UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
ESPECIALIZAÇÃO EM INFORMÁTICA: ÁREA DE CONCENTRAÇÃO: GESTÃO DE TECNOLOGIA DA INFORMAÇÃO

Everton Batista Dos Santos

APRIMORAR A PESQUISA DE ITENS NO CATÁLOGO DE MATERIAIS E SERVIÇOS DO SIASG

Brasília
2019

EVERTON BATISTA DOS SANTOS

APRIMORAR A PESQUISA DE ITENS NO CATÁLOGO DE MATERIAIS E SERVIÇOS DO SIASG

Monografia apresentada ao Curso de Especialização em Informática do Departamento de Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do grau de Especialista em Informática.

Área de Concentração: Gestão de Tecnologia da Informação

Orientador: Wagner Meira Junior

Brasília
2019

Ficha catalográfica elaborada pela Biblioteca do ICEx – UFMG

Santos, Everton Batista dos

S237a Aprimorar a pesquisa de itens no catálogo de materiais e serviços do SIASG / Everton Batista dos Santos – Brasília, 2019.
xi, 181 f., il.

Monografia (especialização) – Universidade Federal de Minas Gerais. Departamento de Ciência da Computação.

Orientador: Wagner Meira Junior

1.Computação – Monografias. 2. Recuperação da Informação. 3. Administração Pública. I. Orientador. II. Título

CDU 519.6*

UNIVERSIDADE FEDERAL DE MINAS GERAIS

INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
ESPECIALIZAÇÃO EM INFORMÁTICA: ÁREA DE CONCENTRAÇÃO GESTÃO EM TECNOLOGIA DA INFORMAÇÃO

Qualificação do Catálogo de Materiais e Serviços dos Sistemas de Compras Governamentais

EVERTON BATISTA DOS SANTOS

Monografia apresentada aos Senhores:

Prof. Wagner Meira Júnior
Orientador
DCC - ICEx – UFMG

Prof. José Nagib Cotrim Árabe
DCC - ICEx - UFMG

Prof. José Marcos Silva Nogueira
DCC - ICEx - UFMG

Belo Horizonte, 14 de março de 2019

Dedicatória

Gostaria de dedicar esse trabalho a Deus por ser tão presente e essencial em minha vida, o autor do meu destino, meu guia que nunca me abandonou.

A todos os professores que ministraram disciplinas nesta pós-graduação, dos quais foram fundamentais e de grande importância na evolução da minha vida profissional.

Em especial, ao professor Wagner Meira Jr., pela sua paciência, conselhos e ensinamentos que foram essenciais para o desenvolvimento do TCC.

Dedico este trabalho também à minha família e amigos que sempre estiveram presentes nos momentos de minha formação, aos colegas de classe que tanto contribuíram para o meu desenvolvimento.

Agradecimentos

De forma especial aos professores pela atenção dispensada na concepção dos conteúdos a serem abordados diante do pouco conhecimento sobre os rumos que o curso deveria permear, pelas eventuais disparidades entre as expectativas da UFMG, expectativas da ENAP e de nós alunos.

Nada obstante, quero agradecer a cada um dos colegas de classe que, como eu, alternavam seus a fazeres diários nos órgãos, suas respectivas responsabilidades pessoais e os estudos e dedicação na conclusão dos trabalhos avaliativos.

Conseguimos, diante de tantas dificuldades logísticas, pessoais e laborais chegar ao final deste curso com a esperança aflorada em utilizar boa parte das tecnologias abordadas no dia a dia dos processos de trabalho, bem como nos serviços disponíveis à sociedade.

Convivemos por cerca de 18 meses, com missões diferentes nos órgãos lotados, porém sempre nos dedicamos ao bem público, preocupados em servir, que na sua essencialidade originam nosso título de servidor público.

Resumo

Este trabalho tem como premissa o aprimoramento da pesquisa de materiais e serviços catalogados no sistema SIASG. Esses itens são peças fundamentais na construção e concepção do processo licitatório e inclusive fomenta e balisa a pesquisa de preços a fim de definir os valores estimados a serem adquiridos. Atualmente as existem 3 formas de consulta (SIASG, SIASGNet e Comprasnet) que apresentam os itens catalogados, porém são extremamentes rudimentares e pouco intuitivas, além de truncar características importantes que especificam os materiais e serviços. Por meio do uso da ciência da Recuperação da Informação, pretendemos sugerir requisitos e comportamentos que a nova ferramenta deverá contemplar com a finalidade de simplificar a busca de itens. Outro benefício proposto será a acurácia dos valores mais próximos da realidade efetivamente realizada, através do painel de preços,

Palavras-chave: Recuperação da Informação, Administração Pública, SIASG, SIASGnet, Comprasnet, pesquisa de materiais e serviços, catálogo.

Abstract

This work has as premise the improvement of the research of materials and services cataloged in the SIASG system. These items are essential pieces in the construction and conception of the bidding process and even foments and balises the price survey in order to define the estimated values to be acquired. Currently there are 3 ways of consulting (SIASG, SIASGNet and Comprasnet) that present the cataloged items, but are extremely rudimentary and not intuitive, and truncate important characteristics that specify the materials and services. Through the use of information retrieval science, we intend to suggest requirements and behaviors that the new tool should contemplate in order to simplify the search for items. Another proposed benefit will be the accuracy of the values closest to reality actually realized, through the price panel.

Keyword: Information Retrieval, Public Administration, SIASG, SIASGnet, Comprasnet, materials and services search, catalog.

Lista de ilustrações

Lista de abreviaturas e siglas

ABNT	Associação Brasileira de Normas Técnicas
CATMAT	Catálogo de Materiais do SIASG
CATSER	Catálogo de Serviços do SIASG
IDF	Frequência Inversa Documento
RI	Recuperação da Informação
SEGES/MP	Secretaria de Gestão do então Ministério do Planejamento, Desenvolvimento e Gestão
SIASG	Sistema Integrado de Administração de Serviços Gerais
SISG	Sistema de Serviços Gerais
TF	Frequência Termo
TF-IDF	Frequência Termo - Frequência Inversa Documento
UND	Unidade

Sumário

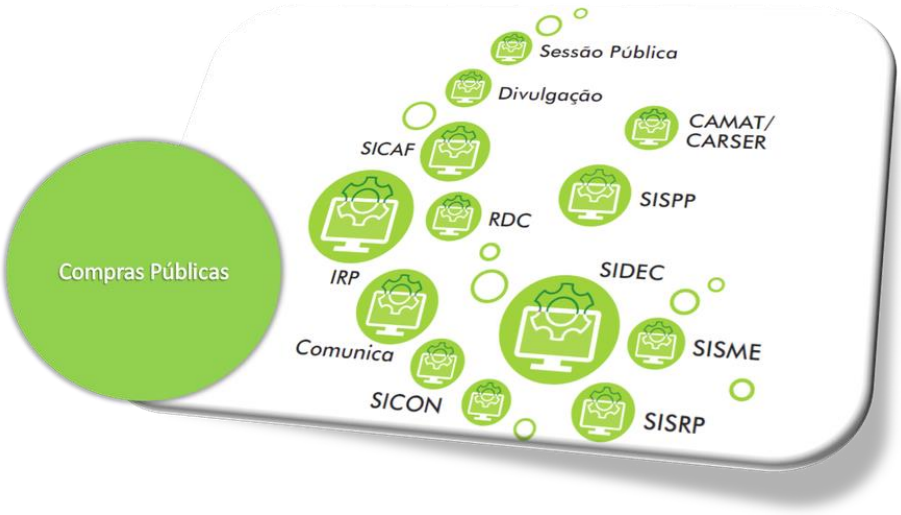
1 Introdução

Em 1994, o Decreto n.º 1.094 (1) regulamentou os arts. 30 e 31 do Decreto-Lei n.º 200, de 25 de fevereiro de 1967 (2), e instituiu o Sistema de Serviços Gerais - SISG. Este sistema, em sentido lato, surge como parte integrante de um sistema administrativo orgânico que engloba toda a Administração Pública Federal, num esforço de coordenação das atribuições de logística pública com vistas a maior eficiência. Trata-se de um entre os vários sistemas auxiliares da Administração, responsáveis pela execução de atividades de cunho transversal.

A Secretaria de Gestão do então Ministério do Planejamento, Desenvolvimento e Gestão (SEGES/MP), órgão central do Sistema de Serviços Gerais (SISG), têm conduzido ações estruturantes visando estabelecer um modelo de atuação mais moderno e dinâmico, garantido, dessa forma, a coordenação e convergência entre os diversos órgãos e entidades que integram este Sistema. (3)

Diferentemente dos entendimentos pretéritos e diante dessa nova releitura do SISG, o SIASG deixa de ser visto apenas como os submódulos de compras governamentais - cadastro de fornecedores, o catálogo de materiais e serviços, o módulo de divulgação eletrônica de licitações, o de registro de preços praticados, o sistema de gestão de contratos, o módulo de emissão empenhos, o pregão eletrônico, a cotação eletrônica e um extrator de dados estatísticos (Datawarehouse) – e ganha relevância estratégica, passando a ser visto como um instrumento de apoio, transparência e controle na execução das atividades do SISG, por meio da informatização e operacionalização do conjunto de suas atividades, bem como no gerenciamento de todos os seus processos.

Figura 1 – Família Sistemas Compras Governamentais

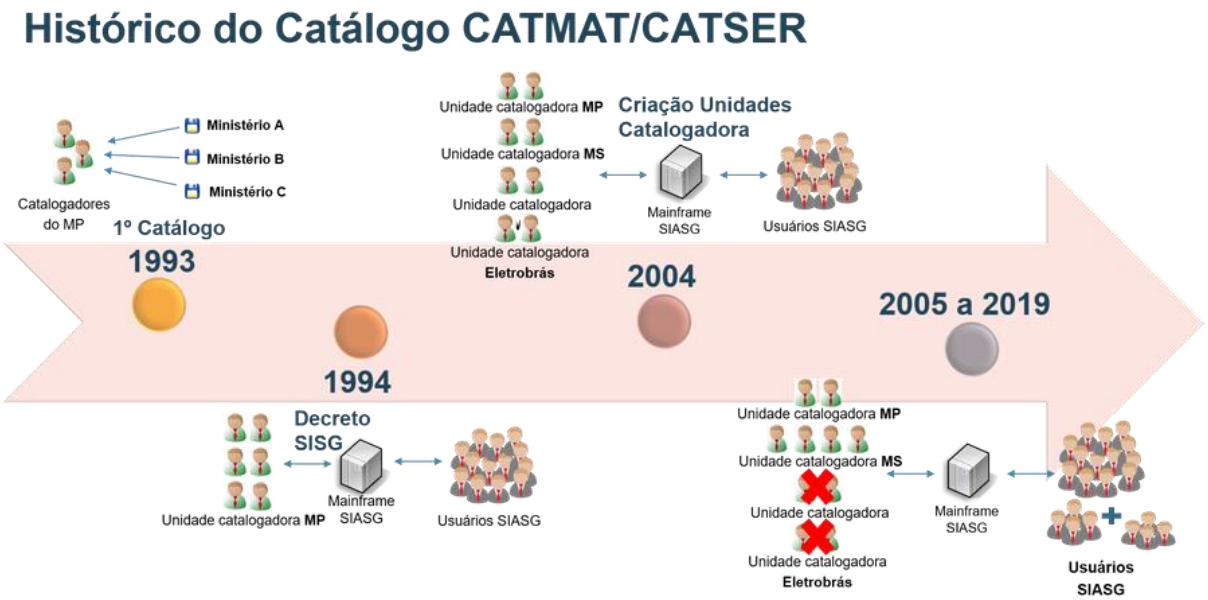


Fonte: Referencial de Governança e Gestão do Sistema de Serviços Gerais – SISG

O Catálogo de Materiais e Serviços (CATMAT/CATSER) do SIASG representa a base de registros que identificam todos os bens e serviços a serem licitados e contratados pela Administração Pública Federal. Todas as operações realizadas por meio dos sistemas de Compras Governamentais utilizam esse catálogo para definir os objetos das respectivas licitações e contratações.

Historicamente esse módulo foi renegado dentre as prioridades e projetos a serem reformulados no âmbito do processo de compras - abstraio aqui o debate sobre o mérito deste direcionamento -, que causou e ainda causa o enfraquecimento e obsolescência em contraponto aos outros módulos que compõem os sistemas de compras governamentais. Abaixo, apresento na figura, alguns fatores que corroboram com a afirmação supracitada.

Figura 2 – Histórico do Catálogo



Fonte: Apresentação Coordenação-Geral de Sistemas de Compras Governamentais

Nota-se então que o catálogo de materiais e serviços é uma preocupação contundente dos órgãos de controle conforme consta no relatório de acompanhamento, produzido pela Secretaria de Fiscalização em Tecnologia da Informação - SEFTI/TCU, referente às contratações públicas operadas nos sistemas SIASG e Comprasnet, que originou o Acórdão 2593/2017 - TCU - Plenário, cujas determinações concernentes ao então Ministério do Planejamento, transcrevemos abaixo: (4)

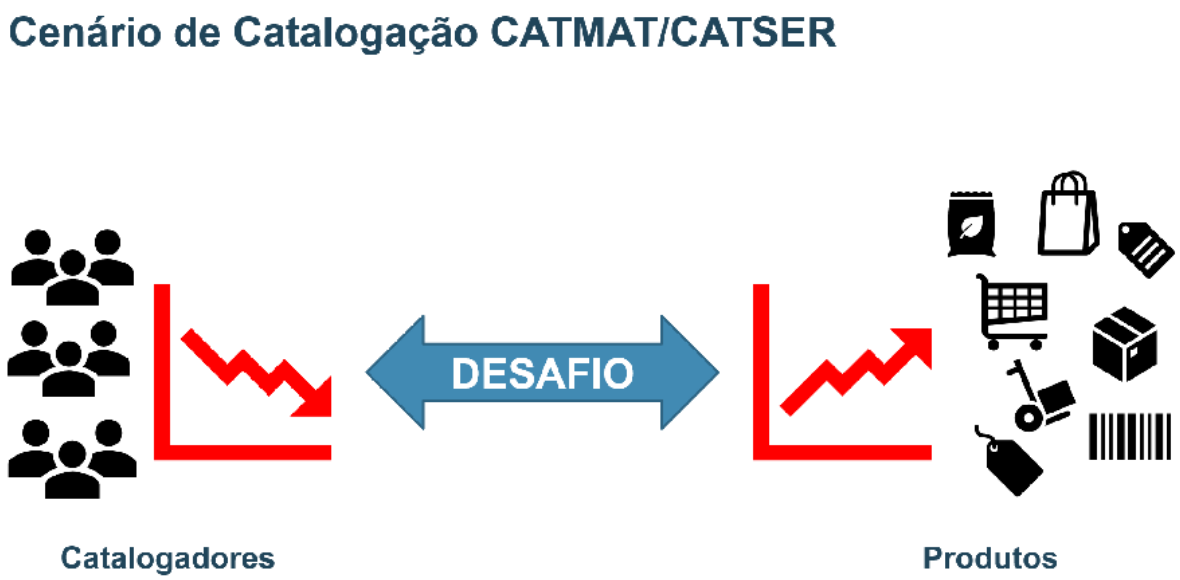
> VISTOS, relatados e discutidos estes autos de relatório de acompanhamento destinado a promover o acompanhamento das contratações públicas operadas nos sistemas Sidec, Siasg e Comprasnet, utilizando procedimentos de auditoria contínua e aplicando técnicas de análises de dados, a fim de propiciar a construção de painel eletrônico de contratações (Dashboard),
>
> ACORDAM os Ministros do Tribunal de Contas da União, reunidos em sessão do Plenário, ante as razões expostas pelo Relator, em:

> 9.1 determinar ao Ministério do Planejamento, Desenvolvimento e Gestão, que:
>
> 9.1.1. com fulcro no art. 8º da Lei 12.527/2011 c/c art.8º, § 1º, do Decreto 8.777/2016, que mantenha atualizado o repositório de informações sobre as contratações públicas no portal dados abertos do Governo Federal;
>
> 9.1.2. no prazo de 180 (cento e oitenta) dias, depure a base de dados do painel de preços (http://paineldepreços.planejamento.gov.br) e, concomitantemente, crie mecanismos para padronizar os dados nele constantes e a inserção de novas informações, de forma minimizar as divergências observadas pela má-alimentação desse sistema de informação e facilitar a comparação de preços praticados no âmbito da administração-pública.

Sucintamente, o Painel de Preços é a ferramenta que disponibiliza ou exibe, de forma clara, dados e informações de compras públicas registradas no SIASG. Os dados apresentados na ferramenta têm origem em diversos módulos que foram desenvolvidos de forma incremental e em diferentes tecnologias. O banco de dados não relacional e a plataforma alta (mainframe) ainda predominam no SIASG e dificultam sobremaneira o relacionamento entre sistemas.

Esclarece-se que as atividades de “depuração da base de dados do painel de preços” e “criação de mecanismos para padronizar os dados constantes e a inserção de novas informações” são dependentes e não podem ser tratadas de forma isolada. Explica-se: depurar a base de dados sem criar os mecanismos para padronização dos dados no SIASG implicaria em desperdício, pois os dados depurados naturalmente voltariam a ficar desatualizados. Na figura abaixo reproduzo este desafio.

Figura 3 – Desafio de sustentação do Catálogo



Fonte: Apresentação Coordenação-Geral de Sistemas de Compras Governamentais

Além disso, a melhoria na qualidade dos dados do Painel exige evoluções e a padronização dos dados dos módulos CATMAT e CATSER. O principal objetivo desses

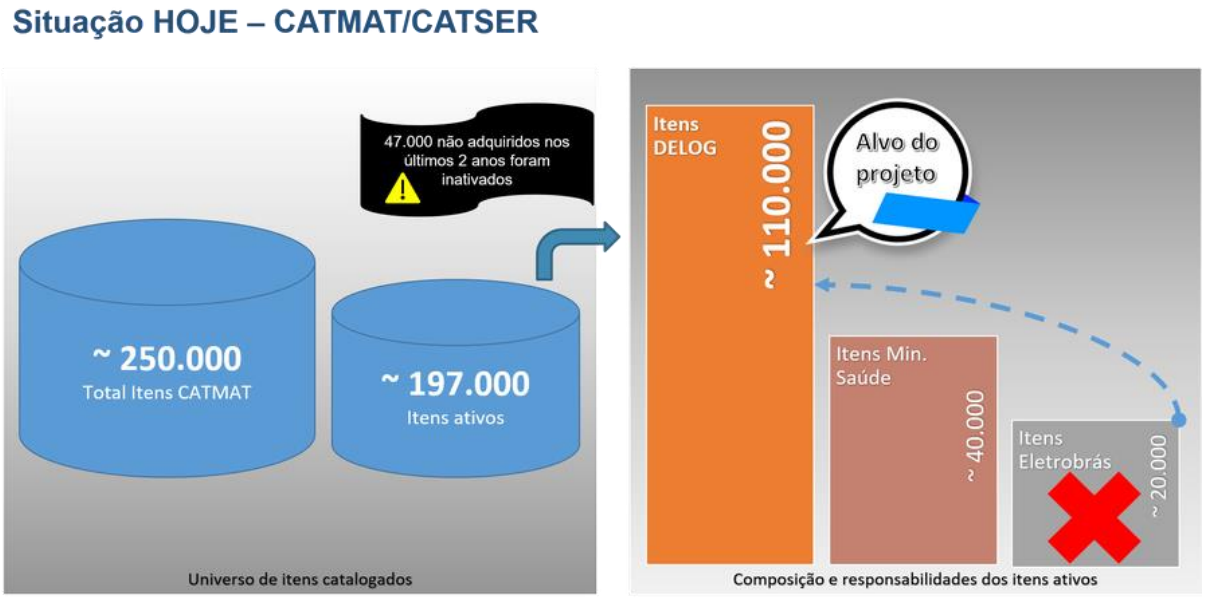
módulos é estabelecer e manter uma linguagem única e padronizada para identificação, codificação e descrição de materiais e serviços adquiridos pelo Governo Federal. A catalogação de materiais e serviços no SIASG segue os critérios adotados na *Federal Supply Classification* desde a década de 90.

Verifica-se portanto, que a organização dos dados nos catálogos tem um impacto direto na qualidade da informação proveniente do SIASG e no cruzamento de informações sobre o gasto público, alimentando os sistemas de informações sobre os itens que serão comprados, com suas respectivas unidades de fornecimento e quantidades. Tais dados deveriam permitir uma rastreabilidade com maior acurácia dos processos de compras, bem como estudos e análises em níveis estratégicos.

Atualmente o catálogo é composto, em números aproximados, por 250.000 itens catalogados, sendo que 197.000 estão ativos. De forma linear, foi inativado 47.000 itens que não foram adquiridos nos últimos 2 anos, com vistas a eliminar materiais dispensáveis ou desnecessários ao interesse público.

Dos itens que permaneceram no catálogo, 110.000 itens ativos estão sob a responsabilidade da unidade catalogadora do extinto Ministério do Planejamento, atual Ministério da Economia e 40.000 sob a responsabilidade da unidade catalogadora do Ministério da Saúde, deixando a unidade catalogadora da Eletrobrás de exercer esta atividade, conforme mostra a figura abaixo.

Figura 4 – Situação atual do Catálogo



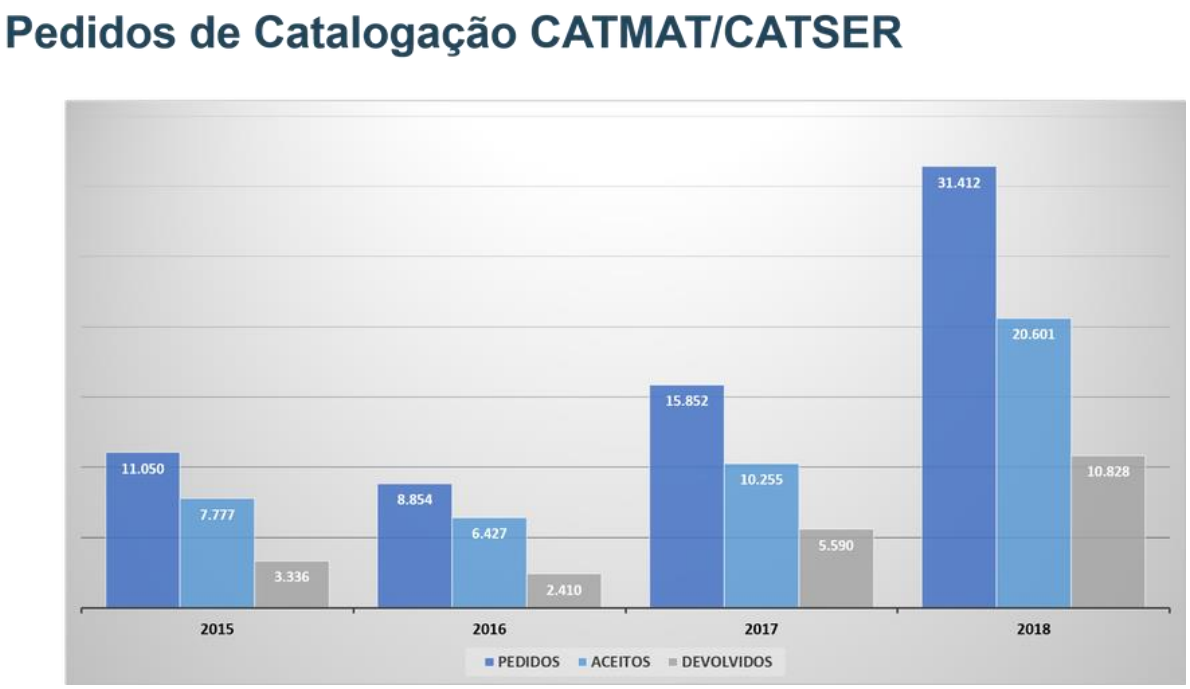
Fonte: Apresentação Coordenação-Geral de Sistemas de Compras Governamentais

Esta descentralização de atividades de catalogação por classes de materiais possibilita a eficiência no atendimento aos pedidos de inclusão, alteração de características, unidades de fornecimento, natureza de despesas com a dinamicidade requerida.

Propicia a atuação qualitativa, desempenhado por profissionais especialistas formando um time multidisciplinar.

Informações obtidas através do sistema SIASG, mais precisamente das rotinas de pedidos de catalogação de materiais e serviços, permitiram-me aferir a quantidade de solicitações realizadas nos últimos 4 anos, apresentado na figura abaixo.

Figura 5 – Quantidade de pedidos de catalogação



Fonte: Apresentação Coordenação-Geral de Sistemas de Compras Governamentais

Pode-se inferir que possivelmente o aumento de pedidos de catalogação teve como causa a inativação dos 47 mil itens citados na figura - Situação atual do Catálogo. Outra possibilidade vislumbrada tem como causa raiz a dificuldade em se consultar os itens pretendidos com as devidas especificações técnicas ou características que mais se assemelham com o produto a ser adquirido.

Ainda sobre o gráfico acima, depreendemos que cerca de 1/3 dos pedidos de catalogação realizados são devolvidos. Isso corrobora com a afirmação do último período do parágrafo anterior, onde a dificuldade e diversidade encontrado ao escolher o item desejado, força o usuário a demandar requerimentos indevidos às unidades catalogadoras. Graficamente exponho o cenário descrito na figura abaixo.

Figura 6 – Cenário dos pedidos de catalogação

Fonte: Apresentação Coordenação-Geral de Sistemas de Compras Governamentais

No exercício de 2017, tomado como recorte para esta análise, o Ministério da Saúde que trata cerca de 27% do catálogo, recebendo cerca de 51% dos pedidos de catalogação, gerou apenas 16% de itens novos no catálogo. Em contraponto o Ministério da Economia que se ocupa de cerca de 73% do catálogo, recebendo cerca de 7.697 solicitações de catalogação, gerou cerca de 82% de itens novos no catálogo.

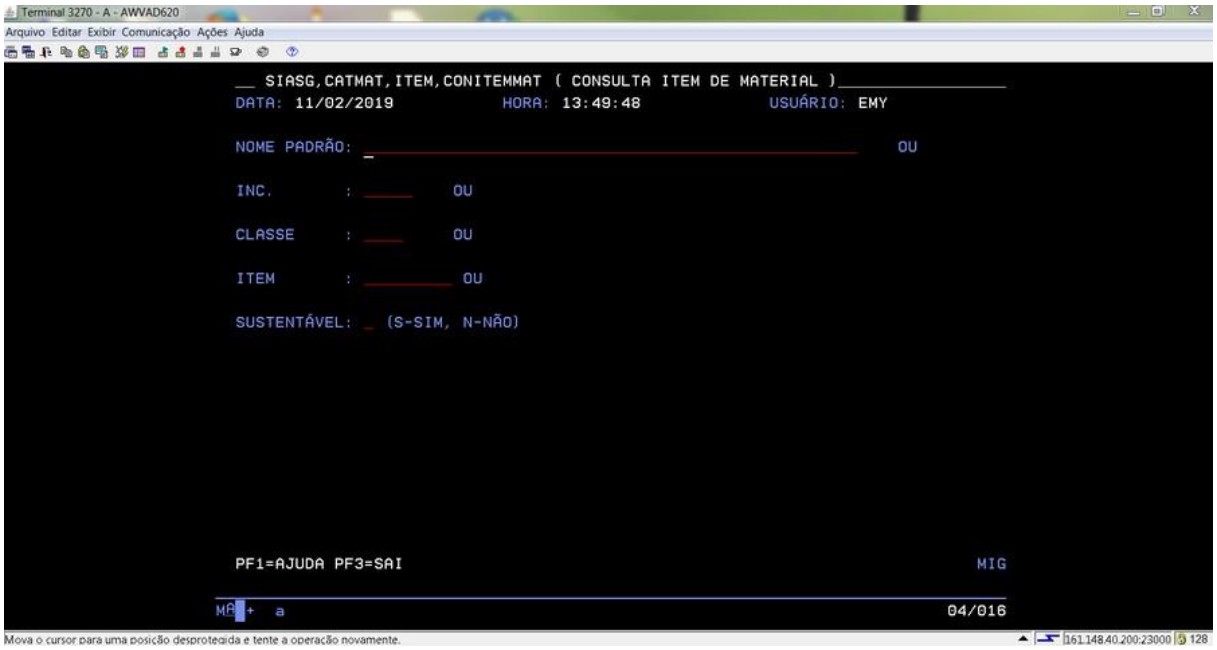
Depreende-se então a existência de fatores divergentes que podem influenciar neste panorama. Porém, existem fatores convergentes que poderiam diminuir consideravelmente as solicitações e esforço das unidades catalogadoras e do processo envolvido.

Vislumbra-se então que um desses fatores convergentes seria a deficitária e ineficiente pesquisa no catálogo. Estas ferramentas se comportam de maneiras distintas, apresentam informações desconexas ou omitem atributos e características necessárias ao usuário, bem como aparências totalmente díspares.

A pesquisa aos itens catalogados é precária e arcaica. Atualmente existem três funcionalidades ou ferramentas de busca no contexto do SIASG com comportamentos diferentes, múltiplas tecnologias em locais diversos e que apresentam informações desiguais.

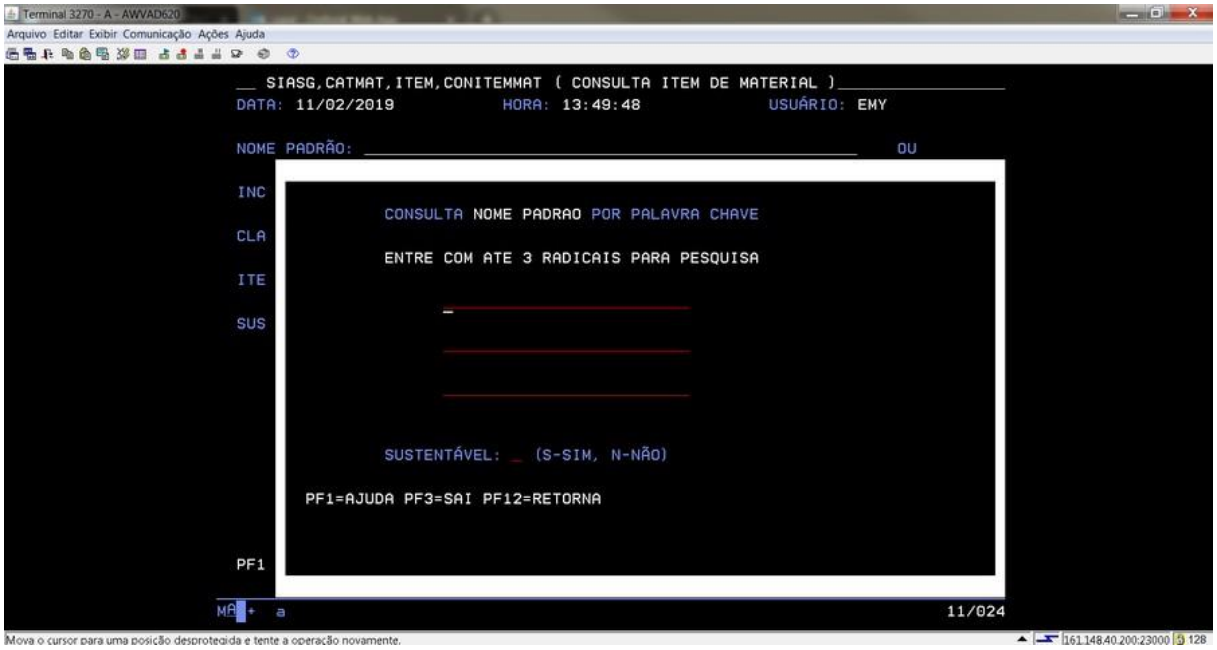
A primeira é por meio da transação consulta item dentro do SIASG em plataforma baixa (mainframe) que utiliza radicais em linhas separadas como parâmetros. Esta consulta além de não ser intuitiva, necessita que o usuário tenha se autenticado para obter acesso à funcionalidade e ainda seja autorizado.

Figura 7 – Pesquisa item catálogo grande porte



Fonte: https://acesso.serpro.gov.br -> SIASG -> CONITEMMAT

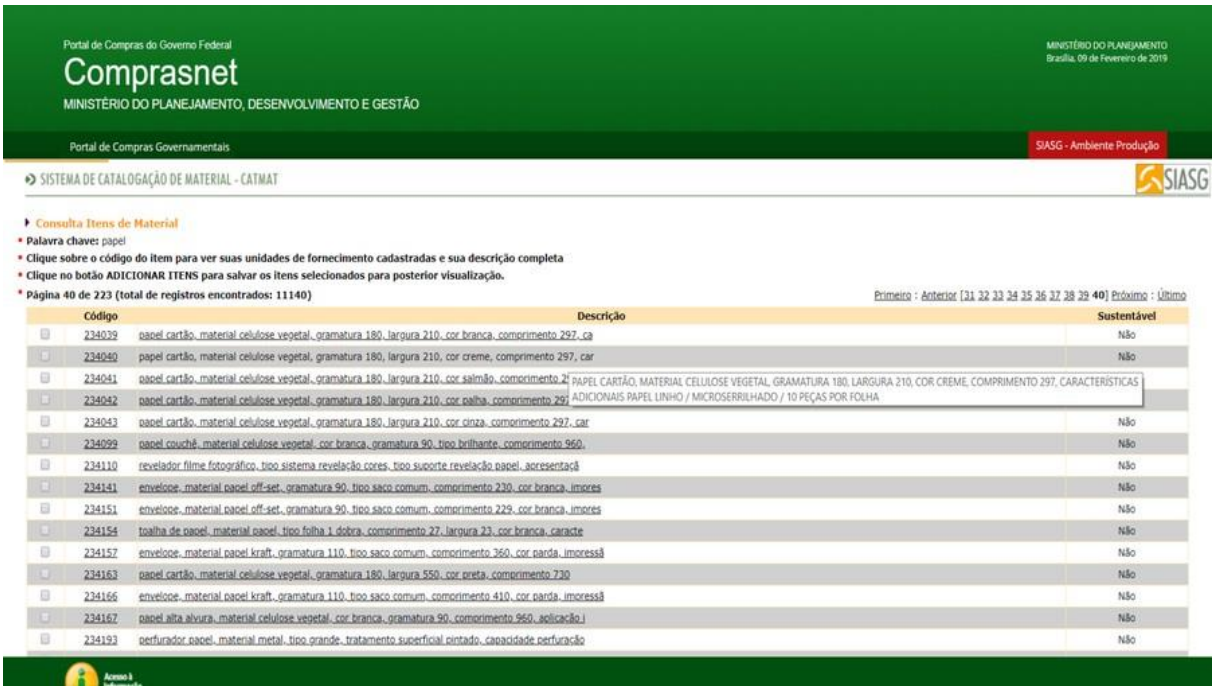
Figura 8 – Pesquisa item catálogo grande porte com radicais



Fonte: https://acesso.serpro.gov.br -> SIASG -> CONITEMMAT

A segunda, oferece a busca aberta ao público em plataforma alta (asp), sem necessidade de autenticação, porém com informações truncadas, quase sempre apresentadas equivocadamente ou, pior ainda, que não deveriam serem apresentadas, conforme demonstrado na figura abaixo:

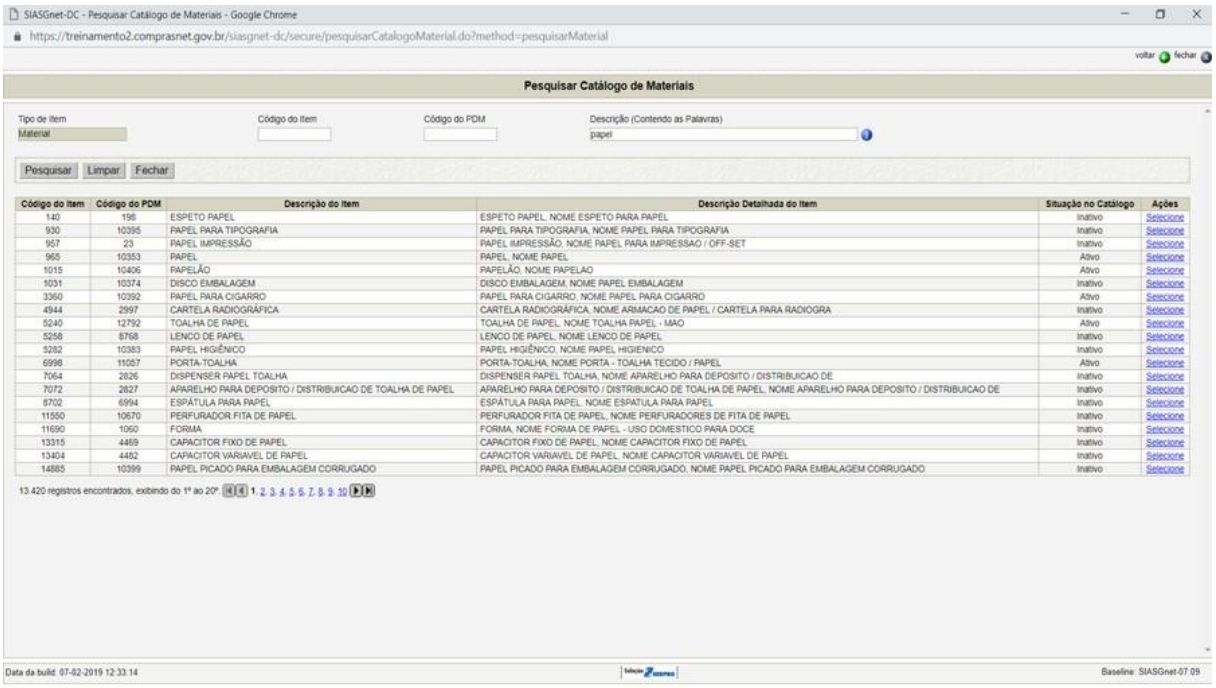
Figura 9 – Pesquisa item aberta ao público



Fonte: http://comprasnet.gov.br/acesso.asp?url=/Livre/Catmat/Conitemmat1.asp

A terceira e que exige autenticação, denominada divulgação de licitação, desenvolvida em plataforma alta (java), apresenta poucas informações capazes de fornecer aos usuários a escolha mais adequada, e que influenciam diretamente no processo licitatório, pois é nela que o usuário divulga o certame, objeto e itens a serem adquiridos.

Figura 10 – Pesquisa item no cadastro da licitação



Fonte: https://www.comprasnet.gov.br/seguro/loginPortal.asp -> Divulgação de Compras

O catálogo atualmente não está padronizado, gerando informações redundantes e descompassadas, trazendo algumas dificuldades em relação aos quesitos acima.

Como exemplo podemos citar a existência de milhares de itens de materiais com descrições similares com anuências mínimas.

Aliado ao problema acima relatado, os usuários encontram ferramentas de pesquisa nada intuitivas ou cognitivas que propiciariam o melhor enquadramento dos materiais ou serviços a serem adquiridos.

Outro exemplo é a existência de unidades de fornecimento e/ou unidade de medidas semelhantes na grafia que se refere a mesma medida. Ex: “Litro, L, 1000 mil mililitros; “Kg, quilo, quilograma, Kilo”; “UN, UND, UNIDADE, UNID, UM,”. Isto fortalece ainda mais o prejuízo semântico que a busca poderia promover aos usuários.

Nesse sentido, esse projeto visa sugerir melhorias aos usuários na consulta aos itens catalogados através da ciência de recuperação de informações, padronização no resultado da busca, sugerir itens no que tange a similaridade do termo pesquisado.

1.1 Modernização do Catálogo de materiais e serviços

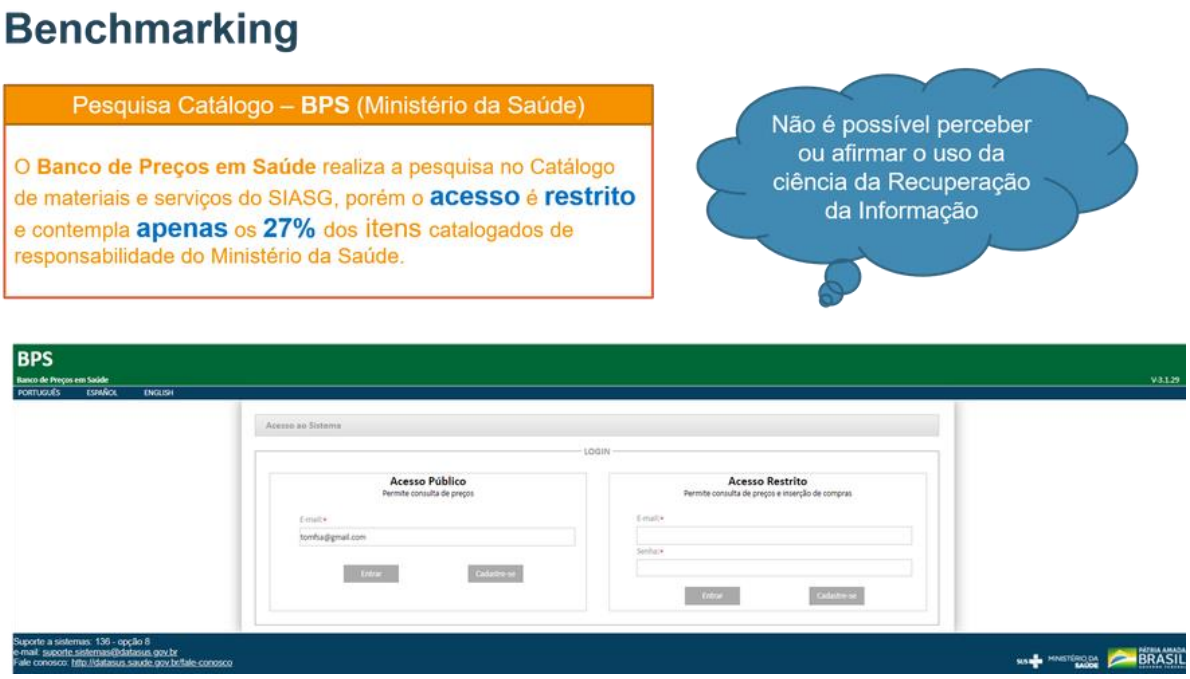
Aprofundando neste tema, vislumbro a possibilidade de automação de parte das tarefas, pois ainda sim são necessárias atividades supervisionadas, isto é, que demanda de análise crítica pelo capital humano. Destaco a busca de materiais e serviços no catálogo como uma atividade a ser automatizada, bem como a inserção dos itens padronizados no sistema.

Em contraponto, a atividade de definição dos Padrões de Descrição de Materiais – PDMs, exige um conhecimento holístico acerca do material ou serviço estudado. Isto porque envolvem preceitos legais, industriais, econômicos e de logística de compras públicas, que estabelecem um conjunto mínimo de características que especificam de forma a não produzir ambiguidade e discrepância de preços, em busca da vantajosidade para a Administração Pública.

Em busca de iniciativas relacionadas a busca facilitada no catálogo de materiais e serviços, encontramos o Banco de Preços em Saúde do Ministério da Saúde. Esta ferramenta possibilita a pesquisa apenas nas classes de materiais sob responsabilidade daquele Ministério, deixando por incompleto a abrangência de todo o catálogo.

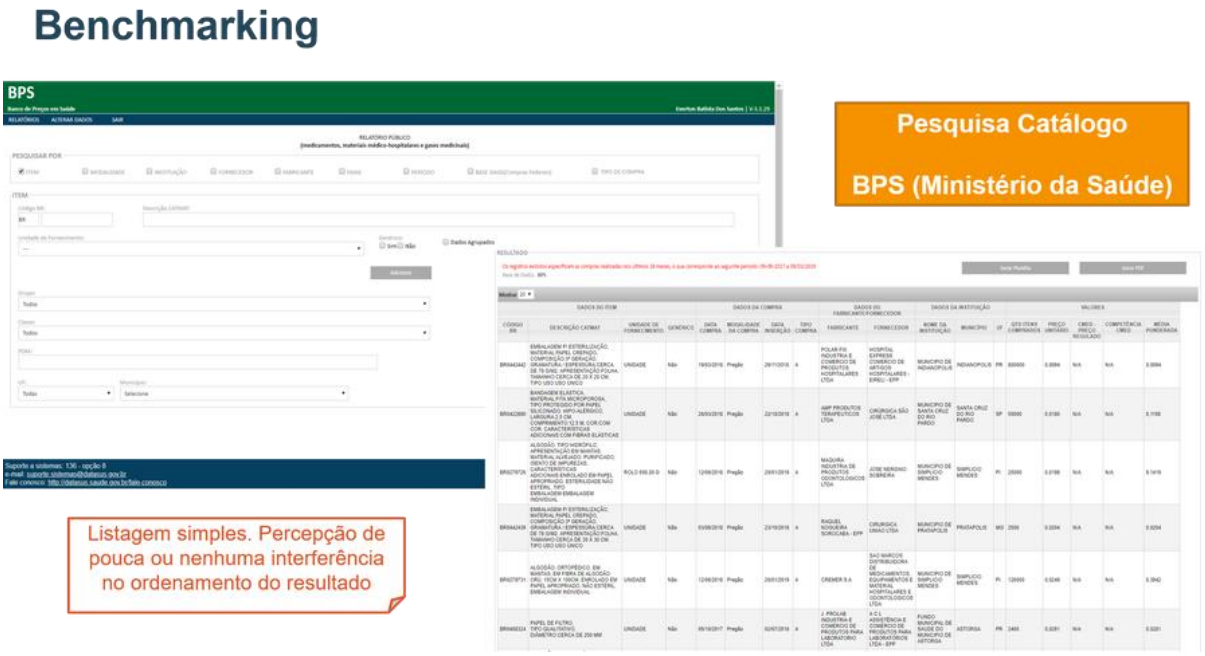
Abaixo apresento o resultado do estudo de benchmarking de soluções existentes no governo federal. O Banco de Preços em Saúde do Ministério da Saúde se mostra um tanto quanto restrito e atende uma percentual pouco expressivo diante do universo de itens existentes no catálogo de materiais e serviços do SIASG. Corrobora ainda a impossibilidade de afirmar com exatidão a presença da ciência da Recuperação da Informação nesta ferramenta.

Figura 11 – Banco de Preços em Saúde - BPS



Fonte: Apresentação Coordenação-Geral de Sistemas de Compras Governamentais

Figura 12 – Resultado busca no BPS



Fonte: Apresentação Coordenação-Geral de Sistemas de Compras Governamentais

Ao realizar uma busca nesta ferramenta, conforme exposto na figura acima, apresenta o resultado como uma listagem simples com pouca ou nenhuma percepção do uso de relevâncias ou ponderação de atributos no ordenamento dos itens listados. Ainda sob essa ferramenta, não podemos constatar o uso de ações cognitivas e

intuitivas disponíveis nos atuais mecanismos de buscas existentes, tais como o google, bing entre outros.

Neste contexto e em consonância com o estudo a ser abordado (RI), a ferramenta de pesquisa no catálogo de materiais e serviços do SIASG é requisito indispensável para melhorar a instrução do processo licitatório, especificamente na fase interna ou seja no planejamento da contratação, podendo dar segurança e qualificar as compras para o gestor público adquirente.

Nos próximos capítulos iremos abordar como o uso da ciência de Recuperação da Informação poderá auxiliar no aprimoramento da Ferramenta de Pesquisa de Itens no Catálogo de materiais e serviços do SIASG.

No capítulo 2 esboçamos as especificações estudadas da ciência de Recuperação da Informação. O capítulo 3 demonstra como utilizar essas especificações vinculadas ao escopo deste estudo de caso. No capítulo 4 apresento os resultados alcançados e elucido sobre as estratégias de indexação, de recuperação da informação e de ranqueamento abordadas. Em síntese, explicito no capítulo 5 o motivo da utilização das estratégias com foco no aprimoramento da Ferramenta de Pesquisa de Itens no Catálogo de materiais e serviços do SIASG.

2 Recuperação da Informação

Neste trabalho iremos tratar do uso da ciência da Recuperação da Informação (RI) diante da literatura existente e suas perspectivas quanto ao uso no aprimoramento na pesquisa de itens previamente catalogados no catálogo de materiais e serviços, módulo do Sistema de Administração de Serviços Gerais - SIASG.

Esta ciência se concentra essencialmente na oferta facilitada e simplificada, de informações de seu interesse, considerando as ações cognitivas dos serem humanos. Esta área pode ser estudada sob dois pontos de vistas: um centrado no computador e outro centrado no usuário. Enquanto a primeira consiste na construção de índices eficientes e algoritmos de ranqueamento, a segunda prima em estudar o comportamento e principais necessidades do usuário.

Dada a consulta do usuário, o objetivo maior do sistema de RI é apresentar informações que sejam úteis ou relevantes a ele. A ênfase está na recuperação de informações e não na simples recuperação de dados. Daí o problema da RI reside em recuperar todos os documentos relevantes e, ao mesmo tempo, descartar os menos irrelevantes.

Um ponto que merece destaque é a subjetividade da relevância dos documentos por ser um julgamento pessoal que depende do contexto envolvido. A dificuldade não está em extrair a informação dos documentos, mas sim a relevância que este documento representa para este usuário de acordo com sua necessidade.

Traduzir uma consulta na linguagem de RI, como uma máquina de busca, geralmente exige um conjunto de palavras que reproduz semântica para essa necessidade. Notamos então que os usuários ou estão buscando assuntos de seu interesse ou ainda que estejam navegando em assuntos correlatos. Iremos neste estudo preconizar apenas a tarefa do usuário em buscar ou consultar assuntos de seu interesse. Abaixo apresento os objetivos da Recuperação da Informação.

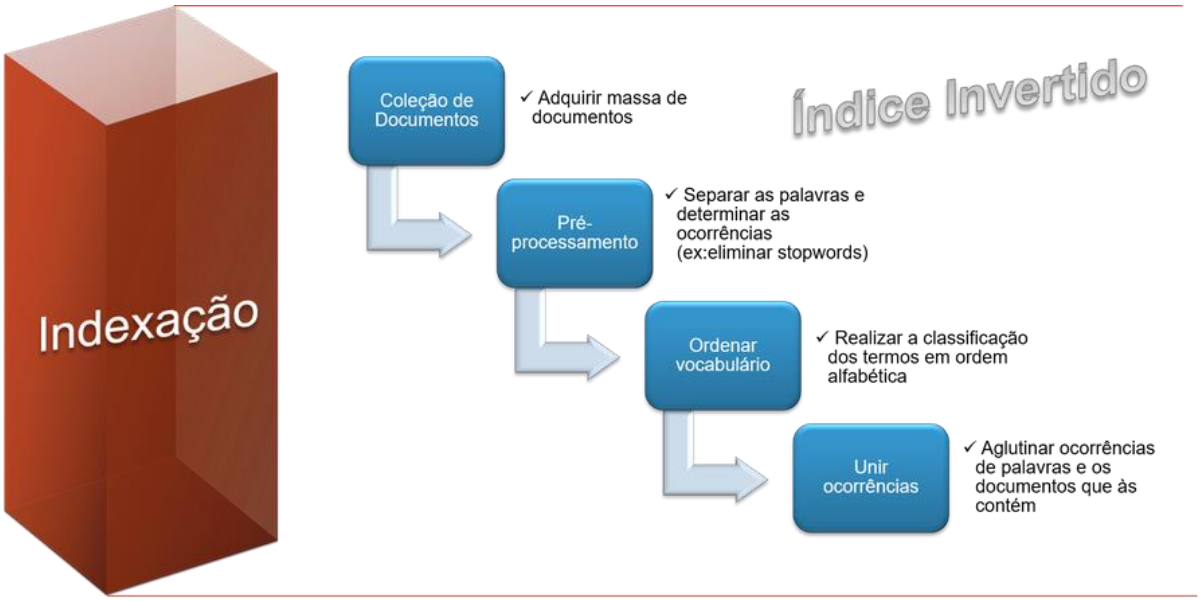
Figura 13 – Objetivos da Recuperação da Informação

Fonte: Apresentação Monografia

Dado o escopo supracitado, inicialmente irei explicitar a diferença entre recuperação de informação e recuperação de dados. Este último consiste na identificação de quais documentos contêm as palavras-chave da consulta do usuário, o que nem sempre é suficiente para satisfazer a necessidade do usuário. Já a recuperação da informação lida com a linguagem natural que quase sempre é não estruturada, mas que tem total sentido semântico com as palavras-chave buscadas.

Compondo a arquitetura de um sistema RI, temos que primeiramente obter a coleção de documentos e que, posteriormente é armazenada em um repositório central. Estes documentos precisam ser indexados para que os processos de recuperação e de ranqueamento sejam executados rapidamente. A estrutura mais utilizada na indexação é o índice invertido composto por todas as palavras distintas da coleção, e para cada palavra, os documentos que a contém, conforme esquematizado abaixo.

Figura 14 – Índice invertido esquematizado



Fonte: Apresentação Monografia

Esta coleção uma vez indexada, o próximo processo a ser iniciado é a recuperação. Este processo traz à tona e apresenta documentos que satisfaçam uma determinada consulta. A seguir esta consulta é analisada sintaticamente e expandida nas mais diversas variações das palavras-chave. Esta consulta expandida é então processada e utilizando-se do índice exibe um subconjunto de documentos.

Em seguida, este resultado dos documentos recuperados é então ranqueado e aqueles que constam no topo são retornados aos usuários. Este processo constitui-se da identificação dos documentos que têm a maior probabilidade de serem relevantes à consulta realizada. Esta tarefa, portanto, é a mais crítica de um sistema de RI, por ser inerentemente subjetiva a questão da relevância como já citamos anteriormente.

Avaliar a qualidade do conjunto-resposta é a chave para melhoria do sistema de RI. Exercer sistematicamente esta avaliação permite evoluir continuamente o algoritmo de ranqueamento e consequentemente a qualidade dos resultados apresentados aos usuários. Destaco que o procedimento avaliativo deve-se considerar não apenas máquinas ou algoritmos, mas também os sentimentos humanos através de, por exemplo, opiniões e comentários.

2.1 Processo de Indexação

Nesta seção iremos abordar o aspecto da eficiência da RI por meio do uso racional de recursos computacionais capazes de processar essas consultas. Como dito anteriormente, a RI tem como principal propósito ajudar usuários a encontrar informações de seu interesse, isto é, atingir a eficácia.

Diante da imensidão de terabytes de documentos existentes na internet é

imprescindível a utilização de processos de indexação mais eficientes, uma vez que as técnicas e algoritmos de busca evoluem e se sofisticam constantemente.

Mas o que é um índice? Um índice é uma estrutura construída para acelerar as buscas. É uma forma de obter tempos de respostas satisfatórias ou a níveis preconcebidos aceitáveis, porém sua sustentação é consideravelmente complexa. Nesta seara existem desafios que envolvem:

a) Tempo de indexação: necessário para criar o índice;

b) Espaço para indexação: utilizado durante a geração do índice;

c) Armazenamento do índice: espaço que armazena o índice após ser gerado;

d) Latência de consulta: intervalo entre a requisição e a geração da resposta; e

e) Taxa de transferência da consulta: número de consultas processadas por segundo.

Sobre os modelos e processos de construção deste processo, cito a técnica mais usual utilizada: O Índice Invertido. Ele é empregado mais frequentemente no modelo vetorial e booleano, sendo este modelo o mais frequentemente em sistemas de RI. Também não se afasta a possibilidade de ser implementado no modelo probabilístico e na maioria dos outros modelos que se alicerçam no ranqueamento dos documentos de acordo com as frequências das palavras.(5, p.339-40)

Outra forma de indexação, temos os Arquivos de Assinatura, cuja maior vantagem está no ganho de velocidade e no menor número de falsos positivos na busca ou recuperação de documentos. Sua estrutura está baseada em vetores binários, onde cada palavra no vocabulário da base de documentos é mapeada em uma função *hash*, ou seja, sua assinatura.

A assinatura de cada documento pode ser obtida com base nas assinaturas das suas palavras aplicando o operador <*or*> às assinaturas dos termos que aparecem no documento. Esta estrutura na maioria das vezes encontra aplicações em ambientes distribuídos presentes em aplicações para determinar se um índice remoto inclui respostas a uma dada consulta.

2.1.1 Índice Invertido

Este índice é um mecanismo orientado por palavras distintas para organizar uma coleção de textos. Ele é estruturado em dois elementos: o vocabulário e as ocorrências. O vocabulário ou dicionário representa o conjunto de todas as palavras diferentes do texto e para cada palavra o índice armazena os documentos que os contêm.

Com isso, podemos reconstruir o texto através do índice e por esse motivo é denominado *índice invertido*. Já as ocorrências representam a associação dos documentos a uma lista ou conjunto das listas de palavras existentes no vocabulário, sua posição no texto, frequência entre outras.(6, p.104)

Considera-se ainda que um campo ou atributo pode ter um conjunto relativamente pequeno de valores. Então uma zona ou bloco pode ser pensado como uma quantidade arbitrária e ilimitada de texto. Por exemplo, títulos de documentos e resumos são geralmente tratados como zonas ou blocos. Pode-se construir um índice invertido separado para cada zona ou bloco de um documento, para melhor sustentar as consultas.

A existência de parâmetros para cada campo, por exemplo, data de criação, nos permite selecionar ou restringir apenas os documentos correspondentes aos critérios estabelecidos na consulta. Alguns campos podem assumir valores ordenados entre outros mecanismos de pesquisa, bem como informações adicionais sobre determinado campo do índice que propiciarão eficiência neste processo.

Entre os métodos de construção deste índice temos:

1) Baseado em ordenação por blocos (BSBI): este método analisa os documentos em pares (termID, docID) até que um bloco esteja cheio, em seguida classifica e armazena na memória. Posteriormente mescla todos os blocos em um índice final.

2) Baseado em memória: este método adiciona o endereço diretamente na lista de ocorrências, ao invés de buscar todos os pares de (termID, docID) e depois ordená-los como no BSBI.

3) Distribuída: este método descentraliza o processo de indexação em várias máquinas. Geralmente aplicado em coleções de documentos de grandes dimensões, o que não se consegue, com eficiência, utilizando uma única máquina.

4) Dinâmica: este método baseia-se na necessidade que os vocabulários e as ocorrências estejam constantemente atualizados. Para satisfazer esta necessidade, podemos simplesmente reconstruí-lo novamente desde o início. Caso exija que novos documentos sejam incluídos rapidamente, uma solução é manter dois índices: um principal que armazena todo o índice sem estes documentos e outro auxiliar que armazena os novos.

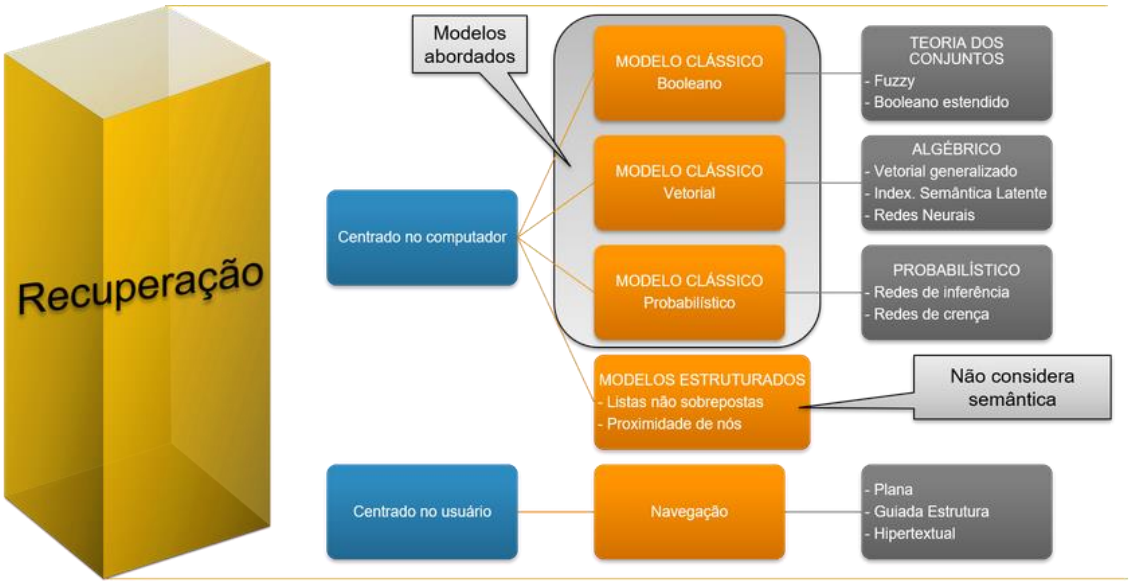
2.2 Processo de Recuperação

Nesta seção irei abordar os modelos clássicos de RI, denominados Booleano, vetorial e probabilístico. O primeiro, dizemos que se origina da teoria dos conjuntos e não associa peso algum aos termos de indexação, sendo simplesmente elementos de um conjunto.

O modelo vetorial é representado como vetores em um espaço de n dimensões. O modelo probabilístico basea-se na teoria das probabilidades para representações dos documentos e consultas. Sob esses dois modelos se aplicam nos termos de indexação pesos associados com o objetivo de melhorar o ordenamento dos documentos. (5, p.26)

Estes termos de indexação citados acima são palavras ou grupo de palavras consecutivas em um documento. São essas palavras pré-selecionadas que evidenciam os conceitos-chaves em um documento. São essencialmente substantivos, por possuírem um significado próprio. Adjetivos, advérbios e conectores geralmente funcionam como complementos. A figura abaixo ilustra a taxonomia de modelos de RI de forma abrangente, porém iremos abordar apenas os modelos clássicos supracitados.

Figura 15 – Taxonomia modelos de RI



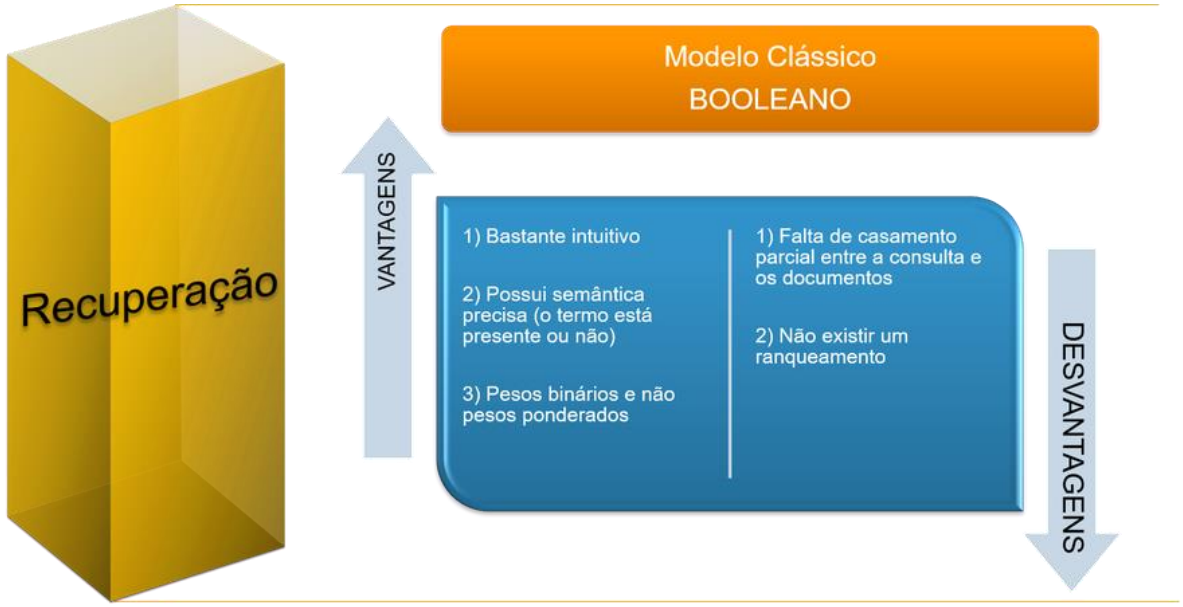
Fonte: Apresentação Monografia

Indo ao detalhe do modelo booleano observamos que este retrata a teoria da álgebra booleana, tendo como estratégia sedimentada no critério binário, ou seja, o documento é ou não é relevante à consulta sem graduação de importância. Além desta estratégia as palavras ou termos podem ser combinadas com os operadores booleanos *<and>, <or> e <not>*.

Os termos chamados de folhas, são aqueles isolados e os nós internos são denominados operadores booleanos, conforme citado acima. O algoritmo percorre a

árvore da consulta a partir das folhas por palavras isoladas no arquivo invertido e os nós internos definem os conjuntos de documentos recuperados. O Operador <*and*> faz a intersecção dos resultados restringindo o universo de resultados, enquanto o operador <*or*> realiza a união, ou seja, amplia estes resultados. Enquanto isso o operador <not> exclui dos resultados o termo referido na consulta.

Figura 16 – Modelo Booleano



Fonte: Apresentação Monografia

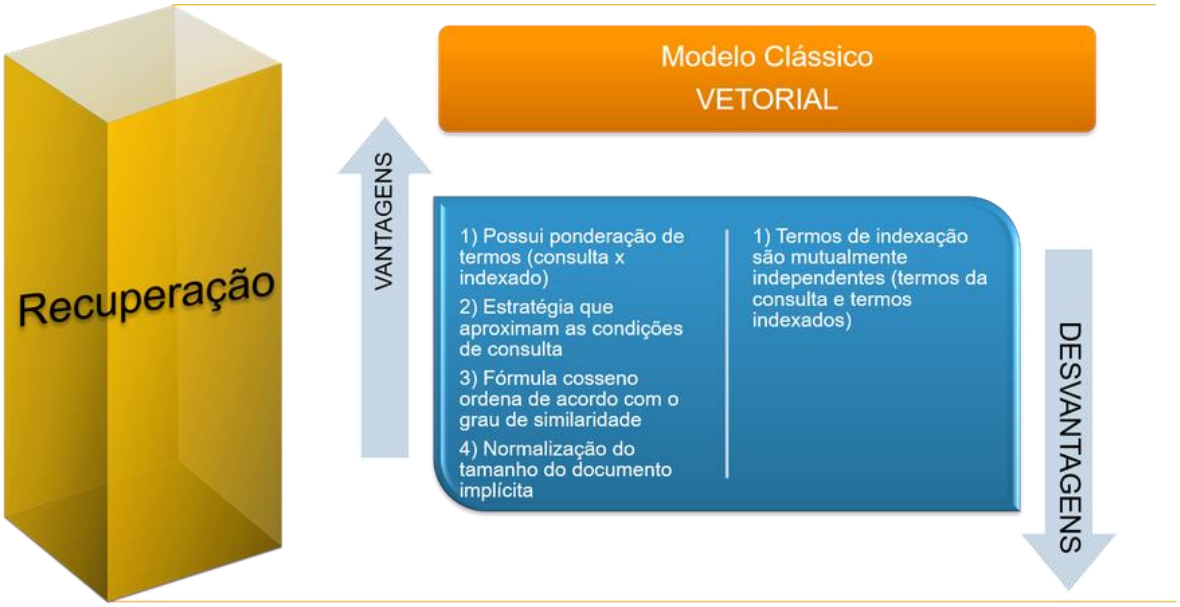
Conforme exposto na figura acima, observamos que a grande desvantagem está na inexistência de um raqueamento. Como a falta de congruência entre a consulta aos documentos é parcial pode ser relativamente pormenorizada àquela.

No modelo vetorial os termos são eixos no espaço, enquanto os documentos são pontos neste espaço. O propósito é ranquear documentos de acordo com seu grau de proximidade ou similaridade à consulta, bem como posicionar melhor os documentos relevantes dos não relevantes. Esta distância de similaridade é ordenada pelo o ângulo entre o documento e a consulta. Quanto mais próximo de zero melhor, pois este valor revela a máxima similaridade.

Quanto aos pesos utilizados no modelo vetorial, basicamente são os utilizados no esquema de ponderação TF-IDF. Este esquema é fundamentado na frequência do termo (TF) e a frequência inversa de documentos (IDF). O TF atribui um peso proporcional à frequência do termo, isto é, quanto mais frequente este termo ocorre no texto do documento, maior será seu valor.

Já a IDF fundamenta-se sob as noções das propriedades de exaustividade e de especificidade dos termos. Exaustividade pode ser interpretada como a abrangência que provê aos tópicos principais de um documento. A especificidade refere-se ao grau de riqueza que este termo descreve os tópicos deste documento.

Figura 17 – Modelo Vetorial



Fonte: Apresentação Monografia

Podemos observar que a combinação de ambos os esquemas de ponderação supracitados apresenta comportamento de equilíbrio entre um ao outro. Altos valores de TF tendem a estar associados e baixos valores de IDF, enquanto valores baixos de TF tendem a ser associados a valores altos de IDF. Em outras palavras, os termos mais discriminativos não são àqueles com maior valor de IDF, isto é, termos comuns como *stopwords (como exemplo: e, da, de, na, no entre outras)*, palavras estrangeiras e erros de grafia não são de grande valia para o ranqueamento. (5, p.43)

O modelo probabilístico aborda ou explora, sob a ótica da teoria da probabilidade, de modo a estimar a hipótese ou plausibilidade de um documento ser relevante para uma consulta. Chamamos esta tarefa de conjunto de documentos de resposta ideal, do qual se possa extrair ou inferir especificações das propriedades deste conjunto. O problema é que o universo destas propriedades não é conhecido e tudo que obtemos são termos de indexação cuja semântica às simbolizam.

Figura 18 – Modelo Probabilístico

Fonte: Apresentação Monografia

Para se conseguir uma estimativa inicial destas propriedades uma interação com o usuário é iniciada fomentando estímulos preciosos com intuito de melhorar a descrição probabilística do conjunto de resposta ideal.

Vale ressaltar a existência de duas categorias principais: modelos probabilísticos e modelos inferenciais. Os modelos probabilísticos, são baseados na evidência sobre quais documentos são relevantes para uma determinada consulta. Os modelos inferenciais aplicam conceitos e técnicas originadas de áreas tais como lógica e Inteligência Artificial.

No modelo probabilístico generalizado (Okapi system) aplica-se o Teorema de Bayes que é usado para o cálculo da probabilidade de um evento dado que outro evento já ocorreu, o que é chamado de probabilidade condicional. Mas para que esse teorema tenha eficácia é necessário ter alguma informação anterior, ou seja, preciso saber que um determinado evento já ocorreu e qual a probabilidade desse evento.

2.2.1 Processo de Ranqueamento

Este processo permite ou é equivalente a ordenar os resultados de busca de acordo com a sua probabilidade de relevância diante das necessidades dos usuários.

Se os documentos recuperados são ranqueados de forma decrescente de acordo com a probabilidade de relevância, então a efetividade do ranking será a melhor possível daquela coleção de documentos.

Para isso a utilização de índices parametrizados serve a dois propósitos. O primeiro nos permite indexar e recuperar documentos por metadados na linguagem

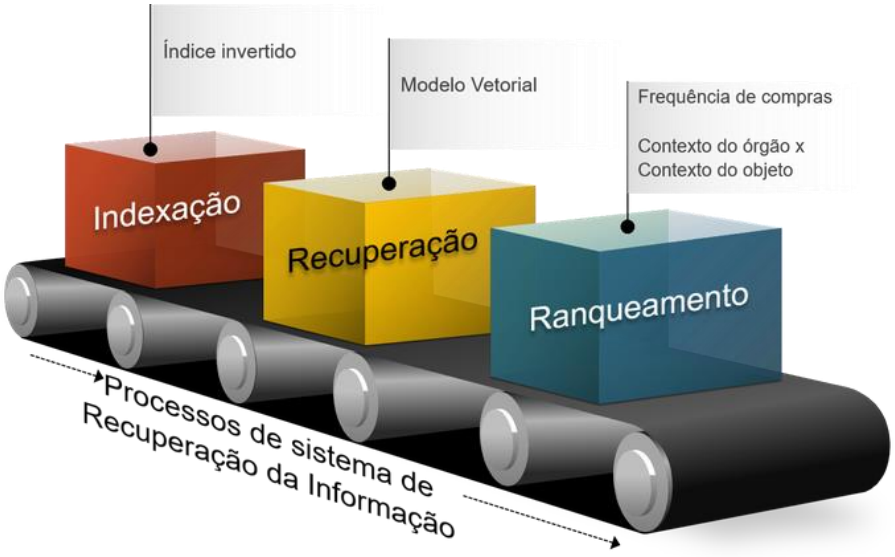
em que foi gerado. O segundo proporciona um meio simples para marcar e, assim, classificar os documentos em resposta a uma consulta.

A ideia de ponderar um termo buscado em um documento com base nas estatísticas de ocorrência do termo é de enorme importância. Visualizar cada documento como um vetor de pesos, permite calcular uma pontuação entre uma consulta e cada documento.

Alguns recursos de ranqueamento, tais como a função de ordenação BM25 são dependentes das consultas. Esse modelo, amplamente difundido, incorpora frequência de termos e normalização de tamanho. A pontuação mais simples para o documento *X* é somente os pesos IDF dos termos de consultas presentes no documento. Melhoramos o termo IDF ao fatorar na frequência de termos e tamanho do documento.

De maneira holística, abaixo apresento uma visão abrangente de um sistema de Recuperação da Informação, contendo alguns dos processos supracitados.(6, p.147)

Figura 19 – Visão holística de um RI



Fonte: Apresentação Monografia

3 Aprimoramento da Ferramenta de pesquisa

Atualmente o armazenamento dos materiais e serviços do SIASG estão estruturadas através do sistema gerenciador de banco de dados *ADABAS,* que consequentemente têm objetos e fatos bem definidos.

Como as categorias de objetos e fatos são bem definidos e mantidas de forma estruturada, consideramos atendida uma das propriedades que esta ferramenta deverá estar sedimentada, a representação das informações armazenadas.

Outras propriedades como o método de responder a uma solicitação de informação, a relação entre a consulta formulada e a satisfação do usuário e a definição para um sistema eficiente em relação à recuperação de dados e recuperação de informação que também devem ser contempladas.

O método de resposta a uma consulta de informação é direto, ou seja, existe ou não existe. Há determinados cenários em que os prováveis dados requeridos e outros em que a forma engessada de solicitação restringe sua pesquisa. Nesta conjuntura essa propriedade deve ter comportamentos de ajuda mútua, onde uma não exclui a outra, pelo contrário, devem ser utilizados concomitantemente.

A propriedade relação entre a consulta formulada e a satisfação do usuário pode ser determinístico, ou seja, está ou não satisfeito, segundo a recuperação de dados. De acordo com a recuperação da informação a probabilidade do grau de satisfação ser alcançada é o cerne da questão.

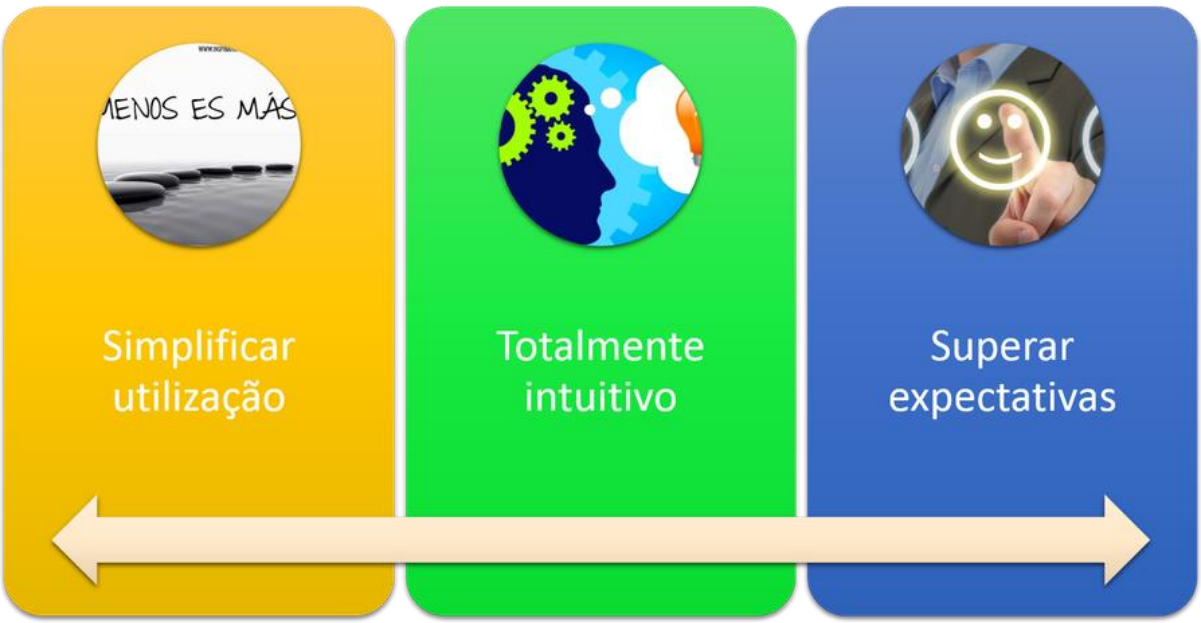
Por fim a eficiência pode ser observada quanto a relevância dos resultados apresentados diante da necessidade do usuário explicitado na consulta, aplicado em ambos os contextos.

Mas como ligar a noção de semântica com o conceito de relevância? A relação entre significado e relevância é demonstrada ao abordar que conteúdo significativo é relevante, e dessa maneira a relevância é uma propriedade da informação.

A semântica, assim como o conceito de relevância, apresenta várias nuances e matrizes teóricas. De maneira geral, pode-se afirmar que a semântica é a ciência da significação, mas sua apropriação na Web é da Semântica Formal, em que o significado é entendido como a relação entre as palavras e o mundo, entre o sentido e a referência.

Objetivando indicações sugestivas na concepção desta ferramenta, ressalto a importância da indexação invertida, explorando as técnicas de modelagem ou estratégias de ranqueamento existentes. Este processo proposto segue três pilares básicos:

Figura 20 – Pilares para aprimoramento da Ferramenta de Pesquisa



Fonte: Apresentação Coordenação-Geral de Sistemas de Compras Governamentais

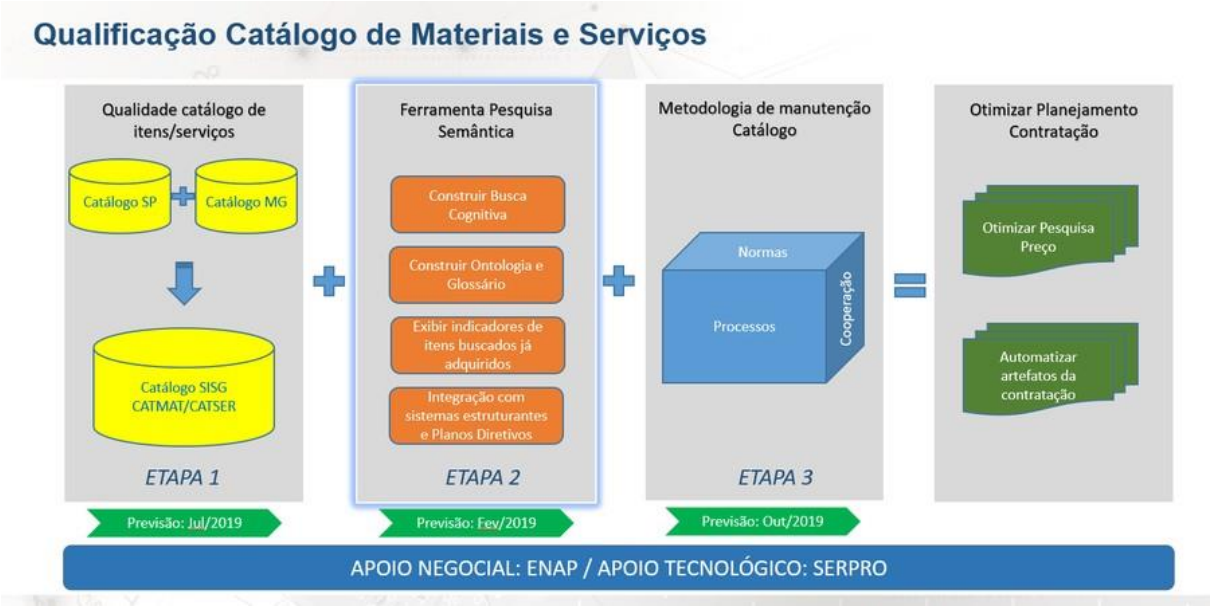
Simplificar utilização implica na busca do vocabulário através das palavras ou padrões presentes na consulta que são pesquisados no vocabulário do arquivo. Variâncias como acentuação, diferenciação entre maiúscula e minúscula, caracteres especiais entre outras não deverão figurar como obstáculos.

Totalmente intuitivo implica na manipulação de ocorrências nas quais as ocorrências são processadas para resolver a consulta. Operações típicas de correção ortográfica e a eliminação de *stopwords* são utilizadas sempre que apropriadas. A consulta expandida e modificada é processada de modo a obter o conjunto de documentos recuperados.

Superar expectativas implica na recuperação de ocorrências que retorna a lista de prováveis ocorrências de todas as palavras ou termos encontrados. Apresentar os bens e serviços mais relevantes dependendo do contexto e do comportamento do usuário pode ser uma alternativa a ser perseguida e um objetivo almejado.

Com o intuito de alcançar os objetivos supracitados, dividimos as possíveis soluções em pacotes conforme abaixo apresentado no *RoadMap* do projeto de modernização do catálogo de materiais e serviços contemplando ações que visam atender às expectativas de partes interessadas na condução do processo licitatório.

Figura 21 – RoadMap do projeto de Modernização do catálogo



Fonte: Apresentação Coordenação-Geral de Sistemas de Compras Governamentais

Na etapa 2 acima apresentada se constituem de várias soluções que de forma integrada irá propiciar ao agente de compras interações cada vez mais intuitivas e indispensáveis para a boa condução dos trabalhos necessários, sendo capaz de dar segurança ao gestor público e qualificar as compras públicas. Como objetivos a serem alcançados com este estudo cito a solução denominada “Construir Busca Cognitiva” cujo uso da ciência de Recuperação de Informação é uma premissa.

Até agora, vemos um documento como uma sequência de termos. Nesta seção iniciamos o estudo da atribuição de uma pontuação para um par (consulta, documento) que consiste em três ideias principais:

1) Introduzir índices paramétricos e de zona que servem dois propósitos. Primeiro, eles nos permitem indexar e recuperar documentos metadados, como o idioma em que um documento é gravado. Segundo, eles nos dão um meio simples para marcar e, assim, classificar documentos em resposta a uma consulta.

2) A seguir, desenvolvemos a ideia de ponderar a importância de um termo em um documento, com base nas estatísticas de ocorrência do termo.

3) Por fim mostramos que ao visualizar cada documento como um vetor de pesos, podemos calcular uma pontuação entre uma consulta e cada documento. Essa visão é conhecida como pontuação de espaço vetorial.

De fato, a maioria dos documentos possuem estruturas adicionais. Documentos digitais geralmente codificam, em forma reconhecível por máquina, certos metadados

associados a cada documento. Por metadados, queremos dizer formas específicas de dados sobre um documento, ou seja, índices paramétricos.

Esses metadados geralmente incluem campos como a data de criação e o formato do documento, como bem o autor e possivelmente o título do documento. Em seguida o processamento de consultas consiste em interseções dos conteúdos destas variáveis.

Existe um índice paramétrico para cada campo, como exemplo a data de criação, que nos permite selecionar apenas os documentos correspondentes a uma data especificada na consulta. Esses campos podem assumir valores ordenados, como no exemplo citado e o mecanismo de pesquisa pode suportar intervalos de consultas para tais valores.

As zonas são semelhantes aos campos, exceto quando o conteúdo de uma zona pode ser texto livre. Considerando que um campo pode ter um conjunto relativamente pequeno de valores, uma zona pode ser pensada como uma quantidade arbitrária e ilimitada de texto. Por exemplo, títulos de documentos e resumos são geralmente tratados como zonas.

Podemos construir um índice invertido separado para cada zona de um documento, para suportar consultas e isso tem o efeito de construir um índice de zona. Tendo em vista que o dicionário para construção deste índice paramétrico vem de um vocabulário fixo, deve-se estruturar este dicionário com qualquer vocabulário decorrente do texto dessa zona.

Explanando a ideia de ponderar a importância de um termo em um documento, com base nas estatísticas de ocorrência do termo, temos como próximos passos um documento ou zona que menciona um termo de consulta mais frequentemente tem mais a ver com essa consulta e, portanto, deve receber uma pontuação maior.

Um mecanismo de pontuação plausível é calcular uma pontuação que é a soma, nos termos da consulta, das pontuações da correspondência entre cada termo de consulta e o documento e atribuímos a cada termo em um documento um peso para esse termo, que depende do número de ocorrências do termo no documento.

A abordagem mais simples para calcular uma pontuação entre um termo de consulta t e um documento d, baseado no peso de t em d é atribuir o peso é igual ao número de ocorrências do termo t no documento d.

Este esquema de ponderação é referido como frequência do termo e é denominado por $tf_{t,d}$, com o termo e o documento subscritos em ordem. Para um documento d, o conjunto de pesos determinado pelos pesos tf acima ou qualquer função de ponderação que mapeie o número de ocorrências de t em d para um valor real positivo pode ser visto como um resumo quantitativo desse documento.

Uma questão relevante é: todas as palavras de um documento são igualmente importantes? Obviamente que não. Daí temos a idéia de palavras de parada ou *stopwords* que nada mais é do que palavras que decidimos não indexar nada, e não contribuem de forma alguma para recuperação e ponderação.

Por último, temos que o cálculo da pontuação entre uma consulta e cada documento, conhecida como pontuação de espaço vetorial, apresenta um passo crucial para as consultas. Mas como quantificar a similaridade entre dois documentos neste espaço vetorial?

Uma primeira tentativa pode considerar a magnitude da diferença do vetor entre dois vetores de documentos. Esta medida sofre de uma desvantagem: dois documentos com conteúdo muito semelhante podem ter uma diferença significativa de vetor simplesmente porque um é muito mais longo que o outro.

Assim, as distribuições relativas de termos podem ser idênticas nos dois documentos, mas as frequências de termos absolutos de um podem ser muito maiores que no outro. Para compensar o efeito do comprimento do documento, o modo padrão de quantificar a similaridade entre dois documentos d1 e d2 é computar a similaridade cosseno de suas representações vetoriais.

Deste arrazoado, dispomos que para aprimorar a Ferramenta de Pesquisa necessitamos de metadados, por possuírem estruturas adicionais dos documentos e classificá-los em resposta a uma consulta. Estes campos poderão ser utilizados na estratégia de ponderação da frequência do termo, com base nas estatísticas de ocorrências no documento. E por fim quantificar a similaridade entre dois documentos através do cálculo do cosseno de suas representações no espaço vetorial.

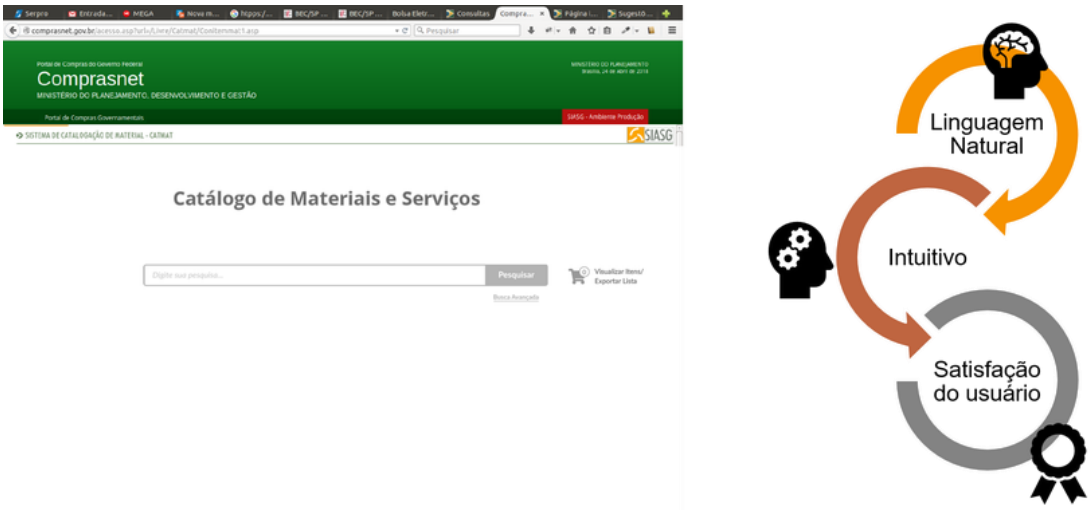
4 Resultados e Discussão

4.1 Resultados

Diante do exposto até o momento apresento um esboço visual do que poderia conter e como poderia ser disposta as informações e objetos quanto a ferramenta de pesquisa do catálogo de materiais e serviços do SIASG.

Primeiro ponto a destacar é a verossimilhança com os sistemas de e-commerce existentes para que o comportamento e a experiência do usuário sejam intuitivos e a utilização desta ferramenta não requeira muito esforço. Destaco também a possibilidade de o usuário manter suas escolhas para posterior instrução processual, requisito cada vez mais exigido das unidades de consultorias jurídica e órgãos de controle externo.

Figura 22 – Proposição da ferramenta



Fonte: http://www.comprasnet.gov.br/catalogo

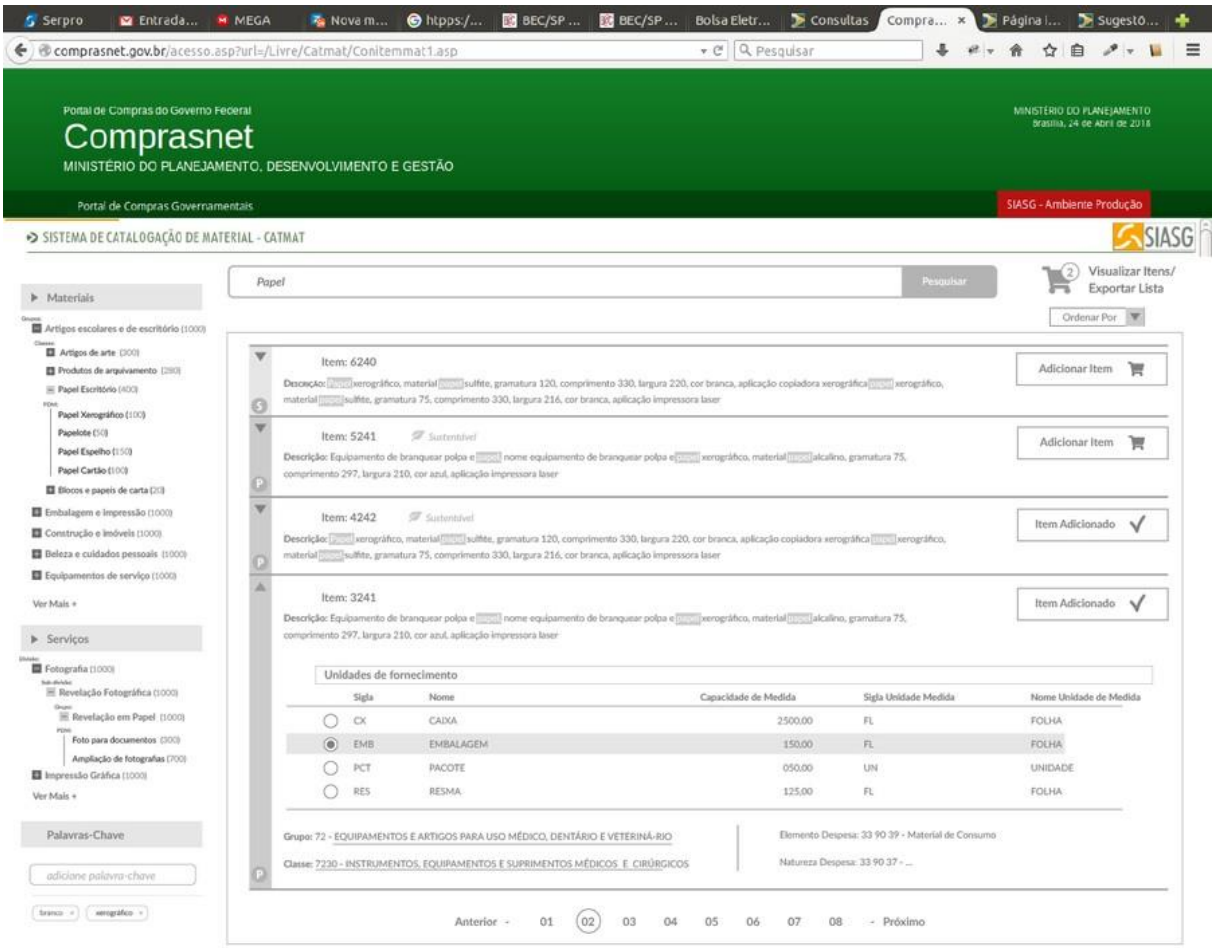
Sugerimos ainda a apresentação da árvore de navegação, pelas categorias estruturais existentes e conhecidas, denominadas classes e PDMs, onde se comportam como filtros. Através destes filtros pré-selecionáveis os usuários poderão com um simples clique restringir o escopo dos resultados apresentado.

Com o processo de categorização podemos classificar documentos em categorias pré-definidas. Sua maior aplicação tem sido para atribuir categorias a documentos e posteriormente utilizar estas categorias para suportar recuperação e filtragem de informação. As categorias são definidas através de um pequeno conjunto de características e tendem a ser mais estáticas que os perfis em filtragem de informação.

Para delimitar ainda mais o escopo dos itens apresentados, dispomos da opção “palavra-chave” onde este termo necessariamente estaria presente nos valores das caraterísticas dos materiais e serviços catalogados. Abaixo apresento o “*wireframe*“

criado para de visualização do proposto.

Figura 23 – Wireframe da ferramenta de pesquisa



Fonte: Apresentação Coordenação-Geral de Sistemas de Compras Governamentais

4.2 Discussão

4.2.1 Estratégia de indexação

Uma abordagem para o problema em questão é varrer a base de dados de compras públicas, buscando palavras chaves e extrair os contextos onde ocorrem tais palavras a informação necessária. Corresponde a precisão com que uma palavra é localizada pelo índice. Esta granularidade pode ser implementada nas seguintes amplitudes:

1. Grossa (*coarse-grained)* que identifica apenas um bloco de texto, requer menos memória, porém, exige da consulta mais processamento em cada bloco e aumenta o número de falsos positivos.

2. Fina (*fine-grained*) que identifica cada palavra, busca por frases e buscas aproximadas são mais eficientes, proximidade pode ser tratada sem ir direto ao texto, porém, requer mais memória (uma ordem de grandeza).

Como neste trabalho não trataremos coleções muito grandes como a "*web*", a indexação distribuída sobre "*clusters*" de computadores com centenas ou milhares de máquinas não é necessária. Apresento abaixo de forma esquematizada esta estratégia.

4.2.2 Estratégia de busca

Para consultas com apenas uma palavra, a busca retorna simplesmente a lista de ocorrência desta palavra. Já para consultas acompanhadas de um contexto se tornam mais complexas. Neste cenário o processo envolve a recuperação não só dos identificadores dos documentos para cada palavra, bem como suas as respectivas posições, além de realizar a intersecção entre todos os documentos que as relaciona. Estas atividades são denominadas de endereçamento em blocos.

Ainda sobre as consultas contextualizadas, temos a busca com proximidade das palavras que se assemelha com o processo supracitado. Inicialmente são selecionados os documentos em que todas as palavras da consulta ocorrem, porém em um contexto que satisfaz as restrições de proximidade definidas no algoritmo. Apesar de sua simplicidade, o modelo vetorial é uma boa estratégia e consegue bons resultados em coleções genéricas, o que não se aplica ao projeto em questão.

Apesar de sua simplicidade, o modelo vetorial é uma boa estratégia e consegue bons resultados em coleções genéricas, o que não se aplica ao projeto em questão. A expansão de consultas é um caminho para solucionar o problema em que nem sempre se usam as mesmas palavras para descrever o mesmo conceito.

Para expandir uma consulta é preciso buscar palavras com significados semelhantes aos termos da busca e acrescentar tais palavras como sugestão ao termo original com o objetivo de melhorar a assertividade da mesma.

Duas abordagens podem ser adotadas: o uso de dicionários de sinônimos e o uso de palavras que co-ocorrem com os termos das consultas em documentos da coleção. No caso de dicionários de sinônimos os resultados obtidos não são, em geral, muito bons. Melhorias significativas foram alcançadas quando se considerou análise automática de termos que co-ocorrem em documentos da coleção.

4.2.3 Estratégia de ranqueamento

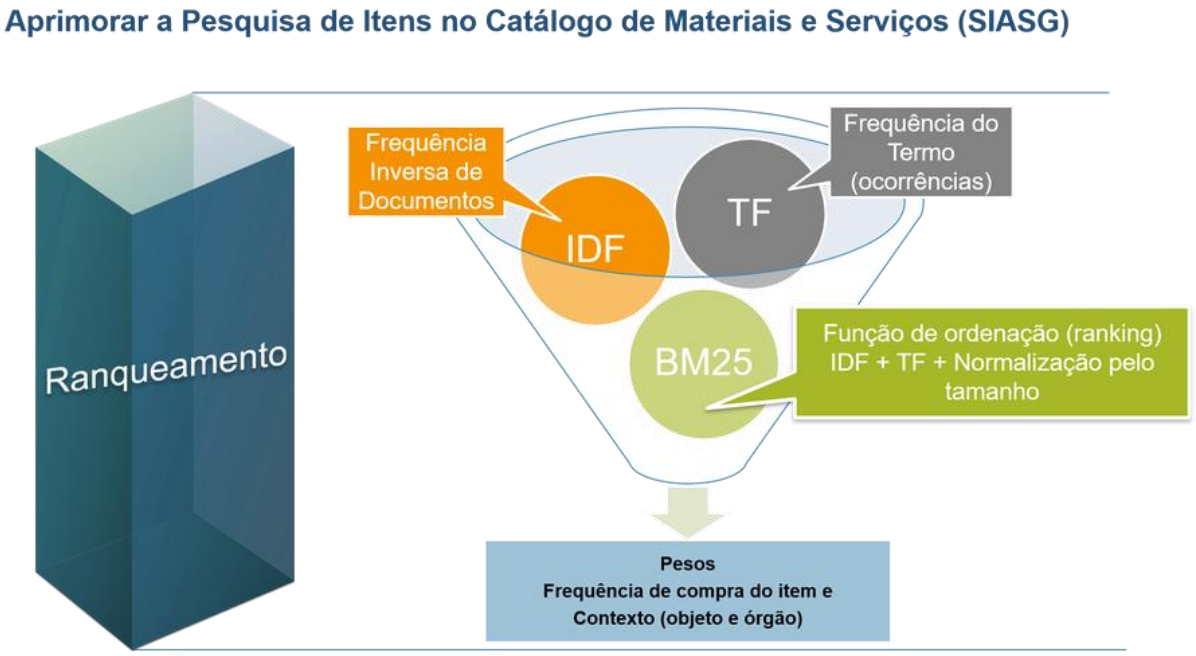
A relevância, sendo algo subjetivo e inerente ao julgamento do usuário, dependerá da interação do mesmo com o sistema e, principalmente, ao que de fato ele espera recuperar em sua busca.

Neste estudo sugerimos a utilização de ranqueamento do contexto em que o licitante se encontra quanto ao órgão e objeto da licitação. Ex: Hospital comprando me-

dicamentos, o mesmo hospital comprando materiais de expediente.

A busca por algo básico e simples, nos remete ao uso do espaço vetorial com pesos TF-IDF. Se quer algo com desempenho excelente, utilize modelos de linguagem ou BM25 com parâmetros bem configurados. No caso em que o meio-termo satisfaça, aplique o BM25 ou modelos de linguagem sem, ou com apenas um parâmetro de configuração. Nossa sugestão neste trabalho sedimenta-se justamente neste último cenário, representado na figura abaixo.

Figura 24 – Proposição de ranqueamento



Fonte: Apresentação Monografia

A distância semântica entre a real necessidade dos usuários e o que ele expressa na consulta formulada é provocada principalmente pelo limitado conhecimento do usuário no universo da pesquisa. Além do problema de formulação da consulta, o grande volume de dados presentes atuais implica na dificuldade em apresentar os resultados para o usuário.

O modelo de probabilístico foi desenvolvido especialmente para combinar pesos de indexação probabilística com pesos baseados nos pesos dos termos das consultas, por exemplo, a relevância da opinião ou comentário. Sua vantagem principal é a sua adequada portabilidade em diferentes esquemas de indexação probabilística.

Existem vários outros modelos sendo desenvolvidos em RI, entretanto pode-se concluir que os Modelos Probabilísticos são mais importantes e satisfatórios em processar informações incertas e imprecisas que é característica dos usuários da RI.

5 Conclusão

Em bases de dados específicas, como no catálogo de materiais e serviços do SIASG, o processo de recuperação de informação poder ser facilmente desenvolvido pautado nos modelos clássicos que melhor harmonize à necessidade e objetivos definidos. Este fato aplica-se por todo o universo de compras públicas ser conhecido e estar acessível seguindo os mesmos padrões.

Temos preocupação, e especial atenção, em recuperar um número suficiente de registros relevantes. Isto implica necessariamente em mitigar a recuperação de registros irrelevantes, evitando ao máximo recuperar um número excessivo e minimizar um número ínfimo de registros.

Com relação à interação com o usuário, expressar uma necessidade de informação é uma tarefa difícil. Existe uma distância semântica entre a real necessidade dos usuários e o que ele expressa na consulta formulada.

O objetivo principal é desenvolver mecanismos para apresentar visualmente os dados ao usuário, bem como permitir que este explore os dados de forma amigável.

Referências

1 BRASIL. DECRETO Nº 1.094, DE 23 DE MARÇO DE 1994. **Dispõe sobre o Sistema de Serviços Gerais (SISG) dos órgãos civis da Administração Federal direta, das autarquias federais e fundações públicas, e dá outras providências**, Brasília, p. 2 – 3, 1994.

2 BRASIL. DECRETO-LEI Nº 200, DE 25 DE FEVEREIRO DE 1967. **Dispõe sobre a organização da Administração Federal, estabelece diretrizes para a Reforma Administrativa e dá outras providências**, Brasília, 1967.

3 MINISTÉRIO DO PLANEJAMENTO, DESENVOLVIMENTO E GESTÃO. **Referencial de Governança e Gestão do Sistema de Serviços Gerais – SISG**, Brasília, 2017.

4 TRIBUNAL DE CONTAS DA UNIÃO. Acórdão 2593/2017. **Acórdão 2593/2017 - Plenário**, Brasília, 2017.

5 BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação**: Conceitos e Tecnologia das Máquinas de Busca. Porto Alegre: Bookman, 2013.

6 MANNING, C. D.; RAGHAVAN, P.; SCHUTZE, H. **Introduction to Information Retrieval**. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719, 9780521865715.