto for embaralhado, gerando um documento randômico totalmente desprovido de semântica, diversas das observações relacionadas aos termos e sua freqüência, anteriormente discutidas, se mantém verdadeiras. Dessa forma, explorar os relacionamentos entre termos, além de intuitivo, consiste numa promissora estratégia de modelagem para CAD.

A fim de prover um melhor entendimento sobre os relacionamentos entre termos, definimos três tipos de relacionamentos, baseado na distinção entre termos kernel e especializados. O primeiro tipo, **relacionamentos sintáticos**, conecta dois termos kernel. Em geral, tais relacionamentos são derivados de regras específicas da linguagem, ou representam conceitos que estabelecem a comunicação. Por exemplo, a interação entre um advérbio e um verbo, tal como *modela precisamente*, representa um relacionamento sintático comumente observado. Devido às características dos termos kernel, esse tipo de relacionamento apresenta o menor número de relacionamentos distintos mas, com maior freqüência de ocorrência e maior variedade de uso, considerando o número de contextos de fala distintos em que ocorrem.

O segundo tipo consiste de **relacionamentos conectivos**, os quais são relacionamentos entre um termo kernel e um termo especializado. Relacionamentos conectivos objetivam definir o significado semântico (i.e., a interpretação) do discurso, bem como permite a construção de sentenças. Por exemplo, relacionamentos entre adjetivos e substantivos (e.g., *grafo direcionado*) ou entre verbos e substantivos ajudam a identificar corretamente as características ou ações relacionadas aos objetos em uma sentença. Além disso, a existência de pronomes, advérbios, dentre outros, entre substantivos permite a construção de sentenças em uma linguagem. Esses relacionamentos possuem uma freqüência de ocorrência menor que relacionamentos sintáticos mas, compreendem um maior número de relacionamentos distintos.

Finalmente, existem os **relacionamentos semânticos**, que conectam dois termos especializados. Relacionamentos semânticos visam definir os conceitos dos contextos de fala. A existência desses relacionamentos é a principal razão para a eficiente identificação do contexto definido pelos termos de um documento. Por exemplo, a co-ocorrência entre dois substantivos (e.g., *algoritmo genético*) pode permitir a identificação do contexto de fala. Tais relacionamentos apresentam uma freqüência de ocorrência menor mas são os mais numerosos, possuindo um grande número de relacionamentos distintos, e restritivos, por ocorrerem em um pequeno número de contextos diferentes.

A fim de ilustrar essa característica restritiva do tipo semântico, comparado aos

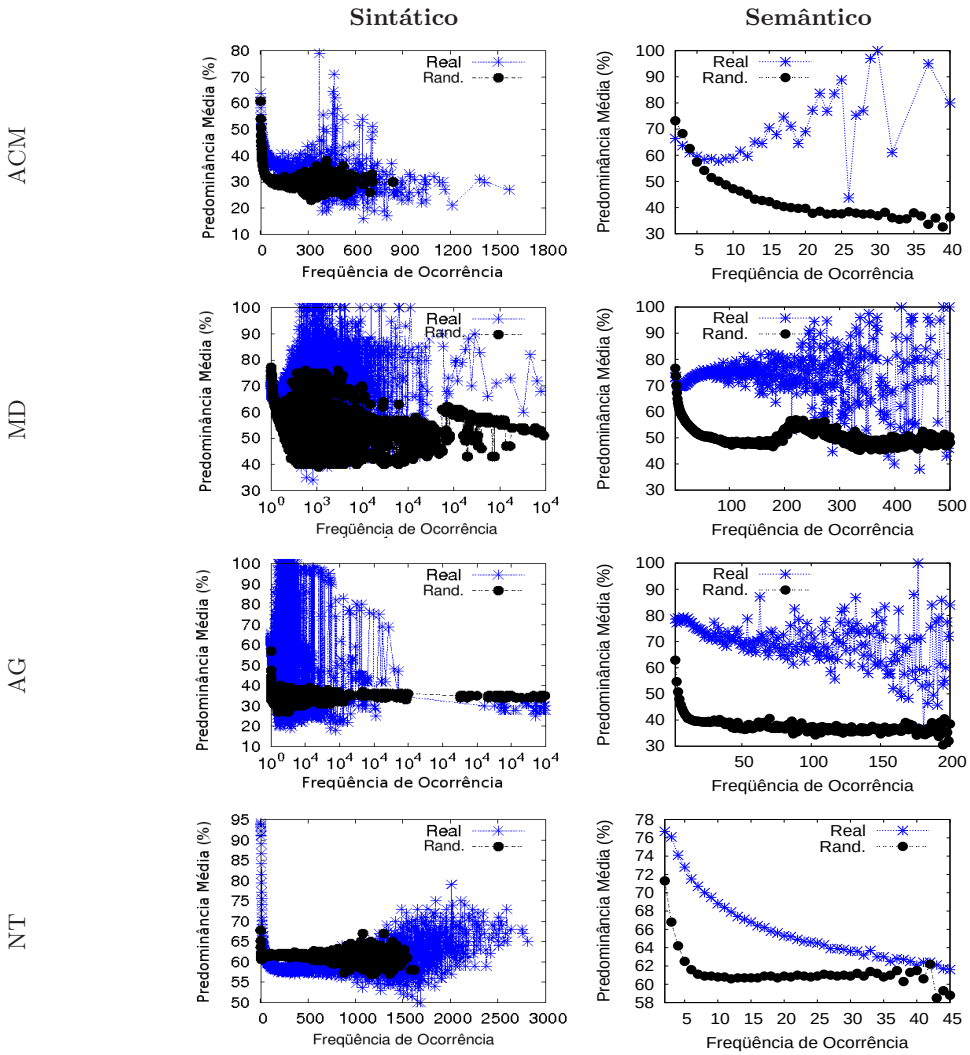


Figura 4.5. *Predominância* por Freqüência

relacionamentos sintáticos, avaliamos novamente a métrica de *Predominância*, associada a cada relacionamento de nossas redes de informação. Visto que relacionamentos semânticos são mais restritivos, tendem a apresentar valores mais elevados de *Predominância* que relacionamentos sintáticos. Entretanto, uma simples comparação entre as distribuições de probabilidade das *Predominâncias* dos dois tipos de relacionamentos pode não ser correto, dado que o conceito de *Predominância* está diretamente relacionado à freqüência de ocorrência. Isto é, como relacionamentos semânticos são menos freqüentes, eles tendem a apresentar Predominâncias mais altas. Assim, calculamos a distribuição de Predominância média por freqüência de ocorrência para cada tipo de relacionamento em nossas redes. Em seguida, contrastamos as distribuições obtidas com distribuições correspondentes em um cenário randômico. Ou seja, geramos uma coleção de documentos randômicos, nos quais os termos de cada documento são selecionados randomicamente dentre todos os termos que aparecem nas coleções originais. Neste cenário, identificamos os relacionamentos semânticos e sintáticos e calculamos

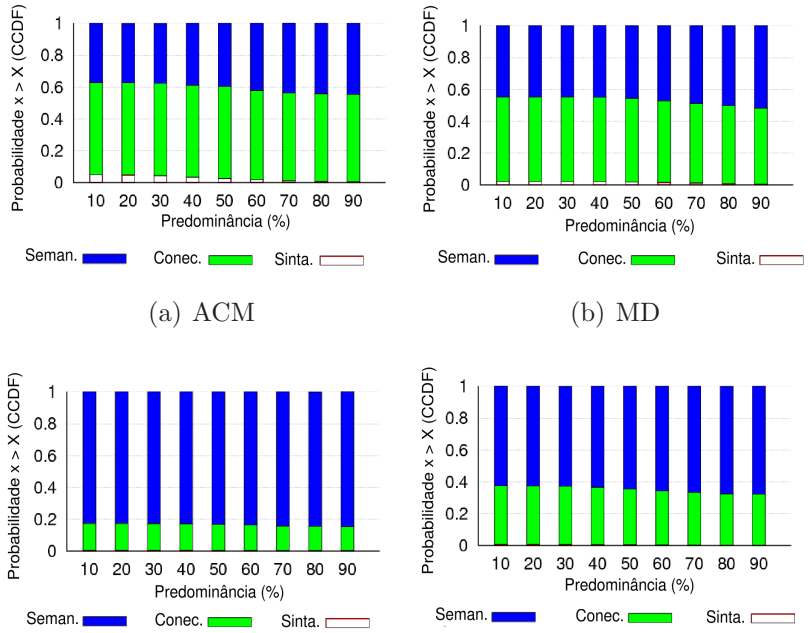
a distribuição de Predominância média por freqüência de ocorrência para cada tipo. Como os relacionamentos são criados de maneira randômica, sua ocorrência em geral é distribuída entre os diversos contextos existentes em cada conjunto de dados. Deste modo, para um tipo de relacionamento ser considerado restritivo, quanto aos contextos, é necessário que sua distribuição de Predominância na rede real apresente valores suficientemente maiores que os observados no cenário randômico. E, quanto maior esta diferença, mais restritivo são os relacionamentos avaliados.

A figura 4.5 contrasta as distribuições de cada tipo de relacionamento nas redes reais (Real) com as versões randômicas (Rand.). Como podemos observar, existe uma diferença maior entre os dois cenários para relacionamentos semânticos. Ou seja, há nas redes reais uma associação entre termos especializados e contextos de fala específicos maior que em um cenário randômico, definindo valores mais altos de Predominância. De forma contrária, essa associação é perto de algo randômico para relacionamentos sintáticos, mostrando que esses relacionamentos, de fato, apresentam um comportamento mais homogêneo, quanto aos diferentes contextos em que ocorrem.

Dessa forma, o uso de relacionamentos semânticos seria uma promissora resposta para a questão de como melhor identificar os contextos de fala dos termos. Seu uso reduz o impacto do problema de reuso de termos, visto que este tipo de relacionamento, usualmente, está associado a um conjunto menor ainda de contextos distintos. Além disso, relacionamentos semânticos provêem uma quantidade maior de informação sobre os documentos, dado que representam as interações entre os termos em cada documento.

Embora o uso de relacionamentos semânticos aborde o problema de reuso, é importante notar que o problema de super-especialização ainda se mantém. Como relacionamentos semânticos ocorrem com uma freqüência ainda menor, o uso deste tipo de relacionamento apenas reforça o problema de super-especialização. Uma forma de abordar a super-especialização seria utilizar os outros dois tipos de relacionamentos. Visto que relacionamentos sintáticos e conectivos apresentam uma freqüência de ocorrência mais elevada, a probabilidade de serem encontrados em novos documentos é maior, provendo uma representação mais ampla de cada contexto. Entretanto, a simples consideração de todos os relacionamentos sintáticos e conectivos degradaria o poder discriminativo dos relacionamentos. Assim, uma maneira mais robusta de definir relacionamentos relevantes para CAD seria selecionar relacionamentos mais *"intensos"*, onde esta intensidade pode ser medida através da Predominância. Para a tarefa de CAD, observamos que quando o valor de Predominância é alto, mesmo relacionamentos sintáticos e conectivos permitem uma melhor diferenciação entre classes, uma vez que sua ocorrência é restrita a poucas classes distintas. Assim, o uso dos três tipos de relacionamentos com alta Predominância endereça ambos problema simultaneamente.

Essa estratégia é particularmente interessante devido a, além de selecionar bons relacionamentos sintáticos e conectivos, garantir o uso de grande parte dos relacionamentos semânticos, que tendem a apresentar valores elevados de Predominância. A figura 4.6 mostra a composição dos tipos de relacionamentos selecionados por cada valor de Predominância. Os relacionamentos semânticos são identificados na figura por *Seman.*, conectivos por *Conec.* e sintáticos por *Sinta.*. Note que, quanto maior o valor da Predominância, maior é a porcentagem de relacionamentos semânticos dentre os selecionados para todas as coleções.



(a) ACM (b) MD

(c) AG (d) NT

Figura 4.6. Composição dos Relacionamentos por Predominância

4.4 Homofilia da Rede

Como exposto em Macskassy & Provost [2004], uma importante observação é que algoritmos relacionais, usualmente, são fortemente enviesados pela premissa de *homofilia*. A *homofilia* é uma propriedade de redes sociais que estabelece que indivíduos similares, com relação a alguma característica relevante, se relacionam mais freqüentemente na rede. Como grande parte dos modelos de classificação relacionais analisam a vizinhança de um nodo, é desejável que essa vizinhança seja composta por indivíduos que apresentem comportamento similar. Em nosso caso, significa que, para a identificação do contexto de cada termo, é necessário que a maior parte de seus vizinhos ocorram em poucos contextos distintos, e que tais contextos sejam os mesmos. Assim,

em nossa rede é desejável que os termos usados na mesma classe estejam relacionados mais freqüentemente entre si.

A identificação da *homofilia* é usualmente desempenhada comparando-se o comportamento apresentado por uma rede real com o observado em uma correspondente versão randômica. Se a probabilidade de relacionamentos entre indivíduos similares é maior na rede real, em comparação com a versão randômica, é dito que a rede apresenta um "desbalanceamento", priorizando relações entre indivíduos similares. Considerando que em nossa rede termos são similares se ocorrem na mesma classe, para identificar a homofilia precisamos mostrar duas coisas distintas. Primeiramente, é preciso mostrar que os vizinhos de um termo tendem a ocorrer em poucas classes distintas. Em segundo lugar, é necessário que estas poucas classes em que cada vizinho é observado sejam as mesmas para todos os vizinhos.

Os gráficos da figura 4.5 mostram exatamente a primeira parte dessa identificação para nossas redes de informação. Como mencionado na seção 4.3, nestes gráficos traçamos a distribuição de Predominância dos relacionamentos entre os termos pela freqüência de ocorrência dos mesmos, para as versões real e randômica das redes. Observando os gráficos, notamos que os relacionamentos entre termos especializados tendem a ocorrer em um número de classes menor que o observado nas redes randômicas, apresentando uma maior Predominância. Quanto a segunda parte, uma análise global da vizinhança de cada termo provavelmente não geraria um comportamento homogêneo. Isso porque termos se relacionam com determinados termos em contextos específicos e, em geral, relacionam-se com subconjuntos distintos de termos para diferentes contextos. Dessa forma, a similaridade da vizinhança de cada termo deve ser avaliada sobre projeções distintas de sua vizinhança global. A forma de se definir tais projeções é que vai tornar uma vizinhança "local" mais ou menos homogênea. Assim, podemos considerar que em redes de termos reais os relacionamentos entre termos mais freqüentemente utilizados em um mesmo contexto podem ser priorizados, definindo funções de projeções adequadas.

4.5 Sumário

Apresentamos neste capítulo uma detalhada discussão sobre algumas propriedades lingüísticas da comunicação humana, que podem beneficiar a tarefa de CAD. Todas as hipóteses e conceitos discutidos foram avaliados em quatro bases de dados reais. Iniciamos nossa discussão descrevendo as coleções de documentos utilizadas para avaliar os conceitos e hipóteses abordados neste trabalho. Em seguida, apresentamos

características intrigantes da linguagem que explicam a eficiente cognição humana na comunicação. Modelando documentos como uma rede de termos, estudos mostraram que tal rede apresenta características de *Mundos Pequenos* e *Livres de Escala*. Tais observações confirmam a intuição de que um restrito conjunto de termos é mais freqüentemente utilizado na comunicação, e que todos termos estão, em geral, próximos na rede.

Outros estudos, mostraram que os termos utilizados na comunicação podem ser divididos em dois grupos distintos, baseado em sua freqüência de utilização, com funções lingüísticas diferentes. O primeiro grupo corresponde aos termos kernel, que são responsáveis pela comunicação geral, sendo utilizados mais freqüentemente em um maior número de contextos de fala distintos. Diferentemente, o segundo grupo consiste nos termos especializados, usados menos freqüentemente porém, em contextos de fala específicos, sendo mais efetivos para a tarefa de CAD. Apesar do potencial de uso dos termos especializados para a classificação, sua simples utilização não é conveniente. Tal estratégia resulta em dois grandes desafios para CAD. O primeiro deles consiste no reuso de termos, dado que por limitações impostas pela linguagem um mesmo termo é utilizado em contextos distintos. Outro fator que contribui para o reuso de termos é a utilização de um mesmo conceito em contextos distintos, decorrente da evolução da própria linguagem. O segundo problema, definido como super-especialização de classes, refere-se à construção de representações restritivas e *ad-hoc* para cada contexto de fala. Como a freqüência de ocorrência dos termos especializados é baixa, a probabilidade deles serem encontrados em novos documentos é baixa, não definindo uma descrição ampla dos contextos.

Assim, através da análise de relacionamentos entre termos, abordamos duas importantes questões, associadas a CAD, que emergem quando consideramos o uso de termos isoladamente: Como podemos identificar o contexto de fala de um termo em um documento específico? Como podemos definir de maneira mais ampla este contexto? Como textos são organizados como sentenças, que são compostas por palavras que interagem entre si, considerar o comportamento relacional dos termos é uma promissora forma de tratar as questões levantadas.

A fim de prover um melhor entendimento sobre os relacionamentos entre termos, definimos três tipos de relacionamentos, baseado na distinção entre termos kernel e especializados. Cada tipo de relacionamento possui características e funções distintas quanto à composição da comunicação. Os dois primeiros tipos de relacionamentos, sintáticos e conectivos, provêem uma descrição mais ampla dos distintos contextos de fala, por possuírem uma freqüência de ocorrência mais elevada. Dessa forma, tais relacionamentos endereçam a questão de definição ampla dos contextos. De forma contrária, os

relacionamentos do terceiro tipo, semânticos, definem os principais conceitos inerentes a cada contexto de fala, por ocorrerem em um restrito conjunto de contextos distintos, abordando a questão de identificação dos contextos. De forma a endereçar ambas questões ao mesmo tempo, propusemos uma forma de selecionar os relacionamentos mais relevantes para CAD, dentre os três tipos, considerando a *"intensidade"* dos relacionamentos. Essa intensidade foi medida através do conceito de Predominância de cada relacionamento, tal como discutido na seção 2.3.

Por fim, mostramos que a rede de termos definida apresenta a importante propriedade de *homofilia*. A *homofilia* estabelece que indivíduos considerados "similares" se relacionam mais freqüentemente na rede. Essa é uma propriedade necessária para o bom desempenho de vários dos algoritmos relacionais de classificação, incluindo os que vamos discutir no próximo capítulo.

Capítulo 5

Algoritmos Relacionais de Classificação

Neste capítulo, apresentamos nosso algoritmo relacional para CAD, baseado na análise de vizinhança da rede de termos. Objetivando melhor distinguir os diferentes passos do algoritmo, apresentamos várias versões incrementais do mesmo, começando da mais simples versão. Executamos um conjunto de experimentos que avaliam as diversas versões do algoritmo sobre as quatro bases de documentos reais descritas na seção 4.1.

Iniciamos nossa discussão pelo mais simples algoritmo relacional baseado em vizinhança, a ser usado como linha de base para comparação com as demais versões. Posteriormente, apresentamos uma segunda versão que objetiva tratar o problema de desbalanceamento de votos entre as classes através de uma função de ponderação dos votos. A terceira versão apresentada, aborda o desafio de selecionar os relacionamentos mais eficazes para CAD. Por fim, comparamos os resultados obtidos por nosso algoritmo com os apresentados por tradicionais algoritmos de classificação de documentos, bem como realizamos um estudo de efetividade e qualidade do algoritmo proposto. Tal estudo aponta inclusive oportunidades de melhorias do nosso algoritmo a serem exploradas.

5.1 MRS: Modelo Relacional Simples

Nessa seção descrevemos e avaliamos o mais simples algoritmo relacional proposto, denominado MRS – Modelo Relacional Simples. O MRS visa servir como linha de base para comparação com as demais versões apresentadas, como argumentado em Macskassy & Provost [2004]. A principal motivação para a construção deste algoritmo é a distinção entre os ganhos associados à informação adicionada apenas pela rede, dos

ganhos provenientes do uso de técnicas mais elaboradas para análise da rede.

A intuição deste algoritmo consiste, basicamente, em explorar a vizinhança de cada nodo na rede para determinar a pontuação de cada classe. De fato, esta consiste na mais popular estratégia de análise de dados relacionais. A principal idéia deste algoritmo é facilmente definida através do seguinte dito popular: *"me digas com quem andas que direi quem és"*. Assim, a pontuação de cada classe corresponde a soma dos votos dados pelos relacionamentos que estão associados a cada classe. Ao fim, a pontuação de todas as classes são retornada. O algoritmo 1 mostra a implementação do MRS para a classificação de documentos baseada em nossa rede de termos. Para cada termo do documento a ser classificado (i.e., termo de teste), definimos uma vizinhança a ser analisada, a qual consiste do conjunto de termos que co-ocorrem com o termo de teste nos documentos do conjunto de treino. Cada termo de teste determina uma pontuação para cada classe que é observada na sua vizinhança. Essa pontuação pode ser definida de diversas formas, tal como a porcentagem de vizinhos conectados a cada termo em cada classe, ou mesmo o número absoluto de vizinhos observados em cada classe. Há apenas uma classe associada a cada relacionamento, que corresponde à classe Dominante. A classe do documento pode ser predita de diversas maneiras, tal como através de uma votação majoritária simples, considerando a classe com mais alta pontuação para a maioria dos termos de teste, ou a classe associada com a mais alta soma de pontuações de todos os termos de teste.

Algoritmo 1 Modelo Baseado em Análise de Vizinhança

```
function NBAnalysis(G, node)
    finalScore[] ← 0
    repeat
        relation ← GetNextNeighbor(node, G)
        class ← relation.class
        score ← DefineScore(class)
        finalScore[class]+ = score
    until (node.neighborhood = ∅)
     return finalScore
```

A versão do MRS que avaliamos aqui usa toda a vizinhança de cada termo de teste e considera cada relacionamento como um voto (i.e., pondera igualmente todos os relacionamentos com o valor um) para a classe do relacionamento. Como conseqüência, cada termo de teste induz uma pontuação para todas as classes observadas em sua vizinhança, definida como o número absoluto de vizinhos conectados ao termo em cada classe. Ao fim, somamos a pontuação dada a cada classe por todos os termos de teste, e a classe com mais alta pontuação final é assinalada ao documento. É importante salientar que, para a definição da pontuação determinada por cada termo, adotamos o valor absoluto de vizinhos que ocorrem em cada classe, antes a porcentagem de vizi-

nhos por exemplo, com o intuito de valorizar termos mais freqüentes em um domínio. Termos mais freqüentemente utilizados em documentos de uma coleção apresentaram um maior grau na nossa rede de termos. Dessa forma, por apresentar uma vizinhança maior, tais termos tendem a ser mais importantes para o MRS. Essa valorização dos termos mais freqüentes visa diferenciá-los dos termos ditos raros, que provêem menos relacionamentos e conseqüentemente menos informações sobre seus contextos de uso. Embora esses termos mais raros possam até ser discriminativos em alguns cenários, a informação provida por eles são, em geral, menos "confiáveis", motivando-nos a penalizá-los por isso.

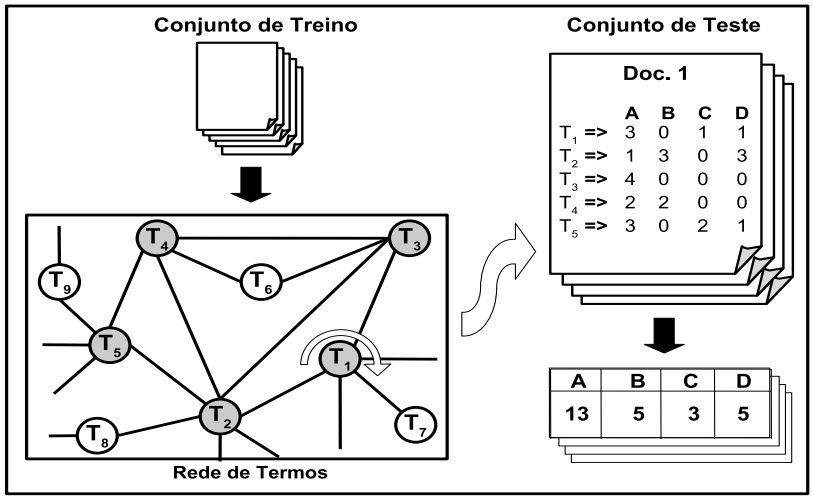


Figura 5.1. Exemplo de Classificação Usando o MRS

A figura 5.1 ilustra o funcionamento do MRS. Inicialmente um conjunto de documentos de treino é utilizado para construir uma grande rede de termos, tal como descrito na seção 2.4. Posteriormente, esta rede é utilizada para classificar cada documento de teste separadamente como segue. Para cada termo de um documento, tal como o *Doc.*1 da figura, cujos termos são apresentados em cinza na rede, contabilizamos o número de relacionamentos que ocorrem em cada classe possível. Finalmente, calculamos a pontuação total de cada classe para o documento, somando-se todos os relacionamentos que ocorrem em cada classe. Repetimos este processo para cada documento a ser classificado. Repare que neste processo não há uma separação clara entre as tradicionais fases de treino e teste, não construindo um modelo de classificação a ser mantido em memória. Além disso, podemos dizer que nosso algoritmo adota uma estratégia postergada [Zhang & Zhou, 2007] de classificação, por classificar cada documento separadamente analisando porções distintas da rede.

Outra importante característica da análise relacional do MRS sobre a rede de termos é a flexibilidade quanto ao contexto de uso de cada termo dos documentos a serem classificados. Diferentemente de algumas estratégias, tal como o *Naïve Bayes* [Salton & McGill, 1986], a rede de termos não assume que um termo pertence a uma classe, na qual ele ocorre mais freqüentemente (i.e., classe Dominante). A classe Dominante está associada ao relacionamento entre dois termos. Com isso, fica evidente a possibilidade de um mesmo termo estar associado a distintos contextos, de acordo com o subconjunto de termos com os quais ele se relaciona.

Aplicando o MRS sobre nossas bases de dados, temos os resultados apresentados na tabela 5.1. Todos os experimentos, tanto deste capítulo quanto dos capítulos seguintes, foram executados usando a estratégia de 10 validações cruzadas (i.e., *10-fold cross-validation*) [Brieman & Spector, 1992], e o resultado final de cada experimento é tido como a média das 10 execuções. As principais métricas utilizadas para avaliar nosso algoritmo, na maioria dos experimentos, foram Acurácia (Ac.) e MacroF1 (MacF1).

Métricas		Coleções			
		ACM	MD	AG	NT
Ac.(%)	MRS	32.84	55.74	30.16	77.41
	MRP	56.00	68.15	52.60	81.75
	ganho	+70.52	+22.24	+74.40	+5.61
	t-t.	▲	▲	▲	▲
$MacF_1$(%)	MRS	17.16	17.31	27.70	21.47
	MRP	37.08	32.16	38.59	40.73
	ganho	+116.1	+85.78	+39.31	+89.70
	t-t.	▲	▲	▲	▲

Tabela 5.1. Resultados do MRS *versus* Resultados do MRP

5.2 MRP: Modelo Relacional Ponderado

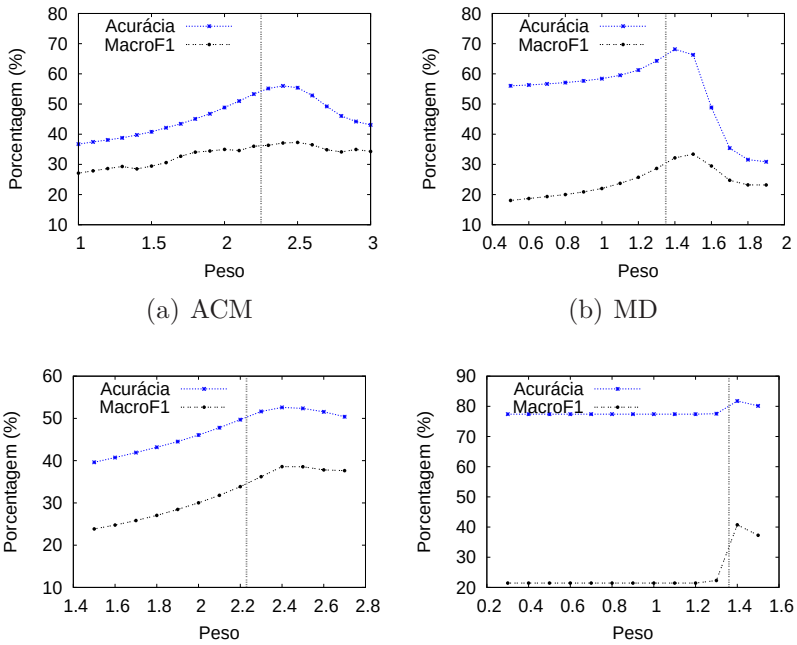
Analisando os resultados do MRS, observamos que a principal razão pelo fraco desempenho apresentado consiste na natural priorização das classes maiores realizada pela algoritmo. Como a ponderação utilizada corresponde a uma função uniforme, classes com um maior número de termos e, conseqüentemente, maior número de relacionamentos, são naturalmente beneficiadas. Como classes maiores usualmente apresentam um número maior de relacionamentos na rede, é mais provável que essas classes recebam mais votos. Uma estratégia para atenuar este problema seria balancear o poder de voto de todas as classes, ponderando os votos de cada classe a fim de tornar o processo de votação mais justo entre as classes.

Uma importante questão relacionada à ponderação dos votos consiste em como reduzir o poder de voto da maior classe mantendo o processo de votação justo. Entende-se por *"justiça"* neste caso, a capacidade do processo de votação atenuar o desbalanceamento entre classes menores e maiores sem desvalorizar em excesso o poder de votação de classes maiores. De maneira geral, essa *"justiça"* estabelece que a importância relativa de cada classe no domínio analisado não deve ser alterada, uma vez que tal alteração implica em distorcer a ordenação natural de representatividade das classes estabelecida em cada domínio.

Podemos definir diversas propostas para tal ponderação. Certamente um importante requisito para propostas de ponderação consiste em uma alta eficiência computacional, usualmente alcançada por estratégias mais simples. Uma das mais simples estratégias seria usar o inverso da probabilidade de ocorrência de cada classe como peso. Assim, quanto maior uma classe, menor seria seu poder de voto. Entretanto, isso poderia super-estimar classes muito pequenas, desobedecendo a restrição de *"justiça"* descrita acima. Por exemplo, na coleção ACM, a probabilidade de ocorrência da menor classe é apenas 0.0024, enquanto a maior classe possui a probabilidade de 0.28. Aplicar a função inverso neste cenário aumentaria os votos relacionados à menor classe por um fator de 416, enquanto que a maior classe teria um pequeno aumento por um fator de 4. Uma outra possibilidade simples seria utilizar a função de complemento da probabilidade de ocorrência. Dessa forma, classes menores teriam um poder de votação maior que classes maiores, e grandes disparidades quanto ao poder de votação seriam eliminadas. No caso da ACM, o complemento para a menor classe seria 0.9976, e para a maior classe seria 0.72. Por outro lado, a diferença entre os valores de peso produzidos pode ser muito pequeno, não sendo suficiente para balancear a votação.

Uma estratégia alternativa seria utilizar a função complemento com um fator C de ajuste. Na prática, o peso de voto w_i, da classe i, passa a ser dado por $w_i = 1 - C \times p_i$, onde p_i é a probabilidade de ocorrência da classe i na coleção de treino. Podemos definir C empiricamente, analisando o desempenho de vários valores para a classificação de um conjunto de teste. Entretanto, tal estratégia pode ser muito cara computacionalmente. Assim, propomos uma heurística para determinar o valor de C. Considerando novamente a restrição de manter a *"justiça"* no processo de votação, podemos definir um limiar teórico para C, que corresponde ao poder de voto da segunda maior classe na base de dados. Se o poder de voto da maior classe se tornar menor que o poder associado à segunda maior classe, o processo de votação deixa de ser justo uma vez que a maior classe deixaria de ser a mais importante na coleção. Assim, uma estratégia seria definir C como um fator que equaliza o poder de voto das duas maiores classes. A figura 5.2 apresenta uma avaliação empírica das porcentagens de Acurácia e

MacroF1 por vários valores de C, para cada uma de nossas bases de dados. Traçamos também em cada gráfico uma linha vertical que representa o fator C definido por nossa heurística. Como podemos observar, a heurística proposta é capaz de encontrar valores de Acurácia e MacroF1 próximos aos melhores resultados encontrados pela busca exaustiva por C, em um dado intervalo, em todas as coleções.



(a) ACM (b) MD

(c) AG (d) NT

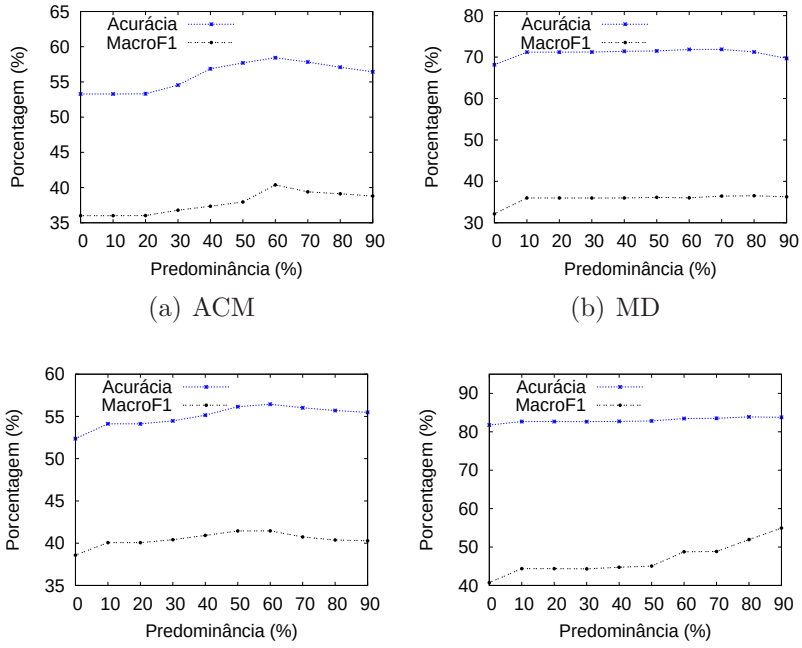
Figura 5.2. Avaliação da Função de Ponderação

O Modelo Relacional Ponderado (MRP) consiste em uma simples extensão do MRS que aplica essa última estratégia de ponderação descrita. Os resultados dessa versão são apresentados, e comparados aos resultados do MRS, na tabela 5.1. Nesta tabela, bem como nas demais discutidas neste trabalho, as linhas com cabeçalho "*ganho*" descrevem a diferença percentual entre os dois algoritmos, as linhas com "*t-t*" descrevem se essa diferença é estatisticamente significante (com ▲ representando uma diferença positiva e ▼ uma diferença negativa) ou não significante (•), dado 99% de confiança em um *2-tailed t-test* [Freund & Simon, 1967]. Percebemos que nossa função de ponderação reduz drasticamente o impacto do desbalanceamento no processo de votação de todas as coleções, especialmente quando consideramos o MacroF1.

5.3 MRP-RV: MRP Com Redução de Vizinhança

Como discutido na seção 4, redes de informação de termos possuem três tipos distintos de relacionamentos, e o uso de relacionamentos semânticos se mostra muito eficaz para

a tarefa de classificação. Entretanto, o uso isolado deste tipo de relacionamento pode resultar em representações restritivas e *ad-hoc* para as classes, tornando o processo de classificação altamente dependente do conjunto de treino. Uma técnica alternativa seria considerar todos os três tipos de relacionamentos, selecionando relacionamentos mais *"intensos"* através da medida de *Predominância*. Ou seja, independente do tipo, é suficiente para a classificação que um relacionamento tenha um alto valor de *Predominância*. Essa estratégia garante a seleção de bons relacionamentos sintáticos e conectivos, provendo uma representação mais ampla de cada classe, aliada ao uso da maioria dos relacionamentos semânticos. Na prática, utilizamos um limiar mínimo de *Predominância* que define quais votos são válidos. Note que o uso da *Predominância* para seleção dos relacionamentos representa para o MRP, que utiliza todos os relacionamentos de cada termo, uma forma de reduzir a vizinhança a ser analisada pelo algoritmo. Aplicando essa estratégia ao algoritmo MRP somos capazes de melhorar os resultados encontrados para todas as coleções, como ilustrado pelos gráficos da figura 5.3.



(a) ACM (b) MD

(c) AG (d) NT

Figura 5.3. Seleção de Relacionamentos usando *Predominância*

Apesar dos ganhos apresentados, podemos ainda identificar o uso de informações paradoxais na análise das redes resultantes. Ou seja, quando consideramos todos os relacionamentos com *Predominância* acima de um limiar, realizamos uma análise global dos relacionamentos de cada termo, os quais podem diferir de comportamentos observados quando consideramos os relacionamentos de um termo com um subconjunto

específico de termos. Dessa forma, consideramos que pode existir uma projeção dos relacionamentos entre termos que melhor define o contexto de fala no qual cada termo ocorre. Identificar tal contexto certamente representa um desafio, uma vez que uma busca exaustiva para cada termo de cada documento a ser classificado é inviável. Analisando nossa definição de rede, notamos que uma promissora heurística de definição desta projeção seria considerar cada documento a ser classificado separadamente. Cada documento representa uma clique, definindo um restrito conjunto de relacionamentos que podem ser mais informativos sobre as características do documento que todos os relacionamentos definidos por cada termo. Assim, cada documento a ser classificado, por definição, gera uma projeção distinta sobre a rede de termos.

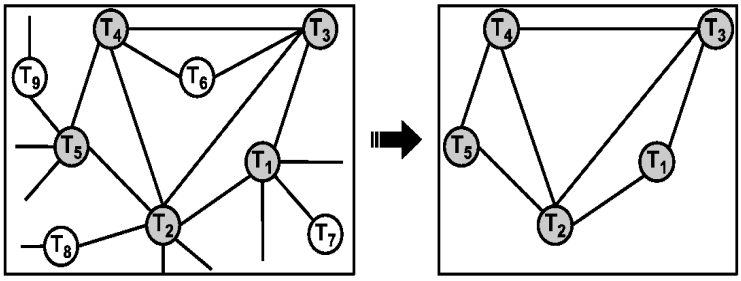


Figura 5.4. Exemplo de Redução de Vizinhança

Baseado nesta observação, definimos uma nova versão do nosso algoritmo, denominado Modelo Relacional Ponderado com Redução de Vizinhança (MRP-RV), onde a vizinhança de cada termo de teste é definido apenas pelos outros termos de teste que co-ocorrem com ele no documento a ser classificado. Ou seja, restringimos a análise para uma "visão local" dos possíveis relacionamentos. A figura 5.4 mostra um exemplo de redução de vizinhança para a rede apresentada no exemplo da figura 5.1. Para cada termo do documento *Doc.*1, ao invés de avaliar todos os seus vizinhos, avaliamos apenas seus relacionamentos com os outros termos que ocorrem neste documento. Tal restrição não apenas reduz o número de relacionamentos a serem avaliados, melhorando o desempenho computacional, mas também permite focar a análise em um conjunto de vizinhos em princípio mais úteis e informativos para a classificação de cada documento específico. Uma vez que a vizinhança é definida, o MRP-RV se comporta exatamente da mesma forma que o MRP a respeito de como classificar um documento. A tabela 5.2 apresenta os resultados de ambos algoritmos considerando a melhor *Predominância* para cada combinação de algoritmo e base de dados, encontrada a partir de uma busca exaustiva sobre um conjunto de validação. É importante enfatizar que os melhores

valores de *Predominância* para o MRP-RV são menores, como conseqüência da ausência de informação que pode resultar da combinação das duas estratégias de seleção de relacionamentos usando altos valores de *Predominância*.

Métricas		Coleções			
		ACM	MD	AG	NT
Ac.(%)	MRP	58.44	71.87	56.45	83.76
	MRP-RV	72.38	82.71	68.27	91.51
	ganho	+23.85	+15.08	+20.93	+9.25
	t-t.	▲	▲	▲	▲
$MacF_1$(%)	MRP	40.38	36.45	41.46	54.94
	MRP-RV	64.22	69.93	59.77	81.28
	ganho	+59.03	+91.85	+44.16	+47.94
	t-t.	▲	▲	▲	▲
Melhor Pred.(%)	MRP	60	70	60	90
	MRP-RV	45	65	55	80

Tabela 5.2. Resultados do MRP *versus* MRP-RV com Predominância

5.4 Análise Experimental

Nesta seção avaliamos a última versão proposta do nosso algoritmo relacional, quanto a qualidade alcançada na classificação, bem como quanto ao custo computacional relacionado à sua execução. Comparamos os resultados obtidos com alguns classificadores tradicionais de RI, de forma a demonstrar sua aplicabilidade em CAD. Além disso, uma análise mais detalhada identificou algumas limitações, e oportunidades de melhorias, relacionadas ao algoritmo, bem como explicações para o competitivo tempo de execução, quanto ao SVM, apresentado.

5.4.1 Análise de Qualidade

Como podemos notar, o algoritmo relacional discutido na seção anterior é muito simples e intuitivo, não necessitando de um amplo e cuidadoso ajuste de valores para diversos parâmetros, a fim de alcançar bons resultados. De fato, o único parâmetro a ser determinado é a *Predominância* mínima, que varia de acordo com características específicas de cada base de dados mas que apresenta um comportamento similar na maioria dos cenários. Isso porque, em geral, valores muito baixos para *Predominância* não são bons para selecionar relacionamentos, uma vez que permitem relacionamentos pouco discriminativos. Por outro lado, valores muito altos não são a melhor opção, dado que podem filtrar importantes relacionamentos. Assim, uma busca pelo melhor valor pode ser desempenhada em um pequeno intervalo de valores possíveis.

Outra importante característica do algoritmo proposto é sua flexibilidade em determinar a classe do documento, a partir das classes preditas para cada termo de teste.

Algoritmo	Coleção							
	ACM		MD		AG		NT	
	$MacF_1$(%)	Ac(%)	$MacF_1$(%)	Ac(%)	$MacF_1$(%)	Ac(%)	$MacF_1$(%)	Ac(%)
Rocchio	56.97	67.95	54.14	69.36	54.14	56.81	72.21	80.80
KNN	56.78	69.8	66.57	79.82	60.85	68.89	74.37	89.13
NB	56.87	74.00	65.64	79.65	61.54	68.31	74.98	84.64
SVM	**60.07**	**73.03**	**72.28**	**83.27**	**65.05**	**72.70**	**83.58**	**92.78**
MRP-RV	**66.06**	**74.37**	**71.49**	**83.35**	**61.62**	**69.01**	**84.02**	**92.24**
ganho	**+9.97**	**+1.83**	**-1.09**	**+0.09**	**-5.27**	**-5.07**	**+0.52**	**-0.58**
t-t.	▲	▲	●	●	▼	▼	●	●

Tabela 5.3. Resultados de Algoritmos para CAD usando informação de *TFxIDF*

Por exemplo, ao invés de somar os votos de cada termo igualmente durante a contagem final dos votos, tal como na implementação corrente, podemos ponderar os votos de cada termo. Essa ponderação pode considerar métricas comumente empregadas em Recuperação de Informação, tal como o $TF{\times}IDF$. Isto é, para cada termo de um documento de teste, definimos uma pontuação para as classes que ocorrem na sua vizinhança, multiplicando o número de relacionamentos em que cada classe é observada na vizinhança pelo valor de $TF{\times}IDF$ do termo.

Os resultados dessa pequena modificação são apresentados na tabela 5.3. Como podemos observar, embora o MRP-RV seja um algoritmo simples, seus resultados são melhores que os encontrados para simples algoritmos *BAG of words*, e até comparáveis aos resultados do algoritmo estado da arte para CAD (i.e., SVM) em nossas coleções. Um desempenho pior que o SVM foi observado no MRP-RV apenas com relação à coleção AG. Acreditamos que tal comportamento está relacionado à composição dos tipos de relacionamentos analisados nesta base. Como mostrado na figura 4.6, para qualquer valor de *Predominância*, a porcentagem de relacionamentos semânticos é muito elevada (acima de 80% dos relacionamentos). Assim, a capacidade dos relacionamentos analisados gerarem uma representação mais ampla das classes é reduzida, tal como discutido no capítulo 4. Com isso, a qualidade da classificação pode degenerar nestes casos. Para suportar nossas afirmações, executamos nossos experimentos com os algoritmos KNN, Rocchio e Naïve Bayes (NB), implementados no *software* de classificação Libbow [McCallum, 1996]. Também utilizamos a implementação SVM-Perf [Joachims, 2006], um pacote que implementa um método SVM linear (técnica *maximum-margin*), que é treinado em tempo linear. Utilizamos a técnica *one-against-all* [Vapnik, 1996], a fim de adaptar o classificador SVM binário para nossas coleções, que possuem mais de duas classes. Para todos os algoritmos usamos o esquema de ponderação através do $TF{\times}IDF$.

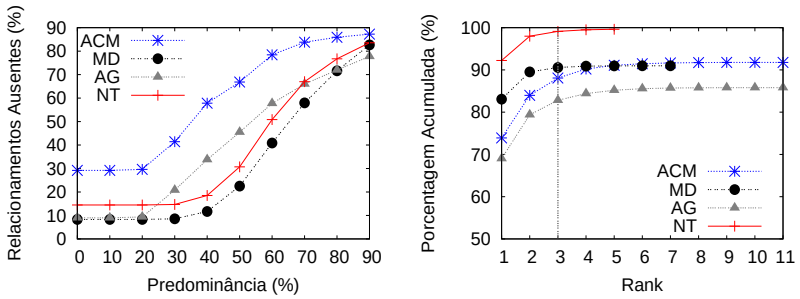
Além disso, uma análise mais detalhada sobre nosso algoritmo mostra que o desenvolvimento de técnicas mais elaboradas pode ser muito promissor. Os resultados

de nossa simples análise de vizinhança estão limitados, sobretudo, por dois fatores. O primeiro deles consiste na confusão inerente à informação presente na rede. Confusão aqui é definida em termos de entropia, isto é, o nível de incerteza ou aleatoriedade associado com o contexto de fala de utilização dos relacionamentos entre termos na rede [Proops, 1987]. Embora o uso de relacionamentos atenue este problema, ele não é completamente resolvido devido à existência de uma confusão natural inerente a restrições da linguagem. Outro aspecto a mencionar é que a evolução da linguagem e do conhecimento podem introduzir ainda mais confusão ao longo do tempo, visto que conceitos usados em certos contextos de fala podem ser empregados em outros, novos termos emergem e outros deixam de ser utilizados. Assim, uma promissora extensão seria tratar esta evolução em nossas redes de informação. Estudos apresentados em Mourão et al. [2008], mostram o impacto dessa evolução no processo de classificação de documentos. Em Rocha et al. [2008], os autores mostraram, inclusive, que é possível melhorar o desempenho da classificação levando em consideração aspectos evolutivos das coleções de documentos.

O segundo fator está relacionado com a quantidade de informação considerada pelo algoritmo. Como impomos algumas filtragens sobre os relacionamentos das redes, o número de relacionamentos resultantes pode ser muito pequeno, gerando escassez de informação. Isso pode ocorrer dado que vários relacionamentos definidos pelos documentos de teste não são observados no conjunto de treino. Esse fenômeno ocorre principalmente em relacionamentos semânticos, uma vez que termos especializados ocorrem menos freqüentemente. A fim de mostrar que este fenômeno pode ocorrer, a figura 5.5 *(a)* mostra a porcentagem de relacionamentos entre termos de teste que não foram encontrados no treino. Podemos perceber que, para altos valores de *Predominância*, a porcentagem de relacionamentos ausentes na rede se torna bastante elevada. Assim, outro importante aspecto a ser estudado é como abordar essa questão de escassez de informação inerente ao nosso algoritmo.

A fim de mostrar os potenciais ganhos em considerar os dois fatores discutidos, realizamos o seguinte experimento. Para cada documento de teste, definimos o *rank* de sua classe verdadeira baseado na pontuação assinalada por nosso algoritmo. O *rank* 1 representa a classe com mais alta pontuação e, portanto, predita por nosso algoritmo para o documento. Posteriormente, calculamos para cada coleção a porcentagem de classes verdadeiras de documentos de teste em cada posição do *rank*. E, finalmente, determinamos a distribuição acumulada dessas porcentagens. Essas distribuições são mostradas na figura 5.5 *(b)*. Como podemos observar, a maioria dos documentos de teste apresentam a classe verdadeira entre as três primeiras posições dos *ranks* mas, não necessariamente na primeira posição. Por exemplo, para a coleção ACM, em mais

de 70% dos casos a classe verdadeira possui *rank* 1, enquanto que em 90% dos casos a classe verdadeira fica entre os 3 primeiros *ranks*, demonstrando que existe uma grande oportunidade para melhorias em nosso algoritmo. Outro importante fator é o valor máximo alcançado pelas distribuições acumuladas. Repare que este valor não é 100% para todas as bases. O valor máximo alcançado representa o máximo de acurácia que o algoritmo pode alcançar para cada base. Como existem documentos de teste cuja classe verdadeira não produz pontuação, devido a não existir relacionamentos entre os termos destes documentos nesta classe, o máximo não é 100%. Para as coleções avaliadas, apenas a AG apresenta um valor máximo abaixo de 90%.



(a) Escassez de Informação (b) Distribuição Acumulada do *Rank*

Figura 5.5. Análise do Algoritmo MRP-RV

5.4.2 Análise de Efetividade

Além de resultados comparáveis ao SVM, e não demandar um complexo ajuste de parâmetros, podemos enfatizar duas importantes características que tornam nosso algoritmo uma alternativa interessante para CAD. Primeiramente, alguns estudos recentes [Cohen & Singer, 1999], mostram a necessidade de algoritmos *model-based* (e.g., SVM, Naïve Bayes) serem inteiramente retreinados periodicamente, à medida que novos documentos surgem, devido a mudanças naturais nas coleções de documentos. O MRP-RV, por outro lado, não necessita ser retreinado, uma vez que trata-se de um algoritmo postergado. A segunda característica consiste na alta eficiência computacional exibida pelo MRP-RV. Embora a rede de termos seja gigantesca, apenas uma pequena parte é utilizada para classificar cada documento. Apenas os termos que ocorrem no documento a ser classificado, e seus vizinhos diretos são analisados, requerendo uma pequena quantidade de memória.

A tabela 5.4 compara o tempo de execução do SVM com o do MRP-RV, onde o cabeçalho *Class.* representa o tempo de classificação, *Treino* representa o tempo de treinamento no SVM, e *CR.* descreve o tempo necessário para construir a rede de termos no MRP-RV. Tais medições de tempo representam médias de cinco execuções

em uma máquina com 2 GHz de CPU e 2 GB de memória RAM. Como podemos observar, o MRP-RV mostrou-se competitivo com o SVM, considerando o tempo de execução, embora apresente uma natureza postergada. Por exemplo, para as coleções ACM e AG os tempos de classificação são 16% e 60% menores que os alcançados pelo SVM, respectivamente. Esse fato é explicado pela técnica *one-against-all* empregada pelo SVM para a classificação multiclasse, devido a ser um classificador inerentemente binário. Mais especificamente, essa estratégia é a técnica mais comum na prática [Manning et al., 2008], e requer a construção de um modelo distinto para cada classe. Para cada modelo construído, a classificação é realizada considerando cada documento de teste. Assim, quanto mais classes temos, maior o tempo de execução. Esse problema é endereçado pelo MRP-RV, uma vez que não existe fase de treinamento, e apenas a classe dominante que ocorre em cada relacionamento é considerada quando assinalamos a classe para cada documento de teste. É importante lembrar também que a classificação multiclasse do SVM requer que cada modelo seja carregado em memória, aumentando o tempo de entrada e saída (i.e., I/O). Um ponto a ser mencionado é que o tempo de execução observado com o MRP-RV para a base NT é o mais alto devido a usarmos todos os termos dos documentos para a construção da rede, antes a utilização apenas dos títulos e dos resumos. Assim, concluímos que definir a rede de termos baseada em todos os termos dos documentos não representa uma estratégia eficiente para o nosso algoritmo.

Tempo de Execução(s)		Coleção			
		ACM	MD	AG	NT
SVM	Treino	134.58	20,567.42	45,630.57	449.30
	Class.	102.51	1,868.38	3,114.49	157.38
MRP-RV	CR.	18.78	697.59	351.42	567.62
	Class.	85.43	2,016.61	1,225.31	2,009.99

Tabela 5.4. Comparação entre os Tempos de Execução

Analisando a complexidade do MRP-RV, podemos entender melhor a eficiência alcançada por nosso algoritmo. Seja T o número de termos distintos que aparecem em média em cada documento de uma coleção. Seja D o número de documentos a serem classificados. Considerando novamente que para o MRP-RV cada documento representa uma clique, cada um dos T termos de um documento possui no máximo $T-1$ vizinhos a serem analisados. Assim, a complexidade do MRP-RV é dada por $O(D \cdot (T^2 - T))$, dado que para cada documento analisamos todos os $T-1$ vizinhos, existentes na rede, dos T termos. Como, para coleções reais[1], $D \gg T$, podemos considerar que nosso algoritmo é linear quanto ao número de documentos a serem classificados. Além da complexidade do algoritmo, é necessário também considerarmos a complexidade de construção da rede

[1]Considerando a Web ou bibliotecas digitais muito grande, com milhões de documentos.

de termos, a qual também é $O(D \cdot (T^2 - T))$. Entretanto, é importante enfatizar que essa etapa de construção é realizada apenas uma vez. À medida que novos documentos chegam, é necessário apenas atualizar alguns relacionamentos já existentes na rede e inserir novos relacionamentos, que aparecem em cada novo documento.

5.5 Sumário

Neste capítulo discutimos em detalhe nosso algoritmo relacional, baseado em vizinhança, para CAD. Objetivando melhor distinguir os diversos passos propostos, apresentamos de forma incremental várias versões do algoritmo. Avaliamos cada uma destas versões através de experimentos sobre as quatro bases de dados reais descritas na seção 4.1. Utilizamos como principais medidas de qualidade para estes experimentos a Acurácia e MacroF1.

A primeira versão apresentada, chamada Modelo Relacional Simples (MRS), consiste em avaliar a vizinhança de cada nodo na rede de termos, a fim de determinar seu comportamento. Essa estratégia é, na verdade, utilizada como linha de base de comparação entre os ganhos provenientes apenas do uso da rede, com os ganhos de utilização de técnicas mais elaboradas para análise da rede. Analisando o desempenho do MRS, observamos que, devido à utilização de uma função uniforme de votação, classes maiores são priorizadas pelo algoritmo. Dessa forma, na segunda versão apresentada, Modelo Relacional Ponderado (MRP), definimos uma função de ponderação dos votos atribuídos a cada classe, tornando o processo de votação mais "justo". Para tanto, utilizamos a função complemento da freqüência de ocorrência das classes, com um fator C de ajuste. O peso de voto w_i, da classe i, passa a ser dado por $w_i = 1 - C \times p_i$, onde p_i é a probabilidade de ocorrência da classe i na coleção de treino. O valor C é definido através de uma heurística que equipara o poder de voto das duas maiores classes de uma coleção. A intuição por trás desta heurística consiste na observação que, para manter o processo de votação "justo", o poder de voto da maior classe pode ser reduzido, no mínimo, ao poder de voto associado à segunda maior classe. Caso contrário a maior classe deixaria de ser a mais importante na coleção. Os resultados do MRP mostram que, de fato, a função de ponderação proposta reduz drasticamente o desbalanceamento de votos entre as classes em todas as bases analisadas.

A terceira versão do algoritmo, MRP com Redução de Vizinhança (MRP-RV), aborda o desafio de selecionar os relacionamentos mais eficazes para a tarefa de classificação. No capítulo anterior, mostramos que selecionar relacionamentos com alta *Predominância* é uma estratégia interessante, não apenas por permitir a utilização da

maioria dos relacionamentos semânticos, mas por capturar relacionamentos sintáticos e conectivos de qualidade. Assim, utilizamos um limiar mínimo de *Predominância* que define quais votos são válidos para análise. Embora esta estratégia seja capaz de melhorar os resultados encontrados para o MRP, percebemos ainda o uso de informações paradoxais na análise das redes resultantes. Ou seja, quando consideramos todos os relacionamentos (com alta Predominância) dos termos, desempenhamos uma análise global de seus relacionamentos, os quais podem diferir de comportamentos observados quando consideramos os relacionamentos com um subconjunto específico de termos (i.e., uma projeção da rede). Analisando nossa definição de rede, notamos que uma heurística promissora de definição desta projeção seria considerar cada documento a ser classificado separadamente. Cada documento representa uma clique, definindo um restrito conjunto de relacionamentos que podem ser mais informativos sobre as características do documento que todos os relacionamentos definidos por cada termo. Utilizando ambas estratégias de seleção de relacionamentos, o MRP-RV foi capaz de melhorar significativamente os resultados do MRP.

Por fim, exploramos a alta flexibilidade do nosso algoritmo em determinar a classe do documento, a partir das classes preditas para cada termo de teste. Ao invés de considerar os votos associados a cada termo de maneira igual, ponderamos os votos de cada termo pelo $TF \times IDF$ do termo no documento de teste. Com esta simples modificação, fomos capazes de obter resultados melhores que os encontrados para simples algoritmos *BAG of words*, e até comparáveis aos resultados do SVM em nossas coleções. Além disso, mostramos empiricamente que há uma grande oportunidade para melhorias em nosso algoritmo. Os resultados de nossa simples e intuitiva análise de vizinhança estão limitados, sobretudo, por dois fatores. O primeiro deles consiste na evolução natural da linguagens e conceitos da comunicação. O uso de conceitos originados em determinados contextos em outros, surgimento de novos termos e desaparecimento de outros ilustram essa evolução. Assim, uma promissora extensão seria tratar a evolução temporal em nossas redes de informação. O segundo fator refere-se à escassez de informação gerada pelas filtragens de relacionamentos imposta pelo algoritmo. Como parte dos relacionamentos presentes nos documentos de teste podem não ocorrer no treino, o número de relacionamentos resultantes para análise pode ser muito pequeno. Dessa forma, outro importante aspecto a ser estudado é como abordar essa questão de escassez de informação.

Além de resultados comparáveis ao SVM, e não demandar um complexo ajuste de parâmetros, enfatizamos duas importantes características que tornam nosso algoritmo uma alternativa interessante para CAD. Primeiramente, trata-se de um algoritmo que adota uma estratégia postergada de classificação, não necessitando ser retreinado à

medida que novos documentos surgem na coleção. De fato, nem há uma fase de treino explícita no algoritmo proposto. Não obstante a sua natureza postergada, como segunda característica, nosso algoritmo ainda mostrou-se competitivo, quanto ao tempo de classificação, com o SVM. Tal eficiência é inclusive explicada pela complexidade linear do nosso algoritmo quanto ao número de documentos a serem classificados.

Capítulo 6

Algoritmos Relacionais Temporais de Classificação

Recentemente a dimensão temporal de análise vem recebendo uma atenção particular não somente em CAD, mas em diversas áreas de estudo. É reconhecida a relevância deste aspecto para a melhoria de métodos de classificação, recomendação e caracterização, dentre outros. Considerando especificamente CAD, sabe-se que modelos relacionais de classificação baseados em todo o histórico da rede podem apresentar um desempenho deteriorado, uma vez que importantes informações sobre mudanças comportamentais da rede são perdidas [Mourão et al., 2009]. Dessa forma, descrevemos neste capítulo a modelagem adotada a fim de capturar tais mudanças temporais na nossa rede de termos. Em seguida, apresentamos e avaliamos três propostas temporais do algoritmo discutido no capítulo 5.

6.1 Dimensão Temporal

Devido à natureza dinâmica de redes reais, relacionamentos usualmente surgem, se modificam, ou simplesmente desaparecem. Assim, um grande desafio na definição de modelos de classificação relacionais consiste em selecionar a granulação temporal mais apropriada a ser considerada para análise. Os modelos de classificação atuais comumente consideram toda a informação disponível para análise de uma rede [Said et al., 2008], agregando as informações presentes ao longo do tempo. Por exemplo, em classificação de documentos, a prática comum é agregar indiscriminadamente todos os documentos disponíveis, independente do seu momento de publicação, para compor o conjunto de treino. Intuitivamente, quanto maior o período observado, mais informações teríamos para definir um bom modelo de classificação. Entretanto, essas infor-

mações podem variar ao longo do tempo, como usualmente é o caso, e a informação agregada resultante pode se tornar imprecisa ou mesmo inconsistente ou contraditória.

Em um estudo recente, quantificamos o impacto de se considerar diferentes escalas temporais para a tarefa de predição em alguns domínios [Mourão et al., 2009; Mourão & Meira Jr., 2008]. Aplicando um modelo de classificação relacional construído em redes definidas sobre diferentes escalas temporais, contrastamos redes constituídas de informações momentâneas com redes que consideram todo o período histórico como uma unidade monolítica. A partir dos experimentos realizados, concluímos que realizar uma agregação histórica de todo o período não é benéfico para a predição dos tipos de relacionamentos entre os objetos nos domínios considerados. Assim, a hipótese que a agregação de informação ao longo do tempo pode ser nociva à tarefa de predição foi confirmada por esses experimentos.

Tal hipótese é igualmente válida para algoritmos tradicionais de classificação de documentos. Estudos apresentados em Mourão et al. [2008], mostram o impacto da evolução da linguagem no processo de classificação de documentos. Como novos termos surgem, outros desaparecem, e alguns passam a ser utilizados em variados contextos, uma "confusão" inerente a este processo degrada o desempenho de classificadores que não consideram tal evolução, como KNN, SVM, Naïve Bayes, dentre outros. Foi verificado, inclusive, a possibilidade de melhorar o desempenho de tais classificadores quando levamos em consideração aspectos evolutivos dos documentos [Rocha et al., 2008].

Dessa forma, a hipótese avaliada neste capítulo é que considerar a evolução da linguagem pode melhorar nosso algoritmo relacional baseado em vizinhança. A fim de verificar tal hipótese, a primeira modificação necessária corresponde à forma como construímos nossa rede de termos. Como descrito na seção 2.4, tal rede é construída considerando todos os documentos presentes em um conjunto de treino, independente do momento de publicação dos mesmos. De forma a capturar mudanças de comportamento dos termos, é necessário definir seus relacionamentos em momentos distintos. Assim, estendemos a definição da rede apresentada previamente para um multigrafo, adicionando informação temporal aos relacionamentos. Ou seja, entre dois termos quaisquer passam a existir M relacionamentos distintos, um para cada momento possível. Cada relacionamento, além de informações sobre a classe dominante e valor de *Predominância* associado a tal classe, conterá informação sobre o momento ao qual se refere. Tal momento representa o momento de publicação do documento onde a co-ocorrência entre dois termos é observada. Nossa proposta é similar a outras existentes na literatura, tais como apresentadas em Liben-Nowell & Kleinberg [2007]; Kossinets & Watts [2006], definidas para outros cenários de estudo.

A figura 6.1 apresenta um multigrafo temporal para a rede de termos das figuras

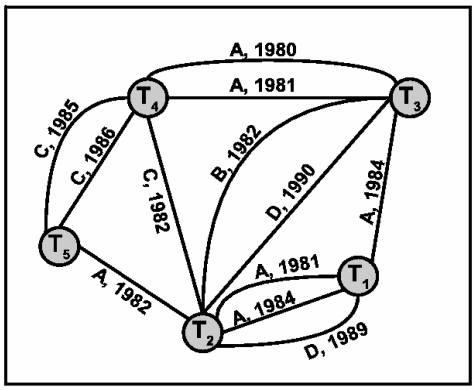


Figura 6.1. Exemplo de Multigrafo Temporal

5.1 e 5.4. Nesta figura são apresentados, para cada relacionamento, apenas a classe dominante e o momento ao qual se refere. Como podemos notar, em momentos distintos é possível haver diferentes classes dominantes nos relacionamentos entre dois termos, como ocorre entre os termos T_2 e T_3. Ou, de forma contrária, pode-se verificar uma mesma classe dominante consistentemente em relacionamentos de diferentes momentos, tal como entre os termos T_3 e T_4.

É importante salientar que os M momentos distintos referem-se à mais comum unidade temporal de criação das entidades em cada domínio de análise. Por exemplo, anos distintos podem ser uma boa escolha para classificação de artigos científicos, visto que eventos científicos são usualmente anuais. Considerando a classificação de notícias, uma unidade diária representaria uma granulação temporal mais apropriada. Dessa forma, para as bases ACM e MD adotamos a unidade anual de discretização temporal, por serem artigos científicos publicados anualmente; para a base AG adotamos a unidade diária, por serem notícias geradas diariamente; e para a base NT utilizamos a granulação semanal, por haver publicações semanais nesta base.

Uma vez definida uma representação que permita capturar o comportamento dos termos em diferentes momentos, o próximo passo consiste em definir a forma como o algoritmo de análise de vizinhança pode utilizar tais informações temporais. Na seção seguinte descrevemos algumas possibilidades de se considerar essas informações.

6.2 Algoritmos Temporais

Podemos considerar diversas maneiras de utilizar as informações temporais contidas nos relacionamentos. De maneira geral, bastariam simples alterações sobre o algoritmo relacional apresentado no algoritmo 1 para isso. Como apresentado pelo algoritmo 2,

basicamente, selecionamos o contexto temporal relevante para análise de cada relacionamento. Este contexto representa o período de tempo no qual os relacionamentos construídos devem ser avaliados. Todos os relacionamentos criados fora deste período são sumariamente desconsiderados. Uma vez selecionado tal contexto, o algoritmo se comporta da mesma forma, identificando os vizinhos existentes e, para cada vizinho, definindo uma pontuação para a classe dominante do relacionamento. Dessa forma, este algoritmo representa um aprimoramento da máxima utilizada por nosso algoritmo de análise de vizinhança, definido como: *"me digas com quem andas,* **quando***, que direi quem és".*

Algoritmo 2 Modelo Temporal-Relacional

```
function TemporalNB(G, node, moment)
    finalScore[] ← 0
    node.neighborhood ← GetNeighborhood(node, G)
    repeat
        relation ← GetNextNeighbor(node, G)
        context ← SelectContext(relation, G, moment)
        class ← relation.class
        score ← DefineScore(class, moment, context)
        finalScore[class]+ = score
    until (node.neighborhood = ∅)
    return finalScore
```

Uma etapa crucial no algoritmo acima consiste na seleção do contexto temporal apropriado. Como discutido em Mourão et al. [2009], esse período deve ser definido de acordo com o momento no qual queremos classificar um dado nodo da rede, uma vez que em diferentes momentos, distintos relacionamentos são mais prováveis de pertencerem a tipos distintos. Chamamos este momento de interesse de **ponto de referência** para a definição do contexto. Uma segunda consideração importante é sobre a duração dos contextos. Longos períodos de tempo podem conduzir a um grande número de tipos de relacionamentos entre dois termos. Por outro lado, períodos muito curtos podem restringir demasiadamente o número de relacionamentos disponíveis para análise, tornando difícil inferir um tipo de padrão. Dessa forma, existe um compromisso entre qualidade e quantidade de informações disponíveis nos relacionamentos selecionados para análise.

Neste trabalho, avaliamos três estratégias distintas para endereçar este compromisso. A mais simples estratégia, a qual denominamos *Análise de Granulação*, consiste em considerar apenas o ponto de referência para análise. Por exemplo, se desejamos classificar um documento científico do ano de 1999, apenas relacionamentos definidos neste ano devem ser considerados. A segunda estratégia, que representa uma extensão da primeira, desconsidera momentos menos relevantes para a classificação de um documento, definindo um período contínuo de tempo, a partir do ponto de referência.

Chamamos esta estratégia de *Análise de janelas temporais*. Considerando o exemplo anterior, utilizar esta estratégia consiste em definir um período entre 1997 a 2003, por exemplo, para classificar um documento de 1999. Finalmente, consideramos a *Análise de ponderação temporal*, em que definimos uma função de ponderação para cada momento baseado em sua relevância estimada para um dado ponto de referência. Frequentemente uma função de decaimento de tempo é utilizada, penalizando momentos mais distantes do ponto de referência considerado. Neste caso, consideraríamos todos os momentos existentes no nosso exemplo, ponderando cada um de forma distinta.

A escolha da estratégia de seleção de contexto interfere diretamente sobre a forma de utilização do contexto para a definição do peso temporal de cada voto, realizada pela função **DefineScore** no algoritmo 2. Cada estratégia estabelece formas distintas de ponderação dos votos. A fim de melhor entender tais formas, bem como o funcionamento de cada uma destas estratégias, as avaliaremos detalhadamente nas subseções seguintes.

6.2.1 Análise de Granulação

Uma maneira simples de considerar a evolução da linguagem para a classificação seria, basicamente, analisar os relacionamentos constituídos dentro da menor granulação temporal de análise. A intuição por trás desta estratégia consiste em isolar as mudanças naturais decorrentes do tempo. Como já discutido, quanto maior o período temporal considerado, mais mudanças podem ocorrer, gerando uma certa "confusão" para os algoritmos de classificação. A menor granulação temporal representa, assim, o período de tempo com menos mudanças, portanto o mais estável e confiável para a tarefa de classificação. Tal granulação é altamente dependente do domínio de estudo, variando por diversas escalas temporais, tais como dia, semana, ano etc.

Dessa forma, nosso primeiro algoritmo temporal, chamado MRP com Redução de Vizinhança Temporal 1 (MRP-RVT1), consiste em utilizar apenas os relacionamentos construídos dentro do ponto de referência temporal, que representa o momento de publicação do documento a ser classificado. É importante observar que, para o nosso algoritmo, tal estratégia de seleção impõe uma função binária de ponderação temporal dos votos, visto que relacionamentos constituídos dentro do período definido pelo ponto de referência possuem o mesmo peso, e todos os demais relacionamentos são desconsiderados (i.e., possuem peso 0). A tabela 6.1 apresenta os resultados do MRP-RVT1, bem como os contrasta com os resultados do MRP-RV. Mais uma vez consideramos informações de $TF \times IDF$ para ambos algoritmos e os melhores valores de *Predominância* para cada base.

Métricas		Coleções			
		ACM	MD	AG	NT
Ac.(%)	MRP-RV	74.37	83.35	69.01	92.24
	MRP-RVT1	60.38	88.22	60.06	78.35
	ganho	-18.81	+5.84	-12.97	-15.15
	t-t.	▼	▲	▼	▼
$MacF_1$(%)	MRP-RV	66.06	71.49	61.62	84.02
	MRP-RVT1	54.46	45.26	53.61	61.05
	ganho	-17.55	-36.42	-12.99	-27.34
	t-t.	▼	▼	▼	▼

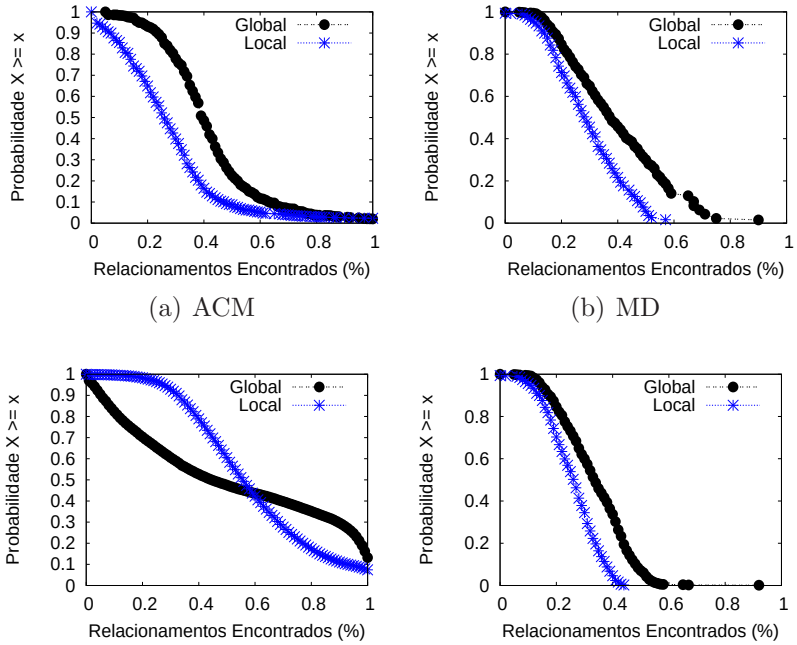
Tabela 6.1. Resultados do MRP-RV *versus* Resultados do MRP-RVT1

Analisando os resultados, observamos uma considerável degradação no desempenho do algoritmo quando utilizamos essa estratégia de seleção temporal. Apenas a Acurácia da base MD apresentou uma melhora mas, por priorizar classes maiores na base e prejudicar as classes menores, justificando o acentuado decaimento de MacroF1 neste caso. Tais resultados decorrem da limitada quantidade de informação utilizada para o processo de classificação, quando consideramos apenas as informações construídas no ponto de referência. Como o MRP-RV realiza novas seleções de relacionamentos, como descrito no capítulo 5, a informação resultante torna-se insuficiente. Este problema, chamado de **Efeito Amostral**, é inclusive abordado em Mourão et al. [2008]. O Efeito Amostral corresponde ao cenário no qual a amostra de dados utilizada para o processo de classificação torna-se insuficiente para capturar as principais características das classes presentes em uma coleção de documentos. Este representa inerentemente um problema estatístico que pode degradar diversos algoritmos de classificação, como KNN ou Naïve Bayes.

De forma a verificar o Efeito Amostral em nossas bases, apresentamos na figura 6.2 a *Complementary Cumulative Distribution Function* ($CCDF$) da porcentagem de relacionamentos, de cada documento a ser classificado, encontrados no treino, quando consideramos todo o histórico da rede (Global) e apenas o ponto de referência (Local). Em todas as bases analisadas, a probabilidade de se encontrar uma porcentagem elevada dos relacionamentos presentes em cada documento a ser classificado é substancialmente menor quando avaliamos apenas o ponto de referência. Dessa forma, a quantidade de informação utilizada para inferir os contextos de uso dos termos avaliados torna-se, de fato, insuficiente para a classificação de diversos documentos.

6.2.2 Análise de Janelas Temporais

Uma extensão natural da primeira estratégia seria definir intervalos de tempo maiores para análise, denominados janelas temporais. Dessa forma, o Efeito Amostral seria amenizado, permitindo alcançar resultados melhores. Entretanto, tais janelas não po-



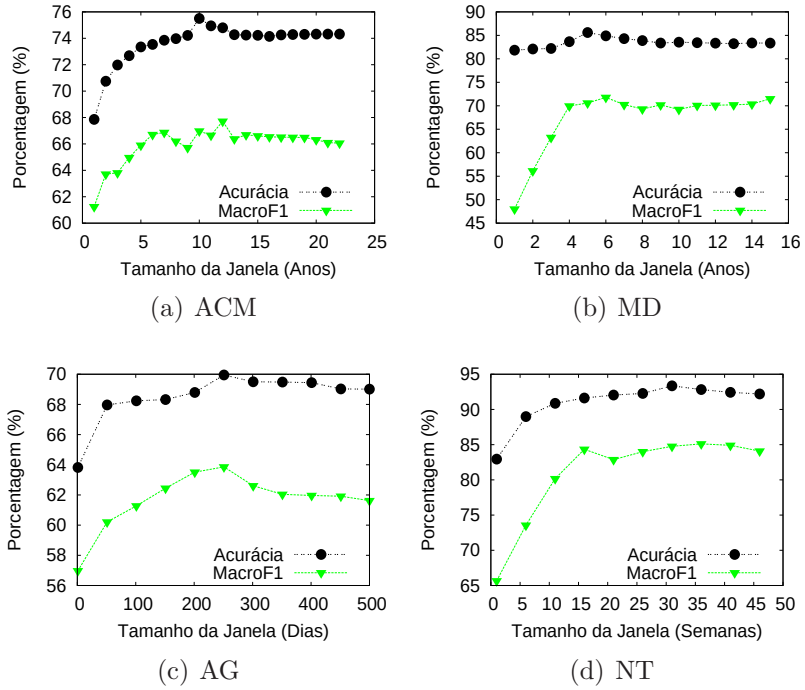
(a) ACM (b) MD

(c) AG (d) NT

Figura 6.2. Avaliação do Efeito Amostral

dem ser muito grandes, uma vez que grandes períodos apresentam maiores variações comportamentais dos relacionamentos, quanto a classe dominante e valores de *Predominância*. De forma a avaliar o desempenho de diversos tamanhos de janelas, a figura 6.3 apresenta resultados de uma análise exaustiva por todos os tamanhos possíveis de janela para cada base. Neste experimento cada tamanho de janela foi utilizado para análise de todos os relacionamentos da rede. Nota-se que os melhores resultados não estão associados às maiores janelas e, tampouco às menores. Ou seja, nossa busca exaustiva mostra que tamanhos intermediários de janelas são os mais apropriados, balanceando o impacto da evolução temporal e o Efeito Amostral.

Cabe salientar que essa busca exaustiva representa, na verdade, uma simplificação do verdadeiro espaço de busca. Como relacionamentos distintos podem possuir tamanhos ideais de janelas diferentes, seria necessário realizar uma busca por todos os tamanhos de janela para cada relacionamento individualmente, permitindo avaliar períodos de tempo distintos para cada relacionamento. Neste cenário, uma busca exaustiva pelo melhor tamanho de janela é computacionalmente inviável em cenários reais. Dessa forma, propomos uma heurística para determinação deste tamanho. Definimos um critério de seleção de janelas temporais para cada relacionamento que consiste em determinar o mais longo período contínuo de tempo, a partir do ponto de referência, em que o nível de "confusão" se mantenha abaixo de um parâmetro C. Tal parâmetro pode ser definido através de diversas métricas. Utilizamos novamente a *Predominância*



(a) ACM (b) MD

(c) AG (d) NT

Figura 6.3. Resultados do MRP-RVT1 por Tamanho de Janela

como parâmetro de seleção temporal, dado que ela mensura o grau de exclusividade dos relacionamentos em cada classe.

De forma a selecionar a janela temporal de análise, procedemos como segue. Para cada relacionamento de cada termo do documento a ser classificado, partimos do ponto de referência e calculamos sua *Predominância*. Em seguida, selecionamos apenas os relacionamentos com *Predominância* acima do limiar C. Para cada relacionamento restante, aumentamos o tamanho da janela em uma unidade temporal para o passado e recalculamos a *Predominância* para o período resultante. Selecionamos os relacionamentos com *Predominância* mínima e aumentamos uma unidade temporal para o futuro (caso possível) e realizamos uma nova filtragem dos relacionamentos. As buscas em cada uma das direções (passado e futuro) são interrompidas na primeira unidade temporal na qual resulta em um valor de *Predominância* abaixo de C. Repetimos este processo até que a janela de todos os relacionamentos seja definida. Chamamos o algoritmo que adota esta estratégia de MRP-RVT2.

É importante observar que o período temporal selecionado pelo MRP-RVT2 é sempre contínuo. Essa característica decorre de dois importantes aspectos. O primeiro consiste no espaço de busca resultante ($O(n^2)$, onde n é o número de momentos distintos), bem inferior ao gerado quando consideramos períodos não contínuos de tempo ($O(2^n)$). O segundo aspecto está associado a uma característica observada na evolução da linguagem [Mourão et al., 2008]. Relacionamentos entre dois termos tendem a

mudar suavemente à medida que aumentamos a distância temporal considerada. Ou seja, usualmente não há mudanças bruscas nos relacionamentos, justificando a adoção de períodos contínuos. Dessa forma, apenas relacionamentos construídos dentro deste período serão considerados para análise, definindo um subconjunto de vizinhos a serem avaliados. Novamente esta representa uma função binária de ponderação dos votos. Neste caso, além do contratempo de se ignorar completamente toda informação exterior às janelas, tal como ocorre na *Análise de Granulação*, há o agravante de todos os relacionamentos internos às janelas terem o mesmo peso. Como dentro de uma janela há relacionamentos de momentos distintos, pode não ser apropriado considerá-los igualmente.

Comparando esta estratégia de seleção temporal com o algoritmo MRP-RV original, temos os resultados apresentados na tabela 6.2. Embora o MRP-RVT2 apresente resultados superiores aos observados para o MRP-RVT1, não são significativamente superiores aos do MRP-RV. Tais resultados devem-se à ponderação dos votos realizada pelo MRP-RVT2, considerando igualmente todos os relacionamentos dentro da janela, e desconsiderando os demais relacionamentos. Isso mostra que, embora distantes temporalmente do ponto de referência, relacionamentos externos às janelas podem conter informações relevantes para a classificação.

Métricas		Coleções			
		ACM	MD	AG	NT
Ac.(%)	MRP-RV	74.37	83.35	69.01	92.24
	MRP-RVT2	75.04	80.95	68.87	91.51
	ganho	+0.90	-2.88	-0.20	-0.79
	t-t.	●	▼	●	●
$MacF_1$(%)	MRP-RV	66.06	71.49	61.62	84.02
	MRP-RVT2	65.49	71.85	61.13	82.07
	ganho	-0.86	+0.50	-0.79	-2.32
	t-t.	●	●	●	▼

Tabela 6.2. Resultados do MRP-RV *versus* Resultados do MRP-RVT2

6.2.3 Análise de Ponderação Temporal

Uma estratégia mais robusta para considerar mudanças temporais nos relacionamentos seria definir uma função de ponderação temporal para os relacionamentos. A idéia consiste em valorizar relacionamentos cujos comportamentos mais se aproximam dos observados no momento de criação do documento que almejamos classificar. Dessa forma, é possível considerar todos os relacionamentos da rede, não descartando nenhuma informação, ao mesmo tempo em que valorizamos relacionamentos potencialmente mais relevantes para cada ponto de referência.

Essa função de ponderação pode levar em consideração aspectos distintos para definir quais relacionamentos são mais relevantes. Por exemplo, assumimos que relacionamentos mais próximos temporalmente ao ponto de referência são mais relevantes por, usualmente, apresentarem comportamentos semelhantes aos observados no ponto de referência, tal como mostrado em Mourão et al. [2008]. Assim, todos os relacionamentos de cada momento M_i são ponderados através do peso p_i, usualmente definidos através de uma função de decaimento temporal, tal como segue:

$$p_i \propto {D_{ir}}^{-a} \tag{6.1}$$

onde D_{ir} representa a distância temporal, medida em unidades temporais definidas para o domínio, entre o momento M_i e o ponto de referência r, e a representa o fator de decaimento da função. Essa função penaliza relacionamentos temporalmente mais distantes do ponto de referência. O desafio neste caso, consiste em definir o valor de a mais apropriado para cada domínio. Assim como um fator de decaimento pequeno pode não melhorar a classificação, um fator muito alto pode inclusive degradar o desempenho encontrado, por penalizar demasiadamente informações temporalmente mais distantes. A fim de definir tal fator de decaimento para nossas bases, para cada distância temporal possível calculamos a porcentagem de relacionamentos observados cuja classe predominante é igual à classe predominante no ponto de referência. Como mostrado na figura 6.4, embora haja um comportamento decrescente em todas as bases, os decaimentos observados são distintos. A base ACM, por exemplo apresenta um fator de decaimento maior, mostrando-se como uma coleção mais dinâmica. Já as bases MD e NT apresentam os menores fatores, apresentando uma dinamicidade mais baixa.

Utilizando tais fatores definidos para cada base, ponderamos os relacionamentos de cada distância temporal de forma a analisar todos os relacionamentos na classificação. Denominamos este algoritmo de MRP-RVT3, e apresentamos seus resultados na tabela 6.3. Comparando tais resultados com os observados para o MRP-RV, observamos uma melhora para duas bases e resultados estatisticamente equivalentes para as outras duas. Apesar destes resultados suportarem a hipótese inicial que abordar a evolução da linguagem poderia melhorar nosso algoritmo relacional, os ganhos obtidos ainda se mostraram tímidos, evidenciando a necessidade de formas mais elaboradas de tratar a evolução temporal.

Embora a função de ponderação apresentada seja simples e intuitiva para incorporação de informações temporais, sabemos que nem sempre corresponde à melhor forma de abordar a evolução temporal. Por exemplo, há cenários de comportamentos sazonais, periodicamente observados em determinados domínios. Este é o caso, por

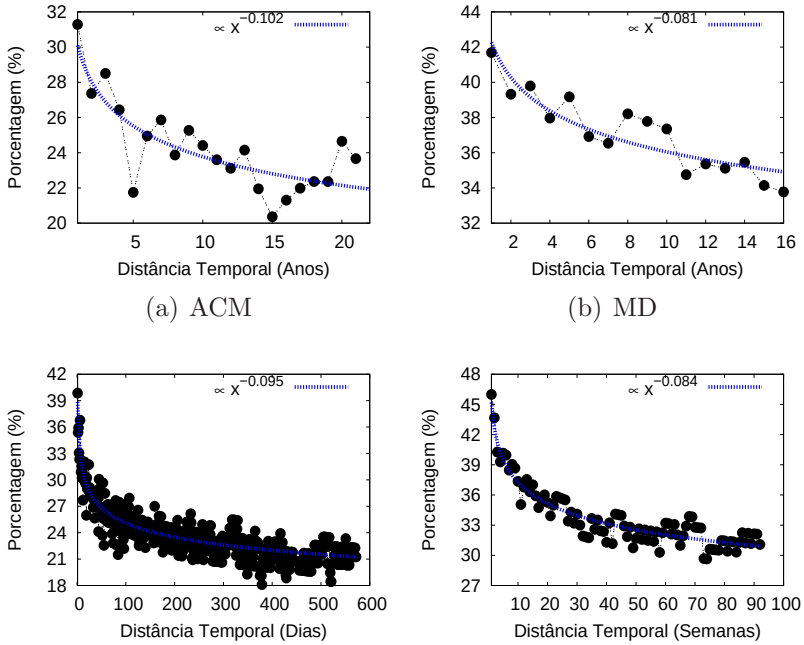


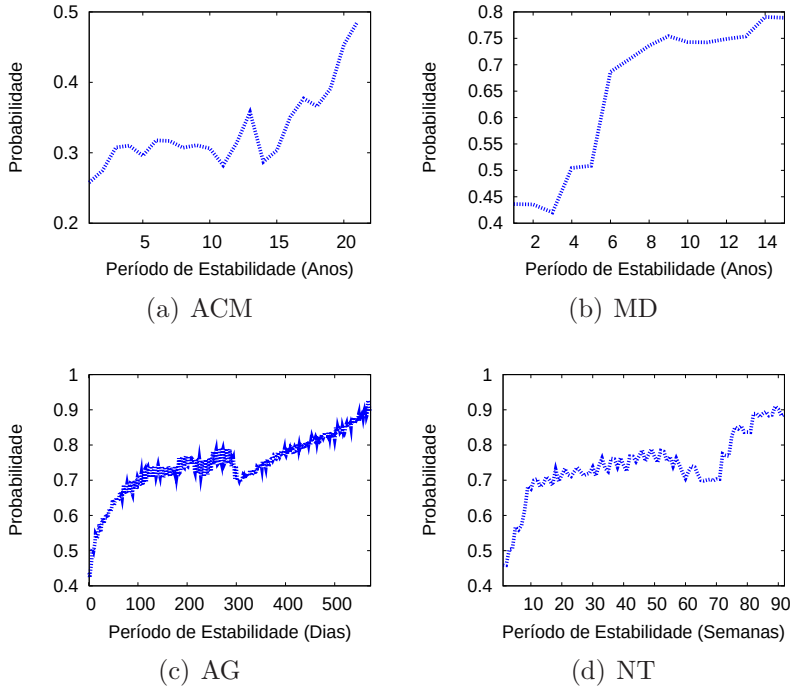
Figura 6.4. Análise de Fator de Decaimento Temporal

Métricas		Coleções			
		ACM	MD	AG	NT
Ac.(%)	MRP-RV	74.37	83.35	69.01	92.24
	MRP-RVT3	75.94	84.89	70.66	92.86
	ganho	+2.11	+1.85	+3.84	+0.67
	t-t.	▲	●	▲	●
$MacF_1$(%)	MRP-RV	66.06	71.49	61.62	84.02
	MRP-RVT3	67.87	66.89	62.98	85.15
	ganho	+2.73	-6.43	+2.21	+1.34
	t-t.	▲	▼	▲	●

Tabela 6.3. Resultados do MRP-RV *versus* Resultados do MRP-RVT3

exemplo, de documentos relacionados a Olimpíadas. Neste caso, distâncias temporais maiores não necessariamente representam relacionamentos menos relevantes para análise. Além disso, a simples penalização de relacionamentos mais distantes do ponto de referência descaracteriza importantes informações providas por relacionamentos **"estáveis"**. Um relacionamento é dito estável por um período X se sua classe dominante permanece inalterada em X. De fato, quanto maior o período de estabilidade dos relacionamentos, mais informativos, e úteis para classificação, se tornam tais relacionamentos. De forma a mostrar isso, traçamos na figura 6.5 a probabilidade da classe dominante dos relacionamentos na distância temporal $D + 1$ ser igual à classe dominante no ponto de referência, pelo período de estabilidade D. Percebemos que, quanto maior o período de estabilidade, mais provável do relacionamento continuar a possuir a

mesma classe dominante. Assim, este tipo de relacionamento a longas distâncias temporais deveria ser beneficiado, antes a penalizado. Uma caracterização detalhada sobre a evolução da rede de termos é de grande importância para um maior entendimento de cenários como estes, permitindo um tratamento mais adequado da evolução temporal em modelos relacionais.



(a) ACM (b) MD

(c) AG (d) NT

Figura 6.5. Análise de Estabilidade dos Relacionamentos

6.3 Sumário

Devido à natureza dinâmica de redes reais, relacionamentos usualmente surgem, se modificam, ou simplesmente desaparecem. Dessa forma, modelos relacionais de classificação baseados em todo o histórico da rede podem apresentar um desempenho deteriorado, uma vez que importantes informações sobre mudanças comportamentais da rede são perdidas [Mourão et al., 2009]. Neste capítulo, avaliamos a hipótese de considerar a evolução da linguagem pode melhorar nosso algoritmo relacional.

A fim de verificar tal hipótese, primeiramente definimos uma representação que permita capturar o comportamento dos termos em momentos distintos. Tal representação estende a definição da rede apresentada previamente para um multigrafo, adicionando informação temporal aos relacionamentos. Ou seja, entre dois termos quaisquer passam a existir M relacionamentos distintos, um para cada momento possível. Es-

tes M momentos distintos referem-se à mais comum unidade temporal de criação das entidades em cada domínio de análise.

Uma vez definida essa representação, avaliamos algumas formas de como o nosso algoritmo relacional pode incorporar tais informações temporais. De maneira geral, modificamos o tradicional algoritmo de análise de vizinhança, selecionando para cada relacionamento um contexto temporal para análise, que representa o período de tempo no qual os relacionamentos construídos devem ser avaliados. Todos os relacionamentos criados fora deste período são desconsiderados. Uma vez selecionado tal contexto, o algoritmo se comporta da mesma forma. Assim, este algoritmo representa um aprimoramento da máxima utilizada por nosso algoritmo de análise de vizinhança, definido como: *"me digas com quem andas,* **quando***, que direi quem és".*

Uma etapa crucial neste algoritmo temporal consiste na seleção do contexto temporal apropriado. Como discutido em Mourão et al. [2009], esse período deve ser definido de acordo com o momento no qual queremos classificar um dado nodo da rede. Uma segunda consideração é sobre a duração dos contextos. Longos períodos de tempo podem conduzir a um grande número de tipos de relacionamentos entre dois termos. Por outro lado, períodos muito curtos podem restringir demasiadamente o número de relacionamentos disponíveis para análise. Dessa forma, há um compromisso entre qualidade e quantidade de informações disponíveis nos relacionamentos selecionados para análise.

Baseado nesta observação, avaliamos três estratégias distintas para endereçar este compromisso. A mais simples estratégia, a qual denominamos *Análise de Granulação*, consiste em considerar apenas o momento que o documento a ser classificado foi publicado (i.e., o ponto de referência). Analisando os resultados desta estratégia, observamos uma considerável degradação no desempenho do algoritmo. Tais resultados decorrem da limitada quantidade de informação utilizada para o processo de classificação. Ou seja, quando consideramos apenas as informações construídas no período definido pelo ponto de referência, um número restrito de relacionamentos é utilizado.

A segunda estratégia, que representa uma extensão da primeira, desconsidera momentos menos relevantes para a classificação de um documento, definindo um período contínuo de tempo, a partir do ponto de referência. Chamamos esta estratégia de *Análise de janelas temporais*. Definimos um critério de seleção de janelas temporais para cada relacionamento que consiste em determinar o mais longo período contínuo de tempo, usando o ponto de referência como base, em que o nível de "confusão" se mantenha abaixo de um valor mínimo C de *Predominância*. Embora esta estratégia apresente resultados superiores aos observados para a primeira, não são significativamente superiores aos do MRP-RV. Tais resultados devem-se à ponderação dos votos realizada,

considerando igualmente todos os relacionamentos dentro da janela, e desconsiderando os demais relacionamentos. Isso mostra que, embora distantes temporalmente do ponto de referência, relacionamentos externos às janelas podem conter informações relevantes para a classificação.

Finalmente, consideramos a *Análise de ponderação temporal*, em que definimos uma função de ponderação para cada momento baseado em sua relevância estimada para um dado ponto de referência. Frequentemente uma função de decaimento de tempo é utilizada, penalizando momentos mais distantes do ponto de referência considerado. Neste caso, consideramos todos os momentos existentes, ponderando cada um de forma distinta. Utilizando esta estratégia, conseguimos verificar a hipótese que considerar a evolução temporal pode beneficiar nosso algoritmo relacional. Embora esta represente uma definição simples e intuitiva para incorporação de informações temporais, sabemos que nem sempre corresponde à melhor definição. Assim, identificar funções capazes de endereçar cenários de comportamentos sazonais, por exemplo, representa uma proeminente direção de estudo.

Capítulo 7

Extensões de Algoritmos Relacionais de Classificação

Além da dimensão temporal de análise, outras dimensões podem ser avaliadas a fim de prover melhorias ao algoritmo relacional proposto. Neste capítulo, avaliamos especificamente extensões do nosso algoritmo que abordam o problema de escassez de informação e o uso de atributos dos relacionamentos em CAD. Iniciando nossa discussão pelo problema de escassez de informação, discutido no capítulo 5, através de uma abordagem simples para este problema, baseada na predição de classe dos relacionamentos, fomos capazes de melhorar o desempenho do nosso algoritmo. Em seguida, discutimos brevemente algumas extensões capazes de incorporar informações de atributos contidos em cada relacionamento da nossa rede de termos.

7.1 Predição de Classe dos Relacionamentos

Naturalmente, diversos termos presentes em novos documentos a serem classificados não ocorrem em um conjunto de treino. Principalmente termos especializados, que usualmente apresentam uma baixa freqüência de ocorrência, possuem baixas probabilidades de serem observados em um conjunto de documentos. Assim, a quantidade de informação disponível na rede de termos para o nosso algoritmo é algumas vezes insuficiente para identificar padrões comportamentais relevantes dos termos. Não obstante a essa constatação, cabe lembrar que o nosso algoritmo realiza sobre os relacionamentos presentes na rede diversas filtragens, a fim de utilizar apenas relacionamentos potencialmente mais úteis e informativos de cada termo de um documento a ser classificado. Dessa forma, uma quantidade ainda menor de relacionamentos é avaliada, configurando um cenário de escassez de informação. O gráfico 5.5 *(a)*, discutido no capítulo

5, demonstra este cenário para as quatro bases estudadas.

Uma das maneiras mais simples de abordar este problema seria tentar inferir a classe dominante de cada relacionamento ausente, no conjunto de treino, do documento a ser classificado. Ou seja, a cada relacionamento não encontrado na rede de termos, construída a partir dos documentos de treino, atribuímos a classe mais provável deste relacionamento ocorrer para a coleção observada. Note que o problema de inferir a classe de um relacionamento é similar ao problema de *Link Prediction*, amplamente estudado em Redes Complexas [Taskar et al., 2003]. Mais formalmente, *Link Prediction* é definido como o problema de identificar relacionamentos não observados na rede, a partir de informações sobre objetos e relacionamentos presentes na rede. Diferentemente, o que objetivamos é identificar a classe, ou tipo, dos relacionamentos não observados na rede.

A identificação dessa classe mais provável pode ser realizada de diversas formas. Adotamos uma forma simples baseada no comportamento relacional de cada termo, definida como segue. Seja T_D o conjunto de termos que ocorrem no documento D a ser classificado. Considere também G como a rede de termos construída sobre um conjunto de treino. A cada relacionamento ausente em G entre os termos t_i e t_j, presentes em T_D, tentamos atribuir uma classe como apresentado no algoritmo 3. A idéia do algoritmo consiste em identificar a classe dominante mais freqüentemente observada, em G, nos relacionamentos entre cada termo t_i e todos os demais termos que ocorrem em D (i.e., todos os demais termos pertencentes a T_D), tal como realizado pela função **getMostCommonRelation** no algoritmo. Uma vez identificada tal classe para ambos termos de cada relacionamento, verificamos se as classes encontradas são iguais. Caso sejam, esta será a classe de saída do algoritmo de predição, caso contrário, geramos como saída a classe com freqüência de ocorrência mais elevada. Caso um dos termos não esteja presente em G, retornamos a classe encontrada para o outro termo. Note que, caso nenhum dos dois termos se encontrem na rede, o algoritmo retorna um valor indefinido, não realizando uma predição. Tal comportamento deve-se à inexistência dos termos na rede implicar em ausência de informação sobre os termos, logo consideramos que seu comportamento, quanto ao contexto de aparição, não deve ser predito.

Apresentamos na figura 7.1 a distribuição de relacionamentos ausentes pelos três casos considerados pelo algoritmo de predição: ambos termos são observados em G ($Ambos$), apenas um dos termos (Um), e nenhum dos dois termos ($Nenhum$) é observado. Como podemos observar, a porcentagem de relacionamentos ausentes que não possuem nenhum dos termos presentes na rede G é baixa em todas as bases. Assim, poucos relacionamentos ausentes são desconsiderados para a classificação, visto que não é feita uma predição para uma pequena porcentagem apenas.

Algoritmo 3 Algoritmo de Predição de Tipos de Relacionamentos

```
function predictRelationClass(t_i, t_j, T_D, G)
    predictedClass ← undef
    c_i ← undef
    c_j ← undef
    if t_i ∈ G then
        c_i ← getMostCommonRelation(t_i, T_D, G)
    if t_j ∈ G then
        c_j ← getMostCommonRelation(t_j, T_D, G)
    if c_i ≠ undef then
        if c_j ≠ undef then
            if c_i == c_j then
                predictedClass ← c_i
        else
            if freq(c_i) > freq(c_j) then
                predictedClass ← c_i
            else
                predictedClass ← c_j
    else
        if c_j ≠ undef then
            predictedClass ← c_j
     return predictedClass
```

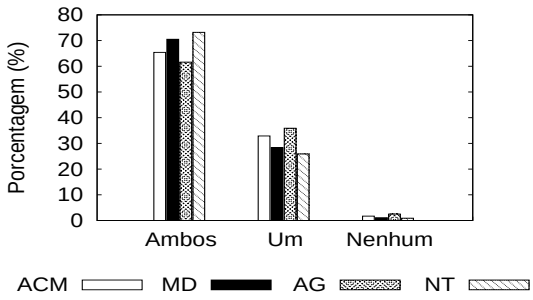


Figura 7.1. Distribuição de Relacionamentos Ausentes

Uma vez preditas as classes dos relacionamentos ausentes, incorporamos tais relacionamentos à rede G, gerando uma rede G'. Em seguida, executamos o algoritmo MRP-RV utilizando G' como entrada. Chamamos esta estratégia de MRP-LP e apresentamos na tabela 7.1 seus resultados, bem como os comparamos com os resultados do MRP-RV executando com a rede G. Observamos, em geral, uma melhoria, principalmente quanto à MacroF1 em todas as coleções. Isso pode ser explicado pelo fato de justamente as classes menores tenderem a apresentar mais relacionamentos ausentes na rede de termos. Tais resultados mostram que técnicas mais elaboradas para abordar o problema de escassez podem ser promissoras. Técnicas capazes de endereçar este problema juntamente com o problema de evolução temporal representam uma direção interessante, porém desafiadora. Embora considerar todo o período não seja benéfico para a predição dos tipos de relacionamentos [Mourão et al., 2009], uma predição considerando apenas momentos distintos pode não ser apropriada. Como relacionamentos ausentes ocorrem, em geral, entre termos que co-ocorrem pouco freqüentemente em um

conjunto de documentos, restringir a análise para um período específico pode tornar a busca por tais termos ainda mais difícil, dada a freqüência ainda menor de aparição dos mesmos em períodos isolados.

Métricas		Coleções			
		ACM	MD	AG	NT
Ac.(%)	MRP-RV	74.37	83.35	69.01	92.24
	MRP-LP	75.95	84.11	70.53	92.89
	ganho	+2.12	+0.91	+2.22	+0.70
	t-t.	▲	•	▲	•
$MacF_1$(%)	MRP-RV	66.06	71.49	61.62	84.02
	MRP-LP	67.89	72.24	62.74	85.86
	ganho	+2.77	+1.05	+1.82	+2.19
	t-t.	▲	•	•	▲

Tabela 7.1. Resultados do MRP-RV *versus* Resultados do MRP-LP

7.2 Intensidade de Relacionamentos

Uma das mais importantes características de redes complexas é sua imensa flexibilidade de modelagem, permitindo capturar facilmente importantes aspectos de cada domínio analisado. Por exemplo, podemos assinalar a cada relacionamento uma medida de intensidade que quantifica a importância do relacionamento, sob algum aspecto de interesse pertinente ao domínio de estudo, para a tarefa de classificação. Relacionamentos mais freqüentes entre dois objetos ilustram este tipo de medida de importância em diversos cenários. Tais medidas, inclusive, são recorrentemente utilizadas em diversos estudos. Em Bernstein et al. [2003] é definido, por exemplo, o conceito de "*strength*"de um relacionamento, que mensura a importância, ou o peso, de cada relacionamento na construção do espaço vetorial relacional. Já em Taskar et al. [2003] a existência de tipos diferentes de relacionamentos é considerada para melhorar modelos que tratam o problema de *Link Prediction*.

A análise de medidas de intensidade é ainda mais enfatizada devido ao sucesso de sua aplicação em outras áreas de pesquisa. Em Recuperação de Informação, por exemplo, o tradicional modelo de ponderação dos termos em documentos $TF \times IDF$ [Salton & Buckley, 1988] não é nada mais que uma medida de intensidade dos termos, em que o TF expressa a freqüência de relacionamento de um termo em um documento e o IDF expressa o grau de exclusividade deste termo no documento (i.e., termos que aparecem em um único documento são totalmente exclusivos, possuindo assim uma intensidade alta e um IDF igual a 1). Já em Mineração de Dados, os con-

ceitos de suporte e confiança aplicados, por exemplo, a regras de associação, ilustram a utilização de medidas de intensidade nesta área de estudo.

Dessa forma, a hipótese levantada nesta seção é que considerar a intensidade dos relacionamentos observados na rede de termos pode beneficiar nosso algoritmo relacional de CAD. Especificamente, avaliamos tal hipótese sobre duas medidas em particular: a freqüência de co-ocorrência dos termos na rede e o valor de *Predominância* da classe dominante de cada relacionamento.

7.2.1 Freqüência de Co-ocorrência

Freqüência é uma das medidas mais populares em estudos de redes complexas. Por exemplo, a freqüência com que os dois pesquisadores publicaram juntos em uma rede de co-autoria pode ajudar a identificar a área de pesquisa mais correlacionada a ambos. A utilização desta medida é baseada na premissa de que quanto mais freqüentes são os relacionamentos, menor a probabilidade do comportamental relacional observado para cada termo representar uma exceção ao comportamento usual dos termos. Considerando nossa rede de termos, quanto maior o número de documentos em que dois termos co-ocorrem mais relevante se torna o relacionamento.

De forma a considerar a informação de freqüência de co-ocorrência no nosso algoritmo relacional temporal, atribuímos a cada relacionamento do multigrafo temporal construído o número de documentos em que os dois termos co-ocorreram em cada momento distinto. Posteriormente, durante o processo de pontuação das classes presentes na vizinhança de cada termo, ao invés de considerar cada relacionamento como um voto de igual valor, ponderamos o voto de cada relacionamento por sua freqüência de co-ocorrência. Com isso, relacionamentos considerados exceções, por ocorrerem pouco freqüentemente, apresentam um peso de voto menor. Aplicando esta modificação ao algoritmo MRP-RVT3, definimos o algoritmo MRP-FC, cujos resultados são apresentados na tabela 7.2. Como podemos perceber, a simples utilização da freqüência de co-ocorrência mostrou-se útil para nosso algoritmo, permitindo alcançar alguns ganhos em relação ao MRP-RVT3 original em todas as coleções.

7.2.2 Valor de Predominância

Como visto, a *Predominância* é uma medida adotada em diversas etapas deste trabalho. A utilizamos para definir a classe dominante de cada relacionamento, utilizamos novamente para selecionar os relacionamentos mais relevantes para a classificação, e posteriormente adotamos a *Predominância* para análise de algumas estratégias de sele-

Métricas		Coleções			
		ACM	MD	AG	NT
Ac.(%)	MRP-RVT3	75.94	84.89	70.66	92.86
	MRP-FC	76.87	85.11	72.08	93.94
	ganho	+1.22	+0.26	+2.01	+1.16
	t-t.	•	•	▲	•
$MacF_1$(%)	MRP-RVT3	67.87	66.89	62.98	85.15
	MRP-FC	68.72	68.33	64.42	86.07
	ganho	+1.25	+2.15	+2.29	+1.08
	t-t.	•	▲	▲	•

Tabela 7.2. Avaliação do uso da Freqüência de Ocorrência dos Relacionamentos

ção temporal de relacionamentos. Entretanto, a utilização desta medida foi sempre com o intuito de selecionar relacionamentos a partir de um valor mínimo de interesse (i.e., um limiar). Ou seja, todos os relacionamentos com *Predominância* mínima acima de um valor C são considerados igualmente, e os demais relacionamentos são descartados.

Uma extensão natural desta estratégia seria considerar a *Predominância* como uma medida de importância de cada relacionamento, representando assim uma ponderação do poder de voto de cada relacionamento. O intuito é que quanto maior a *Predominância* da classe dominante de um relacionamento, mais relevante para a classificação é tal relacionamento. Dessa forma, durante o processo de pontuação das classes presentes na vizinhança de cada termo, além de selecionar os relacionamentos com *Predominância* mínima acima de C, utilizamos a *Predominância* para ponderar o voto de cada relacionamento analisado. Chamamos esta versão do MRP-RVT3, considerando a ponderação descrita, de MRP-PR. A avaliação dessa simples ponderação dos relacionamentos é mostrada na tabela 7.3, e contrastada com os resultados do MRP-RVT3. Observamos que considerar a *Predominância* como uma medida de intensidade também beneficia nosso algoritmo relacional, principalmente quanto à MacroF1. Este comportamento deve-se, sobretudo, a tal ponderação permitir que relacionamentos, com maiores *Predominâncias*, presentes em classes menores, embora menos numerosos, sejam mais importantes para o processo de pontuação das classes.

De maneira geral, as simples medidas de intensidade apresentadas e avaliadas nesta seção ilustram a facilidade de se estender algoritmos relacionais, permitindo incorporar aspectos importantes à classificação. A definição de outras medidas de intensidade, tal como a distância de co-ocorrência dos termos em cada documento, representa uma promissora direção de análise. Além disso, a definição de formas mais elaboradas de se incorporar cada medida de intensidade ao algoritmo, e mesmo formas de se combinar distintas medidas podem ser relevantes.

Métricas		Coleções			
		ACM	MD	AG	NT
Ac.(%)	MRP-RVT3	75.94	84.89	70.66	92.86
	MRP-PR	76.53	84.98	71.91	93.51
	ganho	+0.78	+0.11	+1.77	+0.70
	t-t.	●	●	●	●
$MacF_1$(%)	MRP-RVT3	67.87	66.89	62.98	85.15
	MRP-PR	69.11	67.93	64.26	86.87
	ganho	+1.83	+1.56	+2.04	+2.02
	t-t.	●	●	▲	▲

Tabela 7.3. Avaliação do uso da *Predominância* dos Relacionamentos

7.3 Sumário

Como discutido no capítulo 5, um das limitações do algoritmo proposto é a escassez de informação resultante, dentre outras coisas, das várias filtragens realizadas. Assim, a quantidade de informação disponível na rede de termos para o nosso algoritmo é algumas vezes insuficiente para identificar padrões comportamentais relevantes dos termos. A fim de tratar este problema, neste capítulo, tentamos inferir a classe dominante de cada relacionamento ausente, no conjunto de treino, do documento a ser classificado. Este problema de inferir a classe mais provável de um relacionamento é similar ao problema de *Link Prediction*, definido como o problema de identificar relacionamentos não observados na rede. Para a identificação dessa classe mais provável propomos um algoritmo baseado no comportamento relacional de cada termo. A idéia do algoritmo consiste em identificar a classe dominante mais freqüentemente observada nos relacionamentos entre os termos do documento a ser classificado. A partir dos resultados apresentados por esta simples estratégia, observamos, em geral, uma melhoria, principalmente quanto à MacroF1 em todas as coleções. Isso pode ser explicado pelo fato de justamente as classes menores tenderem a apresentar mais relacionamentos ausentes. Tais resultados mostram que técnicas mais elaboradas para abordar o problema de escassez podem ser promissoras.

Neste capítulo também discutimos brevemente algumas extensões, aplicáveis ao algoritmo relacional proposto, capazes de incorporar informações de atributos contidos em cada relacionamento da nossa rede de termos. Devido à popularidade de tais informações, denominadas aqui de medidas de intensidade dos relacionamentos, avaliamos o impacto de informações que quantificam a importância dos relacionamentos, sob algum aspecto de interesse pertinente ao domínio de estudo, para a tarefa de classificação. Especificamente, avaliamos este impacto considerando duas medidas em particular: a freqüência de co-ocorrência dos termos na rede e o valor de *Predominância* da classe dominante de cada relacionamento. A simples utilização de ambas medidas como forma

de ponderação dos votos de cada relacionamento, para pontuação das classes presentes na vizinhança de cada termo, foram capazes de produzir alguns ganhos sobre nosso algoritmo relacional. Dessa forma, as medidas de intensidade apresentadas e avaliadas ilustram a facilidade de se estender algoritmos relacionais, permitindo incorporar aspectos importantes de cada domínio à classificação.

Capítulo 8

Conclusões e Trabalhos Futuros

Neste capítulo apresentamos um sumário dos principais resultados alcançados nesta dissertação. Além disso, baseado nestes resultados, apresentamos os potenciais e as limitações de uso do nosso algoritmo relacional de classificação, baseado em vizinhança, em CAD. Por fim, discutimos alguns trabalhos futuros pertinentes.

8.1 Conclusões

Avaliamos neste trabalho a utilização de uma rede de termos para representar documentos textuais na tarefa de Classificação Automática de Documentos (CAD). Através de uma discussão apoiada em conceitos lingüísticos, definimos a intuição por trás desta rede, bem como demonstramos sua relevância para CAD. Uma análise detalhada sobre os termos e suas interações, naturalmente observadas na linguagem humana, mostrou que a utilização dos relacionamentos entre termos é bastante útil para identificar o contexto de uso de cada termo. Assim, antes a simples ocorrência de termos, a co-ocorrência entre termos foi vista como informação preponderante para CAD neste trabalho.

A fim de explorar tais relacionamentos, apresentamos um algoritmo relacional, baseado em análise de vizinhança, que utiliza a rede de termos definida. Nosso algoritmo consiste simplesmente no uso da popular máxima: *"me digas com quem andas que direi quem és"*. Foram apresentadas diversas melhorias sobre nosso algoritmo baseadas em características observadas nos relacionamentos e na definição da rede. Avaliações experimentais sobre quatro coleções de documentos reais mostram que nosso algoritmo pode alcançar resultados equiparáveis aos do SVM (i.e., algoritmo *estado-da-arte* para CAD). Tal eficácia, aliada a uma alta eficiência e grande intuição sobre o processo de classificação, são algumas das características que tornam nosso algoritmo uma alterna-

tiva interessante para CAD.

Uma análise detalhada sobre os resultados obtidos mostrou que a evolução temporal dos documentos consiste em uma importante dimensão de análise para o nosso algoritmo. A abordagem de tal dimensão em CAD representa uma forma de obter informações mais precisas sobre o comportamento de cada termo. Ou seja, temos assim um aprimoramento da máxima mencionada para: *"me digas com quem andas,* **quando***, que direi quem és"*. De forma a considerar informações temporais em nossa rede de termos, atribuímos a cada relacionamento informação sobre o momento de sua construção. Simples versões temporais, que avaliam esta nova rede, foram capazes de melhorar o desempenho alcançado por nosso algoritmo relacional. Além disso, avaliações sobre o comportamento dos relacionamentos ao longo do tempo mostram que há espaço para estudos mais detalhados nesta direção.

Por fim, avaliamos nosso algoritmo, também, sob duas outras perspectivas. A primeira delas consiste no seu desempenho em cenários de escassez de informação. Alguns documentos a serem classificados podem não possuir informações suficientes no conjunto de treino, dada a baixa freqüência de ocorrência de muitos termos e as filtragens realizadas. Uma maneira simples de tratar tal problema foi prever a classe dos relacionamentos ausentes na rede, tal como o problema de *Link Prediction*. Através desta estratégia, conseguimos ganhos sobre o desempenho observado para a versão inicial do nosso algoritmo. A segunda perspectiva visa avaliar o uso de atributos dos relacionamentos em nosso algoritmo relacional. A idéia é que cada relacionamento, considerado como voto para definição das classes, pode ser ponderado com base em algum atributo específico. Avaliamos essa ponderação considerando a freqüência de co-ocorrência dos termos e o valor de *Predominância* dos relacionamentos. O uso de ambos atributos mostrou-se útil para o aperfeiçoamento do nosso algoritmo. Dessa forma, podemos observar várias dimensões nas quais o algoritmo relacional proposto pode ser melhorado.

8.2 Potenciais e Limitações

A generalidade das propriedades lingüísticas utilizadas como base neste trabalho, permite-nos acreditar que nossa proposta é potencialmente útil em vários cenários de aplicação de CAD. As características observadas sobre os relacionamentos entre termos são inerentes à forma da comunicação humana, e, pouco provavelmente, apresentam grandes alterações em cenários distintos. Textos são organizados como sentenças, que são compostas por palavras que interagem entre si. Esta interação é parte fundamental

na definição da semântica de cada documento. Além disso, comportamentos distintos, quanto ao uso, entre termos básicos e especializados, tendem a ocorrer em diversos cenários, por limitação própria da linguagem [Chen, 1995]. Dessa forma, a avaliação da nossa proposta junto a outros cenários de classificação de documentos representa um interessante trabalho futuro.

Quanto às limitações de uso do algoritmo proposto, como discutido no capítulo 5, a eficiência de execução pode ser crítica para classificação de documentos inteiros. Como a construção da nossa rede de termos não faz restrição quanto a distância de co-ocorrência entre os termos de um documento, gerando uma clique para cada documento, documentos com muitos termos distintos geram redes gigantescas. Neste caso, o tempo de avaliação dos vizinhos de cada termo pode se tornar muito alto. Dessa forma, o uso deste algoritmo torna-se mais eficiente para documentos que apresentem poucos termos. Documentos científicos, ou que possuem um pequeno texto de alta qualidade semântica capaz de sumarizar todo o documento, tal como um resumo, representam cenários ideais de uso do nosso algoritmo. Para a classificação de documentos inteiros, seria interessante avaliar o impacto de se restringir a distância de co-ocorrência entre termos nos documentos, gerando redes menores para análise.

8.3 Trabalhos Futuros

Durante a elaboração e execução deste trabalho, pudemos vislumbrar um conjunto grande de oportunidades de trabalhos que podem ser exploradas a partir das contribuições e resultados alcançados. Nas subseções seguintes descrevemos brevemente três dessas oportunidades consideradas mais relevantes.

8.3.1 Análise de Outros Modelos Relacionais

Embora modelos colaborativos baseados em análise de vizinhança constituam os mais populares modelos de análise relacional, não são os únicos. Uma outra técnica muito popular de análise de relacionamentos entre objetos de um domínio são os chamados **Modelos de Fatores Latentes** (MFL) [Koren, 2009]. A análise de fatores consiste em um método estatístico usado para descrever a variabilidade e relação entre as variáveis observadas em um domínio e outras variáveis não observadas, denominadas fatores. As variáveis observadas são usualmente modeladas como uma combinação linear dos fatores. Um dos mais populares representantes de MFLs são os métodos de SVD (*Singular Value Decomposition*). MFLs são, inclusive, amplamente utilizados para a tarefa de recomendação, apresentando alta eficácia em diversos cenários. Sua utilidade tanto em

recomendação, mas potencialmente para CAD, reside em sua capacidade de identificar as características mais importantes do domínio de análise, identificando, dentre outras coisas, combinações mais relevantes entre os objetos. Dessa forma, consideramos como uma direção de pesquisa altamente relevante a aplicação dos conceitos aqui discutidos em algoritmos relacionais baseados em fatores latentes.

Alguns estudos apontam, inclusive, que modelos baseados em análise de vizinhança e baseados em fatores latentes são, na verdade, complementares. Enquanto os primeiros são bons para identificar e explorar correlações locais entre subgrupos de objetos, os MFLs conseguem identificar fatores globalmente importantes em um domínio, permitindo identificar as características mais relevantes de análise. Baseado nisso, tal como feito em cenários de recomendação, uma estratégia interessante seria estudar modelos híbridos de classificação relacional, capazes de explorar tanto o conhecimento local dos termos quanto o conhecimento global do domínio.

8.3.2 Definição de Funções Temporais

Uma análise mais aprofundada sobre a evolução temporal das coleções de documentos, como argumentado no capítulo 6, representa outra frente de pesquisa promissora para este trabalho. Fenômenos sazonais, ou periódicos, tornam a definição de uma função de ponderação temporal robusta uma tarefa desafiadora. Além disso, vimos que relacionamentos consistentes por longos períodos tendem a ser mais estáveis e, portanto, deveriam ser beneficiados, antes a serem penalizados por uma função de ponderação. Este tipo de observação sugere que funções de ponderação individuais para cada relacionamento modelariam melhor estas que chamamos de *micro-evoluções*. Como argumentado em Koren [2009], há além de tendências globais de um domínio, compartilhada por grande maioria dos objetos, tendências individuais distintas. Em geral, objetos não evoluem na mesma velocidade e para a mesma direção. Há tendências locais que podem modelar melhor um domínio de estudo. Certamente modelar e considerar tais tendências locais representa um grande desafio pertinente não somente a CAD, mas a diversas áreas de estudo.

8.3.3 Escassez de Informação

Diversas formas de abordar o problema de escassez de informação podem ser avaliadas. Além da definição de formas mais elaboradas de se realizar a predição dos tipos dos relacionamentos, podemos utilizar tesauros, por exemplo, para tratar este problema. Um tesauro consiste, basicamente, de uma lista de palavras com significados semelhantes,

dentro de um domínio específico de conhecimento. A construção de um tesauro permite identificar palavras tipicamente utilizadas em um mesmo contexto de fala, e que, conseqüentemente, podem substituir umas às outras em diversos cenários. A utilidade dos tesauros para a escassez de informação reside justamente nesta capacidade de encontrar termos semanticamente similares. Caso um termo presente em um documento a ser classificado não seja encontrado no conjunto de treino, uma possibilidade seria verificar os relacionamentos de outros termos similares a ele observados no treino. Com isso, aumentamos as informações disponíveis acerca de cada documento a ser classificado, permitindo alcançar melhores resultados de classificação.

Referências Bibliográficas

Adamic, L. A. (2000). Zipf, power-laws, and pareto-a ranking tutorial. `http://ginger.hpl.hp.com/shl/papers/ranking/ranking.html`. Published by Xerox Palo Alto Research Center, Palo Alto, CA.

Aizawa, A. (2001). Linguistic techniques to improve the performance of automatic text categorization. In *Proc. of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS)*, pp. 307--314.

Albert, R. & Barabasi, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47.

Albert, R.; Jeong, H. & Barabasi, A.-L. (1999). The diameter of the world wide web. *Nature*, 401:130.

Alonso, O.; Gertz, M. & Baeza-Yates, R. (2007). On the value of temporal information in information retrieval. *SIGIR Forum*, 41(2):35--41.

Archdeacon, D. (1996). Topological graph theory: A survey. *Congressus Numerantium*, 115:115--120.

Barabasi, A.-L. & Bonabeau, E. (2003). Scale-free networks. *Scientific American*, 288:60--69.

Bernstein, A.; Clearwater, S. & Provost, F. (2003). The relational vector-space model and industry classification. In *In Proc. of the 18th IJCAI*, pp. 8--18. Free Press.

Bigi, B.; Brun, A.; Haton, J. P.; Smaili, K. & Zitouni, I. (2001). A comparative study of topic identification on newspaper and e-mail. In *Proc. of the String Processing and Information Retrieval (SPIRE)*, pp. 238--241.

Bilgic, M.; Namata, G. M. & Getoor, L. (2007). Combining collective classification and link prediction. In *Proc. of the 7th IEEE International Conference on Data Mining Workshops*, pp. 381--386.

Bingham, E.; Kabán, A. & Girolami, M. (2003). Topic identification in dynamical text by complexity pursuit. *Neural Processing Letters*, 17(1):69--83.

Brieman, L. & Spector, P. (1992). Submodel selection and evaluation in regression: The x-random case. *International Statistical Review*, 60:291--319.

Calderón-Benavides, M. L.; González-Caro, C. N.; Pérez-Alcázar, J. J.; García-Díaz, J. C. & Delgado, J. (2004). A comparison of several predictive algorithms for collaborative filtering on multi-valued ratings. In *Proc. of the ACM SAC*, pp. 1033--1039, Nicosia, Cyprus. ACM.

Cancho, R. F. & Sole, R. V. (2001a). The small world of human language. *Royal Society B: Biological Sciences*, 268(1482):2261--2265.

Cancho, R. F. & Sole, R. V. (2001b). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited*. *Journal of Quantitative Linguistics*, 8(3):165--173.

Chakrabarti, S.; Dom, B. & Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. In *Proc. of the SIGMOD*, pp. 307--318, Seattle, Washington, United States. ACM.

Chen, K. (1995). Topic identification in discourse. In *Proc. of the 7th Meeting of the European Association for Computational Linguistics*, pp. 267--271.

Cohen, W. W. & Singer, Y. (1999). Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems*, 17(2):141--173.

Corso, G. M. D.; Gulli, A. & Romani, F. (2005). Ranking a stream of news. In *Proc. of 14th International World Wide Web Conference*, pp. 97--106, Chiba, Japan.

Couto, T.; Cristo, M.; Gonçalves, M. A.; Calado, P.; Ziviani, N.; Moura, E. & Ribeiro-Neto, B. (2006). A comparative study of citations and links in document classification. In *Proc. of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pp. 75--84, Chapel Hill, NC, USA. ACM.

Crandall, D.; Cosley, D.; Huttenlocher, D.; Kleinberg, J. & Suri, S. (2008). Feedback effects between bimilarity and social influence in online communities. In *Proc. of the 14th ACM SIGKDD*, pp. 160--168, Las Vegas, Nevada, USA. ACM.

Denoyer, L.; Zaragoza, H. & Gallinari, P. (2001). Hmm-based passage models for document classification and ranking. In *Proc. of ECIR-01, 23rd European Colloquium on Information Retrieval Research*, pp. 126--135.

Dorogovtsev, S. & Mendes, J. (2002). Evolution of networks. *Advances in Physics*, 51:1079.

Du, N.; Wu, B.; Pei, X.; Wang, B. & Xu, L. (2007). Community detection in large-scale social networks. In *Proc. of the 9th WebKDD and 1st SNA-KDD 2007*, pp. 16--25, San Jose, CA, USA. ACM.

Erdos, P. & Renyi, A. (1960). On the evolution of random graphs. *Publications of Mathematical Institute of the Hungarian Academy of Sciences*, 5:17--61.

Freund, J. E. & Simon, G. A. (1967). *Modern Elementary Statistics*. Prentice-Hall Englewood Cliffs.

Furnkranz, J. (1998). A study using n-gram features for text categorization. *Austrian Research Institute for Artificial Intelligence Technical Report OEFAI-TR-98-30 Schottengasse*, 3.

Furnkranz, J.; Mitchell, T. & Riloff, E. (1998). A case study in using linguistic phrases for text categorization on the www. In *Proc. of the AAAI/ICML Workshop on Learning for Text Categorization*, pp. 5--12. AAAI Press.

Gantz, J. F.; Reinsel, D.; Chute, C.; Gantz, J. F.; Chuta, C.; Manfrediz, A.; Minton, S.; Reinsel, D.; Schlichting, W. & Toncheva, A. (2008). The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011. *IDC. White Paper*.

Gantz, J. F.; Reinsel, D.; Chute, C.; Schlichting, W.; McArthur, J.; Minton, S.; Xheneti, I.; Toncheva, A. & Manfrediz, A. (2007). The expanding digital universe. *A Forecast of Worldwide Information Growth Through 2010*.

Gee, K. R. & Cook, D. J. (2005). Text classification using graph-encoded linguistic elements. In *Proc. of the 18th International FLAIRS Conference*.

Getoor, L. C. (2002). *Learning Statistical Models from Relational Data*. PhD thesis, Stanford University, Stanford, CA, USA. Adviser-Daphne Koller.

Guo, F.; Hanneke, S.; Fu, W. & Xing, E. P. (2007). Recovering temporally rewiring networks: A model-based approach. In *Proc. of the 24th ICML*, pp. 321--328, Corvalis, Oregon, USA. ACM.

Hauser, M. D.; Chomsky, N. & Fitch, W. (2002). The Faculty of Language: What is it, Who has it, and How did it evolve? *Science*, 298(5598):1569.

Hopcroft, J.; Khan, O.; Kulis, B. & Selman, B. (2003). Natural communities in large linked networks. In *Proc. of the 9th ACM SIGKDD*, pp. 541--546, Washington, D.C., USA. ACM.

Hopcroft, J.; Khan, O.; Kulis, B. & Selman, B. (2004). Tracking evolving communities in large linked networks. *PNAS*, 101(1):5249--5253.

Joachims, T. (2006). Training linear svms in linear time. In *Proc. of the 12th ACM SIGKDD Conference*, pp. 217--226, Philadelphia, PA, USA. ACM.

Ke, J. (2007). Complex networks and human language. `http://www.isrl.uiuc.edu/~amag/langev/paper/ke07complexNetworkLanguage.html`.

Koren, Y. (2009). Collaborative filtering with temporal dynamics. In *Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 447--456, New York, NY, USA. ACM.

Kossinets, G.; Kleinberg, J. & Watts, D. (2008). The structure of information pathways in a social communication network. In *Proc. of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 435--443, Las Vegas, Nevada, USA. ACM.

Kossinets, G. & Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, 311(5757):88--90.

Leskovec, J.; Backstrom, L.; Kumar, R. & Tomkins, A. (2008). Microscopic evolution of social networks. In *Proc. of the 11th ACM SIGKDD*, Las Vegas, Nevada, USA. ACM.

Li, C. H. & Park, S. C. (2009). An efficient document classification model using an improved back propagation neural network and singular value decomposition. *Expert Systems With Applications*, 36(2P2):3208--3215.

Liben-Nowell, D. & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal-American Society for Information Science and Technology*, 58(7):1019.

Liebscher, R. A. (2004). Temporal context: Applications and implications for computational linguistics. In *Proc. of the ACL 2004 workshop on Student research*, Morristown, NJ, USA. Association for Computational Linguistics.

Liu, B.; Hsu, W. & Ma, Y. (1998). Integrating classification and association rule mining. *Knowledge Discovery and Data Mining*, pp. 80--86.

Liu, R. & Lu, Y. (2002). Incremental context mining for adaptive document classification. In *Proc. of the 8th ACM SIGKDD conf.*, pp. 599--604, Edmonton, Alberta, Canada. ACM.

Luo, X. & Zincir-Heywood, A. N. (2004). Analyzing the temporal sequences for text categorization. *Lecture Notes in Computer Science*, 3215:498--505.

Macskassy, S. A. & Provost, F. (2003). A simple relational classifier. In *Proc. of the 2nd KDD/MRDM*, pp. 64--76.

Macskassy, S. A. & Provost, F. (2004). Simple models and classification in networked data. *CeDER Working Paper*.

Macskassy, S. A. & Provost, F. (2007). Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, 8:935--983.

Manning, C. D.; Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Markov, A. & Last, M. (2005). A simple structure-sensitive approach for web document classification. *Lecture Notes in Computer Science*, 3528:293--298.

McCallum, A. K. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. `http://www.cs.cmu.edu/~mccallum/bow`.

Mcpherson, M.; Lovin, L. S. & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415--444.

Mendelson, E., editor (2002). *The Complete Works of W. H. Auden: Prose*, volume II. New Jersey: Princeton University Press.

Mittendorf, E. & Schäuble, P. (1994). Document and passage retrieval based on hidden markov models. In *Proc. of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 318--327, New York, NY, USA. Springer-Verlag.

Montejo-Raez, A.; Urena-Lopez, L. A.; Garcia-Cumbreras, M. A. & Perea-Ortega, J. M. (2008). Using linguistic information as features for text categorization. In *Proc. of the MMDSS: Advances in Data Mining, Search, Social Networks and Text Mining, and Their Applications to Security*, Varese, Italy. Ios Press Inc.

Moschitti, A. & Basili, R. (2004). Complex linguistic features for text classification: A comprehensive study. *Lecture Notes in Computer Science*, 2997:181--196.

Mourão, F. & Meira Jr., W. (2008). Entendendo a evolução temporal em redes complexas. In *Proc. of the 1st WIVA*.

Mourão, F.; Rocha, L.; Araújo, R.; Couto, T.; Gonçalves, M. & Meira Jr., W. (2008). Understanding temporal aspects in document classification. In *Proc. of the 1st WSDM*, pp. 159--170, Palo Alto, California, USA. ACM.

Mourão, F.; Rocha, L.; Miranda, L.; A., V. & Meira Jr., W. (2009). Quantifying the impact of information aggregation on complex networks: A temporal perspective. In *Proc. of the 6th Workshop on Algorithms and Models for the Web Graph (WAW)*, Barcelona, Spain.

Neville, J. & Jensen, D. (2007). Relational dependency networks. *Journal of Machine Learning Research*, 8.

Neville, J.; Jensen, D.; Friedland, L. & Hay, M. (2003). Learning relational probability trees. In *Proc. of the 9th ACM SIGKDD*, pp. 625--630, Washington, D.C., USA. ACM.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167--256.

Newman, M. E. J. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323.

Proops, J. L. R. (1987). Entropy, information and confusion in the social sciences. *Journal of Interdisciplinary Economics*, 1:224--242.

Provost, F.; Perlich, C. & Macskassy, S. A. (2003). Relational learning problems and simple models. In *In Proc. of the 18th IJCAI*, pp. 116--120. Free Press.

Quinlan, J. R. (2003). *C4. 5: Programs for Machine Learning*. Morgan Kaufmann.

Rocha, L.; Mourão, F.; Pereira, A.; Gonçalves, M. & Meira Jr, W. (2008). Exploiting temporal contexts in text classification. In *Proc. of the ACM CIKM*, Napa Valley, CA, United States. ACM.

Said, Y. H.; Wegman, E. J.; Sharabati, W. K. & Rigsby, J. T. (2008). Social networks of author-coauthor relationships. *Comput. Stat. Data Analysis*, 52(4):2177--2184.

Salles, T.; Rocha, L.; Mourão, F.; Pappa, G. L.; Cunha, L.; Gonçalves, M. A. & Meira Jr, W. (2009). Classificação automática de documentos robusta temporalmente. In *Proc. of the 24th Brazilian Symposium on Databases*, Fortaleza, CE, Brazil.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513--523.

Salton, G. & McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill Inc., New York, NY, USA.

Schenker, A.; Bunke, H.; Last, M. & Kandel, A. (2004). Building graph-based classifier ensembles by random node selection. *Lecture notes in computer science*, 3077:214--222.

Schenker, A.; Last, M.; Bunke, H. & Kandel, A. (2003). Classification of web documents using a graph model. In *Proc. of the 7th International Conference on Document Analysis and Recognition*, Washington, DC, USA. IEEE Computer Society.

Schmidt, D. C. (2000). Learning probabilistic relational models. *Journal of Relational Data Mining*, pp. 307--333.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1--47.

Sen, P. & Getoor, L. (2007). Link-based classification. Technical report cs-tr-4858, University of Maryland.

Sharan, U. & Neville, J. (2007). Exploiting time-varying relationships in statistical relational models. In *Proc. of the 9th WebKDD and 1st SNA-KDD*, pp. 9--15, San Jose, California, USA. ACM.

Taskar, B.; Chatalbashev, V. & Koller, D. (2004). Learning associative markov networks. In *Proc. of the 21st ICML*, p. 102, Banff, Alberta, Canada. ACM.

Taskar, B.; Segal, E. & Koller, D. (2001). Probabilistic classification and clustering in relational data. In *Proc. of the 7th International Joint Conference on Artificial Intelligence*, pp. 870--878. Lawrence Erlbaum Associates LTD.

Taskar, B.; Wong, M.; Abbeel, P. & Koller, D. (2003). Link prediction in relational data. In *Neural Information Processing Systems*.

Tsoumakas, G. & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1--13.

Tsymbal, A. (2004). The problem of concept drift: Definitions and related work. `https://www.cs.tcd.ie/publications/tech-reports/reports.04/TCD-CS-2004-15.pdf`.

Vapnik, V. (1996). Structure of statistical learning theory. *Computational Learning and Probabilistic Reasoning*.

Vazquez, A.; Flammini, A.; Maritan., A. & Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, 21(6):697--700.

Veloso, A.; Meira Jr, W. & Zaki, M. J. (2006). Lazy associative classification. In *Proc. of the International Conference on Data Mining (ICDM)*, pp. 645--654.

Watts, D. J. (1999). *Small worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press.

Widmer, G. & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23(1):69--101.

Wilson, E. O. (1998). *Consilience: The Unity of Knowledge*. Knopf.

Zhang, M. L. & Zhou, Z. H. (2007). Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038--2048.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An introduction to Human Ecology*. Addison-wesley press.