Universidade Católica de Brasília

Pró-Reitoria de Pós-Graduação e Pesquisa

Programa de Pós-Graduação *Strictu Sensu* em Gestão do Conhecimento e Tecnologia da Informação

Sérgio Dagnino Falcão

Proposição, Aplicação e Avaliação de um Método de Classificação Temática em Bases de Dados Textuais Indexadas com Auxílio de Vocabulários Controlados

Brasília-DF, 2003

Sérgio Dagnino Falcão

Proposição, Aplicação e Avaliação de um Método de Classificação Temática em Bases de Dados Textuais Indexadas com Auxílio de Vocabulários Controlados

Dissertação apresentada ao Programa de Pós-Graduação *Strictu Sensu* em Gestão do Conhecimento e Tecnologia da Informação da Universidade Católica de Brasília como requisito parcial para obtenção do Título de Mestre em Gestão do Conhecimento e Tecnologia da Informação.

Prof. Dr. Paulo Sérgio Vilches Fresneda
Orientador

Brasília-DF, 2003

Ficha Catalográfica

Falcão, Sérgio Dagnino

Proposição, Aplicação e Avaliação de um Método de Classificação Temática em Bases de Dados Textuais

Indexadas com Auxílio de Vocabulários Controlados. Brasília: UCB, 2003.

125 f.: il.

Dissertação (mestrado) – Universidade Católica de Brasília. Programa de Pós-Graduação em Gestão do Conhecimento e Tecnologia da Informação, Brasília, BR – DF, 2003. Orientador: Fresneda , Paulo Sergio V.

1. Gestão do Conhecimento 2. Classificação temática. 3. Descoberta de Conhecimentos.

Folha de Aprovação

Sérgio Dagnino Falcão

Proposição, Aplicação e Avaliação de um Método de Classificação Temática em Bases de Dados Textuais Indexadas com Auxílio de Vocabulários Controlados

Esta Dissertação foi considerada aprovada para obtenção do grau de Mestre em Gestão do Conhecimento e Tecnologia da Informação no Programa de Pós-Graduação - Mestrado *Strictu Sensu* em Gestão do Conhecimento e Tecnologia da Informação da Universidade Católica de Brasília.

Brasília, 25 de fevereiro de 2003

_____

Professor Doutor Paulo Sergio Vilches Fresneda – Orientador
Universidade Católica de Brasília

_____

Professor Doutor Gentil José de Lucena Filho
Universidade Católica de Brasília

_____

Professor Doutor. Jaime Robredo
Universidade de Brasília

Dedicatória

Para a pequena e bela Tatiana.

Agradecimentos

Aos meus pais, Jayme e Ana Maria, que sempre procuraram me transmitir elevados valores e que ofereceram apoio e incentivo em todos os momentos da vida,

À minha esposa Susana, que soube compreender a importância deste desafio que agora se encerra, demonstrando afeto e compreensão,

Ao meu amigo Carlos, companheiro de todas as horas, que me incentivou a embarcar junto nesta aventura,

Aos meus colegas da Câmara dos Deputados que se mostraram disponíveis e prontos a colaborar, especialmente Vilma Pereira, que acompanhou todos os passos do desenvolvimento do método de classificação temática desenvolvido neste trabalho, Ricardo Oliveira dos Santos, que me auxiliou na extração dos dados para a base de dados laboratório e todos os demais integrantes da equipe Sileg,

Ao meu orientador Paulo Sergio Vilches Fresneda, que soube conciliar suas inúmeras atividades com a condução deste trabalho, compartilhando sua experiência e seus conhecimentos.

Epígrafe

Nas favelas, no Senado,

Sujeira para todo lado.

Ninguém respeita a Constituição,

Mas todos acreditam no futuro da Nação.

Que País é este?

Renato Russo

Resumo

A falta de informações foi freqüentemente apontada como fator limitante para a tomada de decisões de forma racional. No entanto, vivemos hoje a Era da Informação e há diversos exemplos de perda de eficiência e de níveis de produtividade abaixo do esperado por parte dos trabalhadores do conhecimento, devido, em parte, ao excesso de informações a que são expostos cotidianamente.

Este trabalho descreve a proposição, a aplicação e a avaliação de um método de classificação temática em uma base de dados com discursos proferidos por deputados federais no Plenário da Câmara dos Deputados da República Federativa do Brasil entre outubro de 2000 e outubro de 2002 e que foi indexada com auxílio de um vocabulário controlado. O método desenvolvido utiliza os recursos de um banco de dados relacional para atribuir temas aos discursos, por meio da análise dos descritores utilizados na indexação. Os 10.627 discursos foram agrupados em 14 temas; foram feitas subdivisões da classificação temática por região geográfica e por partido político do orador, também ao longo do tempo e a possível correlação entre os temas.

A aplicação do método de classificação temática foi avaliada por 36 funcionários da Câmara dos Deputados, envolvidos profissionalmente com o assunto da base dados analisada, os quais foram entrevistados por meio de questionário. Verificou-se que a aplicação do método proposto permite contextualizar as informações armazenadas, agregando-lhes valor através da atribuição de significado e propósito. Constatou-se ainda que a aplicação do método possibilita a

descoberta de conhecimentos através da identificação de padrões válidos, novos e potencialmente úteis nas informações armazenadas. Foram também relacionadas possíveis utilizações da aplicação do método na base de dados completa ou em outras bases de dados.

Palavras-chave

Gestão do Conhecimento, Sobrecarga de Informações, Descoberta de Conhecimento em Textos, Classificação Temática.

Abstract

The lack of information has been often appointed as limiting factor with respect to the decision making process in rational ways. However, we live today in the Information Age and there are many examples of loss of efficiency and levels of productivity below expected by the so called knowledge workers, due to, in part, the information overload they are exposed to.

This work describes the proposal, deployment and evaluation of a thematic classification method in a full text database containing speeches pronounced by representatives in the plenary assembly of the Chamber of Deputies of the Federative Republic of Brazil, between October 2000 and October 2002, and that was indexed with aid of a controlled vocabulary. The method developed uses the features of a relational data base to assign subjects to the speeches, through the analysis of the terms used in the indexation field. The 10,627 speeches had been grouped into 14 main subject categories and it had been made subdivisions of thematic classification by geographic region and political party of the speaker, and also by the date of the speech as well as the possible correlation between subjects.

The deployment of the method of thematic classification was evaluated by 36 employees of the Chamber of Deputies, professionally involved with the subject of the database being analyzed, who had been interviewed by means of questionnaires. It was verified that the deployment of the proposed method allows the contextualization of the stored information, adding value to it through the assignment of meaning and intention. It was also evidenced that the deployment of

the proposed method makes possible the process of knowledge discovery through identification of valid, new and potentially useful patterns in the stored data. Possible uses of the deployment of the method in the complete database or other databases had been related.

Keywords

Knowledge Management, Information Overload, Knowledge Discovery in Texts, Subject Classification.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

CEDI	Centro de Documentação e Informação (Câmara dos Deputados)
CENIN	Centro de Informática (Câmara dos Deputados)
CONLE	Consultoria Legislativa (Câmara dos Deputados)
DETAQ	Departamento de Taquigrafia, Revisão e Redação (Câmara dos Deputados)
DILEG	Diretoria Legislativa (Câmara dos Deputados)
GE	Grande Expediente
KDD	*Knowledge Discovery in Databases*
MP3	Formato de áudio digital, que permite grande compactação, definido pelo Moving Picture Experts Group.
PE	Pequeno Expediente
RI	Recuperação de Informações
SGM	Secretaria-Geral da Mesa (Câmara dos Deputados)
SITAQ II	Sistema Informatizado de Registro de Notas Taquigráficas
TDM	*Text Data Mining*
WWW	*World Wide Web*

Proposição, Aplicação e Avaliação de um Método de Classificação Temática em Bases de Dados Textuais Indexadas com Auxílio de Vocabulários Controlados

Sumário

1. INTRODUÇÃO

"O New York Times de domingo contém mais informações factuais numa única edição do que todo o material escrito com que contavam os leitores do Séc. XV".
Thomas Davenport e John Beck, *A Economia da Atenção* (DAVENPORT; BECK, 2001)

"Estamos nos afogando em informação, mas sedentos de conhecimento".
John Naisbit, autor do livro *Megatrends* (BAX; SOUZA, 2001)

Um fato inerente à época em que vivemos é o rápido crescimento do volume de informações disponíveis, notadamente as armazenadas em formato digital e, portanto, manipuláveis por computadores (LYMAN, 2000).

Desde meados do século passado, os sistemas de informação implantados com o apoio da informática tiveram como foco principal o tratamento das informações estruturadas, normalmente armazenadas sob a forma de valores numéricos ou alfanuméricos de tamanho predefinido. Com o passar do tempo, tornou-se mais difundido o armazenamento de informações não estruturadas como textos, sons e imagens estáticas ou em movimento (DAVENPORT, 2000). O aparecimento e a popularização da Internet e, particularmente da World Wide Web (WWW) na segunda metade da década de 1990, ocasionaram um crescimento explosivo do volume de informações não estruturadas, armazenadas e disponibilizadas para qualquer pessoa com acesso à Rede Mundial. (DAVENPORT; BECK, 2001).

Apesar de essas informações serem armazenadas em computadores cada vez mais velozes, e de surgirem poderosas ferramentas de pesquisas, como as de alguns sites de busca, muitas pessoas ainda sentem grande dificuldade ao tentar localizar as

informações que buscam dentro deste universo em expansão, que é a WWW. Essa mesma dificuldade pode ocorrer no âmbito das organizações ao se procurarem informações relevantes, estejam elas na forma de mensagens trocadas via correio eletrônico, ou armazenadas em bases de dados de textos integrais (DAVENPORT, 2000).

Para tentar solucionar essa questão, novos processos e novas técnicas foram propostos e estão sendo aperfeiçoadas com o objetivo de auxiliar a pesquisa de informações não estruturadas, bem como a descoberta de conhecimentos, em bases de dados textuais.

Este trabalho se propõe a sugerir e a aplicar uma técnica de descoberta de conhecimentos em textos, através da classificação temática, e também a avaliar se essa técnica permite a contextualização das informações armazenadas através da atribuição de significado e propósito.

1.1.Definição do Tema

Em seu livro *The Social Life of Information,* John Seely Brown e Paul Duguid afirmam que "a escassez de informação sempre pareceu ser um dos problemas fundamentais da humanidade. A falta de informações foi freqüentemente apontada como fator limitante para a tomada de decisões de forma racional" (BROWN; DUGUID, 2000).

Hoje vivemos na Era da Informação (STEWART, 1998). O que, um dia, foi fome se transformou em indigestão. Se o desafio era reter ou ter acesso ao maior

volume possível de informações, hoje esse desafio se transformou em preocupação de como lidar com o volume crescente a que se tem acesso atualmente (BROWN; DUGUID, 2000).

. Se antes o senso comum dizia que "informação é poder", hoje se sabe que isto só é verdade na medida em que se cria novos **significados** a esta informação, distinguindo-se o que realmente é importante e significativo na enxurrada de informações. Bill Jensen [JENSEN] defende a idéia a importância da simplicidade, que é a arte de se tornar claro o que era complexo. Só a simplicidade poderia aumentar a eficiência do trabalhador do conhecimento. As tarefas do trabalhador manual têm sido estudadas e otimizadas nestes últimos cem anos por pessoas como Ford e Taylor, porém a produtividade de um trabalhador do conhecimento que é solicitado freqüentemente a fazer escolhas e a tomar decisões só tem

Certamente a rápida popularização da Internet ocorrida a partir de 1995, especialmente do uso do correio eletrônico e da WWW, colaborou de forma decisiva para o grande aumento do volume de informações publicadas e disponíveis no mundo, tornando a Rede Mundial um acervo aparentemente infinito de textos, sons e imagens. Entretanto, justamente devido ao tamanho deste acervo, a experiência de pesquisar na Internet através dos sites de busca pode trazer sentimentos semelhantes aos dos cientistas do projeto SETI[1] da Universidade da Califórnia em Berkeley, que tentam (até agora sem sucesso) localizar algum sinal de vida extraterrestre inteligente procurando padrões a partir da análise de um infindável fluxo de ruídos eletromagnéticos vindos do espaço (BROWN; DUGUID, 2000).

[1] SETI: Search for Extraterrestrial Intelligence (setiathome.ssl.berkeley.edu).

Os exemplos relacionados a seguir retratam de forma clara a velocidade explosiva de crescimento das informações disponíveis, principalmente sob a forma de páginas WWW e mensagens de correio eletrônico, e alguns graves inconvenientes causados por este aparente excesso de oferta de informações:

- No estudo *How Much Information?* os pesquisadores Peter Lyman e Hal Variant da Universidade da Califórnia em Berkeley procuraram medir o volume de informações publicado a cada ano em todo o mundo, nas diversas mídias como papel, filmes, meios ópticos e magnéticos. De acordo com suas estimativas, apenas em 1999 foram produzidos cerca de 1,5 Exabytes (10^{18} bytes), ou 1,5 milhão de Terabytes (10^{12} bytes), a uma taxa anual de crescimento de 50%. Cada Terabyte representa o equivalente ao texto de um milhão de livros. Considerando-se a população mundial de 6 bilhões de habitantes, foram publicados 250 Megabytes (10^{6} bytes) por habitante do planeta naquele ano. Deste volume de informações, 93% estavam armazenadas em formato digital (LYMAN, 2000).
- O site *Search Engine Watch*[2], especializado na análise de sites de busca, estima que o portal *Google*[3] atingiu a marca de 1 bilhão de páginas Web indexadas em junho de 2001. Em dezembro do mesmo ano, este site havia chegado à marca 1,5 bilhão de páginas e, no final de 2002, mais de 3 bilhões de páginas já estavam disponíveis para

[2] searchenginewatch.com/reports/sizes.html
[3] www.google.com

busca, além de mais de 700 milhões de mensagens de grupos de discussão e 330 milhões de imagens.

- Em 2001 a Intel, fabricante de microprocessadores, estimou que seus servidores internos de correio eletrônico tinham que lidar com cerca de três milhões de mensagens por dia, e que alguns funcionários recebiam cerca de 300 mensagens a cada 24 horas, o que lhes consumia cerca de 2h 30min diárias para tratar este volume de informações (OVERHOLT, 2001).
- Segundo artigo publicado na revista *Oracle Magazine*, algumas empresas de biotecnologia, como a *Celera Genomics*, gerenciam um volume de informações estimado em 10 Terabytes, o que é comparável ao tamanho do acervo completo da maior biblioteca do mundo, a Biblioteca do Congresso Norte-Americano, avaliado em 12 Terabytes (SPICER, 2003).

Esses são apenas alguns exemplos que comprovam o enorme volume de informações a que as pessoas estão expostas. Por outro lado, a crescente disponibilização de informações é bem-vinda e não deve ser vista, por si só, como algo indesejável.

Os especialistas em Gestão do Conhecimento Thomas Davenport e Laurence Prusak, no livro *Conhecimento Empresarial*, afirmam que "a codificação das informações (registro de forma escrita) converte o conhecimento em formatos acessíveis, úteis e aplicáveis em novas situações". Esses autores citam como exemplo o sistema legal de um país, no qual as leis e decisões que criam

jurisprudência são organizadas em forma de textos e publicadas em papel e também, mais recentemente, em meios eletrônicos. Esse acervo textual representa apenas parte daquilo que é a lei e de como ela é praticada, ou seja, o conhecimento explícito; ele não abrange os conhecimentos tácitos de advogados e juízes. Todavia, esse material codificado incorpora e torna acessível significativa parcela do conhecimento legal articulado (DAVENPORT; PRUSAK, 1998p. << >>).

Por determinação expressa no art. 37 da Constituição Federal do Brasil de 1988 (Princípio da Publicidade)[4], muitas informações geradas ou mantidas em órgãos governamentais têm caráter público e, cada vez mais, são armazenadas em bases de dados disponíveis para consulta pública. No entanto, com o passar do tempo, essa democratização do acesso à informação pode frustrar cidadãos. À medida que cresce o volume de informações tornadas públicas, aumenta o grau de dificuldade de localizar aquelas consideradas relevantes pelos pesquisadores, pois as ferramentas tradicionais de recuperação de informação, baseadas em palavras-chave e operadores booleanos, se mostram inadequadas na maioria das vezes (BAX; SOUZA, 2001).

Dessa forma, novas técnicas e ferramentas para localização e contextualização de informações se fazem necessárias.

Uma boa pista para novas técnicas a serem propostas é dada por Brown e Duguid: "O fato de observarmos a informação muito de perto, pode ofuscar seu **contexto** social, que ajuda as pessoas a compreenderem seus possíveis **significados** e **porque ela é importante**" (BROWN; DUGUID, 2000). É preciso

[4] "Art. 37. A administração pública [...] obedecerá aos princípios de legalidade, impessoalidade, moralidade, **publicidade** e eficiência e, também, ao seguinte:...".

então oferecer mecanismos que auxiliem a contextualização da informação armazenada, para que ela se torne de fato útil.

1.2.Objetivos

Este trabalho tem como objetivos a proposição, a aplicação e a avaliação de um método de classificação temática em bases de dados textuais indexadas com uso de vocabulário controlado.

O método a proposto deverá demonstrar praticidade e desempenho suficientes para ser utilizado em grandes bases de dados e será considerado bem sucedido se auxiliar a **contextualização** e atribuição de **significado** e **propósito** às informações armazenadas, permitindo a descoberta de conhecimentos.

O **significado** terá sido atribuído às informações armazenadas, se os usuários que as utilizam reconhecerem aumento na compreensão do conteúdo, permitindo-lhes a identificação de tendências e relacionamentos antes ocultos.

Terá sido dado **propósito** às informações armazenadas se os usuários consultados declararem que a aplicação do método é útil para a realização de suas tarefas ou à de outras pessoas ou instituições. Essa característica é particularmente importante para aferir a geração de novos conhecimentos, já que eles estão intrinsecamente ligados a alguma ação.

1.3. Hipótese

A classificação temática de bases de dados textuais através da atribuição de temas é uma forma de **contextualização** das informações armazenadas e agrega

significado e **propósito** a elas, possibilitando assim a geração de novos conhecimentos.

1.4.Organização do Documento

No Capítulo 2 serão apresentados a revisão da literatura e o estado da arte referentes à descoberta de conhecimentos em bases de dados textuais.

No Capítulo 3 serão descritos os passos que foram seguidos para a seleção dos dados que formaram uma base de dados laboratório, bem como, a escolha, descrição e comentários a respeito do método de classificação temática utilizado.

No Capítulo 4, serão apresentados os resultados da aplicação do método de classificação temática na base de dados laboratório.

No Capítulo 5, as respostas ao questionário utilizado como avaliação do método de classificação temática são apresentadas e comentadas.

As conclusões finais e sugestões de trabalhos futuros são apresentadas no Capítulo 6.

2. REVISÃO DA LITERATURA

"A relevância é muito mais importante que a plenitude".

Patrícia Seaman, Hoffman-LaRoche (DAVENPORT; PRUSAK, 1998)

"If only we knew what we know at HP".

Lew Platt, ex-CEO da HP (BROWN; DUGUID, 2000)

"O Excesso de informações (information overload) é um problema sério que pode causar o declínio da produtividade".

José Cláudio Terra, *Gestão do Conhecimento* (TERRA, 2000)

Como foi mencionado no capítulo anterior, o volume de informações disponíveis, incluindo as de natureza textual em formato digital, aumenta a cada dia, ocasionando uma sobrecarga de informações para as pessoas. Dessa forma, é imprescindível estudar novos processos e ferramentas que possibilitem a contextualização dos crescentes volumes de informação disponíveis. Inicialmente serão abordados os aspectos relativos à natureza do trabalho de usuários da informação e depois serão analisados processos e técnicas de localização de informações e de descoberta de conhecimentos.

2.1. O Trabalhador do Conhecimento

Em seu livro *Landmarks of Tomorrow*, publicado em 1959, Peter Drucker utilizou pela primeira vez o termo **trabalhador do conhecimento**, em oposição ao trabalhador manual, envolvido em atividades de manufatura, comum nas empresas tradicionais surgidas após a Revolução Industrial (DRUCKER, 1999).

Rafael Echeverría, em seu livro *A Empresa Emergente,* ao comentar as idéias propostas por Drucker, afirma que o surgimento do trabalhador do conhecimento representa uma mudança na natureza do trabalho desempenhado no interior da nova organização que começa a surgir e que é substancialmente diversa da empresa tradicional do século XX. Enquanto o trabalho manual se baseia na destreza física, permitindo sua desagregação em tempos e movimentos, o trabalho não manual sustenta-se na informação e no conhecimento (ECHEVERRÍA, 2001).

Segundo Drucker, a mais importante contribuição da Administração no século XX foi o aumento da produtividade do trabalhador manual em cerca de 50 vezes em um período de aproximadamente cem anos. Este aumento brutal da produtividade foi obtido graças ao estudo à aplicação das idéias de pessoas como Frederick Taylor (estudos de tempos e movimentos e Administração Científica), Henry Ford (linha de montagem) e W. Edwards Deming (Qualidade Total). Através do aumento de produtividade foi possível a melhora substancial da qualidade de vida de grande parte dos trabalhadores dos países industrializados, por meio da redução da jornada de trabalho e do aumento de sua remuneração. Drucker defende ainda que a aplicação sistemática dos princípios relacionados com a Administração Científica deu aos Estados Unidos a capacidade para vencer alemães e japoneses nos campos de batalha da Segunda Guerra Mundial, e também para superá-los em capacidade de produção industrial por várias ordens de magnitude (DRUCKER 1999).

Ainda segundo Drucker, o estudo da produtividade do trabalhador do conhecimento está apenas começando e "a mais importante contribuição que a Administração precisa fazer no século XXI é, analogamente, elevar a produtividade

do trabalhador do conhecimento". Essa produtividade, ao contrário do que ocorria com o trabalhador manual, é mais uma questão de qualidade do que de quantidade. Ele sugere que a próxima revolução da informação trate a matéria-prima do trabalhador do conhecimento – a informação – e responda às questões: qual o **significado** da informação e qual o seu **propósito**, como forma de elevar a produtividade do trabalhador do conhecimento (DRUCKER, 1994, 1999).

Portanto, torna-se mais clara a necessidade do emprego de mecanismos que permitam a **contextualização** do crescente volume de informações disponíveis e a agregação de valor através da atribuição de **significado** e **propósito** e que precisam ser tão automatizados quanto possível para não recair no problema original, ou seja, a dificuldade de se lidar com grandes volumes de informação.

Um estudo realizado pelo professor Jean Moscarola, da Universidade de Savoie na França, identificou dois problemas decorrentes da sobrecarga de informações: um está relacionado com a localização das fontes de informações relevantes e o outro, com a identificação e extração de conhecimentos presentes nas informações relevantes encontradas (MOSCAROLA, 1998). Esses problemas identificados bem como algumas abordagens que se propõem a auxiliar sua solução serão analisados a seguir.

2.2.Localização de Informações Relevantes

Um dos mecanismos mais difundidos para localizar informações relevantes em uma coleção de documentos textuais é o sistema de **Recuperação de Informação** (RI), ou *Information Retrieval,* que funciona da seguinte forma: o usuário informa ao sistema as palavras-chave (às vezes sob a forma de uma expressão booleana) que, segundo seu entendimento, descrevem o assunto de seu interesse. O sistema, após realizar uma pesquisa das palavras solicitadas, exibe um resultado na forma de uma lista de referências a documentos que contém essas palavras, ordenada por critérios de relevância internos. A única forma do usuário avaliar a pertinência do resultado da pesquisa é através da inspeção minuciosa da lista resultado.

Segundo Marti Hearst, pesquisadora da *School of Information Management and Systems* - SIMS da Universidade da Califórnia em Berkeley, considera RI um processo de busca de informações que já sejam de conhecimento prévio e que tenham sido inseridas em uma base de dados pelo autor. Ela define RI como "um processo de extrair de um repositório todos os documentos que são de interesse do pesquisador e descartar todos os demais" (LUCAS, 1999).

Esse método costuma ser eficaz quando se procuram documentos específicos, dos quais se conhecem algumas informações referenciais. Nesse caso, o consulente pode utilizar metadados do documento, como autor, idioma do texto, título ou data de publicação, tornando a busca mais direcionada e, portanto, com maior chance de sucesso. No entanto, em uma pesquisa aberta por assunto, os sistemas de RI podem não funcionar tão bem. Pesquisadores que já experimentaram consultar grandes bases de dados bibliográficas ou a própria WWW através de sites de busca que

utilizam sistemas de RI, como o *Google*, sabem como é difícil descrever e delimitar com precisão o assunto de seu interesse (FALCÃO et al., 1999).

Isso ocorre porque o grau de sucesso na utilização dessa forma de pesquisa depende inteiramente da capacidade de o usuário escolher as palavras-chave de forma que expressem inequivocamente o universo de documentos buscados, atingindo o máximo de precisão (todos os documentos relacionados no resultado são relevantes) e de revocação (todos os documentos relevantes do universo pesquisado estão no resultado).

Segundo os pesquisadores Marcelo Bax e Renato Souza, um problema inerente às ferramentas baseadas em RI é "o fato de que, em maior ou menor grau, essas ferramentas ignoram os contextos onde as palavras-chave aparecem e à que outras palavras ou conceitos estão relacionadas". Os sistemas de pesquisa baseados em RI apresentam limitações por depender fortemente da linguagem natural, que admite ambigüidades, como polissemia[5] e sinonímia[6] (BAX; SOUZA, 2001).

Com o objetivo de reduzir as ambigüidades intrínsecas à linguagem natural, os bibliotecários passaram a utilizar linguagens de indexação controladas para exprimir o conteúdo semântico de documentos. Estes vocabulários podem ter a forma de listas de termos autorizados na indexação e, nesse caso, são denominados **vocabulários controlados**.

Se os termos desses vocabulários possuírem relacionamentos que indiquem, por exemplo, uma hierarquia de termos genéricos ou específicos em relação aos demais, então o vocabulário é denominado **tesauro**. A estrutura dos tesauros é

[5] Multiplicidade de sentidos de uma palavra ou locução (Fonte: Dicionário Houaiss Eletrônico, v. 1.0 Dez.2001).
[6] Relação de sentido entre dois vocábulos que têm significação muito próxima, permitindo que um seja escolhido pelo outro em alguns contextos, sem alterar o sentido literal da sentença como um todo (idem).

controlada por padrões internacionais que estão entre os mais influentes já desenvolvidos para a área de Biblioteconomia e Ciência da Informação. Três padrões são a Norma ISO 2788:1986 que define os relacionamentos a serem usados entre termos de um tesauro monolíngue, a Norma ISO 5964:1985, que define os relacionamentos adicionais para tesauros multilíngues e a ISO 5963:1985, que padroniza métodos para exame de documentos, determinação de seus assuntos e seleção de termos de indexação (WILSON; MATTHEWS, 2002).

Formalmente, segundo a Norma ISO 2788:1986, tesauro monilíngue é "o vocabulário formalmente organizado de uma linguagem de indexação controlada, de forma que relacionamentos definidos *a priori* são tornados explícitos".

Os tesauros auxiliam a indexação dos documentos de uma base de dados ao padronizar a linguagem de indexação e tentar eliminar as ambigüidades da linguagem natural. Uma indexação padronizada e precisa é uma forma de melhorar a localização de informações relevantes em um acervo de documentos, desde que o usuário tenha conhecimento prévio da linguagem de indexação utilizada ou tenha acesso a ele durante a pesquisa.

2.3.Identificação e Extração de Conhecimentos

Após a constatação do crescimento contínuo do volume de informação atualmente disponibilizado bem como das dificuldades inerentes às ferramentas de RI para tratá-lo adequadamente, novos mecanismos estão sendo aprimorados.

Segundo M. Norton, um dos primeiros pesquisadores a propor uma nova abordagem para o tratamento de grandes volumes de informação foi Usama Fayyad,

então na Microsoft Research (NORTON, 1999). Em seu artigo *Data Mining and Knowledge Discovery*, publicado em 1996, Fayyad cunhou a expressão Descoberta de Conhecimentos em Bases de Dados, ou ***Knowledge Discovery in Databases*** (KDD), definido por ele como o "processo não trivial de identificação de padrões nos dados que sejam válidos, novos, potencialmente úteis e inteligíveis". Na visão de Fayyad, a técnica de *data mining* é uma das fases do processo de KDD e pode ter como objetivos a classificação, a regressão, o agrupamento, o resumo, a modelagem de dependências ou detecção de mudanças nas informações analisadas (FAYYAD, 1996).

A ênfase dada por Fayyad, quando enfatiza que KDD é um processo, tem a intenção de tornar claro que a busca de novos conhecimentos em coleções de dados envolve aspectos intelectuais e tecnológicos projetados para procurar conhecimentos e não para simplesmente manipular dados. A descoberta de conhecimentos envolve experimentação, iteração, interação do usuário e muitas decisões de projeto e personalizações em cada caso (FAYYAD, 1996) e (NORTON, 1999).

Ao estudar a aplicação de KDD em informações textuais, Marti Hearst propôs o conceito de ***Text Data Mining*** (TDM). Segundo ela, a TDM é uma forma de examinar uma coleção de documentos e descobrir informações que não pertencem a nenhum documento específico da coleção. Ao utilizar TDM, o pesquisador procura novas informações que até então não eram do conhecimento de ninguém (LUCAS, 1999).

Com a TDM, o pesquisador procura relacionamentos entre o conteúdo de múltiplos textos e posteriormente tenta conectar esta informação para formar uma

hipótese comprovável a cerca da nova informação. A literatura sobre pesquisas médicas se mostra um alvo promissor para a TDM, pois grande quantidade de artigos publicados em jornais de medicina está disponível hoje em meio digital. Pelo menos em teoria, a aplicação de TDM deveria ser capaz de auxiliar os pesquisadores na tarefa de apontar elos entre os resultados de pesquisas publicados (por exemplo, conexões entre causa e efeitos de doenças), mesmo entre disciplinas diferentes (LUCAS, 1999).

Ainda segundo Hearst, para se realizar a TDM, pode-se utilizar seqüencialmente as abordagens de agrupamento (*clustering*) e categorização. O agrupamento se baseia nas similaridades percebidas durante a análise computacional dos documentos de uma coleção. Depois dessa análise, um especialista, adapta os grupos identificados em um sistema de categorias personalizado para a coleção de documentos analisada (LUCAS, 1999).

Tanto Fayyad quanto Hearst ressaltam o fato de que processos de descoberta de conhecimentos são necessariamente interativos e só ocorrem com um laço de *feedback* entre as ferramentas baseadas em computadores e um usuário especialista. As ferramentas automatizadas auxiliam o especialista a navegar em grandes volumes de informação, e ele, por sua vez, direciona o processo filtrando resultados espúrios e analisando quão significativos são os padrões localizados pelo computador (FAYYAD, 1996) e (LUCAS, 1999).

2.4.A Web Semântica

Complementando as abordagens de RI e descoberta de conhecimentos, uma nova idéia foi proposta por Tim Berners-Lee, considerado o inventor da WWW. Ele propôs uma extensão da Web atual, denominada Web Semântica, na qual é atribuído significado precisamente definido à informação que compõe os documentos, permitindo que computadores e pessoas trabalhem de forma cooperativa (BERNERS-LEE, 2002).

A idéia da Web Semântica envolve a inclusão de informações nos documentos utilizando-se uma linguagem de marcação semântica, ou seja, uma marcação que não interfere na visualização do documento (como HTML[7]), mas expressa o conteúdo do documento. Essa característica é descrita como "uma Web para as máquinas" em oposição a "uma Web para ser lida pelas pessoas".

A Web Semântica será construída a partir de novos padrões como *Universal Resource Identifiers* (URI), *Resource Description Framework* (RDF) e *Web Ontology Language* (WOL) que estão sendo propostos pelo *World Wide Web Consortium*[8] (W3C) para viabilizar a ligação semântica entre recursos como documentos, imagens, pessoas e conceitos (eg, *ÉAautorDe*, *TrabalhaPara*, *DependeDe*, *EstáLocalizadoEm,* etc.) tornando explícitos relacionamentos contextuais que são implícitos na Web atual (BERNERS-LEE, 2002).

Ao contrário da Inteligência Artificial, a Web Semântica não tem uma conotação antropomórfica, mas pretende complementar os esforços humanos em áreas em que as pessoas têm dificuldades, como lidar com grandes volumes de informação em um intervalo limitado de tempo. A Web Semântica foi planejada para

[7] Hipertext Markup Language - Linguagem de marcação usada para formatar páginas Web.
[8] www.w3c.org//2001/sw

auxiliar a tarefas de agentes de softwares que realizariam tarefas complexas de forma relativamente autônomas (EUZENAT, 2002).

Outra possibilidade da Web Semântica seria a representação de ontologias para a construção de vocabulários de descrição de recursos, baseados em WOL e RDF. Michael Wilson e Brian Mattthews propõem que novos recursos podem ser construídos para substituir o papel hoje desempenhado pelos tesauros para representar ontologias, com pelo menos duas vantagens: não estariam limitados pelos poucos relacionamentos possíveis nos tesauros (definidos por normas internacionais) e teriam maior facilidade de intercâmbio de informações por estarem baseados em padrões RDF e XML (WILSON; MATTHEWS, 2002).

No restante deste trabalho, será proposto um método de descoberta de conhecimentos, baseado na classificação temática, bem como sua aplicação em uma base de dados textual indexada *a priori*, com base em um vocabulário controlado. A aplicação do método nessa base de dados foi avaliada por uma amostra formada por usuários relacionados profissionalmente com o assunto da base de dados.

3. MATERIAIS E MÉTODOS

Neste capítulo serão descritos os passos que foram seguidos para a seleção dos dados que formaram a base de dados analisada, chamada base de dados laboratório, bem como a escolha, a descrição e considerações a respeito do método de classificação temática utilizado.

3.1. Escolha da Base de Dados Analisada

Optou-se por utilizar como objeto de análise uma base de dados que contém discursos proferidos por deputados em sessões da Câmara dos Deputados da República Federativa do Brasil. Nessa escolha foram levados em conta os seguintes fatores:

- Os dados armazenados são públicos, livres de sigilo e direitos autorais;
- A temática abordada e o formato dos textos sofrem menos restrições que os documentos de outras bases de dados disponíveis na Câmara dos Deputados, como a de Proposições em Tramitação e a de Legislação. Nada impede, no entanto, que essas bases de dados também sejam estudadas no futuro;
- Os dados estavam organizados e armazenados em uma base de dados relacional de fácil acesso ao autor.

A base de dados analisada contém discursos proferidos por Deputados Federais e é atualizada diariamente pelo Departamento de Taquigrafia, Revisão e

Redação (DETAQ)[9] da Câmara dos Deputados através do Sistema Informatizado de Registro de Notas Taquigráficas (SITAQ II). Esse Sistema gerencia o fluxo de trabalho que gera a redação final dos registros taquigráficos das reuniões ocorridas na Câmara dos Deputados e é composto por cinco etapas principais. As três primeiras - taquigrafia (apanhamento e digitação dos textos), sumarização e revisão - são realizadas no mesmo dia em que os discursos são proferidos em Plenário. No dia seguinte, ocorre a quarta etapa – a supervisão, que consiste em uma revisão mais cuidadosa dos textos dos discursos e sumários. Nos dias seguintes, ocorre a quinta etapa - a indexação dos discursos, realizada por especialistas com auxílio de um vocabulário controlado.

O SITAQ II, atualmente em sua segunda geração tecnológica, segue a arquitetura de duas camadas. A versão corrente do Sistema foi implantada em outubro de 2000 e desde então tem recebido algumas melhorias e novas funcionalidades. Atualmente está sendo integrado ao Sistema de Áudio que gerencia e armazena o áudio das reuniões em formato digital e compactado, MP3.

A base de dados do SITAQ II é mantida pelo gerenciador Microsoft SQL Server 2000 e ultrapassa 8 Gigabytes. Contém tanto as informações relativas ao fluxo dos trabalhos quanto os textos dos discursos em suas fases intermediárias e final, em formato *Rich Text Format* (RTF). O editor de textos utilizado é o Microsoft Word 97, incorporado à camada cliente da aplicação, escrita em linguagem Microsoft Visual Basic 6.0.

[9] As atribuições do DETAQ estão regulamentadas no Art. 118 da Resolução 20/1971 da Câmara dos Deputados.

3.2.Descrição da Base de Dados Analisada

Entre a data da implantação do SITAQ II (10/10/2000) e a data da extração dos textos (10/10/2002), haviam sido registrados 29.526 discursos em sua base de dados. Destes, foram selecionados para análise 10.627 discursos que, na data da extração, haviam sido indexados e estavam de acordo com os seguintes critérios:

- Terem sido proferidos por deputados federais durante sessões da Câmara dos Deputados, excluindo-se aqueles que ocorreram durante sessões do Congresso Nacional, as quais reúnem deputados e senadores;
- Terem sido proferidos durante as fases do Grande Expediente (GE) e do Pequeno Expediente (PE)[10]. Os discursos dessas fases normalmente são preparados com alguma antecedência, com o auxílio de assessores e costumam estar vinculados a fatos considerados relevantes à corrente política do parlamentar ou do partido a que pertence.

As falas da Ordem do Dia, durante a qual ocorrem os debates e as votações, foram descartadas por serem curtas e freqüentemente interrompidas. Também foram excluídas as falas do tipo Questão de Ordem, em que são feitos questionamentos à Mesa Diretora, com base em dispositivos do Regimento Interno ou da Constituição Federal, que normalmente se referem ao andamento da sessão e ao processo de votação. Foram ainda excluídas as fases de Abertura e Encerramento da Sessão, e as demais falas protocolares do Presidente.

[10] De acordo com o Regimento Interno da Câmara dos Deputados, os discursos do Grande Expediente e do Pequeno Expediente têm duração respectivamente de 25min e 5min e temática livre (RICD, Arts. 87, 75 e 66).

Os discursos selecionados foram importados para uma base de dados laboratório onde seriam analisados pelo processo de classificação temática. A descrição dessa base encontra-se no ANEXO A – Descrição da Base de Dados Laboratório.

3.3.Escolha do Método de Classificação Temática

Inicialmente foi realizada uma pesquisa das ferramentas de classificação temática disponíveis no mercado. Após análise dos produtos relacionados no site do Centro de Referência em Inteligência Empresarial (CRIE)[11] da COPPE - UFRJ, verificou-se que os produtos são, em sua maioria, preparados para lidar apenas com a língua inglesa e necessitam de adaptação para lidar corretamente com textos na língua portuguesa. Verificou-se ainda que os fabricantes da maioria dos produtos relacionados no site não disponibilizavam versões funcionais de demonstração. Na data da pesquisa (junho/2002), os produtos listados no site eram os seguintes: Megaputer, SemioMAP, MultiCentrix, IBM Intelligent Miner for Text, Smart TextMiner, DataSet for MSWord, SMIsC e dtSearch.

Após a avaliação e o descarte da possibilidade de usar as ferramentas comerciais, cogitou-se a utilização do software Eureka desenvolvida por Leandro Krug Wives apresentado em sua dissertação de mestrado em Ciências da Computação na Universidade Federal do Rio Grande do Sul em abril de 1999 (WIVES). Essa ferramenta se propõe a analisar um conjunto de textos não formatados e a identificar e agrupar aqueles considerados semelhantes

[11] kmtools.crie.ufrj.br/index.html

semanticamente. Dentre as facilidades oferecidas pela ferramenta estão a possibilidade de configuração de listas de *stopwords* e da escolha de um entre quatro algoritmos de agrupamento: Best Star, Cliques, Full Star e Star.

Para avaliar a ferramenta Eureka, foi preparada uma amostra inicial contendo 300 textos extraídos da base de dados laboratório. Para cada um deles foram exportados os campos contendo o texto integral do discurso (txAsciiDiscurso)[12], o texto do sumário do discurso (txAscii) e a indexação do discurso (txIndexacao e txCatalogo). A fase de análise e identificação de similaridades, para essa amostra, durou cerca de 7min 22seg. A fase seguinte, de agrupamento dos textos, depende do algoritmo utilizado e do coeficiente de similaridade adotado. Foram feitas diversas experiências com a amostra de textos utilizando-se diferentes algoritmos e coeficientes de similaridade. Um problema não solucionado foi o fato de a grande maioria dos textos não serem enquadrados em nenhum agrupamento. A Tabela 1 - Resumo dos Resultados Obtidos com a Ferramenta Eureka - apresenta os melhores resultados obtidos.

Tabela 1 - Resumo dos Resultados Obtidos com a Ferramenta Eureka

Algoritmo Utilizado	**Tempo de Processamento (agrupamento)**	**Clusters Identificados**	**Textos não Agrupados** (n=300)
Best Star	10 a 12 seg	11	288
Cliques	10 a 12 seg	11	288
Full Star	4 seg	10	280
Star	11 seg	11	278

[12] A descrição da estrutura da base de dados laboratório encontra-se no ANEXO A – Descrição da Base de Dados Laboratório.

Resultados semelhantes foram obtidos quando os textos da amostra não continham o texto integral do discurso. O objetivo desse novo teste foi verificar o comportamento da ferramenta ao tratar textos menores, mas semanticamente semelhantes, já que os campos de indexação exprimiam o assunto principal do discurso. Nesse teste foram obtidos resultados semelhantes, porém houve considerável diminuição do tempo de processamento devido ao tamanho menor dos textos.

Outro teste foi realizado com o objetivo de avaliar o comportamento da ferramenta com amostras contendo quantidades significativamente maiores de textos. Ao analisar uma amostra de cerca de 6.000 textos, a ferramenta apresentou falha de execução por falta de memória RAM, após algumas horas de processamento.

Os testes foram executados com a versão 3.0.1 beta da ferramenta Eureka no mesmo computador que hospedou a base de dados laboratório e que contava com a seguinte configuração: processador Intel Pentium III 1GHz, 256MB de memória RAM e sistema operacional Microsoft Windows XP Professional.

Edilberto Silva, que também utilizou a ferramenta Eureka, estima em 562 horas de processamento o tempo necessário para realizar o agrupamento de 5.800 textos com o uso da ferramenta Eureka em computador com dois processadores Pentium Xeon 1GHz e 2GB de memória RAM (SILVA, 2002).

Wives faz diversas análises e considerações sobre a performance dessa ferramenta em sua dissertação. Em um computador Pentium 133 MHz com 96 MB de

memória RAM, foram analisadas duas coleções. A primeira, com 11 mensagens de correio eletrônico, demorou 23min 30seg para ser analisada. A segunda, com 120 textos, levou 6h 23min 15seg para ser analisada, quando não foi utilizado o pré-processamento, que trunca o texto e considera apenas as primeiras palavras (WIVES, 1999).

Com base nos resultados obtidos nos testes realizados, considerou-se que a ferramenta Eureka não seria adequada para este trabalho por não agrupar os textos da amostra adequadamente nem oferecer desempenho apropriado, no equipamento disponível.

Tendo em vista o objetivo primordial deste trabalho – avaliar a importância da classificação temática em quantidades variáveis de textos, decidiu-se construir uma ferramenta que fosse mais adequada ao estudo da base de dados de discursos.

A base de dados analisada tem uma característica importante com a qual as ferramentas disponíveis, a princípio, não contam: ser indexada com base em um vocabulário controlado[13]. Isso significa que especialistas realizaram previamente um trabalho intelectual que identificou os assuntos predominantes em cada discurso e atribuíram a eles descritores a partir de uma lista predefinida. Os descritores de um vocabulário controlado, ao contrário do que ocorre em um tesauro, não estão relacionados entre si segundo uma estrutura hierárquica, que permitiria a identificação automática do tema. Dessa forma, para se associar um tema a um discurso, deve-se antes relacionar cada descritor utilizado na indexação com um

[13] A base de dados analisada foi indexada com base no vocabulário denominado VCB, disponível para consulta em http://recreio.senado.gov.br:4505/ALEPH/-/start/sen10.

tema, para então estabelecer a ligação entre os discursos e os temas. Para isso pode-se utilizar relacionamentos implementados de forma natural pelo banco de dados que hospeda a base de dados laboratório.

Assim, decidiu-se construir um método de classificação temática que fosse aplicável a bases de dados indexadas previamente, de construção relativamente simples, baseado nos recursos disponíveis em um banco de dados relacional e que apresentasse escalabilidade suficiente para ser aplicado em volumes de dados significativamente maiores que a utilizada neste trabalho. Esse método é descrito a seguir.

3.4.Descrição do Método de Classificação Temática

A definição do método de classificação temática de documentos desenvolvido e aplicado neste trabalho contou com a colaboração e avaliação da especialista Vilma Pereira[14], funcionária do Departamento de Taquigrafia, Revisão e Redação da Câmara dos Deputados.

O método de classificação temática desenvolvido agrupa os discursos em categorias temáticas, denominadas temas, através da associação dos descritores usados na indexação dos discursos com estas categorias.

Para se atribuírem temas aos discursos, o seguinte procedimento é executado na base de dados laboratório:

[14] A Sra.Vilma Pereira é bibliotecária e trabalha desde 1987 na indexação dos discursos de parlamentares. Atualmente é chefe da Coordenação de Histórico de Debates, onde lidera uma equipe de 15 pessoas responsáveis pela indexação e pesquisa da base de dados do SITAQ II, de onde foram extraídos os dados da base de dados laboratório.

1. São definidos os temas e armazenados na tabela Temas;
2. São extraídos os descritores utilizados para indexar os discursos e armazenados na tabela TermosTemas com os respectivos temas;
3. Para cada descritor da tabela TermosTemas são selecionados os respectivos discursos e atribuídos os temas a eles através da tabela DiscursosTemas.
4. Aos discursos que não tenham sido associados a nenhum tema é atribuído o tema "Outros".

O código fonte em linguagem *Visual Basic for Applications* (VBA), que implementa esse procedimento, está listado no ANEXO C – Código Fonte do Método de Classificação Temática.

Inicialmente o método de classificação temática foi projetado para atribuir apenas um tema a cada discurso, tendo em vista o critério da simplicidade. Porém, ao analisar os resultados obtidos, a especialista consultada recomendou que o método fosse modificado, para permitir a atribuição de mais de um tema por discurso.

Cada tema atribuído tem igual peso na análise realizada, de forma que, se um discurso tiver, por exemplo, três descritores, sendo os dois primeiros ligados ao tema A e o terceiro ao tema B, o discurso seria associado uma vez ao tema A e uma vez ao tema B. Ou seja, cada tema atribuído tem igual peso na descrição do discurso. Após a aplicação do método, constatou-se que, em média há 1,6 tema associado a cada discurso.

A escolha inicial dos temas foi realizada com a ajuda da especialista e visou a identificar os 10 temas mais freqüentes na base de dados. Durante o processo de sucessivas aplicações do método o conjunto de temas foi refinado com a junção de temas pouco expressivos e a divisão de temas muito abrangentes. Ao final, chegou-se a um conjunto de 14 temas, relacionados na Tabela 2 - Temas Utilizados na Classificação Temática dos Discursos.

Tabela 2 - Temas Utilizados na Classificação Temática dos Discursos

Tema	Inclui
Administração Federal	Executivo Federal e Poder Judiciário
Agricultura	Política Rural e Reforma Agrária
Corrupção	Ética
Economia	Finanças Públicas, Comércio e Indústria
Educação	Ciência, Tecnologia, Informática, Cultura, Religião, Meios de Comunicação e Esportes
Infra-estrutura	Transporte, Energia e Telecomunicações
Meio Ambiente	Recursos Hídricos
Outros	
Política Externa	Defesa Nacional e Inteligência
Política Interna	Congresso Nacional, Câmara dos Deputados, Senado Federal
Política Regional	Administração Estadual e Administração Municipal
Política Social	Trabalho, Emprego, Minorias, Previdência, Seguridade Social
Saúde Pública	
Segurança Pública	

A associação dos descritores utilizados na indexação dos discursos aos temas também contou com a intensa colaboração da especialista consultada e teve como referência inicial a freqüência dos descritores na base de dados. Com a sucessiva aplicação do método, a lista foi refinada, observando-se se a classificação temática atribuída correspondia à realidade do discurso. Também se buscou reduzir ao máximo os discursos classificados na categoria "Outros". No final chegou-se a uma

lista com 607 descritores, relacionada no ANEXO B - Descritores Utilizados para a Classificação Temática. A Figura 1, a seguir, representa a aplicação do método de classificação temática na base de dados laboratório, que partiu de 10.627 discursos até agrupá-los em 14 temas.

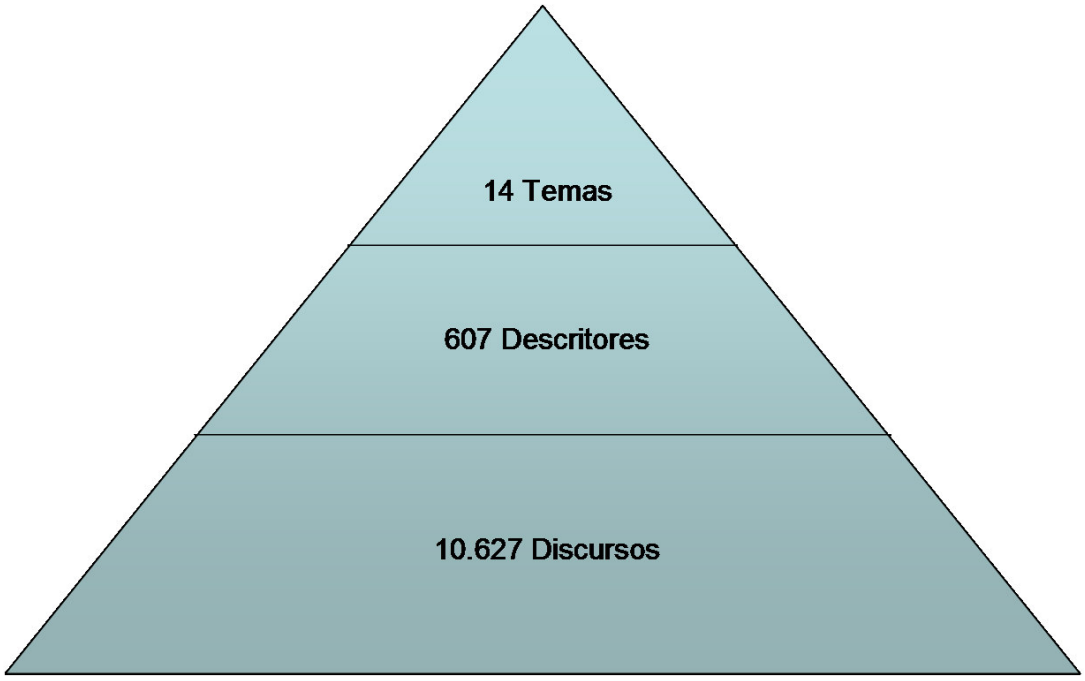

Figura 1 - Representação da Aplicação do Método de Classificação Temática Utilizado

3.5.Considerações sobre o Método de Classificação Temática

O método de classificação temática apresentado neste trabalho foi desenvolvido com o objetivo primordial de viabilizar sua aplicação e posterior avaliação em uma base dados existente. Dessa forma, procura utilizar a seu favor uma característica importante da base de dados analisada, que é a de ser indexada. Apesar de a indexação não estar presente nas coleções analisadas por Wives e Silva, não se trata

de uma particularidade da base de dados, mas, ao contrário, se aplica a diversas outras bases de dados mantidas na Câmara dos Deputados. É o caso de bases de dados relevantes como, por exemplo, as de Proposições em Tramitação, de Legislação, de Questões de Ordem, de Processos Administrativos e de Trabalhos da Consultoria Legislativa, algumas das quais utilizam o mesmo vocabulário controlado da base de Discursos. Ressalte-se que o hábito de indexar bases de dados tampouco é exclusivo da Câmara dos Deputados, pois existem bases de dados textuais indexadas em diversos outros órgãos públicos como o Senado Federal (Pronunciamentos, Acervo da Biblioteca), o Tribunal de Contas da União (Processos em Tramitação), o Supremo Tribunal Federal (Jurisprudência), Instituto Brasileiro de Informação em Ciência e Tecnologia (Teses e Dissertações).

O fato do método desenvolvido neste trabalho se basear na indexação dos textos em vez do no texto integral como a ferramenta Eureka (que tem propósitos diferentes) possibilitou uma implementação simples e baseada em ferramentas comuns como o banco de dados relacional. Assim, a performance e a escalabilidade atingidas indicam que esse método pode ser aplicado em bases de dados substancialmente maiores que a base de dados laboratório. A Tabela 3 - Tempos Necessários para a Execução do Método de Classificação Temática indica os tempos decorridos (*elapsed time*) para aplicar a classificação temática em 4 amostras de tamanhos diferentes da base laboratório. Foram medidos os tempos utilizados pelas duas versões do método, uma mais rápida, mas que só atribui um tema por discurso e outra, que atribui mais de um tema a cada discurso. O computador utilizado

contava com processador Pentium 1GHz e 256MB de memória RAM e a base de dados estava hospedada no Microsoft Access XP.

Tabela 3 - Tempos Necessários para a Execução do Método de Classificação Temática

Número de Discursos	1 Tema por Discurso	Mais de 1 Tema por Discurso
1.000	9 seg	20 seg
2.000	16 seg	39 seg
5.000	40 seg	4 min 0 seg
10.000	2 min 17 seg	10 min 26 seg

É importante observar que a implementação do método não levou em conta a otimização do código nem o modelo de dados para melhorar o desempenho. Algumas mudanças tecnológicas que poderiam aumentar substancialmente a performance do método são:

- Adoção de arquitetura cliente/servidor com o banco de dados hospedado em um ambiente de porte adequado (hardware e software gerenciador de banco de dados) à amostra analisada;
- Utilização de linguagem compilada (em vez de interpretada) para executar o método de classificação temática;
- Utilização de gerenciador de banco de dados com recurso de pesquisa textual para evitar o uso da cláusula 'LIKE' na atribuição de temas aos discursos.

Os resultados da aplicação do método de classificação temática na base de dados laboratório serão detalhados a seguir.

4. RESULTADOS DA APLICAÇÃO DO MÉTODO DE CLASSIFICAÇÃO TEMÁTICA

A aplicação do método de classificação temática associou cada discurso da base de dados laboratórios a um ou mais temas. Como os registros (discursos) importados para a base de dados laboratório continham outras informações, como data em que o discurso foi proferido, bem como nome, partido e unidade da federação do orador, foi possível comparar a classificação temática com essas informações. Ao todo foram realizados quatro tipos de classificação temática e uma correlação entre categorias:

- Classificação temática geral;
- Classificação temática por região geográfica do orador;
- Classificação temática por partido político do orador;
- Classificação temática por mês em que o discurso foi proferido;
- Correlação entre categorias temáticas.

Esses resultados são apresentados a seguir.

4.1. Classificação Temática Geral

A Tabela 4 - Classificação Temática dos Discursos em Ordem Decrescente de Freqüência relaciona as quantidades e os percentuais dos discursos associados a cada uma das categorias temáticas utilizadas, em ordem decrescente de freqüência:

Tabela 4 - Classificação Temática dos Discursos em Ordem Decrescente de Freqüência

Tema	Inclui	Discursos Qtd.	Discursos (%)
1. Política Social	Trabalho, Emprego, Minorias, Previdência e Seguridade Social	2.342	13,7

2.	Política Regional	Administração Estadual e Administração Municipal	2.294	13,4
3.	Economia	Finanças Públicas, Comércio e Indústria	1.783	10,4
4.	Infra-estrutura	Transporte, Energia e Telecomunicações	1.630	9,5
5.	Política Interna	Congresso Nacional, Câmara dos Deputados e Senado Federal	1.624	9,5
6.	Educação	Ciência, Tecnologia, Informática, Cultura, Religião, Meios de Comunicação e Esportes	1.466	8,6
7.	Administração Federal	Executivo Federal e Judiciário	1.151	6,7
8.	Segurança Pública		896	5,2
9.	Agricultura	Política Rural e Reforma Agrária	797	4,7
10.	Saúde Pública		789	4,6
11.	Outros		733	4,3
12.	Meio Ambiente	Recursos Hídricos	599	3,5
13.	Política Externa	Defesa Nacional e Inteligência	501	2,9
14.	Corrupção	Ética	484	2,8

4.2. Classificação Temática por Região Geográfica do Orador

Uma vez identificadas as categorias temáticas de um discurso, pode-se comparar essa distribuição com outros atributos do discurso, como, por exemplo, a região geográfica da unidade da federação pela qual o parlamentar foi eleito. Antes, porém, deve-se avaliar o tamanho das bancadas dos partidos políticos de cada uma das regiões.

A Tabela 5 - Distribuição de Deputados e Discursos por Região Geográfica relaciona as composições das bancadas regionais, as respectivas quantidades de discursos proferidos, em termos absolutos e proporcionais, e a média de discursos por deputado.

Tabela 5 - Distribuição de Deputados e Discursos por Região Geográfica

Região	Bancadas[15]		Discursos		Discurso /
	Deputados	(%)	Quantidade	(%)	Deputado
SE	**179 (1)**	34,9	3.081 (2)	29,0	17,2 (5)
NE	151 (2)	29,4	**3.236 (1)**	30,5	21,4 (4)
S	77 (3)	15,0	1.880 (3)	17,7	**24,4 (1)**
N	65 (4)	12,7	1.434 (4)	13,5	22,1 (3)
CO	41 (5)	8,0	996 (5)	9,4	24,3 (2)
Total	**513**	**100,0**	**10.627**	**100,0**	**20,7**

Observa-se que os deputados da região Sudeste proferiram proporcionalmente menos discursos que os parlamentares das demais regiões geográficas com a média de 17,2 discursos por deputado. A maior proporção foi obtida pelos parlamentares da região Sul que proferiram 24,4 discursos por deputado.

A seguir são exibidas as Tabelas 6 e 7 contendo as distribuição dos temas dos discursos por região geográfica. São relacionados os dados quantitativos, os índices e *rankings* dos temas por região.

Tabela 6 - Distribuição Temática de Discursos por Região Geográfica (total de discursos)

Tema	CO	N	NE	S	SE	Total
Política Social	255	288	608	513	677	**2.341**
Política Regional	225	450	800	341	477	**2.293**
Economia	173	215	488	375	531	**1.782**
Infra-estrutura	161	209	481	254	525	**1.630**
Política Interna	130	177	564	277	476	**1.624**
Educação	140	170	447	289	420	**1.466**
Administração Federal	92	144	382	204	329	**1.151**
Segurança Pública	116	122	245	90	323	**896**
Agricultura	84	154	212	199	148	**797**
Saúde Pública	69	105	206	153	255	**788**
Outros	61	84	231	128	229	**733**
Meio Ambiente	44	104	291	45	115	**599**
Política Externa	58	93	109	81	160	**501**
Corrupção	44	77	175	78	110	**484**
Total	**1.652**	**2.392**	**5.239**	**3.027**	**4.775**	**17.085**

[15] Fonte: SILEG: Sistema de Informações Legislativas da Câmara dos Deputados.

Tabela 7 - Distribuição Temática de Discursos por Região Geográfica (índice e *ranking*)

Tema	CO	N	NE	S	SE
Política Social	1,13 (2)	0,88 (4)	0,85 (5)	**1,24 (1)**	1,03 (3)
Política Regional	1,01 (3)	**1,40 (1)**	1,14 (2)	0,84 (4)	0,74 (5)
Economia	1,00 (3)	0,86 (5)	0,89 (4)	**1,19 (1)**	1,07 (2)
Infra-estrutura	1,02 (2)	0,92 (4)	0,96 (3)	0,88 (5)	**1,15 (1)**
Política Interna	0,83 (4)	0,78 (5)	**1,13 (1)**	0,96 (3)	1,05 (2)
Educação	0,99 (4)	0,83 (5)	0,99 (3)	**1,11 (1)**	1,03 (2)
Administração Federal	0,83 (5)	0,89 (4)	**1,08 (1)**	1,00 (3)	1,02 (2)
Segurança Pública	**1,34 (1)**	0,97 (3)	0,89 (4)	0,57 (5)	1,29 (2)
Agricultura	1,09 (3)	1,38 (2)	0,87 (4)	**1,41 (1)**	0,66 (5)
Saúde Pública	0,91 (4)	0,95 (3)	0,85 (5)	1,10 (2)	**1,16 (1)**
Outros	0,86 (4)	0,82 (5)	1,03 (2)	0,99 (3)	1,12 (1)
Meio Ambiente	0,76 (3)	1,24 (2)	**1,58 (1)**	0,42 (5)	0,69 (4)
Política Externa	1,20 (2)	**1,33 (1)**	0,71 (4)	0,91 (5)	1,14 (3)
Corrupção	0,94 (3)	1,14 (2)	**1,18 (1)**	0,91 (4)	0,81 (5)

Na tabela 7 exibe o índice e o *ranking* da distribuição temática por região geográfica. Os índices[16] representam o quanto cada valor se distancia de uma distribuição perfeitamente homogênea, caso em que todos os valores seriam iguais a 1. Valores acima de 1 significam maior concentração que a média (levado-se em conta as linhas e as colunas) e, inversamente, valores menores entre 0 e 1 significam ocorrências menores que a média.

O *ranking*, exibido entre parênteses, representa a posição do valor em ordem decrescente na coluna.

A Tabela 7 - Distribuição Temática de Discursos por Região Geográfica (índice e ranking) mostra que, na amostra analisada, é possível fazer as seguintes associações entre regiões geográficas dos parlamentares e temas, comparando-se a distribuição dos temas dos discursos por região:

[16] O índice é calculado por meio da seguinte fórmula: $\frac{\text{((valor da célula) x (total geral))}}{\text{((total da linha) x (total da coluna))}}$

Norte	Política Regional e Política Externa
Nordeste	Política Interna, Administração Federal, Meio Ambiente e Corrupção
Centro-Oeste	Segurança Pública
Sudeste	Infra-estrutura e Saúde Pública
Sul	Política Social, Economia, Educação e Agricultura

4.3.Classificação Temática por Partido Político do Orador

Antes de analisar a distribuição temática dos discursos por partidos, deve-se conhecer o tamanho das respectivas bancadas. A Tabela 8 - Bancadas dos Partidos Políticos relaciona o tamanho das bancadas partidárias, os respectivos números de discursos proferidos e a média de discursos por deputado de cada partido. É importante notar que a composição dos partidos retrata a situação em novembro de 2002, de acordo com o Sistema de Informações Legislativas da Câmara dos Deputados (SILEG), enquanto a quantidade de discursos proferidos considera o partido ao qual o deputado estava filiado na data em que proferiu o discurso. Por isso, o PV e o PSDC, atualmente sem representantes, contabilizaram juntos 32 discursos.

Tabela 8 - Bancadas dos Partidos Políticos

Partido	Bancadas[17]		Discursos		Discurso /
	Deputados	(%)	Quantidade	(%)	Deputado
PFL	**96 (1)**	18,7	1.267 (4)	11,9	13,2
PSDB	94 (2)	18,3	1.475 (3)	13,9	15,7
PMDB	88 (3)	17,2	1.854 (2)	17,4	21,1
PT	58 (4)	11,3	**2.522 (1)**	23,7	43,5
PPB	53 (5)	10,3	535 (6)	5,0	10,1
PTB	33 (6)	6,4	455 (8)	4,3	13,8
PL	23 (7)	4,5	229 (11)	2,2	10,0

[17] Fonte: SILEG: Sistema de Informações Legislativas da Câmara dos Deputados. Situação em nov/2002.

PDT	16 (8)	3,1	672 (5)	6,3	42,0
PSB	16 (9)	3,1	453 (9)	4,3	28,3
PPS	12 (10)	2,3	486 (7)	4,6	40,5
PC do B	10 (11)	1,9	310 (10)	2,9	31,0
PST	6 (12)	1,2	70 (13)	0,7	11,7
PSL	5 (13)	1,0	220 (12)	2,1	44,0
PTN	1 (14)	0,2	1 (18)	0,0	1,0
PHS	1 (15)	0,2	8 (17)	0,1	8,0
Sem Partido	1 (16)	0,2	38 (14)	0,4	38,0
PV			10 (16)	0,1	
PSDC			22 (15)	0,2	
Total	**513**	**100,0**	**10.627**	**100,0**	

Observa-se que, na amostra analisada, os deputados filiados a partidos de perfil oposicionista proferiram proporcionalmente mais discursos do que os de tendência governista. Por exemplo, os deputados filiados ao PT, apesar de formarem apenas a quarta maior bancada (11,3% do total de deputados), proferiram o maior número de discursos (23,7% do total). Inversamente, o PFL, a despeito de possuir a maior bancada (18,7% do total de deputados), foi o quarto partido no total de discursos proferidos (11,9%), na amostra analisada.

A seguir são exibidas as Tabelas 9 e 10 contendo as distribuições temáticas dos discursos por partido político. Com o objetivo de facilitar a visualização dos resultados, foram tabulados apenas os discursos dos cinco partidos cujos parlamentares proferiram mais discursos na amostra analisada, ou seja, PT, PSDB, PMDB, PFL e PDT. Os discursos somados dessas cinco bancadas representam 73,3% do total de discursos da amostra analisada. As Tabelas 9 e 10 exibem os dados quantitativos e índices de temas por partido.

Tabela 9 - Distribuição Temática de Discursos por Partido Político (total de discursos)

Tema	PDT	PFL	PMDB	PSDB	PT	Total

Política Social	180	195	341	227	809	**1.752**
Política Regional	125	294	437	361	486	**1.703**
Economia	128	205	317	239	392	**1.281**
Política Interna	91	180	280	179	470	**1.200**
Infra-estrutura	95	217	233	227	387	**1.159**
Educação	83	175	262	232	343	**1.095**
Administração Federal	73	123	163	143	341	**843**
Segurança Pública	49	114	177	86	203	**629**
Agricultura	54	77	152	102	219	**604**
Saúde Pública	73	82	122	150	156	**583**
Outros	42	100	151	105	131	**529**
Meio Ambiente	25	84	115	107	109	**440**
Corrupção	32	34	50	44	202	**362**
Política Externa	19	68	65	68	119	**339**
Total	**1.069**	**1.948**	**2.865**	**2.270**	**4.367**	**12.519**

Tabela 10 - Distribuição Temática de Discursos por Partido Político (índice e *ranking*)

Tema	PDT	PFL	PMDB	PSDB	PT
Política Social	1,20 (2)	0,72 (4)	0,85 (3)	0,71 (5)	**1,32 (1)**
Política Regional	0,86 (4)	1,11 (3)	1,12 (2)	**1,17 (1)**	0,82 (5)
Economia	**1,17 (1)**	1,03 (4)	1,08 (2)	1,03 (3)	0,88 (5)
Política Interna	0,89 (4)	0,96 (3)	1,02 (2)	0,82 (5)	**1,12 (1)**
Infra-estrutura	0,96 (3)	**1,20 (1)**	0,88 (5)	1,08 (2)	0,96 (4)
Educação	0,89 (5)	1,03 (3)	1,05 (2)	**1,17 (1)**	0,90 (4)
Administração Federal	1,01 (2)	0,94 (4)	0,84 (5)	0,94 (3)	**1,16 (1)**
Segurança Pública	0,91 (4)	1,16 (2)	**1,23 (1)**	0,75 (5)	0,93 (3)
Agricultura	1,05 (2)	0,82 (5)	**1,10 (1)**	0,93 (4)	1,04 (3)
Saúde Pública	**1,47 (1)**	0,90 (4)	0,91 (3)	1,42 (2)	0,77 (5)
Outros	0,93 (4)	1,21 (2)	1,25 (1)	1,09 (3)	0,71 (5)
Meio Ambiente	0,67 (5)	1,23 (2)	1,14 (3)	**1,34 (1)**	0,71 (4)
Corrupção	1,04 (2)	0,60 (5)	0,60 (4)	0,67 (3)	**1,60 (1)**
Política Externa	0,66 (5)	**1,29 (1)**	0,84 (4)	1,11 (2)	1,01 (3)

A Tabela 10 mostra que, na amostra analisada, é possível fazer as seguintes associações entre partidos e temas, comparando-se a distribuição dos temas dos discursos por partido:

PT	Corrupção, Política Social, Administração Federal e Política Interna.
PMDB	Segurança Pública e Agricultura.
PSDB	Meio Ambiente, Política Regional e Educação.
PFL	Infra-estrutura e Política Externa.
PDT	Saúde Pública e Economia.

4.4.Classificação Temática por Mês em que o Discurso foi Proferido

A seguir são apresentadas as Tabelas 11 e 12 contendo as distribuições temáticas dos discursos agrupados por mês em que foram proferidos. As tabelas exibem os dados quantitativos, os índices e os *rankings* de temas por mês. Alguns meses com número muito reduzido de discursos foram omitidos.

Tabela 11 - Distribuição Temática de Discursos por Mês (total de discursos)

	2000	2001											2002					Total
Tema	**10**	**11**	**12**	**1**	**2**	**3**	**4**	**5**	**6**	**8**	**9**	**10**	**11**	**12**	**2**	**3**	**4**	
Política Social	81	116	80	13	100	172	136	143	81	177	116	209	338	192	92	265	31	**2.342**
Política Regional	127	140	76	5	86	147	119	152	108	216	171	250	184	148	94	238	33	**2.294**
Economia	54	91	63	12	122	90	98	96	61	136	99	139	193	182	78	241	28	**1.783**
Infra-estrutura	27	57	51	2	52	106	97	247	138	145	86	127	118	112	65	177	23	**1.630**
Política Interna	22	72	57	10	103	84	71	71	48	91	61	78	295	172	101	242	45	**1.624**
Educação	41	58	52	5	55	69	83	90	99	165	70	147	168	100	67	171	25	**1.465**
Administração Federal	12	31	31	2	46	34	25	31	43	55	39	45	237	158	90	234	36	**1.150**
Segurança Pública	18	44	27	3	46	55	60	61	46	78	54	48	82	55	67	136	16	**896**
Agricultura	18	33	20	3	36	45	51	63	49	73	52	69	77	61	19	111	16	**797**
Saúde Pública	22	32	30	2	30	58	67	78	46	68	39	67	84	45	39	72	10	**789**
Outros	12	14	29	1	22	40	63	71	53	93	55	109	47	44	19	52	9	**733**
Meio Ambiente	9	38	16	2	16	47	38	60	75	74	27	36	48	30	23	57	3	**599**
Política Externa	13	18	16	6	23	22	32	28	11	35	64	56	51	47	17	46	16	**501**
Corrupção	3	8	11	1	8	64	106	112	21	28	9	11	24	22	17	37	2	**484**
Total	**459**	**752**	**559**	**67**	**745**	**1.033**	**1.046**	**1.303**	**879**	**1.434**	**942**	**1.391**	**1.946**	**1.368**	**788**	**2.079**	**293**	**17.087**

Tabela 12 - Distribuição Temática de Discursos por Mês (índice e *ranking*)

	2000			2001											2002		
Tema	**10**	**11**	**12**	**1**	**2**	**3**	**4**	**5**	**6**	**8**	**9**	**10**	**11**	**12**	**2**	**3**	**4**
Política Social	**1,29**	1,13	1,04	**1,42**	0,98	1,21	0,95	0,80	0,67	0,90	0,90	1,10	1,27	1,02	0,85	0,93	0,77
Política Regional	**2,06**	**1,39**	1,01	0,56	0,86	1,06	0,85	0,87	0,92	1,12	1,35	1,34	0,70	0,81	0,89	0,85	0,84
Economia	1,13	1,16	1,08	**1,72**	**1,57**	0,83	0,90	0,71	0,67	0,91	1,01	0,96	0,95	1,27	0,95	1,11	0,92
Infra-estrutura	0,62	0,79	0,96	0,31	0,73	1,08	0,97	**1,99**	**1,65**	1,06	0,96	0,96	0,64	0,86	0,86	0,89	0,82
Política Interna	0,50	1,01	1,07	1,57	1,45	0,86	0,71	0,57	0,57	0,67	0,68	0,59	**1,59**	1,32	1,35	1,22	**1,62**
Educação	1,04	0,90	1,08	0,87	0,86	0,78	0,93	0,81	1,31	1,34	0,87	1,23	1,01	0,85	0,99	0,96	1,00
Administração Federal	0,39	0,61	0,82	0,44	0,92	0,49	0,36	0,35	0,73	0,57	0,62	0,48	1,81	1,72	**1,70**	**1,67**	**1,83**
Segurança Pública	0,75	1,12	0,92	0,85	1,18	1,02	1,09	0,89	1,00	1,04	1,09	0,66	0,80	0,77	**1,62**	**1,25**	1,04
Agricultura	0,84	0,94	0,77	0,96	1,04	0,93	1,05	1,04	1,20	1,09	1,18	1,06	0,85	0,96	0,52	1,14	1,17
Saúde Pública	1,04	0,92	**1,16**	0,65	0,87	1,22	1,39	1,30	1,13	1,03	0,90	1,04	0,93	0,71	1,07	0,75	0,74
Outros	0,61	0,43	1,21	0,35	0,69	0,90	1,40	1,27	1,41	1,51	1,36	1,83	0,56	0,75	0,56	0,58	0,72
Meio Ambiente	0,56	1,44	0,82	0,85	0,61	1,30	1,04	1,31	**2,43**	**1,47**	0,82	0,74	0,70	0,63	0,83	0,78	0,29
Política Externa	0,97	0,82	0,98	**3,05**	1,05	0,73	1,04	0,73	0,43	0,83	**2,32**	1,37	0,89	1,17	0,74	0,75	1,86
Corrupção	0,23	0,38	0,69	0,53	0,38	**2,19**	**3,58**	**3,03**	0,84	0,69	0,34	0,28	0,44	0,57	0,76	0,63	0,24

Analisando-se a Tabela 12, observa-se, por exemplo, o repentino aumento do número de discursos associados ao tema "Infra-estrutura" nos meses de maio e junho de 2001, quando o País sofria restrições no abastecimento de energia elétrica. Logo depois, esse tema retornou ao seu patamar usual, sugerindo que houve perda de interesse pelo assunto. No entanto, entre os meses de junho e agosto do mesmo ano, nota-se um expressivo aumento da freqüência do tema "Meio Ambiente", que engloba recursos hídricos. Depreende-se então que a forma de abordar o assunto passou da preocupação do racionamento em si para a observação do regime de chuvas e dos níveis dos reservatórios das usinas hidroelétricas, que poderiam determinar ou não o final do período de racionamento.

4.5.Correlação entre Categorias Temáticas

O método de classificação temática utilizado permite a associação de mais de um tema por discurso. Dessa forma, dado um tema, pode-se avaliar com quais outros temas ele está relacionado.

Para exemplificar essa possibilidade, foram verificados os dados relativos aos 484 discursos associados ao tema "Corrupção" para descobrir com quais outros temas ele estaria relacionado. A tabela Tabela 13 - Outros Temas Associados a "Corrupção" lista esses outros temas em ordem decrescente de ocorrência na amostra analisada.

Tabela 13 - Outros Temas Associados a "Corrupção"

Temas	Discursos	
	Quantidade	(%)
Política Interna	116	24,0
Política Regional	94	19,4
Administração Federal	74	15,3
Economia	64	13,2
Infra-estrutura	52	10,7
Política Social	47	9,7
Segurança Pública	42	8,7
Educação	37	7,6
Agricultura	15	3,1
Saúde Pública	14	2,9
Política Externa	6	1,2
Meio Ambiente	6	1,2

A correlação entre os temas "Corrupção" e "Política Interna" pode ser parcialmente explicada pelo número de discursos (33) em que são mencionadas CPIs

(descritor ligado ao tema Política Interna) criadas para apurar denúncias ou fatos ligados ao tema "Corrupção".

Já a ligação entre os temas "Corrupção" e "Política Regional" se explica pela menção a governos estaduais e municipais, a deputados estaduais e vereadores, bem como a organismos de cunho regional como SUDAM e SUDENE.

4.6.Comentários sobre a Aplicação do Método Proposto

A aplicação do método de classificação temática permitiu que quatro tipos de análise fossem realizadas, além do estudo da possível correlação entre temas. As análises realizadas foram: por categoria temática geral, por região geográfica e por partido do orador e por mês em que o discurso foi proferido. Informações de duas classes distintas surgiram baseadas nas observações dos resultados obtidos.

A primeira delas é de natureza quantitativa, como o fato de os deputados da região Sul proferirem, em média, 40% mais discursos que seus colegas da região Sudeste, ou que os parlamentares filiados ao PT utilizarem a tribuna 3,3 vezes mais que os do PFL Também foram identificados, no período analisado, os meses em que os parlamentares mais proferiram discursos como março de 2002, novembro e agosto de 2001.

A outra forma é de natureza qualitativa e associa temas a bancadas regionais ou partidárias ou a períodos do ano. Identificou-se, por exemplo, a afinidade das bancadas da região Sul e do PT com temas sociais, do Centro-Oeste e do PMDB com o tema Segurança.

A validade e a utilidade destes resultados foram avaliadas por uma amostra de funcionários da Câmara dos Deputados e serão mostradas a seguir.

5. AVALIAÇÃO DA APLICAÇÃO DO MÉTODO DE CLASSIFICAÇÃO TEMÁTICA

O método de classificação temática foi avaliado por um grupo representativo de servidores (amostra) do quadro permanente da Câmara dos Deputados ligados à atividade parlamentar (população). Optou-se pelo uso de técnicas de amostragem por ser de difícil operacionalização a análise de toda a população através da realização de um censo por motivos como: a existência constante de funcionários de férias ou licença bem como a necessidade de compatibilizar a agenda de um grande número de pessoas, que deveriam se afastar, ainda que por algumas horas, de suas atividades usuais.

Foram utilizados alguns critérios para determinar a população alvo da qual se extraiu a amostra que participou da avaliação. Esses critérios foram escolhidos de forma a abranger o maior número possível de funcionários da Câmara dos Deputados que têm contato permanente, e de cunho profissional, com o conteúdo da base de dados estudada, ou seja, discursos de parlamentares. Assim, chegou-se aos seguintes critérios:

- Ser funcionário do quadro efetivo da Câmara dos Deputados;
- Ocupar função de nível superior;
- Estar lotado na Secretaria-Geral da Mesa ou em órgãos ligados à Diretoria Legislativa;
- Ter mais de três anos de efetivo exercício da função.

A Tabela 14 - População e Amostra Utilizada nas Entrevistas resume a distribuição da população e da amostra de funcionários entrevistados.

Tabela 14 - População e Amostra Utilizada nas Entrevistas

Órgão[18]	População[19]		Amostra	
	Funcionários	(%)	Funcionários	(%)
DILEG	5	0,8	1	2,8
CONLE	162	25,1	5	13,9
CEDI	147	22,8	12	33,3
DETAQ	171	26,5	9	25,0
CENIN	135	20,9	4	11,1
SGM	26	4,0	5	13,9
Total	**646**	**100,0**	**36**	**100,0**

A técnica de amostragem utilizada foi a amostragem aleatória simples, tendo como base para identificação da população alvo o cadastro do Sistema Integrado de Gestão de Pessoal da Câmara dos Deputados (SIGESP) da Câmara dos Deputados.

A amostra foi composta por 36 funcionários, ou 5,62% da população alvo. Eles foram escolhidos e convidados a responder ao questionário de avaliação pelos respectivos diretores após uma reunião em que foram expostos os propósitos deste trabalho.

Os dados foram coletados por meio de questionário, respondido pelos integrantes da amostra, após apresentação de cerca de uma hora, com exposição de slides e distribuição de material impresso. A reunião foi dividida em três partes, além do tempo necessário para responder ao questionário:

[18] DILEG: Diretoria Legislativa, CONLE: Consultoria Legislativa, CEDI: Centro de Documentação e Informação, DETAQ: Departamento de Taquigrafia, Revisão e Redação CENIN: Centro de Informática, SGM: Secretaria-Geral da Mesa.
[19] Fonte: SIGESP: Sistema Integrado de Gestão de Pessoal da Câmara dos Deputados.

1. Exposição dos objetivos da reunião, apresentação dos principais conceitos envolvidos para contextualizar o trabalho e nivelar os critérios de julgamento e explicação do método de classificação temática;
2. Apresentação dos resultados obtidos com a aplicação do método de classificação temática com uso de tabelas e gráficos;
3. Esclarecimento de dúvidas.

O questionário utilizado, reproduzido no ANEXO D - Questionário, foi dividido em quatro partes. A primeira (questões 1 a 4) procurou identificar o perfil do entrevistado, a segunda parte (questões 5 a 22) permitiu aos entrevistados avaliarem a aplicação do método e seus eventuais usos. A quarta parte (questão 23) avaliou a possibilidade de generalização do método e na última parte (questão 24) os entrevistados tinham um espaço livre para registrarem impressões, críticas e sugestões sobre o método apresentado.

O questionário foi validado por três pessoas da população alvo antes de ser aplicado. O formato do questionário seguiu as recomendações do Tribunal de Contas da União no documento Técnica de Entrevistas para Auditorias[20]. As folhas de respostas dos questionários eram preenchidas e devolvidas de forma anônima.

As respostas dos questionários foram digitadas em um banco de dados Microsoft Access e analisadas com o auxílio da planilha Microsoft Excel. Os resultados obtidos serão relacionados e analisados a seguir.

[20] Disponível em www.tcu.gov.br/SAUDI/Download/Entrevista.exe

5.1.Análise das Respostas do Questionário de Avaliação

A seguir serão analisadas as respostas do questionário de avaliação da aplicação do método de classificação temática.

Questões 1 e 2

Perguntas: 1. Você é servidor do quadro efetivo da Câmara dos Deputados? Há quantos anos?

2. Você ocupa cargo ou função de nível superior?

Objetivo: Identificar o perfil do entrevistado, verificando se ele se enquadra no perfil desejado da amostra.

Resultados: A amostra estudada foi composta por 36 funcionários da Câmara dos Deputados que responderam ao questionário, dos quais,

- 35 (97,2%) são funcionários do quadro efetivo;
- 32 (88,9%) ocupam função de nível superior.

Os funcionários entrevistados trabalham na Câmara dos Deputados há 11,5 anos, em média, sendo o menor valor 1 ano, o maior, 36 anos e o desvio padrão de 8,6.

Comentários A amostra de funcionários que respondeu ao questionário é formada majoritariamente por funcionários efetivos, de nível superior e com grande vivência na Câmara dos Deputados.

Questão 3

Pergunta: 3. Em qual órgão da Câmara dos Deputados você trabalha?

Objetivo: Identificar em que órgãos os funcionários entrevistados trabalham.

Resultados:

Órgão	Entrevistados	
	Quantidade	(%)
SGM – Secretaria-Geral da Mesa	5	13,9
DILEG - Diretoria Legislativa	1	2,8
CEDI – Centro de Documentação e Informação	12	33,3
DETAQ – Departamento de Taquigrafia, Revisão e Redação	9	25,0
DECOM – Departamento de Comissões	0	0,0
CONLE – Consultoria Legislativa	5	13,9
CENIN – Centro de Informática	4	11,1
Total	**36**	**100,0**

Comentários Foram entrevistados funcionários lotados em órgãos da área legislativa, e excluíram-se os da área administrativa.

A Secretaria-Geral da Mesa é a responsável, dentre outras tarefas, pela organização das sessões que ocorrem no Plenário.

A Diretoria Legislativa coordena as atividades dos demais órgãos da sua área de atuação.

O Centro de Documentação e Informação tem em seus quadros bibliotecários que atuam no tratamento da informação relacionada com o processo legislativo, incluindo a coleta, o registro, a indexação, a pesquisa e a disseminação de informações.

Os funcionários do Centro de Informática entrevistados trabalham na Coordenação do Sistema Eletrônico de Votação e estão entre os responsáveis pela operação do Painel Eletrônico de Votação do Plenário.

A Consultoria Legislativa é responsável pela redação de muitos dos discursos proferidos pelos parlamentares.

O Departamento de Taquigrafia é o responsável pelo registro, em forma textual, dos discursos proferidos pelos parlamentares, bem como a alimentação, indexação e pesquisa na base de dados analisada.

Pode-se concluir que a amostra é composta por funcionários com contato de cunho profissional com o conteúdo da base de dados estudada.

Questão 4

Pergunta: 4. Com que freqüência você utiliza a base de dados de discursos ou acompanha os discursos proferidos pelos Srs. Deputados?

Objetivo: Identificar o grau de familiaridade do entrevistado com o assunto da base de dados.

Resultados:

Freqüência	**Entrevistados**	
	Quantidade	**(%)**
Nenhuma	2	5,6
Raramente	6	16,7
Eventualmente: algumas vezes por mês	10	27,8
Freqüentemente: algumas vezes por semana	8	22,2
Diariamente	10	27,8
Total	**36**	**100,0**

Comentários A maioria dos entrevistados (77,8%) mantém pelo menos contato eventual (algumas vezes por mês) com os discursos dos deputados, demonstrando afinidade profissional com as informações que foram objeto do método de classificação temática apresentado e avaliado.

Nas questões 5 a 22, foram formuladas afirmações às quais os entrevistados foram instruídos a responder, utilizando uma escala: desde se concordam integralmente (5) até se discordam integralmente (1) com as afirmações apresentadas. Caso não soubessem responder, deveriam deixar a questão em branco.

Questão 5

Afirmação: 5. A divisão temática apresentada é adequada.

Objetivo: Avaliar se a escolha das categorias temáticas foi adequada.

Resultados:

Respostas	36
Média	4,1
Desvio Padrão	0,7

Comentários Os entrevistados concordaram parcial ou integralmente com a escolha das categorias temáticas utilizadas. Algumas ressalvas, registradas oralmente ou na Questão 24, foram:

- O nome mais apropriado para a categoria "Política Interna" poderia ser "Poder Legislativo";
- Para aumentar o interesse e a efetividade da aplicação do método de classificação temática em cada uma nas diversas unidades organizacionais, seria interessante a adequação das categorias temáticas utilizadas ao interesse de cada unidade. Por exemplo, na CONLE os temas poderiam seguir a divisão em áreas temáticas desse órgão.

Questão 6

Afirmação: 6. A classificação temática apresentada corresponde à realidade.

Objetivo: Avaliar o grau de fidelidade do método de classificação temática apresentado com a realidade.

Resultados:

Respostas	35
Média	4,1
Desvio Padrão	0,7

Comentários Os entrevistados concordaram parcial ou integralmente que o método apresentado retrata a realidade temática dos discursos, permitindo a análise de seus resultados e conseqüências.

Questões 7 e 8

Afirmações:
7. A classificação temática apresentada é uma forma de atribuir significado às informações armazenadas na base de dados.
8. Depois de conhecer a classificação temática apresentada, consigo descrever melhor o conteúdo da base de dados para outras pessoas.

Objetivo: Avaliar se o método de classificação temática aplicado atribuiu **significado** à informação armazenada na base de dados analisada.

Resultados:

Questão	**7**	**8**
Respostas	36	36
Média	4,7	4,5
Desvio Padrão	0,7	0,7

Comentários As duas questões tiveram resultados semelhantes e próximos a cinco, comprovando que o método de classificação temática é capaz de atribuir significado às informações armazenadas na base de dados analisada.

Questões 9 a 11

Afirmações A classificação temática apresentada possibilita a **identificação de padrões** nas informações armazenadas na base de dados que sejam:

9. Válidos.
10. Novos, ou anteriormente desconhecidos.
11. Potencialmente úteis.

Objetivo: Avaliar se o método de classificação temática aplicado é KDD (*Knowledge Discovery in Databases*) de acordo com a definição de Fayyad[21].

Resultados:

Questões	**9**	**10**	**11**
Respostas	35	32	35
Média	4,7	4,5	4,6
Desvio Padrão	0,5	0,8	0,6

Comentários As respostas a essas questões indicam que o método de classificação temática apresentado tem características que o torna uma ferramenta de KDD, segundo a definição de Fayyad.

[21] KDD é um processo não trivial de identificação de padrões válidos, novos e potencialmente úteis nos dados analisados (Fayyad, 1996).

Questões 12 a 17

Afirmações A atribuição de temas apresentada possibilita às pessoas interessadas em analisar a atuação parlamentar (por exemplo: assessores parlamentares, jornalistas, cientistas políticos, sociólogos) identificar:

12. A distribuição temática dos discursos proferidos pelos Deputados.
13. Os temas predominantes **por região**.
14. Os temas predominantes **por partido**.
15. A variação dos temas **ao longo do tempo**.
16. A correlação entre temas.
17. Outros (favor especificar).

Objetivo: Avaliar a utilidade potencial do cruzamento de cada uma das dimensões relacionadas com as categorias temáticas (se o método atribui **propósito** à informação armazenada).

Resultados:

Questão	**12**	**13**	**14**	**15**	**16**	**17**
Respostas	34	36	36	36	33	4
Média	4,7	4,7	4,8	4,6	4,1	4,3
Desvio Padrão	0,6	0,6	0,8	0,8	1,1	1,0

Comentários Os entrevistados consideraram mais relevantes, em ordem decrescente, as classificações temáticas:

- por partido;
- por região e geral;
- ao longo do tempo (mês a mês);
- correlação entre categorias.

Alguns entrevistados afirmaram que a classificação temática ao longo do tempo seria considerada mais relevante, desde que abrangesse períodos maiores que o analisado no presente trabalho, de modo que permitisse a identificação de tendências de aumento ou decréscimo consistentes de categorias temáticas.

Questões 18 a 22

Afirmações: A atribuição de temas apresentada fornece indicadores que podem ser úteis para:

18. Os Srs. Deputados.
19. A Câmara dos Deputados.
20. Minha unidade organizacional.
21. A Sociedade (público externo).
22. Outras pessoas ou instituições (favor especificar).

Objetivo: Avaliar a utilidade potencial do método de classificação temática para alguns perfis de usuários, ou seja, aferir se o método proposto atribui **propósito** às informações armazenadas.

Resultados:

Questões	**18**	**19**	**20**	**21**	**22**
Respostas	36	35	36	35	14
Média:	4,6	4,4	4,0	4,7	4,8
Desvio Padrão:	0,7	0,7	1,2	0,5	0,6

Comentários

As respostas a essas questões demonstram que o método de classificação temática é capaz de **atribuir propósito** às informações armazenadas na base de dados analisada.

Para os entrevistados, os maiores beneficiados com a aplicação do método e a análise dos resultados, seriam pessoas externas à Câmara dos Deputados, principalmente entidades que avaliam a atividade parlamentar como o DIAP[22] e as ONGs, as assessorias parlamentares e os grupos de pressão.

Curiosamente, a menor utilidade potencial identificada pelos entrevistados foi a relacionada com suas respectivas unidades organizacionais. Talvez isso ocorra por não haver hoje processos que procurem analisar os resultados obtidos e transformem em ações que contribuam para o melhor desempenho da unidade organizacional. Esta possibilidade poderia ser testada em futuros estudos.

[22] DIAP – Departamento de Intersindical de Estudos Parlamentares (www.diap.org.br).

Questão 23

Pergunta 23. Você considera válida a aplicação deste método de classificação temática em outras bases de dados da Câmara dos Deputados que tenham sido indexadas com auxílio de vocabulários controlados (como Proposições ou Legislação)?

Objetivo: Avaliar a possibilidade de generalização do método de classificação temática considerando sua possível aplicação em outras bases de dados da Câmara dos Deputados.

Resultados: Os 36 funcionários entrevistados responderam *sim* a esta pergunta, sendo que, destes, 11 (30,6%) fizeram a ressalva de que o método poderia ser aprimorado.

Comentários Os entrevistados consideraram válida a generalização do método para outras bases de dados igualmente indexadas e que dizem respeito à atividade parlamentar.

Questão 24

Pergunta 24. Por favor, relate aqui suas impressões, críticas e sugestões sobre a apresentação e sobre o método apresentado e possíveis utilizações.

Objetivo: Deixar os entrevistados livres para expressar tópicos não cobertos no questionário, bem como fazer críticas e propor sugestões à apresentação e ao método.

Resultados: 26 entrevistados (72,2%) registraram algum comentário.

Comentários

Os comentários mais freqüentes foram aqueles em que os entrevistados se revelaram surpresos com o tipo de análise sugerida com base na classificação temática apresentada.

Muitos entrevistados demonstraram grande interesse em que o método apresentado sofresse alguma adequação nas categorias utilizadas para que fosse aplicado nas respectivas unidades organizacionais.

Algumas respostas salientaram, ainda, a importância do uso sistemático do método apresentado como ferramenta para descrever a parte da atuação parlamentar relacionada com os discursos proferidos.

Alguns entrevistados demonstraram interesse e curiosidade na utilização do método de classificação temática para analisar:

- As possíveis mudanças temáticas nos discursos dos partidos na legislatura seguinte (2003-2006), quando os atuais partidos de situação e de oposição terão seus papéis invertidos;
- A possível influência do início da transmissão dos discursos dos Deputados pela TV Câmara a partir de 1998;
- A comparação da temática dos discursos em períodos bem diferentes, como legislaturas ou décadas.

6. CONCLUSÃO

Este trabalho teve como objetivos centrais a proposição, a aplicação e a avaliação de um método de classificação temática em bases de dados textuais indexadas com uso de vocabulário controlado. No entendimento do autor, esses objetivos foram plenamente atingidos.

O método proposto se baseia na indexação da base de dados analisada, que representa um trabalho intelectual realizado por especialistas ao atribuir descritores aos documentos. Ao ser posto em prática, o método comprovou sua praticidade e performance quando comparado a uma ferramenta de agrupamento não dependente da indexação.

A aplicação do método de classificação temática em uma base de dados laboratório, alimentada com discursos proferidos por deputados federais, possibilitou a identificação dos temas mais freqüentes no período analisado. A comparação desses temas com a unidade da Federação pela qual o parlamentar foi eleito, bem como de seu partido político identificou os temas mais freqüentes por região geográfica e por partido, respectivamente. A Análise da classificação temática por mês em que o discurso foi proferido apontou tendências de variação das freqüências dos temas ao longo do tempo. Finalmente, a correlação entre categorias temáticas mostrou como um tema pode estar associado a outros temas, já que o método de classificação temática permite a atribuição de mais de um tema a cada discurso.

A avaliação da aplicação do método de classificação temática foi realizada através de questionários respondidos por funcionários da Câmara dos Deputados

experientes e envolvidos profissionalmente com o assunto da base de dados analisada aos quais foram apresentados os resultados da classificação temática proposta neste trabalho. Em suas respostas, os entrevistados validaram a divisão temática apresentada ao considerarem que ela é adequada e corresponde à realidade. Também consideraram a classificação temática apresentada atribui significado às informações armazenadas e possibilita a identificação de padrões válidos, novos e potencialmente úteis, ou seja, permite a descoberta de conhecimentos.

Portanto, a hipótese formulada nesse trabalho, ou seja, "a classificação temática de bases de dados textuais através da atribuição de temas é uma forma de contextualização das informações armazenadas e agrega significado e propósito a elas, possibilitando assim a geração de novos conhecimentos" foi comprovada através das respostas dos entrevistados que formaram a amostra avaliada. Deve-se ressaltar, no entanto, que essa comprovação baseou-se em uma aplicação do método proposto em um subconjunto da base de dados de discursos, disponível na Câmara dos Deputados.

Em suas respostas ao questionário, os entrevistados consideraram válida a aplicação do método proposto em outras bases de dados, também indexadas com auxílio de vocabulário controlado. Para facilitar essa aplicação, em bases de dados de maior porte que a base de dados laboratório, algumas providências poderiam ser adotadas para aumentar o desempenho do software que implementa o método proposto, tais como:

- Adoção de arquitetura de duas (cliente/servidor) ou mais camadas;

- Otimização do código utilizado para implementar o método e utilização de linguagem compilada;
- Utilização de gerenciador de banco de dados com recurso de indexação textual.

Além dos aspectos tecnológicos, a utilização de um tesauro em vez de um vocabulário controlado na indexação da base de dados poderia simplificar significativamente a aplicação do método por associar cada descritor a um tema. É Importante salientar que a Câmara dos Deputados criou um grupo de trabalho encarregado de definir e atualizar um Tesauro que abranja os temas relativos à sua área de atuação.

6.1.Sugestões para Trabalhos Futuros

O fato de o método de classificação temática proposto depender de uma indexação prévia da base de dados analisada restringe, de certa forma, sua generalização. Porém, existem diversas outras bases de dados relevantes e disponíveis não só na Câmara dos Deputados, como também em outros órgãos da Administração Pública Federal brasileira. Dessa forma, futuros trabalhos a serem realizados por pesquisadores das áreas de Gestão do Conhecimento, Biblioteconomia, Ciência da Informação, História, Sociologia, Lingüística, Comunicação e Ciências Políticas poderiam incluir as seguintes linhas:

- Análise temática dos discursos proferidos por períodos históricos através da aplicação do método proposto (ou outro equivalente), em toda a base

de dados de Discursos, que engloba o período de 1945 até os dias de hoje;

- Estudos comparativos da temática da base de dados de Discursos com outras bases de dados também disponíveis na Câmara dos Deputados e indexadas com uso do mesmo vocabulário controlado. São exemplos de bases de dados com este perfil as de Proposições em Tramitação, de Legislação e de Questões de Ordem. Este estudo talvez possa estabelecer uma relação entre o discurso e a prática dos parlamentares;
- Utilização de ferramentas de indexação automáticas ou semi-automáticas (com revisão humana) para possibilitar o emprego deste método (ou outro equivalente) em outras bases de dados de grande porte e que não tenham sido indexadas. Se for comprovada a eficácia dessas ferramentas para esta finalidade, então talvez o método proposto poderá ser considerado de aplicação genérica tanto em relação ao tamanho das bases de dados como também em relação ao fato de serem ou não previamente indexadas. Outra possível vantagem do uso dessas ferramentas seria a eliminação dos efeitos decorrentes de mudanças dos critérios de indexação que certamente ocorreram no decorrer longos de períodos de tempo como o abrangido pela base de dados de Discursos.

6.2.Considerações Finais

Espera-se, como fruto deste trabalho, ter-se colaborado para o melhor entendimento de como se desenvolver um processo de descoberta de conhecimentos

a partir de acervos de informações textuais armazenadas em bancos de dados, disponíveis em muitas organizações. Para que esse processo seja bem sucedido, é imprescindível o auxílio de especialistas que direcionam o processo para atingir seus objetivos e avaliam se esses foram atingidos.

Na avaliação do autor, a aplicação de processos de descoberta de conhecimentos pode contribuir para a redução de alguns dos efeitos negativos da sobrecarga de informações e para o aumento da produtividade do trabalhador do conhecimento.

Os novos conhecimentos obtidos com a aplicação do processo de descoberta de conhecimentos apresentado nesse trabalho poderão ser compartilhados com a sociedade, possibilitando que a Câmara dos Deputados cumpra melhor suas atribuições institucionais, com o aumento da transparência das atividades parlamentares e melhor atendimento dos anseios dos cidadãos representados pelos deputados federais.

REFERÊNCIAS BIBLIOGRÁFICAS[23]

BAX, Marcelo, SOUZA, Renato. *Uma Proposta de Uso de Agentes e Mapas Conceituais para Representação de Conhecimentos Altamente Contextualizados.* 4o. Simpósio Internacional de Gestão do Conhecimento / Gestão de Documentos - ISKM/DM, 2001, Curitiba. Disponível em: http://cuba.eci.ufmg.br/Bax/Publis/agentes

BERNERS-LEE, Tim, MILLER, Eric. The Semantic Web Lifts Off. *ERCIM News*. n. 51, p. 9-11. Oct. 2002.
Disponível em: http://www.ercim.org/publication/Ercim_News/enw51/EN51.pdf

BRASIL. Tribunal de Contas da União. Técnicas de Apresentação de Dados. Brasília: junho 2001.
Disponível em: http://www.tcu.gov.br/Download/Relatorios/Tec.Ap.Dados(Roteiro).pdf

BRASIL. Tribunal de Contas da União. Técnicas de Amostragem para Auditorias. Brasília: março 2002.
Disponível em: http://www.tcu.gov.br/Download/Relatorios/Tec.Amostragem.pdf

BRASIL. Tribunal de Contas da União. Técnica de Entrevistas para Auditorias. Brasília: abril 1998.
Disponível em: http://www.tcu.gov.br/SAUDI/Download/Entrevista.exe

BRASIL. Câmara dos Deputados. Regimento Interno da Câmara dos Deputados. Disponível em: http://www.camara.gov.br/Internet/Regimento/default.asp

BROWN, John Seely, DUGUID, Paul. *The Social Life of Information*. Boston: Harvard Business School Press, 2000.

DAVENPORT, Thomas. *Ecologia da Informação: Porque a Tecnologia não Basta para o Sucesso na Era da Informação*. 3ª. ed. São Paulo: Futura, 2000.

DAVENPORT, Thomas, BECK, John. *A Economia da Atenção*. Rio de Janeiro: Campus, 2001.

DAVENPORT, Thomas, PRUSAK, Laurence. *Conhecimento Empresarial: Como as Organizações Gerenciam seu Capital Intelectual*. Rio de Janeiro: Campus, 1998.

DRUCKER, Peter. F. The Age of Social Transformation, *The Atlantic Monthly*, vol. 274, n. 5, p. 53-80, nov. 1994. 168 p.
Disponível em: www.theatlantic.com/atlantic/election/connection/ecbig/soctrans.htm

DRUCKER, Peter. F., *Desafios Gerenciais para o Século XXI*, 1ª ed. São Paulo: Pioneira, 1999.

ECHEVERRÍA, Rafael. *A Empresa Emergente: A confiança e os desafios da transformação.* Brasília: Universa, 2001.

[23] Todas as referências a sites Web foram re-visitadas em dezembro de 2002.

EUZENAT, Jerome. A Few Words About the Semantic Web and its Development in the ECRIM Institutes. *ERCIM News*. n. 51, p. 7-8. Oct. 2002.
Disponível em: http://www.ercim.org/publication/Ercim_News/enw51/EN51.pdf

FALCÃO, Sergio D. HERNANDES, Carlos A. M. e SANTANA, Roberto A. Sites de Busca na Internet. *Revista da III Jornada de Produção Científica das Universidades Católicas do Centro-Oeste*. Set. 1999. vol. I p. 136-140.

FAYYAD, Usama M. Data Mining and Knowledge Discovery: Making Sense out of data. *IEEE Expert* Oct. 1996, p. 20-25.

LUCAS, Marty. Mining in Textual Mountains. *Mappa.Mundi Magazine.* Jan. 1999. Disponível em: http://mappa.mundi.net/trip-m/hearst.

LYMAN, Peter, VARIAN Hal R. *How Much Information?* School of Information Management and Systems, Berkeley University. Oct. 2000.
Disponível em http://www.sims.berkeley.edu/research/projects/how-much-info/

MOSCAROLA, Jean et al. *Technology Watch via textual Data Analysis*. France: Université de Savoie. 1998. 23 p.
Disponível em: http://www.sphinxonline.com/Infos/Technological_watch.PDF.

NORTON, M. J. Knowledge Discovery in Databases. *Library Trends*. vol. 48 n. 1, p 9-21. University of Urbana. EUA: Urbana 1999.

OVERHOLT, Alison, Intel's Got (Too Much) Mail, *Fast Company.com*, n. 44 Mar. 2001, p 56. Disponível em: http://www.fastcompany.com/online/44/intel.html.

SPICER, Jeff, Industrial Revolution, *Oracle Magazine.* Jan 2003. Disponível em http://otn.oracle.com/oramag/oracle/03-jan/o13edit.html.

SILVA, Edilberto M., *Descoberta de Conhecimento com uso de Text Mining: Técnicas para Prover Inteligência Organizacional.* Dissertação de Mestrado em Gestão do Conhecimento e Tecnologia da Informação. Brasília: Universidade Católica de Brasília, 2002.

STEWART, Thomas A., *Capital Intelectual: A Nova Vantagem Competitiva das Empresas.* 5 ed., Rio de Janeiro: Campus, 1998.

TERRA, José Cláudio C. *Gestão do Conhecimento: o grande desafio empresarial: uma abordagem baseada no aprendizado e na criatividade*. São Paulo: Negócio Editora, 2000.

WILSON, Michael, MATTHEWS, Brian, Migrating Thesauri to the Semantic Web. *ERCIM News*. n. 51, p. 28-29. Oct. 2002.
Disponível em: http://www.ercim.org/publication/Ercim_News/enw51/EN51.pdf

WIVES, Leandro K., *Um Estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas Usando Técnicas de "Clustering".* Dissertação de Mestrado em Ciência da Computação. Porto Alegre: Universidade Federal do Rio Grande do Sul, 1999.

SITES DE INTERESSE

Os endereços de sites apresentados a seguir complementam as referências bibliográficas listadas como fonte de pesquisa e apontam para páginas ou listas de discussões relacionadas com a Gestão do Conhecimento em geral, bem como com a descoberta de conhecimento em bases de dados textuais e *text mining*.

Competitive Knowledge – lista de discussão aberta, moderada pela Informal Informática, cujos membros ajudaram a fundar a Sociedade Brasileira de Gestão do Conhecimento.	groups.yahoo.com/group/competitive-knowledge
CRIE - Centro de Referência em Inteligência Empresarial da COPPE – Universidade Federal do Rio de Janeiro	www.crie.coppe.ufrj.br
IBM Lotus Software - Knowledge Management.	www.lotus.com/km
José Cláudio C. Terra: contém sua tese de Doutorado e referências a seus livros.	www.terraforum.com.br/
KDnuggets - Data Mining, Knowledge Discovery, Genomic Mining, Web Mining.	www.kdnuggets.com/
KM Iniciatives – Lista de discussão	groups.yahoo.com/group/KMinitiatives/
Página da Web Semântica no W3C. Contém referências a padrões e projetos de pesquisa	www.w3c.org/2001/sw
Palo Alto Research Center (PARC)	www.parc.xerox.com/
Recursos de *text mining* (página organizada por Leandro Krug Wives).	www.inf.ufrgs.br/~wives/portugues/textmining.html
Search Engine Watch – disponibiliza extenso material sobre sites de busca.	searchenginewatch.com/
Sociedade Brasileira de Gestão do Conhecimento	www.sbgc.org.br/

ANEXO A – Descrição da Base de Dados Laboratório

A base de dados laboratório possui informações importadas do SITAQ II armazenadas em ambiente Microsoft Access XP.

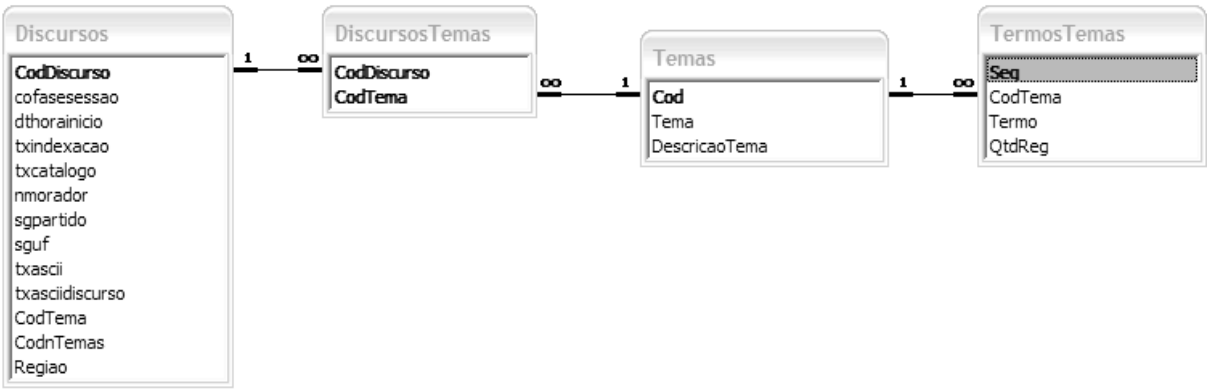

Figura 2 - Estrutura do Banco de Dados Laboratório

A seguir são descritas cada uma das tabelas utilizadas:

Tabela **Discursos**: Contém os discursos analisados.

Campo	Descrição
codDiscurso	Número seqüencial, chave primária da tabela
coFaseSessao	Código da fase da sessão (selecionados Grande Expediente e Pequeno Expediente)
dtHoraInicio	Data e hora do início do discurso
txIndexacao	Indexação do discurso
txCatalogo	Indexação do discurso (assunto principal)
nmOrador	Nome parlamentar do deputado que proferiu o discurso
sgPartido	Sigla do partido do deputado que proferiu o discurso
sgUF	Sigla da unidade da federação pela qual o deputado foi eleito
txAscii	Texto sem formatação do sumário do discurso
txAsciiDiscurso	Texto sem formatação da íntegra do discurso
Codtema	Identificador do tema atribuído ao discurso (usado no algoritmo que atribui apenas um tema por discurso)
Codntemas	Identificados dos temas atribuídos ao discurso (usado no algoritmo que atribui mais de um tema por discurso)
Regiao	Identificador da região geográfica da UF do deputado

Tabela **Temas:** Contém os temas utilizados na classificação temática.

Campo	Descrição
Cód	Identificador do tema
Tema	Descrição do tema

Tabela **DiscursosTemas:** Relacionamento entre discursos e temas

Campo	**Descrição**
CodDiscurso	Identificador do discurso
CodTema	Identificador do tema

Tabela **Termostemas**: Relacionamento entre os descritores usados para indexar os discursos e os temas.

Campo	**Descrição**
Seq	Número seqüencial, chave primária da tabela
Cód	Código do tema associado a um Descritor
Termo	Descritor que está associado a um tema
QtdReg	Quantidade de registros na tabela Discursos que possuem este descritor

ANEXO B - Descritores Utilizados para a Classificação Temática

Tema *Administração Federal*

Cod 14

DescricaoTema Executivo Federal, Judiciário

QtdReg	Termo
373	GOVERNO FEDERAL
146	PRESIDENTE DA REPÚBLICA
84	FERNANDO HENRIQUE CARDOSO
82	MEDIDA PROVISÓRIA
80	EXECUTIVO
64	PRESIDÊNCIA DA REPÚBLICA
60	EMENDA CONSTITUCIONAL
56	MINISTRO DE ESTADO
49	JUDICIÁRIO
35	MINISTÉRIO PÚBLICO
27	STF
16	CONSTITUIÇÃO FEDERAL
9	REELEIÇÃO
8	TCU
7	JUIZADO ESPECIAL
4	REFORMA ELEITORAL
3	ADMINISTRAÇÃO FEDERAL
2	PROCURADOR DA REPÚBLICA
1	DIREITO PENAL
1	ELEIÇÃO FEDERAL
1	MEDIDA PROVISÓRA
1	REFORMA CONSTITUCIONAL
1	STM

Tema *Agricultura*

Cod 25

DescricaoTema Política Rural, Reforma Agrária

QtdReg	Termo
206	RURAL
70	AGROPECUÁRIA
64	POLÍTICA AGRÍCOLA
62	REFORMA AGRÁRIA
60	AGRÍCOLA
54	PECUÁRIA
35	AGRÁRIA
33	MST
24	ALIMENTOS
24	IRRIGAÇÃO
24	LEITE
18	EMBRAPA
15	PESCA
15	POLÍTICA AGRÁRIA
14	AGRICULTOR
12	CACAU
12	CAFÉ
11	AGROTÓXICO
11	GRILAGEM
11	LAVOURA
11	SOJA
9	AGROPECUÁRIO
9	COOPERATIVISMO
8	CANA DE AÇÚCAR
7	FINANCIAMENTO AGRÍCOLA
7	POLÍTICA FUNDIÁRIA
6	SEM-TERRA

5 LATICÍNIO

4 AGRÔNOMO

4 CAJU

4 DOENÇA ANIMAL

4 PECUARISTA

3 ALGODÃO

3 COLONIZAÇÃO

3 EUCALIPTO

3 MILHO

2 ESTATUTO DA TERRA

1 AMENDOIM

1 CANA-DE-AÇÚCAR

1 COCO

1 DESAPROPRIADO

1 ERVA MATE

Tema *Corrupção*

Cod 17

DescricaoTema

QtdReg	Termo
274	CORRUPÇÃO
1	CORRRUPÇÃO

Tema *Economia*

Cod 27

DescricaoTema Finanças Públicas, Comércio e Indústria

QtdReg	Termo
176	COMÉRCIO
138	POLÍTICA ECONÔMICO FINANCEIRA
120	INDÚSTRIA
116	ORÇAMENTO
94	PRIVATIZAÇÃO
92	RECURSOS ORÇAMENTÁRIOS
78	IMPOSTO
74	RECURSOS PÚBLICOS
68	FINANCIAMENTO
65	ECONOMIA
62	TURISMO
50	EXPORTAÇÃO
47	REFORMA TRIBUTÁRIA
39	EMPRESÁRIO
37	INDUSTRIA
36	IMPORTAÇÃO
35	DESENVOLVIMENTO ECONÔMICO
35	RECURSOS FINANCEIROS
34	ORÇAMENTÁRIA
28	CPMF
27	POLÍTICA FISCAL
27	RESPONSABILIDADE FISCAL
26	JUROS
24	TRIBUTAÇÃO
24	TRIBUTOS
19	GLOBALIZAÇÃO
18	MODELO ECONÔMICO

18 SISTEMA FINANCEIRO

17 POLÍTICA ECONOMICO FINANCEIRA

15 (CEF)

14 CRESCIMENTO ECONÔMICO

14 DÍVIDA PÚBLICA

14 INSTITUIÇÃO FINANCEIRA

14 MICROEMPRESA

14 POLÍTICA ORÇAMENTÁRIA

13 DÍVIDA EXTERNA

12 EMPRESA NACIONAL

11 BANCO DO BRASIL

11 POLÍTICA SÓCIO ECONÔMICA

11 SONEGAÇÃO FISCAL

9 CONSTRUÇÃO CIVIL

9 FINANÇAS

7 DESENVOLVIMENTO NACIONAL

7 FÁBRICA

7 INCENTIVO FISCAL

7 SISTEMA TRIBUTÁRIO NACIONAL

6 MERCADO FINANCEIRO

6 POLÍTICA DE CRÉDITO

6 SERVIÇO BANCÁRIO

5 DESENVOLVIMENTO INDUSTRIAL

5 FALÊNCIA

5 MOEDA

5 SISTEMA BANCÁRIO NACIONAL

4 (ICMS)

4 CARTÃO DE CRÉDITO

4 COMERCIO

4 EMBRAER

4 VEÍCULO AUTOMOTOR

3 (BANESPA)

3 (BANESTADO)

3 (BNDES)

2 (IPI)

2 COMPLEXO INDUSTRIAL

2 CUSTO DE VIDA

2 DÉBITO FISCAL

2 EVASÃO FISCAL

2 LEGISLAÇÃO TRIBUTÁRIA

2 SIGILO BANCÁRIO

1 (BACEN)

1 (BASA)

1 (USIMINAS)

1 ANÁLISE FISCAL.

1 BANCO DE DESENVOLVIMENTO

1 CONCORDATA

1 CONSTRUÇÃO NAVAL

1 CRÉDITO ORÇAMENTÁRIO

1 CRÉDITO SUPLEMENTAR

1 DESNACIONALIZAÇÃO

1 ÍNDÚSTRIA

1 POLITICA ECONOMICO FINANCEIRA

1 POLÍTICA FINANCEIRA

1 RECURSOS FINANCIEROS

Tema *Educação*

Cod 20

DescricaoTema Ciência, Tecnologia, Informática, Cultura e Esportes

QtdReg	Termo
231	CULTURA
219	EDUCAÇÃO
159	ENSINO
123	JORNAL
107	PROFESSOR
90	UNIVERSIDADE
57	ESCOLA
54	MEIOS DE COMUNICAÇÃO
44	TELEVISÃO
38	EDUCACIONAL
33	ESPORTE
28	CIÊNCIA
28	FUTEBOL
27	ESTUDANTE
27	RÁDIO
24	ESCRITOR
21	TECNOLOGIA
18	CIENTÍFICA
18	ESCOLAR
15	FACULDADE
14	DESENVOLVIMENTO TECNOLÓGICO
14	INTERNET
14	POLÍTICA EDUCACIONAL
13	LIVRO
12	RÁDIO COMUNITÁRIA
11	INFORMÁTICA
10	ATLETA

9 BOLSA ESCOLA

8 ESTUDANTIL

7 CRÉDITO EDUCATIVO

6 CURSO SUPERIOR

6 MERENDA ESCOLAR

6 POETA

6 REITOR

5 CINEMA

5 FUNDEF

2 (MEC)

2 ALFABETIZAÇÃO

2 CIENTISTA

2 PATRIMONIO HISTÓRICO

2 PEDAGOGO

1 (CNPQ)

1 (IPHAN)

1 (UFPR)

1 (UNESCO)

1 ARTE POPULAR

1 CREDITO EDUCATIVO

1 PATRIMÔNIO ARQUEOLÓGICO

Tema *Infra-estrutura*

Cod 26

DescricaoTema Transporte, Energia e Telecomunicações

QtdReg	Termo
196	ENERGIA
191	RODOVIA
190	POLÍTICA ENERGÉTICA
59	TRANSPORTE
39	PETROBRAS
37	PETROBRÁS
37	PETRÓLEO
28	RODOVIÁRIO
27	TRÂNSITO
25	TELECOMUNICAÇÃO
24	FERROVIÁRIO
23	PORTO
18	DNER
18	HIDROELÉTRICA
17	FERROVIA
16	TELECOMUNICAÇÕES
15	GÁS NATURAL
15	RFFSA
14	TELEFONIA
13	AEROPORTO
13	HORÁRIO DE VERÃO
13	RADIODIFUSÃO
11	CÓDIGO DE TRÂNSITO BRASILEIRO
11	TERMOELÉTRICA
10	COMBUSTÍVEL
10	SISTEMA ELÉTRICO
9	PONTE

9 TRANSPORTES

8 INFRAERO

7 ALCOOL

7 ÁLCOOL

7 HIDROVIA

7 MINERAÇÃO

6 ANEEL

6 NUCLEAR

6 TELEFONE

5 PEDÁGIO

5 REDE RODOVIÁRIA

4 (ECT)

4 GASOLINA

4 RECURSOS ENERGÉTICOS

3 MATRIZ ENERGÉTICA

2 (CELG)

2 (MME)

2 FERROVIÁRIA

1 (CNT)

1 (DNPM)

1 (FURNAS)

1 (FUST)

1 (GLP)

1 CODIGO DE TRANSITO BRASILEIRO

1 ESTRADAS VICINAIS

1 HIDROELETRICA

1 HIDROVIÁRIO

1 METRÔ

1 PETROQUÍMICA

1 PISTA DE POUSO

1 VIADUTO

Tema *Meio Ambiente*

Cod 23

DescricaoTema Recursos Hídricos

QtdReg	Termo
97	MEIO AMBIENTE
54	SECA
50	ÁGUA
44	RECURSOS HÍDRICOS
29	FLORESTAL
28	RIO SÃO FRANCISCO
25	ABASTECIMENTO DE ÁGUA
16	FLORESTA
11	CHUVA
11	ECOLOGIA
10	LIXO
9	CODEVASF
9	IMPACTO AMBIENTAL
7	PARQUE NACIONAL
6	(DNOCS)
4	POLÍTICA AMBIENTAL
3	FAUNA
3	RESERVA ECOLÓGICA
2	DESERTIFICAÇÃO
1	(IBAMA)
1	ATMOSFERA
1	BIOSSEGURANÇA
1	FLORA
1	PULUIÇÃO
1	QUEIMADA

Tema *Política Externa*

Cod 29

DescricaoTema Defesa Nacional

QtdReg	Termo
115	PAÍS ESTRANGEIRO
52	(EUA)
45	TERRORISMO
40	ALCA
31	EXÉRCITO
21	PALESTINA
20	POLÍTICA EXTERNA
19	ORIENTE MÉDIO
17	FORÇAS ARMADAS
16	POLÍTICA INTERNACIONAL
15	ARGENTINA
15	DEFESA NACIONAL
13	ACORDO INTERNACIONAL
13	MERCOSUL
11	RELAÇÕES INTERNACIONAIS
11	SOBERANIA
10	SOBERANIA NACIONAL
9	ORGANISMO INTERNACIONAL
9	RELAÇÕES DIPLOMÁTICAS
8	(ABIN)
6	AMÉRICA LATINA
6	SIVAM
5	CONFERÊNCIA INTERNACIONAL
5	DIPLOMACIA
3	AERONÁUTICA
3	MARINHA
3	SEGURANÇA NACIONAL

2 (ONU)

1 (FAB)

1 COMANDO MILITAR

1 DIPLOMATA

1 EMBAIXADA ESTRANGEIRA

1 SOLDADO

1 TIMOR LESTE

Tema *Política Interna*

Cod 16

DescricaoTema Congresso Nacional, Câmara dos Deputados, Senado Federal

QtdReg	Termo
389	PROJETO DE LEI
292	PARTIDO POLÍTICO
214	DEPUTADO FEDERAL
170	CÂMARA DOS DEPUTADOS
105	SENADO
71	ORDEM DO DIA
59	CONGRESSO NACIONAL
54	LEGISLATIVO
47	REFORMA POLÍTICA
39	MESA DIRETORA
28	ATUAÇÃO PARLAMENTAR
27	IMUNIDADE PARLAMENTAR
23	CONGRESSISTA
19	OBSTRUÇÃO PARLAMENTAR
18	(CPI)
18	MANDATO
17	EX DEPUTADO
13	SESSÃO SOLENE
11	DECORO PARLAMENTAR
9	FIDELIDADE PARTIDÁRIA
7	COMISSÃO PERMANENTE
4	SESSÃO ORDINÁRIA
4	VOTAÇÃO SECRETA
3	PARLAMENTARISMO
3	POLÍTICA PARTIDÁRIA
1	PEQUENO EXPEDIENTE

Tema *Política Regional*

Cod 28

DescricaoTema Administração Estadual, Administração Municipal

QtdReg	Termo
333	MUNICÍPIO
229	ESTADUAL
220	GOVERNO ESTADUAL
197	MUNICIPAL
190	PREFEITO
142	GOVERNADOR
141	REGIONAL
139	MUNICÍPIOS
109	CIDADE
78	REGIÃO NORDESTE
60	AMAZÔNICA
47	EMANCIPAÇÃO POLÍTICA
35	SUDAM
33	VEREADOR
31	SUDENE
24	(RJ)
24	(SP)
19	(CE)
18	(AC)
17	(PI)
15	(RO)
13	(BA)
13	(PE)
13	(RS)
12	(PR)
10	(GO)
10	(MA)

10 (SC)

9 (MTS)

7 (ES)

6 (AM)

6 (DF)

6 (FCO)

6 (RR)

6 REGIÃO NORTE

5 (AL)

5 (PB)

5 (SE)

4 (MG)

4 (PA)

4 AMAZÔNIA LEGAL

3 (AP)

3 (MT)

3 (TO)

2 (GDF)

1 (RN)

1 PROCURADOR GERAL DE ESTADO

Tema *Política Social*

Cod 19

DescricaoTema Trabalho, Emprego, Minorias, Previdência, Securidade Social, Religião

QtdReg	Termo
262	TRABALHO
237	POLÍTICA SOCIAL
217	CLT
199	SALÁRIO
158	GREVE
134	POLÍTICA SALARIAL
123	MULHER
112	DISCRIMINAÇÃO
86	SERVIDOR PÚBLICO CIVIL
81	EMPREGO
71	DESIGUALDADE SOCIAL
63	NEGRO
59	POBREZA
56	FGTS
55	DIREITOS HUMANOS
54	CRIANÇA
52	DESEMPREGO
52	IGREJA
47	IDOSO
45	APOSENTADO
44	PREVIDÊNCIA SOCIAL
43	FOME
43	HABITAÇÃO
42	PREVIDÊNCIA
40	POLÍTICA HABITACIONAL
38	DESENVOLVIMENTO SOCIAL
32	INDÍGENA

29 ÍNDIO

27 DEFESA DO CONSUMIDOR

26 ASSISTÊNCIA SOCIAL

25 POLICIAL

24 PLANO DE CARREIRA

22 ADOLESCENTE

21 HABITACIONAL

20 LEGISLAÇÃO TRABALHISTA

19 MERCADO DE TRABALHO

15 CONCURSO PÚBLICO

13 CAMPANHA DA FRATERNIDADE

13 DEFICIENTE FÍSICO

13 SEGURIDADE SOCIAL

12 HOMOSSEXUAL

12 PENSIONISTA

10 DIREITOS SOCIAIS

10 EMPREGADO

10 PREVIDENCIÁRIO

9 CONTRIBUIÇÃO PREVIDENCIÁRIA

9 FORMAÇÃO PROFISSIONAL

9 JUSTIÇA SOCIAL

6 (FAT)

6 PESSOA DEFICIENTE

6 POLÍTICA PREVIDENCIÁRIA

5 ACIDENTE DO TRABALHO

5 MUTUÁRIO

5 TORTURA

4 CATEGORIA PROFISSIONAL

4 FAVELA

4 PROVENTOS

3 BEM ESTAR SOCIAL

3 BOLSA ALIMENTAÇÃO

2 (PIS)

2 EXPECTATIVA DE VIDA

2 INDIO

2 REAJUSTE SALARIAL

2 SEGREGAÇÃO RACIAL

2 SERVIDOR PÚBLICO FEDERAL

2 SITUAÇÃO SOCIAL

1 DIREITOS E GARANTIAS INDIVIDUAIS

1 FEBEM

1 ISONOMIA SALARIAL

1 PERSEGUIÇÃO RELIGIOSA

1 SERVIÇO SOCIAL

Tema *Saúde Pública*

Cod 22

DescricaoTema

QtdReg	Termo
259	SAÚDE
99	SANEAMENTO
58	MEDICAMENTOS
57	MÉDICO
49	DOENÇA
40	HOSPITAL
38	DENGUE
36	MEDICAMENTO
29	INSS
25	AIDS
19	FEBRE
13	MEDICINA
12	TABAGISMO
10	CÂNCER
9	VIGILÂNCIA SANITÁRIA
7	ABORTO
6	DEFESA SANITÁRIA
5	(SUS)
5	ENFERMAGEM
4	DOENÇA ENDÊMICA
4	DOENÇA TRANSMISSÍVEL
4	FUMO
4	POLÍTICA SANITÁRIA
3	DEPENDÊNCIA QUÍMICA
3	MORTALIDADE MATERNA.
2	ACUPUNTURA
2	ENDEMIA

2 ÓRGÃO HUMANO

1 AMAMENTAÇÃO

1 DEPENDÊNCIA, QUÍMICA

1 HANSENÍASE

Tema *Segurança Pública*

Cod 18

DescricaoTema

QtdReg	Termo
272	SEGURANÇA PÚBLICA
175	VIOLÊNCIA
171	POLÍCIA
82	MILITAR
80	CRIME
51	HOMICÍDIO
44	DROGA
35	TRÁFICO
22	PENITENCIÁRIO
15	FRONTEIRA
15	TRÁFICO INTERNACIONAL
12	PENITENCIÁRIA
12	PRESÍDIO
10	DELINQÜÊNCIA JUVENIL
7	ROUBO
6	PORTE DE ARMA
4	DROGAS
2	GRUPO DE EXTERMÍNIO
1	POLICIAIS
1	SISTEMA PENITENCÁRIO
1	TRATAMENTO DE PRESO

ANEXO C – Código Fonte do Método de Classificação Temática

A listagem abaixo é o código fonte do procedimento que atribui temas aos discursos da base de dados laboratório. Foi utilizada a linguagem VBA – Visual Basic for Applications em ambiente Microsoft Access XP.

```
Private Sub atribuiVariosTemas()
    Dim db        As Database   '-- Base de dados Discursos.
    Dim RsTermos  As Recordset  '-- RS da Tabela TermosTemas.
    Dim RsDisc    As Recordset  '-- RS da tabela Discursos.
    Dim strSQL    As String     '-- String para comandos SQL.
    Dim NumTermos As Long       '-- Contador de registros em TermosTemas.
    Dim Inicio    As Date       '-- Horáro de início do procedimento.
    Dim Duracao   As Date       '-- Duração do procedimento.

    Me.lst1TemaDisc.Visible = False
    Me.lstResultado.Visible = False
    InicializaTemas (nTemasXDisc)
    Screen.MousePointer = 11    '-- cursor = ampulheta
    Set db = CurrentDb          '-- abre a base de dados de Discursos
    Me.PgrProgresso.Visible = True

    Inicio = Now()              '-- começa a contar o tempo
    Set RsTermos = db.OpenRecordset("SELECT Termo, CodTema FROM TermosTemas")
    RsTermos.MoveLast
    RsTermos.MoveFirst
    While Not RsTermos.EOF      '-- Para cada Termo na Tabela TermosTemas ...
         strSQL = "          SELECT CodDiscurso, Codntemas FROM Discursos "
         strSQL = strSQL & "WHERE(txCatalogo LIKE '*"&RsTermos.Fields("Termo") & "*') "
         strSQL = strSQL & " OR((TxCatalogo IS NULL)AND(txIndexacao LIKE '*" &
                           RsTermos.Fields("Termo") & "*'))"
         Set RsDisc = db.OpenRecordset(strSQL)
         While Not RsDisc.EOF
            RsDisc.Edit
            RsDisc.Fields("CodnTemas") = RsDisc.Fields("CodnTemas") &
                        RsTermos.Fields("Codtema") & ","
            RsDisc.Update
            db.Execute "INSERT INTO DiscursosTemas VALUES (" &
               RsDisc.Fields("CodDiscurso") & "," & RsTermos.Fields("Codtema") & ")"
            RsDisc.MoveNext
         Wend
         Me.PgrProgresso.Value = RsTermos.PercentPosition
         DoEvents
         RsTermos.MoveNext
    Wend '-- Próximo termo em TermosTemas

    '-- atribui o tema "Outros" aos discursos ainda não classificados, mas indexados

    strSQL = "         SELECT CodDiscurso FROM Discursos "
    strSQL = strSQL & "WHERE CodnTemas is NULL AND((txCatalogo IS NOT NULL) OR
                           (txindexacao IS NOT NULL))"
    Set RsDisc = db.OpenRecordset(strSQL)
    While Not RsDisc.EOF
        db.Execute "INSERT INTO DiscursosTemas VALUES (" & RsDisc.Fields("CodDiscurso")
                           & "," & CodTemaOutros & ")"
        RsDisc.MoveNext
    Wend
    Me.PgrProgresso.Visible = False
    Screen.MousePointer = 0
    Duracao = Now() - Inicio
    MsgBox "Duração: " & Minute(Duracao) & " min e " & Second(Duracao) & " seg",
            vbInformation, "Final de Atualização"
    RsTermos.Close
```

```
    Set RsTermos = Nothing
    db.Close
    Set db = Nothing
    Me.lstResultado.Requery
    Me.lstResultado.Visible = True

End Sub
```

ANEXO D - Questionário

Finalidade
Este questionário tem por objetivo obter a avaliação, por parte de servidores da Câmara dos Deputados convidados, sobre a importância de realizar a classificação temática de discursos proferidos pelos senhores Deputados, tendo como referência uma base de dados do SITAQ II – Sistema de Registro de Notas Taquigráficas.

Instruções
Por favor, responda às questões abaixo com base na sua experiência profissional e na apresentação a que você acabou de assistir.

Perfil do entrevistado

1. Você é servidor do quadro efetivo da Câmara dos Deputados?

( ) Não. ( ) Sim, há _____ anos.

2. Você ocupa cargo ou função de nível superior?

( ) Não. ( ) Sim.

3. Em qual órgão da Câmara dos Deputados você trabalha?

( ) SGM
( ) DILEG
( ) CEDI
( ) DETAQ
( ) DECOM
( ) CONLE
( ) CENIN
( ) Outro: _____

4. Com que freqüência você utiliza a base de dados de discursos ou acompanha os discursos proferidos pelos Srs. Deputados?

( ) Nenhuma: não utilizo a base de dados nem acompanho os discursos.
( ) Raramente.
( ) Eventualmente: algumas vezes por mês.
( ) Freqüentemente: algumas vezes por semana.
( ) Diariamente.

Avaliação do método proposto

Nas questões abaixo, assinale com um X a alternativa que melhor representa o seu julgamento, baseado na apresentação a que você assistiu. Para isso, utilize a seguinte escala:

concordo integralmente	concordo parcialmente	indiferente	discordo parcialmente	discordo integralmente
5	4	3	2	1

Caso não saiba ou não queira responder, deixe a questão em branco. Por favor, só utilize este recurso em último caso.

	5	4	3	2	1
5. A **divisão temática** apresentada é adequada.					
6. A classificação temática apresentada **corresponde à realidade**.					
7. A classificação temática apresentada é uma forma de atribuir **significado** às informações armazenadas na base de dados.					
8. Depois de conhecer a classificação temática apresentada, **consigo descrever melhor** o conteúdo da base de dados para outras pessoas.					

A classificação temática apresentada possibilita a **identificação de padrões** nas informações armazenadas na base de dados que sejam:

9. Válidos.					
10. Novos, ou anteriormente desconhecidos.					
11. Potencialmente úteis.					

A atribuição de temas apresentada possibilita às pessoas interessadas em analisar a atuação parlamentar (por exemplo: assessores parlamentares, jornalistas, cientistas políticos, sociólogos) identificar:

12. A distribuição temática dos discursos proferidos pelos Deputados.					
13. O perfil das **bancadas regionais.**					
14. O perfil das **bancadas partidárias.**					
15. A variação dos temas **ao longo do tempo.**					
16. A **correlação** entre categorias temáticas.					
17. Outros. (favor especificar):					

A atribuição de temas apresentada fornece indicadores que podem ser úteis para:

18. Os **Srs. Deputados e os Srs. Líderes** avaliarem e aprimorarem sua atuação parlamentar**.**					
19. A **Câmara dos Deputados** realizar suas atribuições de forma mais eficiente ou eficaz.					
20. A **minha unidade organizacional** realizar suas atribuições de forma mais eficiente ou eficaz.					

21. A **sociedade** (público externo) conhecer melhor a atuação parlamentar de seus representantes.					
22. Outras pessoas ou instituições (favor especificar):					

Possibilidade de generalização da aplicação do método

23. Você considera válida a aplicação deste método de classificação temática em outras bases de dados da Câmara dos Deputados que tenham sido indexadas com auxílio de vocabulários controlados (como Proposições ou Legislação)?

( ) Sim.
( ) Sim, desde que o método fosse aprimorado.
( ) Não, mas talvez outro método.
( ) Não, método nenhum.
( ) Não sei dizer.

Comentários

24. Por favor, relate aqui suas impressões, críticas e sugestões sobre a apresentação e sobre o método apresentado e possíveis utilizações.

Desejo receber os resultados desta pesquisa. Meu endereço de e-mail é:

_____

Muito obrigado por ter comparecido e respondido ao questionário.