COPPE/UFRJ

ESTUDO DE UMA METODOLOGIA DE MINERAÇÃO DE TEXTOS CIENTÍFICOS EM LÍNGUA PORTUGUESA

Ingrid Martins de Oliveira

Tese de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Civil.

Orientador: Nelson Francisco Favilla Ebecken

Rio de Janeiro
Julho de 2009

ESTUDO DE UMA METODOLOGIA DE MINERAÇÃO DE TEXTOS CIENTÍFICOS EM LÍNGUA PORTUGUESA

Ingrid Martins de Oliveira

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA CIVIL.

Aprovada por:

_____
Prof. Nelson Francisco Favilla Ebecken, D.Sc.

_____
Prof. Beatriz de Souza Leite Pires de Lima, D.Sc.

_____
Prof. Valeria Menezes Bastos, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
JULHO DE 2009

Oliveira, Ingrid Martins de

Estudo de uma Metodologia de Mineração de Textos Científicos em Língua Portuguesa / Ingrid Martins de Oliveira. – Rio de Janeiro: UFRJ/COPPE, 2009.

VIII, 65 p.: il.; 29,7 cm.

Orientador: Nelson Francisco Favilla Ebecken

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia Civil, 2009.

Referências Bibliográficas: p. 57-59.

1. Mineração de Texto. 2. Clusterização. 3. Extração de Conhecimento. I. Ebecken, Nelson Francisco Favilla II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Titulo.

Dedico este trabalho aos profissionais de informação, que com os avanços tecnológicos, precisam lidar diariamente com enormes massas de dados, organizando e disseminando informações.

AGRADECIMENTOS

Acima de tudo, agradeço a Deus por tudo que me proporciona e por se fazer mais que presente, mostrando-me muitas vezes que durante a angústia e o cansaço, ainda vale a pena. Obrigada Senhor pelas chuvas de bençãos que derrama todos os dias sobre a minha vida e sobre a minha casa.

À Fundação Biblioteca Nacional pelo suporte financeiro viabilizando o desenvolvimento desta pesquisa.

Ao meu orientador, prof. Nelson Francisco Favilla Ebecken pelo apoio e incentivo no desenvolvimento da dissertação.

Agradeço especialmente a grande incentivadora do meu mestrado, Valéria Bastos. Obrigada por toda ajuda, dedicação, atenção, orientação e apoio. E acima de tudo, pela amizade demonstrada.

Agradeço à minha família, em especial aos meus pais, Dulce e Nelson, por toda dedicação, apoio e compreensão nos tempos difíceis. Delle, Gí e Edinho, obrigada pelo estímulo e carinho demonstrados.

Agradeço a Well, Môrmeu da minha vida, por toda paciência, incentivo, amor e compreensão. Obrigada por estar sempre ao meu lado.

Agradeço a todos os professores do curso por todas as dicas e ensinamentos.

A todos os colegas do mestrado pelas trocas, pela amizade e companheirismo. Agradeço em especial ao Renan de Souza e ao Bruno Vilela .

À *Cortex Intelligence* pela confiança e compreensão, principalmente nesta reta final e pela disponibilização do Cortex e outras ferramentas. Agradeço especialmente ao Christian Aranha, Daniel Chada, Carolina Monte e Lucas Porto pela ajuda e pelas boas idéias.

À equipe do Sistema Maxwell por disponibilizar a base de teses para o desenvolvimento da pesquisa.

A todos que me incentivaram e apoiaram de forma direta e indireta no decorrer do curso e principalmente no final. Obrigada Joana pela ajuda e preocupação.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

ESTUDO DE UMA METODOLOGIA DE MINERAÇÃO DE TEXTOS CIENTÍFICOS EM LÍNGUA PORTUGUESA

Ingrid Martins de Oliveira

Julho / 2009

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

Este trabalho destina-se à extração de conhecimento de textos científicos e/ou literários em língua portuguesa através das análises estatística e contextual. Propõe-se desenvolver uma metodologia de Mineração de Textos, aplicando a técnica de *Clustering* na coleção de teses digitais da PUC-Rio, disponível no Sistema Maxwell. Esta técnica possibilita o agrupamento de textos em português, segundo a similaridade dos conteúdos para auxiliar a distribuição de documentos para determinados perfis de usuários ou pesquisas. Com a validação da metodologia desenvolvida, esta poderá ser aplicada em outros conjuntos de documentos em português, seja científico ou literário, por ser considerado um estudo genérico. Apresenta como trabalhos futuros, a aplicação da metodologia no acervo da Fundação Biblioteca Nacional e a utilização de um classificador para complementar a otimização do processo de indexação de documentos.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

STUDY OF A TEXT MINING METHODOLOGY FROM SCIENTIFIC DOCUMENTS IN PORTUGUESE LANGUAGE

Ingrid Martins de Oliveira

July / 2009

Advisor: Nelson Francisco Favilla Ebecken

Department: Civil Engineering

This work is applied to the knowledge extraction of scientific and/or literary texts in Portuguese language through the contextual and statistical analysis. The purpose is to develop a Text Mining methodology, applying Clustering techniques in the digital thesis collection, available in the PUC-Rio University Library. This technique makes possible grouping texts in Portuguese, according to contents similarity, allowing the document distribution for users' or research's profiles. With the validation of the developed methodology, this could be applied in other set of documents in Portuguese, either scientific or literary, for being considered a generic study. As future work, it can be suggested the application of this methodology in the digital documents of National Library Foundation and the use of a classifier to complement the document indexing optimization process.

Sumário

1 INTRODUÇÃO 1

1.1 MOTIVAÇÃO 2
1.2 OBJETIVOS DA PESQUISA 3
1.3 LINHAS GERAIS DA DISSERTAÇÃO 3

2 MINERAÇÃO DE TEXTOS 5

3 FERRAMENTAS DE MINERAÇÃO DE TEXTOS 10

3.1 CORTEX 10
3.2 REFVIZ 13
3.3 MCL SYSTEM 16

4 AMBIENTES DE APLICAÇÃO DA METODOLOGIA 19

4.1 SISTEMA MAXWELL 19
4.2 FUNDAÇÃO BIBLIOTECA NACIONAL 21

5 DESENVOLVIMENTO DA METODOLOGIA 23

5.1 COLETA DOS DADOS 24
5.2 PRÉ-PROCESSAMENTO 25
5.3 PROCESSAMENTO 27
5.4 ANÁLISE DE RESULTADOS 28

6 ESTUDOS DE CASO 29

6.1 MCL SYSTEM 29
6.1.1 Com o uso do Cortex 30
6.1.2 Sem o uso do Cortex 35
6.2 REFVIZ 43
6.2.1 RefViz com Stoplist 44
6.2.2 RefViz sem Stoplist 48
6.2.3 RefViz com Cortex 49

7 CONCLUSÕES 55

REFERÊNCIAS 57

APÊNDICE A - STOPLIST 60

APÊNDICE B – AMOSTRA DO THESAURUS CONSTRUÍDO PARA O REFVIZ 62

1 Introdução

Métodos de recuperação de textos sempre foram utilizados para organizar documentos. Porém, com o crescimento contínuo de informações e documentos em meio eletrônico, principalmente, pela digitalização e disponibilização de documentos e pela Internet, que técnicas automáticas de tratamento de textos e extração de conhecimento tornam-se cada vez mais necessárias para valorizar a gigantesca quantidade de dados armazenada nos sistemas de informação. Para solucionar esses problemas surge uma nova linha de pesquisa, a Mineração de Textos (*Text Mining)*.

O cotidiano das pessoas e das grandes organizações é caracterizado por uma massa crescente de dados e informações armazenados em meio eletrônico. Inúmeras páginas compostas de textos são lançadas todos os dias na Internet. Diversos tipos de documentos, como atas, comunicações internas, relatórios, currículos, entre outros, são gerados periodicamente e são constantemente atualizados. Hoje, com o advento da tecnologia de Mineração de Textos, todos os tipos de documentos produzidos pelas organizações podem ser explorados e assim, fornecer significativas vantagens para as organizações como um todo e conseqüentemente para o desempenho de seus funcionários. A extração de informações em textos se tornou possível e o conhecimento extraído passou a servir como suporte à tomada de decisões e ainda como indicador de sucesso ou fracasso. Ou seja, as aplicações desta tecnologia podem fornecer novas dimensões das informações disponíveis nas organizações.

É de conhecimento dos pesquisadores que ao surgir uma nova área de pesquisa, são necessários muitos estudos e discussões dentro da academia para que alguns padrões sejam estabelecidos. De acordo com este procedimento, o desenvolvimento de técnicas específicas de Mineração de Textos ganhou grande importância nos últimos anos, visto que as técnicas de Mineração de Dados (*Data Mining*), já desenvolvidas e estudadas há mais tempo, são voltadas para dados estruturados.

Algumas definições de Mineração de Textos são propostas por vários pesquisadores. Algumas encontradas na literatura levantada podem ser analisadas a seguir. Segundo Chen (2001), *Text Mining* realiza várias funções de busca, análise lingüística e categorização. Ressalta que mecanismos de busca se restringem à Internet.

Já para Sullivan (2000), *Text Mining* é o estudo e a prática de extrair informação de textos usando os princípios da lingüística computacional. Biggs (2005) diz que *Text Mining* é ideal para inspecionar mudanças no mercado, ou para identificar idéias. Outro autor encontrado foi Tan (1999), que define Descoberta de Conhecimento em Textos (KDT) ou *Text Mining* como sendo o processo de extrair padrões ou conhecimento, interessantes e não-triviais, a partir de documentos textuais. Mais três definições bastante interessantes encontradas no levantamento da literatura da área são as seguintes. Para Lucas (2000), *Text Mining* é uma forma de examinar uma coleção de documentos e descobrir informação não contida em nenhum dos documentos. Na visão de Hearst (1999), *Text Mining,* como análise de dados exploratória, é um método para apoiar pesquisadores a derivar novas e relevantes informações de uma grande coleção de textos. É um processo parcialmente automatizado onde o pesquisador ainda está envolvido, interagindo com o sistema. Por fim, Thuraisingham (1999) visualiza *Text Mining* como sendo *Data Mining* em dados textuais. *Text Mining* tem como objetivo extrair padrões e associações desconhecidas de um grande banco de dados textual. Para a Cortex Intelligence (2007), *Text Mining* pode ser definido como um processo que utiliza métodos para navegar, organizar, achar e descobrir informação em bases textuais escritas em linguagem natural. A empresa ressalta que é possível manipular mais facilmente informações não estruturadas como notícias, textos em web sites, blogs e documentos em geral, utilizando a tecnologia.

Como se pode observar na literatura científica, Mineração de Textos possui várias funções para atender a diferentes demandas. Algumas demandas interessantes para o desenvolvimento desta pesquisa são a busca e recuperação de informação (*information retrieval*), clusterização (*clustering*) e classificação de textos. A estruturação lógica de bases textuais, através das tarefas e ferramentas que a mineração de textos dispõe, é uma tarefa desafiadora.

1.1 Motivação

Apesar de estudos sobre o tema da pesquisa já estarem bastante avançados, percebe-se na literatura científica, a escassez de aplicações de técnicas de mineração em textos da língua portuguesa. Tal escassez se revela pela grande maioria de documentos disponíveis em meio eletrônico se encontrarem na língua inglesa. Isto evidencia o grau de interesse e de estudos na área. Para tratar os conteúdos dos

arquivos, é necessário que pelo menos uma parte da abordagem esteja ligada ao idioma. O fato de os dados se apresentarem em formato de texto, o que já exige um trabalho adicional, a língua em que os mesmos se encontram influencia a análise em vários momentos.

Além disso, pode-se definir Mineração de Textos como um conjunto de técnicas e processos que se presta a descobrir conhecimento inovador nos textos. Esta tecnologia vem sendo empregada em diversas áreas do conhecimento humano. De acordo com os estudos e pesquisas realizados na área, percebem-se os benefícios provenientes dessas técnicas. Assim, a motivação central para o desenvolvimento do presente trabalho é a aplicação de tais técnicas em grandes acervos ou coleções de documentos relevantes.

A pesquisa se destina a extração de conhecimento de textos científicos e/ou literários em língua portuguesa através de análises estatísticas, análise contextual baseada em técnicas de Processamento de Linguagem Natural (PLN). Desta forma, o resultado do desenvolvimento deste trabalho poderá ser aplicado em qualquer conjunto de documentos em português, seja científico ou literário, pelo fato de ser considerado um estudo genérico.

1.2 Objetivos da Pesquisa

Agrupar os textos em português, segundo a similaridade dos conteúdos, de forma que facilite a interpretação dos resultados e auxilie a distribuição de documentos para determinados perfis de usuários.

1.3 Linhas Gerais da Dissertação

Esta dissertação é composta de uma primeira parte que mostra a literatura levantada, a motivação para o desenvolvimento da pesquisa e seus objetivos.

No segundo capítulo é apresentada a definição geral da tecnologia utilizada para o desenvolvimento da metodologia.

O terceiro capítulo é composto pelas ferramentas utilizadas nas etapas de pré-processamento e processamento.

A quarta parte do trabalho é composta pelos ambientes propostos para a aplicação da metodologia desenvolvida.

No quinto capítulo está o desenvolvimento da metodologia propriamente dito. As etapas do processo estão descritas detalhadamente.

No sexto são apresentados os estudos de caso, com os resultados obtidos a partir dos processamentos realizados.

O sétimo e último capítulo apresenta um resumos das análises realizadas, as conclusões alcançadas e os trabalhos futuros.

Em seguida estão disponíveis as referências e os apêndices, respectivamente.

2 Mineração de Textos

De maneira geral um processo de mineração de textos contém cinco grandes etapas: coleta, pré-processamento, indexação, processamento ou mineração e pós-processamento ou análise da informação. A pesquisa contemplará as cinco etapas do processo, pois como é um estudo genérico, tem como proposta a utilização da metodologia desenvolvida em qualquer base de textos relevante.

As Figuras 1 e 2 mostram de maneiras diferentes, as etapas ou camadas da mineração de textos.

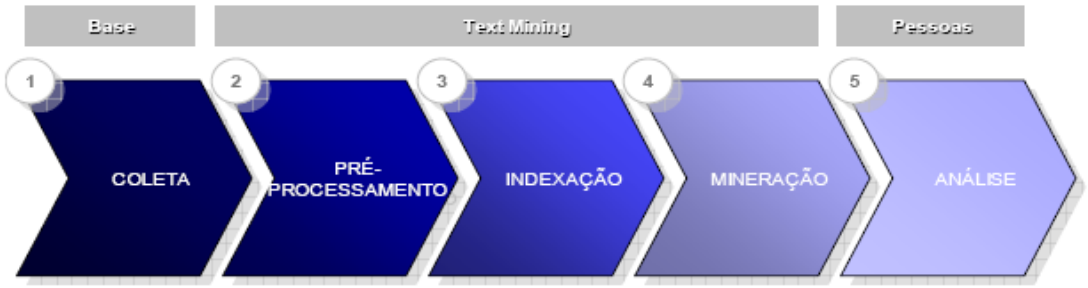

Figura 1: Diagrama de camadas da Mineração de Textos (Cortex Intelligence, 2008)

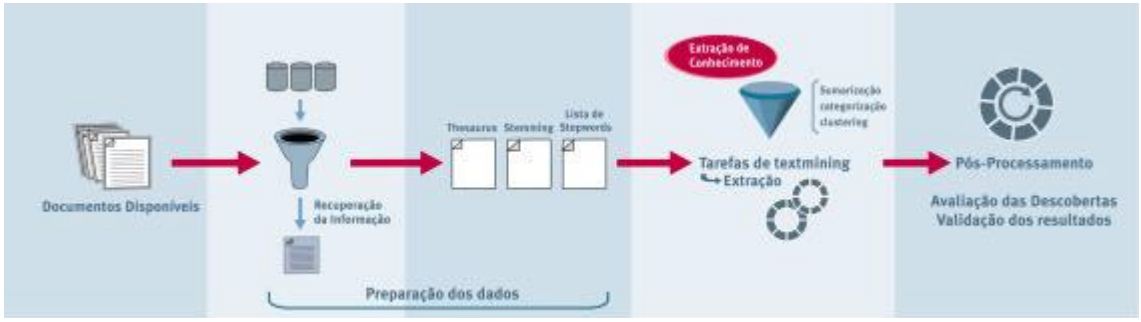

Figura 2: Etapas da Mineração de Textos (LOPES, 2004, p.7)

Na primeira etapa realiza-se a **coleta das informações** que vão compor a base de textos do trabalho. Para isto, é necessário determinar o universo de aplicação das técnicas de mineração de texto. A base selecionada, também é conhecida na literatura como *Corpus* ou *Corpora.* Segundo Carrilho (2008), a coleta pode ocorrer de várias maneiras, porém todas demandam grande esforço para que se consiga material de qualidade para aquisição de conhecimento.

Na segunda etapa, denominada de **pré-processamento**, é necessário transformar os documentos em formato adequado para serem submetidos aos

algoritmos de extração de conhecimento. É a etapa responsável por obter uma representação estruturada dos documentos, o que requer cuidadosa atenção, tornando-se bastante custosa. É uma etapa que necessita da aplicação de diversos algoritmos que tem por finalidade melhorar a qualidade dos dados disponíveis e organizá-los, consumindo boa parte de todo o processo de Mineração de Texto. De acordo com Batista (2003), essas transformações consistem em identificar, compactar e tratar dados corrompidos, atributos irrelevantes e valores desconhecidos. Geralmente, o pré-processamento significa aplicar técnicas de *tokenização*, *stop words*, s*temming* e de classificação das palavras segundo a classe gramatical.

A *tokenização* é o primeiro estágio do pré-processamento de um texto. Consiste da identificação e separação dos caracteres que compõem cada símbolo ou palavra no texto, onde cada símbolo é separado por espaços, vírgulas, pontos, etc. Cada grupo de caracteres estabelecidos no primeiro nível é chamado de token e a seqüência de *tokens* forma um *tokenstream.* Tanto os grupos de caracteres, como os delimitadores se tornam *tokens* na nova seqüência, o único caractere dispensado é o espaço em branco. Um exemplo de *tokenização* pode ser visto na sentença a seguir:

[O] [flamengo] [,] [que] [está] [na] [segunda] [divisão] [,] [perdeu] [mais] [um] [jogo]

[contra] [o] [botafogo] [.]

Seja em português, em inglês ou em qualquer língua, o resultado do processo de *tokenização* é uma seqüência de palavras separadas por delimitadores. Com esse primeiro estágio já é possível iniciar um processo de indexação para recuperação de informações entre outros processos.

A redução de s*top words* é a identificação de palavras nos textos que podem ser desconsideradas e assim retiradas, pois não constituem conhecimento. Cria-se ou utiliza-se uma lista de palavras que podem ser descartadas. Essas palavras não apresentam conteúdo semântico que seja significante em todo o contexto. Para Lopes (2004), são palavras auxiliares ou conectivas (como e, para, a eles) que não fornecem nenhuma informação discriminativa na expressão do conteúdo dos textos. Na construção desta lista, que também é conhecida como *stoplist,* são inseridos pronomes, artigos, preposições entre outras palavras auxiliares. São consideradas, também, as palavras que possuem grande incidência no conjunto de documentos.

O *stemming* é uma técnica utilizada para a obtenção dos radicais (*stems*) das palavras. Para Bastos (2006), algumas variantes morfológicas das palavras têm o mesmo significado semântico, e podem ser considerados como equivalentes dentro da

proposta das aplicações de Recuperação de Informações. Assim, os termos de um documento podem ser representados pelo seu radical ou *stem*.

Os algoritmos de *stemming* estão relacionados à língua para qual são desenvolvidos. Isto porque o seu processo consiste na eliminação de sufixos que designam variações indicando plural, flexões verbais ou variantes transformando as palavras em única representação, portanto uma única semântica. Este processo permite uma significativa redução no número de palavras dos documentos, conseqüentemente no espaço de armazenamento e no tempo de processamento.

Todas estas etapas de pré-processamento, principalmente a de eliminação de *stop words* e o processo de *stemming* têm como objetivo reduzir a dimensionalidade do problema, focando no tratamento das palavras (ou *stems*) que concentram a carga semântica do texto.

No entanto, em seu trabalho, Aranha (2007) ressalta que a etapa de pré-processamento vai além das ações citadas acima, pois é necessário transformar os textos em uma representação estruturada adequada para que, a partir disso, os dados possam ser submetidos ao processo como um todo.

Para Martins (2003), durante a transformação dos textos em formato estruturado há a possibilidade da informação que está intrínseca ao conteúdo dos textos seja perdida. Portanto, torna-se um grande desafio obter uma boa representação minimizando a perda de informação.

Um bom pré-processamento é imprescindível para garantir o sucesso da extração de conhecimento. Esta etapa tem papel fundamental para o desempenho de todo o processo.

A etapa seguinte é a de **indexação**. Carrilho (2008) define indexação como **"**processo que organiza todos os termos adquiridos a partir de fontes de dados, facilitando o seu acesso e recuperação. Uma boa estrutura de índices garante rapidez e agilidade ao processo, tal como funciona o índice de um livro". Métodos como este de indexação aumentam a performance do processo.

Após esta etapa, inicia-se a aplicação de fato dos algoritmos de Mineração de Textos, que é a fase do **processamento**. A mineração é responsável pelas inferências e cálculos. O objetivo é a extração de conhecimento e descoberta de padrões úteis e desconhecidos presentes nos textos. Nesta etapa, os algoritmos que serão aplicados aos textos são definidos de acordo com o objetivo da pesquisa. Existem vários algoritmos que se comportam de maneira diferente para cada problema a ser solucionado. Nenhum dos existentes é ótimo para todos os tipos de aplicação. Sabe-

se que muitos estudos empíricos estão sendo realizados com o intuito de relacionar o algoritmo com a aplicação. Enquanto estes estudos não são concluídos, há uma solução que também deve ser analisada: a combinação de resultados de vários algoritmos.

Após o levantamento da literatura na área e de outros estudos realizados, o *clustering* foi selecionado como o algoritmo de mineração para o desenvolvimento da metodologia apresentada nesta pesquisa. Lopes (2004) afirma que o *clustering* de documentos tem sido estudado intensivamente por causa de sua aplicabilidade em áreas tais como *information retrieval*, *web mining*, e análise topológica. Ainda em seu trabalho, destaca o uso de *clustering* para textos em português, e mostra a aplicabilidade e eficiência dessa técnica quando usada em textos.

Bastos (2006) define que o processo de *clustering* consiste em agrupar um conjunto de objetos físicos ou abstratos em classes de objetos similares. Também é conhecido como aprendizado não supervisionado. Vem sendo empregado, mais recentemente, para percorrer coleções de documentos e organizar os resultados retornados após consultas realizadas em mecanismos de busca.

Griffiths, Robinson e Willett (1984) afirmam que o *clustering* também pode servir como um passo do pré-processamento para outros algoritmos de mineração de dados como classificação de documentos.

A utilização de métodos de *clustering* em vários contextos, pelas mais diferentes disciplinas, ressalta a sua grande utilidade na exploração de conhecimento sobre dados. Esta técnica pode ser aplicada em várias áreas, como *Marketing*, *Call Center*, Biologia, Medicina, *Web*, Bibliotecas entre outras.

Outro ponto importante, relatado por Cutting *et al.* (1992), é que este método não requer entradas com marcação semântica, quando aplicado à mineração de textos, e por isso tem sido aplicado em grandes conjuntos de documentos HTML com sucesso.

É interessante ressaltar que as tarefas a serem realizadas nesta etapa estão relacionadas em grande parte aos objetivos a serem alcançados pela análise dos textos. Esta técnica será abordada novamente no capítulo 5, quando será apresentado o processamento realizado para o desenvolvimento da metodologia.

Para finalizar todo o processo, a última etapa que é chamada de **análise da informação** ou **pós-processamento** é a fase de avaliação e interpretação dos resultados. Deve ser executada por pessoas que estão interessadas no conhecimento

extraído e que devem tomar algum tipo de decisão apoiada no processo de Mineração de Texto.

Para Aranha (2007), essa fase envolve todos os participantes. O analista de dados tenta descobrir se o classificador atingiu as expectativas, avaliando os resultados de acordo com algumas métricas tais como taxa de erro, tempo de CPU e complexidade do modelo. O especialista no domínio irá verificar a compatibilidade dos resultados com o conhecimento disponível do domínio. E, por fim, o usuário é responsável por dar julgamento final sobre a aplicabilidade dos resultados.

Nesta etapa, algumas métricas de avaliação de resultados, ferramentas de visualização, e conhecimento de especialistas ajudam a consolidar os resultados.

3 Ferramentas de Mineração de Textos

Este capítulo é dedicado às ferramentas utilizadas para o desenvolvimento da pesquisa. Hoje em dia, existem várias ferramentas aplicadas à mineração de textos. Muitas delas são *free*, bastando o usuário realizar *download* do site onde está disponível e instalá-lo em sua máquina sem pagar nada por isso. Algumas ferramentas classificadas como comerciais são disponibilizadas pelas empresas desenvolvedoras por um determinado período de tempo, geralmente 30 dias, para que o usuário possa testar suas funcionalidades, com limitações, e adquirí-las em seguida caso seja interessante. Outras ferramentas comerciais estão disponíveis para uso só após a compra, não possuindo uma versão limitada para experiência do usuário.

Muitas ferramentas são desenvolvidas para determinados tipos de bases de dados ou com aplicações específicas, geralmente originadas da academia, em pesquisas de mestrado ou doutorado. Outras são usadas de forma geral para obtenção de conhecimento em grandes massas de dados. As ferramentas utilizadas e suas funcionalidades são descritas abaixo.

3.1 Cortex

O Cortex foi desenvolvido pela empresa Cortex Intelligence, que é focada em Inteligência Competitiva. A empresa utiliza tecnologias sofisticadas para tratar as informações de seus clientes de forma inteligente. O Cortex não é uma ferramenta comercial, mas é utilizada no desenvolvimento das plataformas comercializadas pela empresa.

O processo de mineração de textos da Cortex Intelligence é composto pelas cinco etapas descritas no capítulo anterior. Estas etapas foram elaboradas de forma a obter os melhores resultados frente aos desafios encontrados no mercado de tratamento de textos.

A primeira etapa é a coleta de informações, onde os robôs da Cortex Intelligence navegam em qualquer ambiente para captar informações não-estruturadas, seja na Internet ou em bases de dados internas nas empresas. A etapa seguinte consiste no pré-processamento dos textos coletados. É nesta etapa, que o Cortex (ferramenta) atua. O processo de mineração da empresa é diferenciado, pois dedica atenção especial nesta etapa, colocando maior ênfase no conteúdo dos textos. Agentes

inteligentes processam o texto de modo a extrair e identificar entidades, adicionando metadados aos documentos e enriquecendo a base de informações. Estas entidades são ligadas entre si através de relacionamentos semânticos obedecendo a uma ontologia de conhecimento segundo padrões da Web Semântica (Web 3.0). Isso garante ao processo, confiabilidade e qualidade superiores às abordagens baseadas em palavras-chaves ou em métodos puramente estatísticos. A próxima etapa é a indexação, processo indispensável para o tratamento de grandes volumes de dados. Em seguida, a mineração aplica métodos estatísticos de alta dimensionalidade para cada funcionalidade específica, de acordo com a demanda do cliente. Por fim, há a participação do usuário, ficando a seu critério efetuar interpretações dos resultados obtidos, gerar relatórios ou acionar novas minerações.

Foi desenvolvido e implementado um modelo computacional automático de pré-processamento valorizando o conteúdo do texto, ou seja, o modelo baseado em palavras foi transformado em um modelo baseado em lexemas. Assim, o Cortex transformou a abordagem tradicional por conjunto de palavras *(bag-of-words),* bastante utilizada atualmente, para o modelo *bag-of-lexems*.

O modelo de pré-processamento proposto pela ferramenta utiliza conhecimentos das áreas de PLN e Lingüística Computacional para formatar soluções com mais ênfase no conteúdo, visto que o conteúdo de um texto, no entanto, é dependente da língua em que está escrito.

Desta forma, o Cortex inova com um processador de textos na etapa de pré-processamento. O modelo do processador utiliza técnicas de Inteligência Computacional com base em conceitos existentes, como redes neurais, sistemas dinâmicos, e estatística multidimensional. E o modelo de PLN utilizado na etapa de pré-processamento é fortemente baseado em léxico. O léxico utilizado no modelo contém atributos relativos a uma ontologia primitiva. Esse mesmo léxico é ainda atualizado de forma automática. Os algoritmos de aquisição utilizados na sua atualização, assim como os algoritmos de mineração de texto são avaliados segundo medidas de precisão e cobertura, mais conhecidas como *precision* e *recall*.

Outra característica é a autonomia do sistema, de caráter pseudo-supervisionado, que tenta aproveitar as dicas presentes no próprio texto para o aprendizado automático de novos lexemas e novas classificações ontológicas, com o objetivo de minimizar esforços manuais. Gerenciar um léxico é um dos maiores esforços de um modelo de pré-processamento.

A seguir são apresentadas algumas das técnicas de PLN que o Cortex utiliza

em seu processo.

- **Normalização**

É uma técnica para aumentar a Cobertura em virtude das diversas representações de um mesmo conceito. A idéia é esquivar das várias formas de representação de uma palavra associada a um mesmo conceito. O processo de normalização propõe que duas ou mais formas sejam agrupadas em apenas uma, indicando que elas têm o mesmo significado num processo de busca. O problema da normalização é ser uma aproximação de conceitos, ou seja, os lexemas não têm o mesmo significado e sim um alto grau de redundância de significado, que para uma estratégia do tipo Cobertura pode ser pertinente.

Por exemplo, da definição “que cerca ou envolve os seres vivos ou as coisas, por todos os lados” temos a palavra “**ambiente**”, com as representações “**ambiente**” e “**ambientes**”. A normalização propõe que essas duas palavras sejam agrupadas em apenas uma, indicando para a busca que têm o mesmo significado.

Na prática, aumentando a Cobertura com o agrupamento de várias palavras que possuem significados distintos, a Precisão do sistema é bastante prejudicada. Porém essa estratégia reduz o tamanho do léxico, normalmente, apresentando uma maior eficiência quando o objetivo é navegação.

- **Reconhecimento de Entidades Nomeadas**

O pré-processamento comum à maioria das atividades em mineração de textos tem por responsabilidade principal o reconhecimento das entidades mencionadas no texto. Entende-se por entidades, pessoas, lugares, instituições. Porém, para reconhecer essas entidades de forma eficiente faz-se necessário o reconhecimento de todos os objetos do texto. Este é um procedimento natural para humanos, mas mostrou-se uma difícil tarefa para um sistema especialista.

Esse é um dos pontos principais do PLN para inteligência competitiva, pois eles nomeiam os objetos do mundo real de trabalho. Grande parte da informação de uma nova notícia é proveniente de novos nomes, ou relacionamentos entre novas combinações de nomes. Os tipos de relacionamentos são mais finitos que os nomes. Aproximadamente 90% dos novos lexemas a serem aprendidos por um sistema automático são nomes próprios. Sendo assim, é interessante dar especial atenção a tarefa de reconhecimento de entidades.

O processo começa pela avaliação dos candidatos a entidade nomeada. De forma macro, essa avaliação consiste em uma sucessão de filtros. Os candidatos que

persistirem serão agrupados por proximidade e considerados nomes de entidades. Esses filtros algumas vezes utilizam o condicionamento a palavras próximas, comportando-se como um autômato finito.

Um bom marcador utilizado é a letra maiúscula. Ela costuma fornecer uma boa lista de candidatos iniciais. Além disso, um algoritmo específico para avaliar o início de uma frase deve ser utilizado, já que todas as palavras, inclusive os nomes próprios, são marcadas com letra maiúscula no inicio das frases. Uma boa solução para o início de frases é saber se a palavra é um verbo ou um substantivo antes de tornar o *token* candidato a ser um nome próprio. Deve-se considerar também o fato de que um *token* nome próprio é encontrado bastante freqüentemente acompanhado por outro *token* nome próprio.

Esses *tokens* em seqüência, normalmente representam um único lexema e devem ser agrupados. Deve-se levar em conta também, as preposições, mesmo que não sejam marcadas pela letra maiúscula.

Outro padrão recorrente é o uso de siglas no meio do nome como forma de abreviação. Os pontos utilizados nessas abreviações são um grande complicador para o reconhecimento. O computador deve ter regras que auxiliem nesse agrupamento como a raridade de uma frase terminando em sigla, e ainda mais precedido de um nome de pessoa.

Finalmente, esses objetos são passados por um último filtro de datas, religião, localização geográfica e rotulados como nomes de entidades. Esses nomes devem ser catalogados e aprendidos periodicamente para auxiliar as outras tarefas de PLN.

É importante ressaltar que o aprimoramento do sistema é contínuo, pois o algoritmo acumula o conhecimento obtido em processamentos anteriores, utilizando métodos de *Dynamic Learning*.

3.2 RefViz

O *RefViz* é um software com aplicativo para análise, organização e visualização de uma quantidade relativamente grande de referências. Foi desenvolvido para língua inglesa, e registrado pela *The Thomson Corporation.*

As referências organizadas podem ser de livros, revistas, papers, teses, que é o caso deste estudo, entre outros. Por este motivo, o formato de entrada dos textos é composto por vários metadados. São dados que podem existir em diversos documentos que estejam em variados suportes de informação.

Nele, os documentos são organizados por seus conteúdos temáticos e apresentados em visualizações interativas que facilitam a rápida identificação dos principais temas e áreas de interesse. Para isso, usa redução de dimensionalidade. O software oferece algumas ferramentas para a exploração detalhada das referências e de busca on-line de resultados. Permite clusterizar papers de acordo com as palavras-chave que compartilham e visualizar quais as subáreas de uma determinada especialidade que têm recebido mais atenção.

Com *RefViz*, pode-se obter uma visão geral de todas as referências e, em seguida, obter ganho de conhecimento aprofundado dos temas de interesse. Embora ganhando aprofundado conhecimento dos temas, pode-se descobrir o que mais está sendo publicado em determinadas áreas. Também pode-se analisar as tendências ao longo do tempo e perceber novos temas emergentes nos campos de conhecimento.

O processo que o *RefViz* utiliza para dividir um conjunto de referências pode ser comparado a uma leitura dos conteúdos, quando é possível encontrar palavras-chave e suas associações, dividindo o conjunto de trabalhos em grupos baseando-se no assunto. Além dessa divisão, o software também distribui (aproxima ou distancia) os grupos formados de acordo com os temas que os mesmos abordam. Após estes processos, o *RefViz* cria “rótulos” com três palavras-chave para representar cada grupo formado. Esses rótulos facilitam a percepção dos principais temas de uma determinada coleção de referências.

O *RefViz* utiliza algoritmos matemáticos para realizar a divisão do conjunto de referências em grupos baseados em conceitos. Começando com um vocabulário determinado a partir da lista títulos e resumos de cada conjunto de referências, o software utiliza um modelo de estatística para encontrar conceitos-chave. Define os principais temas e conceitos com base no contexto das referências em vez de usar regras pré-estabelecidas.

O resultado é uma análise comparativa mais rica do que uma simples categorização de documentos baseada em contagem de palavras. Fornece dois métodos de visualização que possuem a vantagem da interatividade, além de outras ferramentas que permitem associações entre conceitos e grupos.

- ***Galaxy Visualization***

Esta visualização é um ambiente interativo em que o resultado do processamento se apresenta como um mapa de proximidades, onde as relações entre as referências de um mesmo grupo e as relações entre os grupos podem ser analisadas. A visualização é feita de acordo com as relações conceituais, dando uma

visão geral e rápida dos grupos formados e de todo o conjunto. Nesta visualização, cada quadrado amarelo representa uma referência e cada arquivo representa um grupo, como se observa na Figura 3.

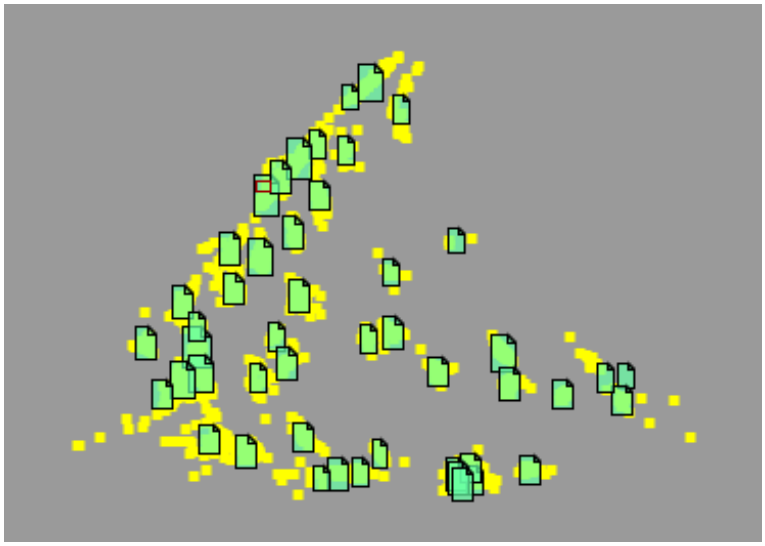

Figura 3: Exemplo da *Galaxy Visualiztion*

- ***Matrix Visualization***

Esta é uma visualização bidimensional da representação das associações entre conceitos e grupos. Pode ser configurado para exibir associações entre os grupos de referência e os seus principais tópicos ou para representar a co-ocorrência de grandes temas com outros grandes temas nas referências.

Por padrão, as linhas representam os mesmos grupos vistos na *Galaxy* e colunas os principais tópicos.

Os rótulos das linhas mostram os números dos grupos e as respectivas palavras que os descrevem e distinguem de outros grupos.

Ao clicar em uma célula, selecionam-se as referências de uma linha (grupo) e as colunas que contém suas palavras representativas. E ao clicar em uma coluna (palavra), selecionam-se todas as referências que contenham tal palavra. Na Figura 4, pode-se verificar esta representação.

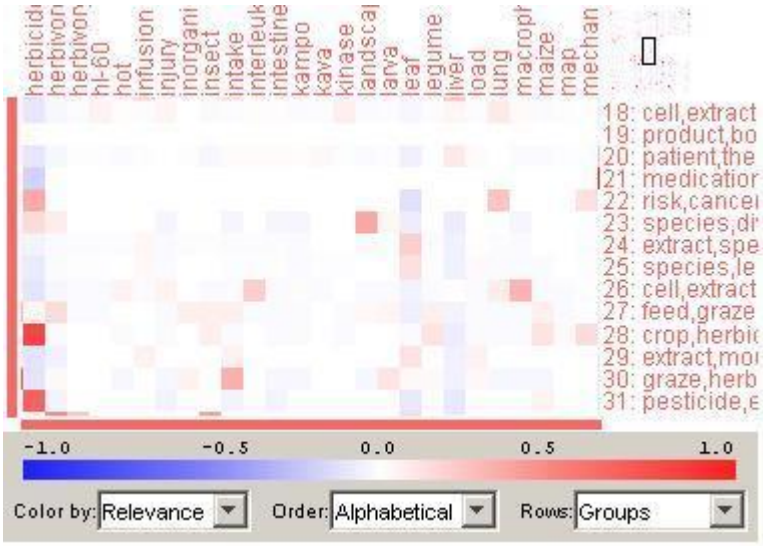

Figura 4: Exemplo da *Matrix Visualization*

3.3 MCL System

Dentre as ferramentas de *clustering* disponíveis, optou-se por uma que está em desenvolvimento, ainda na fase embrionária, mas que pudesse aproveitar o diferencial realizado na etapa do pré-processamento realizado pelo Cortex. Apesar deste método não prover as métricas e funcionalidades de visualização existentes em outros sistemas mencionados, ele foi escolhido pela capacidade de seu algoritmo subjacente e por permitir o aproveitamento da análise semântica e estatística de termos na fase de pré-processamento.

O algoritmo utilizado foi desenvolvido e implementado por Chada (2008) da Cortex Intelligence, baseando-se no *Markov Cluster Algorithm* ou algoritmo MCL. O ambiente de desenvolvimento foi o software livre Eclipse, uma plataforma projetada para fornecer infra-estrutura na construção de programas de alto-desempenho. O algoritmo foi escrito na linguagem Java.

Segundo Van Dongen (2000), o MCL foi concebido dentro do jovem campo de *graph clustering*, que tem fortes similaridades ao campo de partição de grafos. O cerne do MCL está em simular fluxo em um grafo. Para isso, é necessário transformá-lo em um grafo Markov, isto é, um grafo em que, para todo nó, a soma dos pesos das arestas que saem é um. Uma vez realizada esta transformação, a característica de fluxo pode ser exacerbada computando-se a potência da matriz estocástica associada,

que é o funcionamento normal de um processo Markov, chamado de expansão. O que diferencia o método MCL é a inserção de uma nova operação no processo Markov, denominada inflação. O processo gera uma seqüência de matrizes estocásticas através da alternância destas duas operações: expansão e inflação sobre uma matriz inicial.

Estes dois passos são, basicamente, o necessário para 'clusterizar' grafos. Outras tarefas, como a interpretação dos agrupamentos gerados e o estudo das propriedades matemáticas, cabem à implementação do software associado que utiliza o algoritmo. Uma característica relevante do MCL é não conter muitas regras de processamento para a montagem, separação ou junção de grupos, que faz com que a sua passagem da teoria à prática (programa que o aplica) seja fácil e intuitiva. O processo de MCL consiste simplesmente na alternância das operações de expansão e de inflação até se chegar a um estado final.

É relevante, também, fazer a distinção entre o algoritmo e o processo algébrico utilizado pelo algoritmo. Expansão, que é apenas uma multiplicação de matrizes, e pertence à linguagem da álgebra linear. Essa operação tem sido estudada especificamente na configuração das matrizes não negativas. A outra operação, a inflação, é altamente não linear. Isto dificulta o desenvolvimento de ferramentas matemáticas para descrever a sua interação com a expansão.

A simplicidade das suas operações torna o algoritmo MCL intuitivo e elegante. A expansão é usada para exacerbar a característica de fluxo entre nós: arestas com fluxo forte se tornam mais fortes e arestas com fluxo fraco se tornam mais fracas. A Inflação é aplicada então para eliminar arestas com fluxo fraco, desconectando seus nós e, eventualmente, formando agrupamentos de nós.

Outro aspecto importante do método é que o número de *clusters* não precisa ser especificado de forma antecipada pelo usuário, e o algoritmo pode ser ajustado facilmente a diferentes contextos. Estes aspectos tornam o MCL fortemente flexível e diferenciado de outros métodos de *clustering.* Através da variação dos parâmetros iniciais do algoritmo MCL, *clusters* de diferentes granularidades podem ser encontrados. Se os parâmetros forem configurados de forma errada, podemos encontrar uma quantidade de clusters igual à de nós (i.e. cada nó vira um *cluster* unitário), ou então, no outro lado do espectro, podemos encontrar um único *cluster* que engloba todos os nós.

A questão do número de *clusters* que o método encontra não é tratada de forma arbitrária, mas sim por uma lógica interna. A estrutura do grafo a cada iteração direciona o processo de corte de arestas, e os parâmetros de fluxo, escolhidos *a priori*, controlam a granularidade dos *clusters* gerados.

A taxa de convergência do processo MCL, isto é, quão rapidamente o algoritmo chega a um estado de equilíbrio final, também é regulada pela escolha dos parâmetros iniciais.

O limite do processo MCL é muito esparso, e suas iterações são esparsas em sentido ponderado, ou seja, o algoritmo é fortemente escalável. Dadas determinadas condições de dispersão, a poda de arestas pode ser incorporada ao resultado do algoritmo com complexidade $O(Nk^2)$, onde N é o número de nós, e k é o número médio ou máximo de nós vizinhos que poderão existir.

O algoritmo MCL também apresenta algumas limitações. Instâncias do problema nos quais o diâmetro dos clusters não são muito grandes permitem um regime de poda que mantém a qualidade dos clusters recuperados. Por outro lado, se o diâmetro cresce muito, esta abordagem torna-se inviável. Isto limita a aplicabilidade do algoritmo para grafos derivados no modelo vetorial, (grafos de vizinhança).

Foram implementados em *software* os seguintes parâmetros descritos pelo algoritmo:

- ***maxResidual*** – parâmetro que controla o número de operações (expansão e inflação). O algoritmo pára as iterações, quando o vetor de norma máxima for menor que o valor atribuído ao parâmetro. Quanto maior este valor, menor o número de operações realizadas.

- ***pGamma*** – este parâmetro é a potência da matriz. Quanto maior for o seu valor, maior será a diferença entre um ciclo de expansão e um ciclo de inflação.

- ***loopGain*** – corresponde à adição extra de loops ao processo. Antes do início dos ciclos, deve-se adicionar este valor para cada elemento da diagonal principal. Seu valor default é zero e recomenda-se que não seja alterado.

- ***maxZero*** – o algoritmo considera que a matriz alcançou o zero, a partir deste parâmetro. Quanto maior este valor, mais rápido a matriz chegará ao zero. Este parâmetro deve ser relacionado com o maxResidual.

4 Ambientes de Aplicação da Metodologia

A metodologia que está sendo desenvolvida nesta pesquisa, após sua validação poderá ser aplicada em diversos conjuntos de textos, conforme falado anteriormente. Porém, para o seu desenvolvimento foi necessário encontrar uma base de dados em formato texto que estivesse disponível.

Assim, a metodologia de classificação proposta, foi desenvolvida e validada através de testes realizados com o acervo de Teses e Dissertações Eletrônicas da PUC-Rio, disponíveis na Biblioteca Digital do Sistema Maxwell, que será descrito a seguir. Através da mineração dos textos pertencentes ao acervo selecionado, será possível avaliar a eficiência e eficácia de tal metodologia.

Após a validação, havia a proposta de aplicação da metodologia em um conjunto pequeno de documentos, em formato texto, do acervo da Fundação Biblioteca Nacional (FBN), instituição que será detalhada posteriormente. Como não houve tempo hábil para aplicação da metodologia neste acervo, esta constará no item de trabalhos futuros.

Assim, a presente pesquisa apresenta oportunidades futuras de grande importância para o desenvolvimento e crescimento da comunidade acadêmica que se dedica a estudos relacionados ao tema.

4.1 Sistema Maxwell

"O Sistema Maxwell é um Centro Digital de Referência. Ele é a integração do ambiente de ensino assistido por tecnologia de informação baseada na WEB com o ambiente de biblioteca/arquivo/museu digital, recriando-se a associação de uma instituição de ensino/pesquisa com uma biblioteca" (SISTEMA MAXWELL, 2000), desenvolvido pelo Laboratório de Automação de Museus, Bibliotecas Digitais e Arquivos – LAMBDA.

O sistema conta com mais de 10 mil usuários cadastrados provenientes de diferentes categorias. Estes são os alunos, professores e funcionários da PUC-Rio. Desta forma, o sistema possui funções e ambientes diversificados, com níveis de acesso de acordo com o tipo de usuário. Além desses, há um número muito grande de usuários não identificados da Internet que acessam teses, periódicos, entre outros documentos.

A sua biblioteca digital possui mais de 9 mil títulos e mais 35 mil objetos digitais. Os conteúdos são catalogados segundo a sua natureza e tipo. Fazem parte da natureza administrativa, conteúdos como normas, legislações, jurisprudências, atas, entre outros. Conteúdos como textos, artigos, apresentações, monografias e notas de aula são exemplos de conteúdos da natureza autoria. Os conteúdos da natureza docente são os exercícios, testes, gabaritos, trabalhos, provas, entre outros. Há também, os da natureza técnica, que são os manuais, planilhas, simuladores e pré-projetos.

O Maxwell conta ainda, com um ambiente de publicação on-line, contendo tanto artigos avulsos como publicações periódicas, desenvolvidas pelo LAMBDA em conjunto com outros departamentos da própria instituição. Neste ambiente, igualmente, são publicados livros.

A biblioteca digital do Sistema Maxwell é, também, a plataforma de disponibilização das teses e dissertações on-line da PUC-Rio e da UNICAP (Universidade Católica de Pernambuco), com mais de 4 mil teses on-line e na íntegra.

O LAMBDA desenvolve sistemas de informação de maneira abrangente, mas tem como foco principal a aplicação de tecnologia de informação na gestão de coleções digitais, como arquivos, museus e bibliotecas e o seu acesso através de redes de computadores, em particular a INTERNET e/ou INTRANET/EXTRANET.

O LAMBDA está junto a UNESCO na criação da rede internacional de teses e dissertações on-line. O Sistema Maxwell é base integrante da BDTD (Biblioteca Digital de Teses e Dissertações) e da NDLTD (Networked Digital Library of Theses and Dissertations). Abaixo, segue uma relação com mais algumas das bases que indexam o Sistema Maxwell.

- Biblioteca Universia (UNIVERSIA) – internacional, mas na Espanha.
- CyberTesis Net (CyberTesis) – internacional, mas no Chile.
- SCIRUS-Elsevier (SCIRUS) – internacional, mas na Holanda.
- OAISter da Universidade de Michigan (OAISter) – internacional, mas nos Estados Unidos.
- Universidad Nacional de La Plata Servicio de Difusión de la Creación Intelectual (UNLP SeDiCI) – internacional, mas na Argentina.

Hoje, o Sistema Maxwell conta com teses e dissertações de 25 Programas de Pós-Graduação da PUC-Rio e 3 da UNICAP. São mais de 4 mil teses da PUC-Rio e

56 da UNICAP. As teses e dissertações da PUC-Rio foram disponibilizadas para o desenvolvimento desta pesquisa.

4.2 Fundação Biblioteca Nacional

A Biblioteca Nacional do Brasil, também conhecida como Biblioteca Nacional do Rio de Janeiro foi considerada pela UNESCO a oitava biblioteca nacional do mundo, e a maior biblioteca da América Latina. Hoje, chamada de Fundação Biblioteca Nacional, seu acervo está calculado em cerca de nove milhões de itens em diversos suportes, como livros, manuscritos, mapas, estampas, moedas, medalhas entre outros. A FBN é depositária do patrimônio bibliográfico e documental do Brasil e tem a responsabilidade de preservar, divulgar e atualizar a sua coleção, que cresce constantemente a partir de doações, de aquisições e através do Depósito Legal. A lei do Depósito Legal é de 14 de dezembro de 2004. Desde então, a FBN recebe um exemplar de cada publicação realizada no Brasil, se tornando a guardiã da produção intelectual e da memória brasileira, cumprindo a sua finalidade, que é disponibilizar informação cultural em diversas áreas de conhecimento.

Ao longo do tempo, a FBN aperfeiçoou suas atividades e se tornou uma instituição mais diversificada atendendo demandas da comunidade científica. A FBN passou por reformas e buscou acompanhar a evolução tecnológica mundial, adquirindo equipamentos de segurança para a preservação do patrimônio que está sob sua custódia. Além disso, desenvolveu metodologias modernas de catalogação e classificação para seu acervo e adotou novas tecnologias, para ampliar a disseminação da informação e garantir o direito de acesso do cidadão, contribuindo para a sua qualificação.

Através de pesquisas e de consultas à Home Page da instituição, percebeu-se o perfil inovador da FBN, que desenvolve variados projetos e parcerias com outras instituições de renome. A partir desta análise, considerou-se de grande relevância a aplicação da tecnologia de Mineração de Texto em seu acervo. Para tal, verificou-se que diversos tipos de acervos já se encontram digitalizados, porém não se encontram ainda em formato texto. Desta forma, acredita-se que a FBN disponibilizará dados e textos, como os resumos e até mesmo o texto na íntegra, de alguns de seus milhões de documentos em meio eletrônico, em formato texto.

Propõe-se futuramente a aplicação da metodologia desenvolvida nesta pesquisa no acervo da FBN, como forma de descoberta de conhecimento, para auxiliar tanto os usuários comuns e pesquisadores, quanto os profissionais.

5 Desenvolvimento da Metodologia

Primeiramente foi realizada a delimitação do universo a ser abordado na pesquisa. Após o levantamento de bases de dados não-estruturados segundo critérios pré-estabelecidos, o acervo de textos a ser utilizado nos testes foi selecionado. Os seguintes critérios foram levados em consideração:

- Ser do idioma português;
- Textos armazenados em meio eletrônico, em formato texto;
- Fácil acesso ao Banco de Dados.

Após a definição acima, deu-se início ao desenvolvimento da metodologia que está descrita nos subtópicos deste capítulo.

A partir da preparação dos dados e após o processamento, as tarefas de análise estatística e análise contextual baseada no processamento da linguagem natural serão executadas. Para o desenvolvimento da pesquisa, estas duas abordagens serão utilizadas em conjunto para a análise dos dados.

Na análise estatística, os termos considerados relevantes serão basicamente os que possuírem maior ocorrência nos textos. Desta forma, serão abordadas as etapas do aprendizado estatístico dos dados e os modelos de representação de documentos utilizados pelos métodos estatísticos, incluindo os seguintes passos:

- Codificação dos dados;
- Estimativa dos dados;
- Modelos de representação dos documentos.

Ressalta-se que nesta análise não é necessário levar em consideração o idioma dos textos.

Na análise semântica são utilizados fundamentos e técnicas baseadas no processamento de linguagem natural. As técnicas avaliam a seqüência dos termos no contexto da frase para a identificação correta da função de cada termo. Assim, ao contrário da análise anterior, deve-se considerar o idioma dos textos, que neste caso é o português.

Optou-se pela realização, também, da análise semântica, pois sua execução proporciona significativa melhoria na qualidade da mineração de textos.

Para o entendimento da linguagem natural, pelo menos as seguintes divisões do conhecimento devem ser consideradas:

- Conhecimento Morfológico;
- Conhecimento Sintático;
- Conhecimento Semântico;
- Conhecimento Pragmático;
- Conhecimento do Discurso;
- Conhecimento do Mundo.

Para maiores detalhes dos tópicos, recomenda-se a consulta do artigo de (EBECKEN, LOPES E COSTA, 2005).

5.1 Coleta dos Dados

A primeira etapa realizada foi a coleta de dados. Vale ressaltar que, além da disponibilização da base de dados, o sistema gerenciador do banco de dados onde está a base de teses, o IBM DB2, possui um módulo Text Miner, que permite a manipulação de textos. Assim, foi possível implementar funções viabilizando a extração dos dados em formato texto.

Para a realização da coleta de dados, foram analisadas quais as informações mínimas relevantes para o desenvolvimento da pesquisa e que teriam viabilidade para a extração do banco de dados. O número de teses para análise foi delimitado em função do tempo. As teses selecionadas foram defendidas entre 01/08/2002 e 30/07/2006, fechando um período de 4 anos de publicação. Foi solicitado para a analista de sistemas responsável pelo Sistema Maxwell, as seguintes informações das 2140 teses que serão utilizadas:

- ID (nº de identificação da tese no sistema Maxwell);
- Título;
- Subtítulo (se houvesse);
- Resumo em português;
- Palavras-chave;
- Programa de Pós-Graduação;

- Área de concentração;
- Data de defesa.

A extração dos dados foi realizada através de um *select* no banco de dados DB2 com os campos solicitados. Os dados foram exportados para arquivos texto (**.**txt) correspondentes a cada programa de pós-graduação. Os resumos já se encontravam em arquivos texto no banco de dados e apenas foram extraídos.

5.2 Pré-processamento

Nesta etapa foi realizado o pré-processamento dos dados, ou seja, os textos foram preparados para que as tarefas posteriores possam ser realizadas. Apesar de similar ao processo de Mineração de Dados, que trabalha com dados estruturados, o processo de Mineração de Textos difere, principalmente, por trabalhar com dados não estruturados, em formato de texto. Assim, é uma etapa bastante demorada e custosa, visto que os dados textuais devem receber tratamentos diferenciados para posteriormente, serem submetidos aos algoritmos de mineração. Por demandar bastante tempo, dedicação e minuciosa atenção ao conteúdo dos textos, propôs-se a utilização da ferramenta Cortex da empresa Cortex Intelligence. Esta ferramenta inova na fase de pré-processamento, propondo um modelo automático de enriquecimento dos dados para uma análise mais eficiente, conforme visto anteriormente.

Inicialmente a fase do pré-processamento foi realizada em uma amostra da base de dados e foi dividida em 2 etapas:

- Preparação do processo de importação, organização e modelagem da base de dados.
- Pré-processamento via algoritmos de linguagem natural.

A importação utilizando DTS (Data Transformation Service) foi a primeira tentativa, mas como não foi bem sucedida, o processo foi realizado com código Java.

Como a segunda etapa do pré-processamento foi realizada com a utilização de algoritmos de linguagem natural, na primeira etapa foram identificados alguns problemas que trariam conflitos para a etapa subseqüente, como quebra de linha no meio das frases dos resumos e títulos em caixa alta. Assim, após a importação da base para um novo banco esses problemas foram solucionados.

Como a base de dados utilizada é composta basicamente pelos resumos, títulos, palavras-chave e ID, o problema da quebra de linha pôde ser resolvido

rapidamente. Tomou-se como premissa que os resumos são compostos apenas por único parágrafo, conforme as Normas para Apresentação de Trabalhos Acadêmicos da ABNT / NBR-14724. Assim, utilizou-se um algoritmo em Java que concatenava as linhas do texto, retirando as quebras e os espaços, quando houvesse mais de um. Caso a pesquisa fosse desenvolvida com a utilização do texto na íntegra, este problema teria maior complexidade e deveria ser solucionado de outra maneira.

Depois de solucionado o problema das quebras de linha, partiu-se para a resolução do problema da caixa alta no título. Neste caso, foi desenvolvido um algoritmo normalizador que faz uso da plataforma Cortex, baseada em PLN.

Primeiramente, o normalizador utiliza o Cortex no texto corrido, ou seja, nos resumos, para identificar as entidades do texto. O Cortex entende como entidades, organizações, lugares, pessoas, datas. Em seguida, passa todas as palavras do título para caixa baixa. Após este procedimento, o algoritmo faz uma verificação entre o resumo e o título, identificando palavras existentes no título que iniciam com letra maiúscula no resumo. Há o cuidado com as palavras que são iniciadas por letra maiúscula, apenas por estarem em início de frase. Após as identificações, o algoritmo transforma as primeiras letras de tais palavras para maiúscula.

Para exemplificar essas duas questões, apresenta-se o lexema “Santo Agostinho”, que está presente no título e no resumo de um dos documentos (ID: 3736) da base. Neste caso, a quebra de linha entre as duas palavras poderia descaracterizar o lexema, por ser uma palavra composta. Se isso ocorresse, não seriam identificadas e analisadas como sendo um só lexema.

Nos títulos, caso permanecessem todos em caixa alta ou em caixa baixa, o Cortex não identificaria “Santo Agostinho” como um lexema e sim, como duas palavras. Desta forma, não o extrairia como uma entidade.

Estas normalizações foram de grande relevância, visto que alguns termos compostos poderiam ser erroneamente considerados como palavras diferentes, o que acarretaria alterações nos resultados dos processamentos. É interessante ressaltar que estes procedimentos foram necessários devido ao uso desta ferramenta na segunda etapa do pré-processamento dos dados.

Como o processamento foi realizado em dois softwares, outras normalizações foram necessárias, pois as entradas de dados nos mesmos são diferentes. O *RefViz* foi desenvolvido para língua inglesa, assim não reconhece acentos e caracteres com cedilha. Com um algoritmo em Java, todos os acentos e caracteres não reconhecidos pelo inglês foram retirados e os documentos foram colocados no formato de entrada

do software. Pode-se visualizar abaixo na Figura 5, o formato de entrada de dados pela imagem do *SampleView* do *RefViz*.

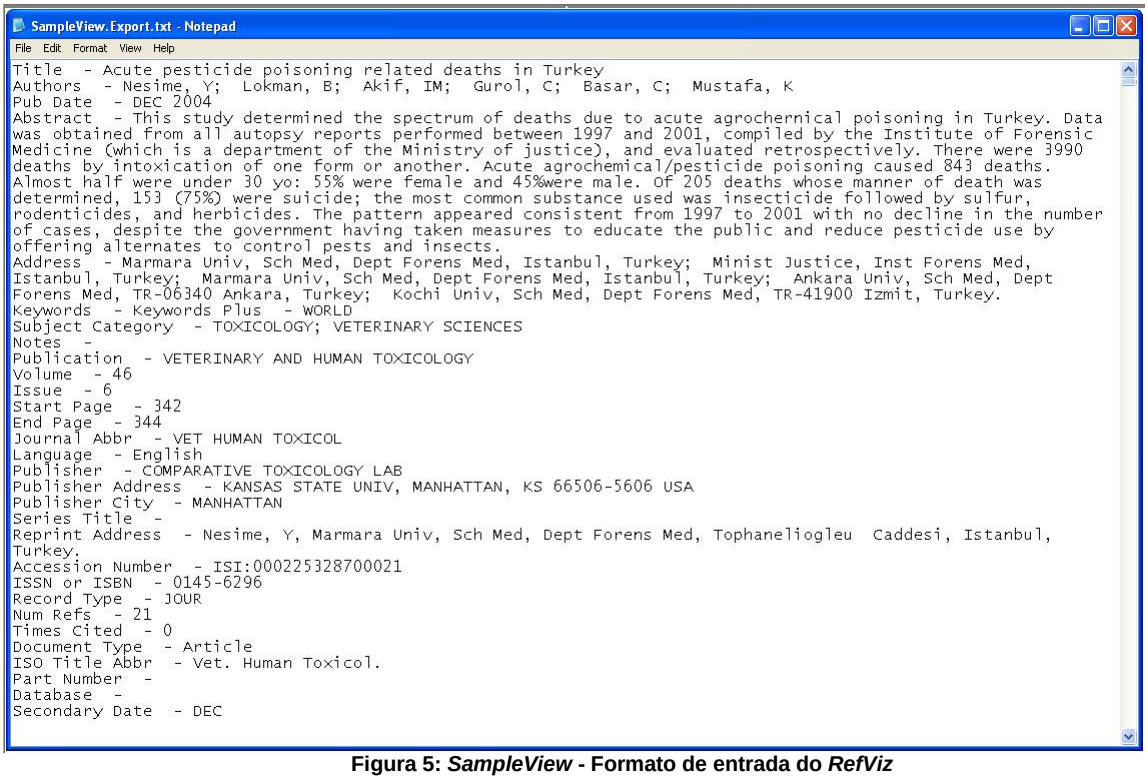

Figura 5: *SampleView* - Formato de entrada do *RefViz*

As três ferramentas utilizadas fazem a redução das s*top words* no momento do processamento. A *stoplist* utilizada foi criada e incrementada ao longo da pesquisa. Ela pode ser consultada no apêndice A.

5.3 Processamento

Na etapa do processamento dos dados, os objetivos do processo de Mineração de Texto devem ser definidos, já que existem diversas tarefas destinadas à extração de conhecimentos em textos, com diferentes propósitos. Algumas das tarefas de mineração que podem ser executadas são o *clustering*, a categorização, a sumarização, a indexação temática entre outras. Goldschimidt e Passos (2005) ressaltam que a dificuldade de escolha de um algoritmo de aprendizado apropriado é intensificada na medida em que surgem novos algoritmos com os mesmo propósitos, aumentando a diversidade de alternativas, mas que em geral, a escolha dos algoritmos se restringe às opções conhecidas pelo analista do processo, que inúmeras vezes deixa de considerar alternativas promissoras.

Conforme anunciado no capítulo 2, o *clustering* foi a técnica selecionada para o desenvolvimento desta metodologia. Assim, antes de serem descritos os processamentos e os resultados obtidos nesta etapa, o *clustering* será rapidamente apresentado.

O processo de *clustering* básico é descrito (LOPES, 2004) da seguinte forma: primeiro, uma descrição simplificada do documento é criada para cada texto que vai sendo adicionado aos *clusters*. A descrição é normalmente um vetor de características, ou uma lista de temas dominantes ou palavras-chave e uma medida de importância relativa de cada tema ou palavra-chave no documento. Em seguida, é determinada a proximidade de dois documentos baseada em seus vetores de características.

Em outras palavras, dado um conjunto de objetos descritos por múltiplos atributos, o algoritmo de *clustering* pretende primeiro, atribuir grupos (*clusters*) homogêneos aos objetos de maneira a maximizar a similaridade de objetos dentro de um mesmo cluster e, minimizar a similaridade de objetos entre clusters distintos. Em segundo, pretende atribuir uma descrição para cada *cluster* formado.

Para medir a distância de documentos numa hierarquia de assuntos, além das métricas de distâncias convencionais já conhecidas, existem as medidas de similaridades conceituais. As medidas de similaridades conceituais são normalmente baseadas numa função de distância entre os tópicos na hierarquia de assuntos e os pesos desses tópicos nos documentos.

Após o cálculo da distância entre os documentos, seja pelas medidas de similaridade padrão ou conceituais, eles podem ser agrupados de várias formas diferentes. A representação dos *clusters* pode ser em forma de grupos, matrizes, árvore hierárquica, entre outros.

É válido lembrar, que os processamentos foram realizados utilizando o mesmo *corpus*, o das 2140 teses e dissertações e serão apresentados no próximo capítulo como os estudos de caso da pesquisa.

5.4 Análise de Resultados

Por fim, existe a etapa do pós-processamento dos dados. Esta fase consiste na validação das descobertas, ou seja, na avaliação da qualidade dos resultados encontrados, para que os mesmos possam ser efetivamente empregados. Algumas métricas de avaliação de resultados, ferramentas de visualização e o próprio conhecimento de especialistas poderão auxiliar na consolidação dos resultados.

6 Estudos de Caso

A apresentação dos processamentos realizados está dividida segundo as ferramentas utilizadas. Primeiro serão apresentados os resultados obtidos a partir do MCL System e em seguida os resultados obtidos a partir do *RefViz*.

6.1 MCL System

Conforme tratado anteriormente, este clusterizador ainda está em desenvolvimento, por isso não possui ferramentas de visualização para auxiliar a análise dos resultados. Outros recursos como dicionário e tesauro também não foram desenvolvidos. Desta forma, os testes foram realizados baseados nas alterações de duas variáveis booleanas que devem ser inicializadas como *true* ou *false* para a realização dos processamentos, são elas:

- *removeStopWordsBeforeIndex = true ou false;*
- *useCortex = true ou false;*

Além dessas variáveis, os parâmetros também foram alterados. Os parâmetros são apresentados novamente a seguir, para que sejam associados aos resultados obtidos.

- ***maxResidual*** – parâmetro que controla o número de operações (expansão e inflação). O algoritmo pára as iterações, quando o vetor de norma máxima for menor que o valor atribuído ao parâmetro.
- ***pGamma*** – este parâmetro é a potência da matriz. Quanto maior for o seu valor, maior será a diferença entre um ciclo de expansão e um ciclo de inflação.
- ***loopGain*** – corresponde à adição extra de loops ao processo.
- ***maxZero*** – o algoritmo considera que a matriz alcançou o zero, a partir deste parâmetro.

Assim, os seis itens relacionados acima foram combinados e ajustados de diversas formas, encontrando variados resultados. Os ajustes foram realizados até a identificação de resultados satisfatórios, que são apresentados nos próximos itens. Os resultados são apresentados em duas etapas, com base na variável *useCortex*.

6.1.1 Com o uso do Cortex

Conforme visto anteriormente, a utilização do Cortex nos processamentos se deve à importância da análise semântica das palavras que compõem os textos. Por tratar da semântica das palavras no pré-processamento, a utilização do Cortex nesta pesquisa passou a ser imprescindível, visto que o processo de mineração de textos se baseia primeiramente na indexação das palavras para posterior mineração.

De acordo com esta premissa, iniciou-se a etapa dos processamentos. A utilização do Cortex gera um custo computacional relativamente alto, visto a sua complexidade. Cada ajuste realizado para testes demanda novo processamento, que pode ser estimado entre 40 e 60 minutos. Vários testes foram realizados, mas uma boa parte dos resultados não foi satisfatória. Os melhores resultados obtidos são apresentados a seguir. Estes foram encontrados através de 2 combinações entre os parâmetros *maxResidual, pGamma e maxZero.* Conforme recomendado, o *loopGain* não foi alterado. Ao encontrar essas 2 combinações, outras foram testadas, mas nenhuma obteve resultado superior aos que serão apresentados. Ainda no início das análises, verificou-se que este clusterizador permite que 1 documento seja inserido em mais de 1 cluster. Ou seja, vários clusters podem compartilhar um mesmo documento. Os clusters serão identificados pelo seu total de documentos, visto que o clusterizador ainda não faz essa indicação.

As Tabelas 1 e 2 dispõem das configurações dos primeiros resultados obtidos. Nas últimas linhas estão os totais de clusters encontrados, respectivamente.

Tabela 1: MCL – 1ª Configuração

Resultados MCL	
maxResidual	0.2
pGamma	0.9
loopGain	0.0
maxZero	0.01
useCortex	TRUE
removeStopWordsBeforeIndex	FALSE
Total de Clusters	**21**

Tabela 2: MCL – 2ª Configuração

Resultados MCL	
maxResidual	1.0
pGamma	0.8
loopGain	0.0
maxZero	0.01
useCortex	TRUE
removeStopWordsBeforeIndex	FALSE
Total de Clusters	**45**

A primeira configuração, com 21 clusters não obteve resultado satisfatório. Apresentou agrupamentos muito genéricos, compostos por muitos documentos, descaracterizando a técnica de *clustering*. Foram formados apenas 7 clusters com menos de 200 documentos e alguns clusters com mais de 1000 documentos. O cluster

7 é composto por 7 documentos de 4 diferentes áreas: Filosofia, Matemática, Engenharia Elétrica e Administração. As palavras-chave atribuídas pelos autores ou pelo Sistema Maxwell aos trabalhos inseridos no cluster 7, podem ser visualizadas na tabela 3, logo abaixo. A freqüência das palavras não foi disponibilizada na tabela, pois todas apresentaram freqüência 1. Percebe-se que há relação entre os termos de um mesmo documento e não entre documentos distintos, o que não caracteriza a formação do cluster.

Tabela 3: Cluster 7 – Palavras-chave

ID documento	Palavras-chave
3280	ACOES POLARES
3280	APLICACOES ISOPARAMETRICAS
3280	APLICACOES TRANSNORMAIS
3280	FOLHEACOES RIEMANNIANAS SINGULARES
3280	HOLONOMIA SINGULAR
3280	SUBVARIEDADES EQUIFOCAIS
3280	SUBVARIEDADES ISOPARAMETRICAS
4294	FIBRADOS DE SEIFERT
4294	ORBIFOLDS
5161	CONTROLE DE ADMISSAO
5161	INTERNET
5161	QUALIDADE DE SERVICO
5161	SERVICOS DIFERENCIADOS
5749	EFICIENCIA DE MERCADO
5749	ESTRUTURA DE CAPITAL
5749	MERCADO DE CAPITAIS
5933	ATRIBUTOS
5933	EXTENSAO
5933	MODOS INFINITOS
6489	ARTE
6489	CONSTRUCAO DE MUNDOS
6489	EXPERIENCIA
6489	INSTAURACAO
6489	LINGUAGEM
6489	NOMINALISMO
6489	REFERENCIA
6489	REPRESENTACAO
6489	SISTEMAS DE SIMBOLOS
9031	CANTOR
9031	DEDEKIND
9031	DEUS
9031	INFINITO
9031	ORDINAL

Os resultados obtidos pela configuração da Tabela 2 foram bons. Foram formados 45 clusters de diversos tamanhos. O maior cluster não passou de 300

documentos. Em análise geral, os clusters formados apresentaram documentos correlacionados. O cluster 18, por exemplo, foi formado por documentos das áreas de Administração, Economia, Engenharia de Produção e Engenharia Elétrica. Pela tabela abaixo se observa que os termos de diferentes documentos estão relacionados.

Tabela 4: Cluster18 – Palavras-chave

ID documento	Palavras-chave	ID documento	Palavras-chave
6990	ANALISE DE DADOS EM PAINEL	9444	GAS NATURAL
5476	APOSENTADORIA	8351	GERENCIAMENTO DE CREDITO
3701	BANCOS	4263	HETEROGENEIDADE PRODUTIVA
8913	BANCOS	9089	IMPORTACAO
6990	BANCOS COM ATIVIDADES DE VAREJO	4324	IMUNIZACAO
4672	BANCOS ESTRANGEIROS	6568	INVESTIMENTO FIXO
6990	BANCOS ESTRANGEIROS	4672	LIBERALIZACAO
6992	BANCOS PRIVADOS	4263	LUCRO
4672	BANCOS PRIVADOS NACIONAIS	9052	MATURIDADE
6992	BANCOS PUBLICOS	4263	MERCADO DE CREDITO
8596	BRASIL	5476	MUDANCAS TECNOLOGICAS
8913	CODIGOS DE ETICA	5476	NIVEL DE QUALIFICACAO
3701	COMPETICAO FORNECEDOR	8266	OPERACIONAL
6992	COMPETITIVIDADE NO MERCADO BANCARIO	9444	OTIMIZACAO ESTOCASTICA
3701	CREDITO COMERCIAL	4263	PEQUENAS EMPRESAS BRASILEIRAS
8596	CURVA DE JUROS	7385	POLITICA MONETARIA
9323	DERIVATIVOS	8596	POLITICA MONETARIA
5476	DESEMPREGO	6992	PRIVATIZACAO
5546	DESENVOLVIMENTO ECONOMICO	9444	PROGRAMACAO LINEAR
7385	DOMINANCIA BANCARIA	6568	REGRESSAO EM PAINEL
5546	EDUCACAO INFANTIL	8351	REGRESSAO LOGISTICA
9089	ELASTICIDADE	5546	RESTRICOES DE CREDITO
9052	EMISSAO DA DIVIDA SOBERANA	6568	RESTRICOES FINANCEIRAS
6568	EMPRESAS BRASILEIRAS	8266	RISCO
9444	ENGENHARIA ELETRICA	9444	RISCO DE CONTRATACAO
5546	ESCOLHA OCUPACIONAL	8351	RISCO DE CREDITO
4263	ESTIMACAO NAO-PARAMETRICA	9052	RISCO DE REFINANCIAMENTO
6990	ESTRATEGIAS DE ALOCACAO DE ATIVOS	4324	RISCO DE TAXA DE JUROS
8596	ESTRUTURA A TERMO	9052	SINALIZACAO
9323	ESTRUTURA A TERMO	8266	SISTEMA FINANCEIRO BRASILEIRO
6990	ESTRUTURA DE FINANCIAMENTO	8596	TAXA DE JUROS
8913	ETICA BANCARIA	3701	TAXA DE JUROS INVARIANTE
8913	ETICA COMERCIAL	5546	TRABALHO INFANTIL
9089	EXPORTACOES	4324	VARIACOES NAO-PARALELAS NA CURVA DE JUROS
7385	FISCALIZACAO BANCARIA	9323	VOLATILIDADE

A configuração da Tabela 5 traz mais um resultado obtido pelo clusterizador, com 24 clusters. Aparentemente é um bom resultado, visto que 24 é o total de áreas do conhecimento das quais as teses pertencem. Como visto inicialmente a coleção de documentos é a uma base de teses digitais correspondentes de 24 Cursos de Pós-Graduação da PUC-Rio. Apesar da redução de *stop words* e da utilização do Cortex, não foi possível encontrar bons resultados. Isto é, o algoritmo não teve bom desempenho para os valores atribuídos aos parâmetros listados. Foram identificados clusters com mais de 2 mil documentos, quase o total da base. Em contrapartida,

apresentou 2 clusters bem pequenos, um com 7 e outro com 3 documentos. Por exemplo, os documentos do cluster 3 pertencem às áreas de Engenharia de Produção, Engenharia Civil e Informática. Pela análise realizada, o clusterizador, também errou ao agrupar esses 3 documentos, pois apesar de suas áreas serem próximas umas das outras, tratam de assuntos distintos. Com a verificação da freqüência das palavras que fazem parte das reais palavras-chave, atribuídas pelos autores, não é possível classificar este cluster. Na Tabela 7, pode-se observar como os documentos estão distantes uns dos outros, através das palavras-chave dos mesmos.

Tabela 5: MCL – 3ª Configuração

Resultados MCL	
maxResidual	0.2
pGamma	0.9
loopGain	0.0
maxZero	0.01
useCortex	TRUE
removeStopWordsBeforeIndex	TRUE
Total de Clusters	**24**

Tabela 6: MCL – 4ª Configuração

Resultados MCL	
maxResidual	1.0
pGamma	0.8
loopGain	0.0
maxZero	0.01
useCortex	TRUE
removeStopWordsBeforeIndex	TRUE
Total de Clusters	**43**

Tabela 7: Palavras-Chave

Ids	Palavras-Chave – Cluster 3
8726	DINAMICA DE SISTEMAS
8726	GESTAO DE PROJETOS SOCIAIS
9502	CARREGAMENTO AUTOMATICO
9502	ESTABILIDADE NAVAL
9502	FORCA DE VENTO
9502	LINHA DE PRAIA
9502	PROGRAMACAO LINEAR
9730	ENGENHARIA SEMIOTICA
9730	FERRAMENTA EPISTEMICA
9730	INTERACAO HUMANO-COMPUTADOR
9730	SISTEMAS COLABORATIVOS

A sexta tabela desta seção apresenta mais uma configuração utilizada para teste. Esta configuração resultou em 43 clusters, que em primeira análise aparentam melhores resultados. Apenas 3 clusters possuem mais de 200 documentos. Para análise, o cluster 20 foi selecionado.

Este cluster está voltado para área de Ciências Humanas, possuindo documentos de Administração, Design, Direito, Economia, Educação, Filosofia,

História, Letras e Psicologia. Na tabela 8 são apresentadas as palavras mais representantes do cluster com suas respectivas freqüências. Por estas palavras, percebe-se que as áreas citadas acima estão sendo bem representadas por tais palavras.

Tabela 8: Classificação do Cluster 20

Palavras Representantes	Freqüência
arte	9
literatura	4
genero	3
contemporanea	3
brasileira	3
texto	2

6.1.2 Sem o uso do Cortex

Foram vários os testes realizados sem o uso do Cortex, o que acarretou resultados baseados na freqüência dos termos. Os primeiros resultados a serem considerados nesta etapa foram encontrados através das seguintes configurações:

Tabela 9: MCL – 1ª Configuração

Resultados MCL	
maxResidual	0.2
pGamma	0.9
loopGain	0.0
maxZero	0.01
useCortex	FALSE
removeStopWordsBeforeIndex	FALSE
Total de Clusters	**46**

Tabela 10: MCL – 2ª Configuração

Resultados MCL	
maxResidual	1.0
pGamma	0.8
loopGain	0.0
maxZero	0.01
useCortex	FALSE
removeStopWordsBeforeIndex	FALSE
Total de Clusters	**82**

As últimas linhas das tabelas acima são os resultados encontrados nas respectivas configurações. É importante ressaltar novamente, que o parâmetro *loopGain* não foi alterado em nenhum processamento, de acordo com suas instruções.

Devido a não remoção das *stop words*, os clusters apresentaram alta freqüência de palavras irrelevantes para a análise dos resultados e para *o clustering*. Esta informação é fácil e rapidamente percebida ao dar início às análises. Nas configurações acima, percebe-se que não houve muitas alterações nos parâmetros.

Dois foram alterados, sendo que apenas o *maxResidual* teve uma significativa mudança. Isto foi o suficiente para o número de clusters ter quase dobrado em relação à primeira configuração. O aumento no valor deste parâmetro resultou em um número menor de iterações, gerando clusters menores e mais específicos.

Para a apresentação dos resultados, alguns clusters foram selecionados aleatoriamente. Como o clusterizador não identifica os clusters por números ou letras, os mesmos serão identificados para a descrição dos resultados, pelo total de documentos contidos no cluster. Quando houver clusters com a mesma quantidade de documentos, serão diferenciados por letras.

A configuração da Tabela 9 resultou em 46 clusters, com um número de documentos relativamente grande em sua maioria. Foram encontrados clusters com mais de mil documentos. Apenas 15 clusters foram formados com menos de 200 documentos. Após a realização da análise, percebeu-se que esta configuração não foi adequada para esta base de documentos, pois a maioria dos clusters não pode ser classificada, devido a grande quantidade de documentos de cada um. Como a base de textos utilizada é relativamente pequena, com 2140 documentos de diversas áreas de conhecimento, um cluster de mil documentos, com quase a metade do total da base, deve ser considerado muito abrangente. No caso desta pesquisa, este resultado não é satisfatório, já que se trata de documentos de uma biblioteca, um ambiente de pesquisa, uma classificação tão abrangente não é interessante.

Para uma análise mais profunda deste resultado, o cluster 201 foi selecionado. Com 201 documentos, este cluster pode ser considerado grande e abrangente como falado anteriormente sobre a grande maioria, pois de acordo com a Tabela 11 e com o gráfico da Figura 6, possui documentos de 14 dos 24 cursos de pós-graduação contidos na coleção completa. Apesar de ser abrangente, o seu foco está entre 4 áreas, representadas pelas freqüências mais altas: Psicologia, Filosofia, História e Letras.

Tabela 11: Áreas de Conhecimento - Cluster 201

Áreas dos Cursos de Pós Graduação - Cluster 201	Total de Documentos/Área
ADMINISTRAÇÃO DE EMPRESAS	1
COMUNICAÇÃO SOCIAL	3
DESIGN	6
DIREITO	4
ECONOMIA	2
EDUCAÇÃO	7
ENGENHARIA ELÉTRICA	1
FILOSOFIA	47
HISTÓRIA	31
INFORMÁTICA	3
LETRAS	30
MATEMÁTICA	1
PSICOLOGIA CLÍNICA	48
RELAÇÕES INTERNACIONAIS	7
SERVIÇO SOCIAL	2
TEOLOGIA	8
Total de Documentos - Cluster 201	**201**

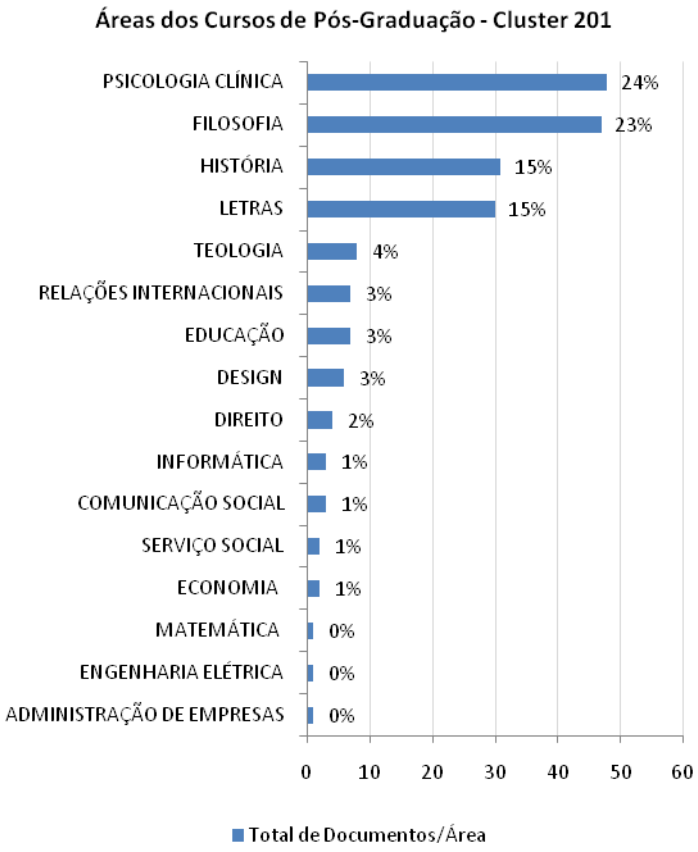

Figura 6: Gráfico de Áreas - Cluster 201

Decidiu-se analisar as palavras representantes do cluster, para concluir se essas 4 áreas, que compõem mais da metade do cluster, conseguem representá-lo através da freqüência de suas palavras. Os documentos dessas 4 áreas tratam de assuntos correlacionados entre si, já que foram agrupados. A Tabela 12 dispõe as palavras representantes do cluster 201, ressaltando que as *stop words* foram desconsideradas nesta tabela. Na Tabela 13 estão as palavras mais freqüentes dentre as palavras-chave atribuídas pelos autores e/ou pela biblioteca digital dos 201 documentos. Sabendo-se que existem 4 áreas bem representadas neste cluster e analisando somente as 2 tabelas abaixo, pode-se considerar que o clusterizador obteve um bom desempenho. Mas levando-se em consideração que as *stop words* não estão nesta tabela, que quase 20% dos documentos foram agrupados erradamente e que a maior parte dos clusters é composta por muitos documentos, considera-se que o desempenho do clusterizador foi ruim, revelando a necessidade de ajustes.

Tabela 12: Classificação do Cluster 201

Palavras Representantes	Freqüência
Memória	36
Moral	22
Relação	20
Vida	20
Filosofia	18
Luto	18

Tabela 13: Palavras-chave do Cluster 201

Palavras-chave	Freqüência
Historia	25
Memória	23
Arte	18
Psicanalise	15
Filosofia	15
Linguagem	11

Os resultados a seguir foram obtidos a partir da configuração da tabela 10, com o total de 82 clusters. Neste resultado, vários documentos estão presentes em 2 ou mais clusters. Por exemplo, dos 18 documentos do cluster 18, apenas 2 são exclusivos. Os outros 16 documentos pertencem, também, a pelo menos mais 1 cluster. Há um documento, de ID 7624, que está presente em mais 8 clusters. Acredita-se que este fato se deve a não remoção das *stop words* neste processamento. Muitos documentos podem ter permanecido juntos por possuírem alta freqüência de palavras em comum entre si, que podem ser as *stop words.* Esta hipótese deve ser validada através das análises desses clusters.

Enfim, os documentos do cluster 18 estão relacionados entre si, tratando de vários assuntos relacionados às engenharias de modo geral, com assuntos de informática e design. As 16 palavras mais freqüentes deste cluster são *stop words* e foram desprezadas na análise. As palavras relevantes e com freqüências significativas

foram consideradas para a classificação do cluster como suas representantes e podem ser analisada na Tabela 14.

O passo seguinte foi comparar tais palavras com as respectivas palavras-chave atribuídas pelos próprios autores ou pelo Sistema Maxwell. Dentre todas as palavras-chave, que são lexemas ou termos, desses documentos, as mais freqüentes estão na Tabela 15.

Tabela 14: Classificação do Cluster 18

Palavras Representantes	Freqüência
sistema	27
diesel	24
gás	23
motores	23
tecnologia	16
natural	13

Tabela 15: Palavras-chave do Cluster 18

Palavras-chave	Freqüência
gas	6
sistema	4
diesel	4
natural	3
motor	3
logística	3

Percebe-se que as palavras mais freqüentes dentre as palavras-chave desses 18 documentos estão bem próximas às palavras representantes do cluster. Considera-se, no geral, que o clusterizador obteve um bom resultado.

De acordo com os resultados obtidos até o momento, nota-se que a redução de *stop words* é indispensável, quando se trata de análise estatística. Como o Cortex não está sendo utilizando nesta etapa, faz-se necessário a utilização de uma *stoplist.* Isso significa que na tentativa de melhores resultados, o *removeStopWordsBeforeIndex* foi ativado. As configurações são as mesmas utilizadas anteriormente, apenas com a utilização da *stoplist* e estão disponíveis nas tabelas abaixo.

Tabela 16: MCL – 3ª Configuração

Resultados MCL	
maxResidual	0.2
pGamma	0.9
loopGain	0.0
maxZero	0.01
useCortex	FALSE
removeStopWordsBeforeIndex	TRUE
Total de Clusters	**83**

Tabela 17: MCL – 4ª Configuração

Resultados MCL	
maxResidual	1.0
pGamma	0.8
loopGain	0.0
maxZero	0.01
useCortex	FALSE
removeStopWordsBeforeIndex	TRUE
Total de Clusters	**153**

O processamento com a configuração da Tabela 16 apresentou 83 clusters em seu resultado. Na análise dos clusters formados, percebeu-se, também, a existência de alguns bastante numerosos, conforme encontrado anteriormente. Alguns clusters foram formados por mais de mil documentos, mas a maioria deles possui um número razoável de documentos. O menor possui apenas 2 documentos, que após verificação, notou-se que o clusterizador errou ao identificá-los como sendo pertencentes ao um mesmo cluster. Um documento é da área da Matemática, enquanto o outro é da Filosofia.

Por ser um cluster tão pequeno, julgou-se importante verificá-lo, pois poderia ser bastante específico e com temas diferenciados do restante dos documentos da coleção. Concluiu-se que os dois documentos poderiam ter sido alocados separadamente em dois clusters já identificados.

O cluster 42 é representado por documentos de 10 diferentes áreas correlacionadas. As palavras mais freqüentes que representam este cluster podem ser visualizadas na Tabela 18, logo abaixo. Já a Tabela 19, disponibiliza as palavras mais freqüentes dentre as palavras-chave atribuídas pelos autores o pelo sistema. Como houve a redução de *stop words* neste processamento, o resultado ficou mais limpo, sem necessidade de desprezar essas palavras na análise e mais consistente, mas ainda assim, é possível obter melhores resultados.

Tabela 18: Classificação do Cluster 42

Palavras Representantes	Freqüência
schopenhauer	9
perspectiva	9
mundo	9
leitura	9
interpretação	9
humano	9

Tabela 19: Palavras-chave do Cluster 42

Palavras-chave	Freqüência
historia	10
nietzsche	5
cultura	4
politica	4
construtivismo	3
design	3

A configuração da Tabela 17 apresentou melhores resultados. Utilizando s*toplist* e os parâmetros descritos na tabela, chegou-se a 153 clusters. Os clusters estão menores e mais específicos. Durante as análises, pode-se perceber alguns erros do clusterizador, mas no geral obteve bons resultados. O cluster 17, por exemplo, possui 4 áreas representadas. São 17 documentos, na sua maioria originados da Engenharia Elétrica. Apenas 2 são de Informática, 1 de Design e 1 de Administração. Para verificação desse resultado, o gráfico da Figura 7 disponibiliza as palavras-chave atribuídas pelos autores dos documentos desse cluster.

É importante ressaltar, que apesar de quase todas as palavras aparecerem apenas em 1 documento como mostra o gráfico, muitas são parecidas e tratam de assuntos relacionados. Algumas estão com a grafia diferente, mas representam o mesmo assunto, como “Qualidade de Serviço” e “Qualidade do Serviço”.

Realizando a análise em relação às freqüências das palavras, conforme vem sendo apresentado, este cluster pode ser classificado de acordo com as palavras das Tabelas 12 e 13, que seguem abaixo.

Tabela 20: Classificação do Cluster 17

Palavras Representantes	Freqüência
redes	35
internet	26
voz	22
ip	21
acesso	19
protocolos	19

Tabela 21: Palavras-chave do Cluster 17

Palavras-chave	Freqüência
redes	11
ip	7
servico	4
qualidade	4
moveis	4
mobilidade	4

É importante destacar novamente, que a classificação está relacionada à *clustering* realizada e a tabela com as palavras-chave é construída a partir da freqüência das palavras atribuídas pelos autores. Percebe-se que as 2 tabelas possuem algumas palavras iguais e que as outras palavras estão relacionadas.

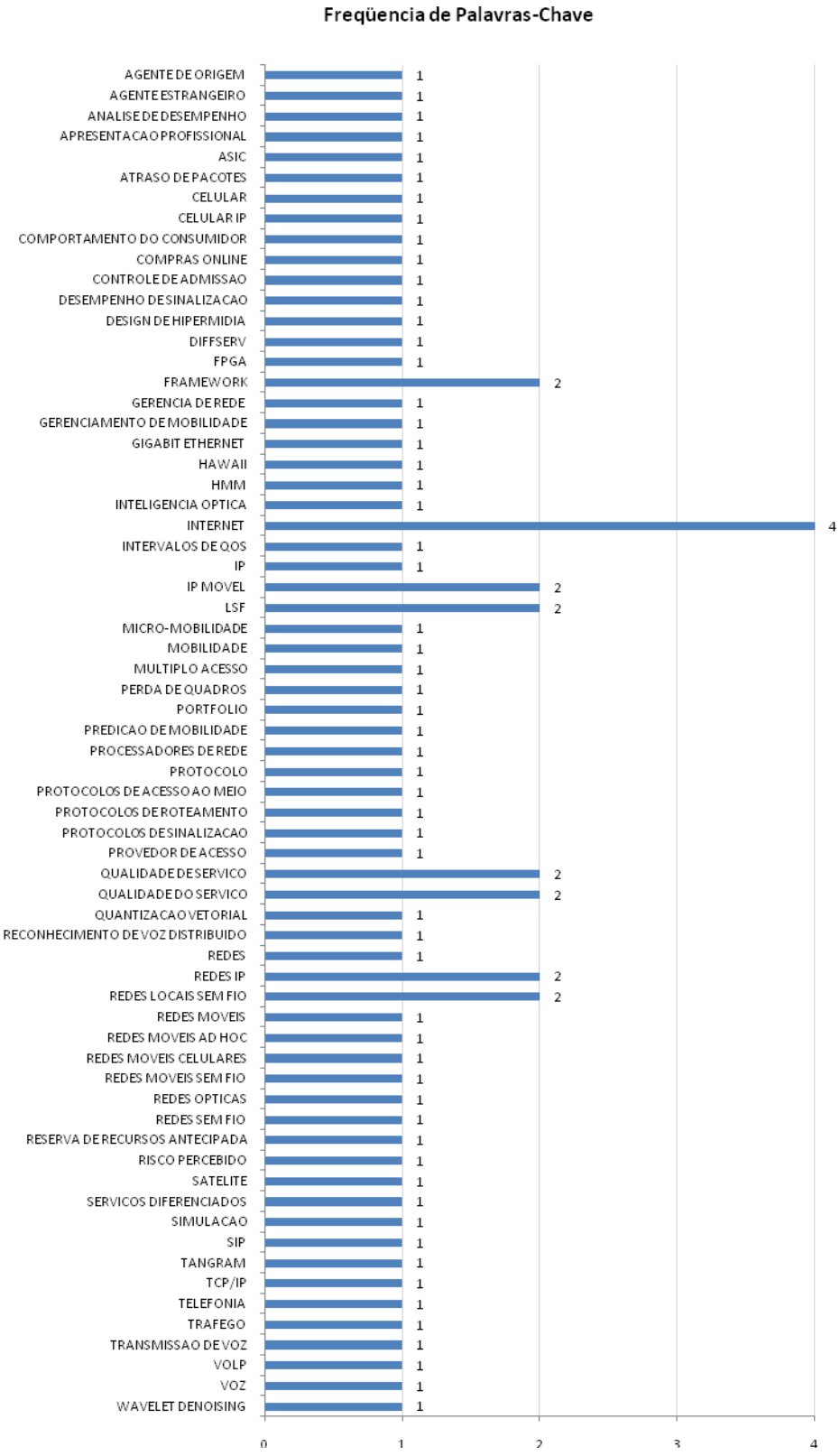

Figura 7: Palavras-chave atribuídas pelos autores – Cluster 17

6.2 RefViz

O software RefViz possui algumas ferramentas para auxiliar o processo e a visualização dos resultados, porém todas estão implementadas em inglês. Desta forma, antes da realização efetiva do processamento, tais ferramentas foram adaptadas para o português.

O software faz a redução das s*top words* no momento do processamento. Assim, a *stoplist* em inglês que o programa possui foi substituída pela *stoplist* em português.

O *Thesaurus* foi a outra ferramenta utilizada, mas também adaptada. A lista de sinônimos em inglês foi substituída pela lista de sinônimos em português. Essa lista foi construída a partir da base de dados utilizada para que não fosse muito abrangente e conseqüentemente mais eficiente. As palavras similares ou que apresentavam alguma redundância de informação foram agrupadas como sinônimos no *Thesaurus* do *RefViz*.

Na construção dessa lista, a *Topics Tool* do sistema foi utilizada. Ao processar um conjunto de referências, o *RefViz* divide as palavras da base de dados processada em 3 grupos, formando a *Topics Tool*. Os grupos formados são os seguintes:

- ***Major Topics*** – estão as palavras determinadas pelo *RefViz*, como os conceitos mais importantes para distinção das referências.
- ***Minor Topics*** – são termos adicionais que são associados ao *Major Topics* para enriquecer os principais conceitos, compreender sinônimos e identificar relacionamentos entre os *Major Topics.*
- ***Other Terms*** – são todas as outras palavras.

Após o primeiro processamento de teste com a base na íntegra, esses três grupos foram analisados. Esta ferramenta possibilita que as palavras sejam movidas de seus grupos originais. Assim, de acordo com a análise dos grupos e de suas palavras realizou-se o tratamento necessário. Palavras que não foram removidas no processamento, por não estarem na *stoplist* e que foram alocadas no *Minor Topics*, por exemplo, foram alteradas para *Other Topics*. Exemplo de palavras deste procedimento são os verbos, que não agregam valor no *clustering*.

Juntamente com este tratamento foi construída a lista de sinônimos mencionada acima. Receberam prioridade, as palavras que foram agrupadas no *Major Topics* e no *Minor Topics*, consecutivamente. A construção e utilização desta lista para

novos processamentos pode ser comparada ao processo de normalização descrito anteriormente. Foi realizada uma aproximação de conceitos, onde conseqüentemente o tamanho do léxico foi reduzido. É importante ressaltar que ao realizar a aproximação dos conceitos, teve-se o cuidado de não causar grandes impactos na precisão. Uma pequena amostra da lista de sinônimos está disponível no Apêndice B e a diferença no tamanho do léxico após o seu uso pode ser observada no gráfico da Figura 8, apresentada abaixo.

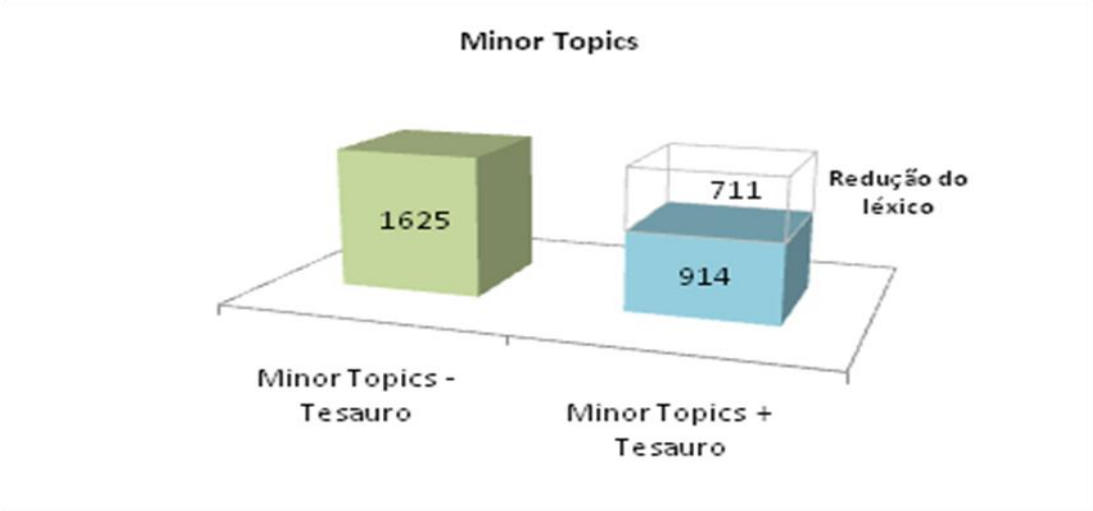

Figura 8: *Minor Topics* com e sem *Thesaurus*

Com todos os dados trabalhados e formatados, conforme a especificação desta ferramenta deu-se início aos processamentos. Foram realizados vários testes utilizando a *stoplist* e outros sem fazer uso dela. Os melhores resultados obtidos são apresentados nos próximos itens.

6.2.1 RefViz com Stoplist

Conforme falado anteriormente, neste processamento a *stoplist* foi utilizada para realizar o *clustering*. O *Thesaurus* também foi utilizado, já que apresentou melhores resultados, como pôde ser observado nos gráficos acima.

Na Figura 9, pode-se analisar a distribuição dos 46 clusters formados a partir deste processamento. Como se pode perceber na *Galaxy Visualization,* as referências apresentam uma distribuição mais aproximada com uma distribuição uniforme, ou seja, estão bem espalhadas ao longo do espaço, caracterizando que algumas áreas

não são altamente representadas. Também, observa-se que alguns clusters são densos, bem definidos, com as referências bem próximas ao seu centro, enquanto outros clusters são mais esparsos e não delimitados, possuindo referências que poderiam ser inseridas em outro cluster.

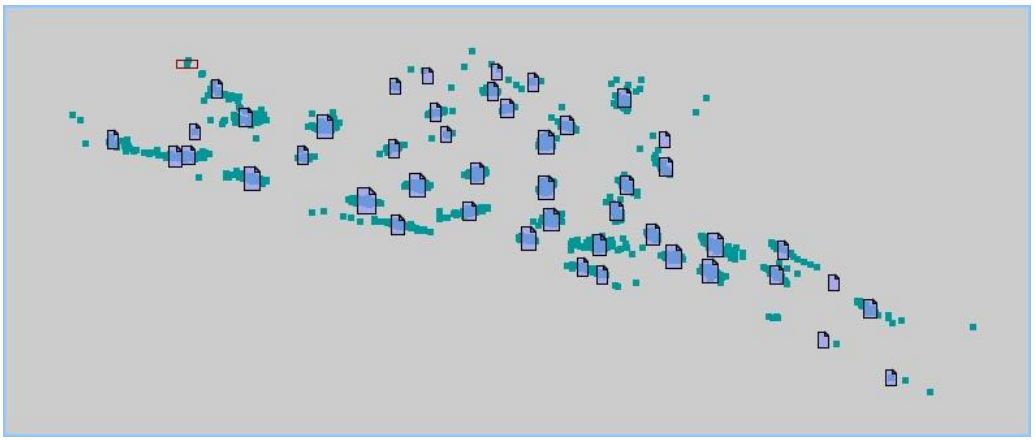

Figura 9: *Galaxy* - 46 clusters e 2140 textos

Figura 10: *Galaxy* - 46 clusters

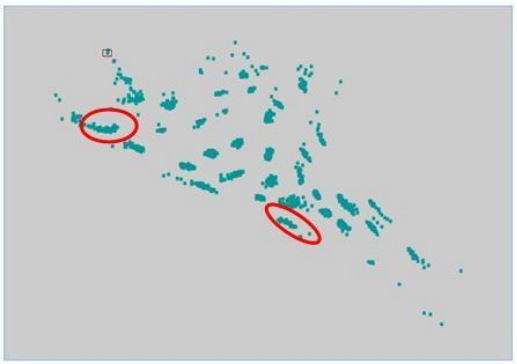

Figura 11: *Galaxy* - 2140 textos

Nas Figuras 10 e 11 pode-se visualizar, separadamente, os clusters e os textos, respectivamente. Nota-se na Figura 10, os clusters com maior proximidade, ou seja, mais similares, inclusive dois que estão sobrepostos. Na Figura 11, os dois clusters sobrepostos aparentam apenas um cluster maior. Esta aparência não ocorre apenas neste caso, mas entre clusters de grande similaridade. A questão da similaridade entre os clusters pode ser melhor analisada com o auxílio da Figura 12, que apresenta a *Matrix Visualization* do resultado obtido.

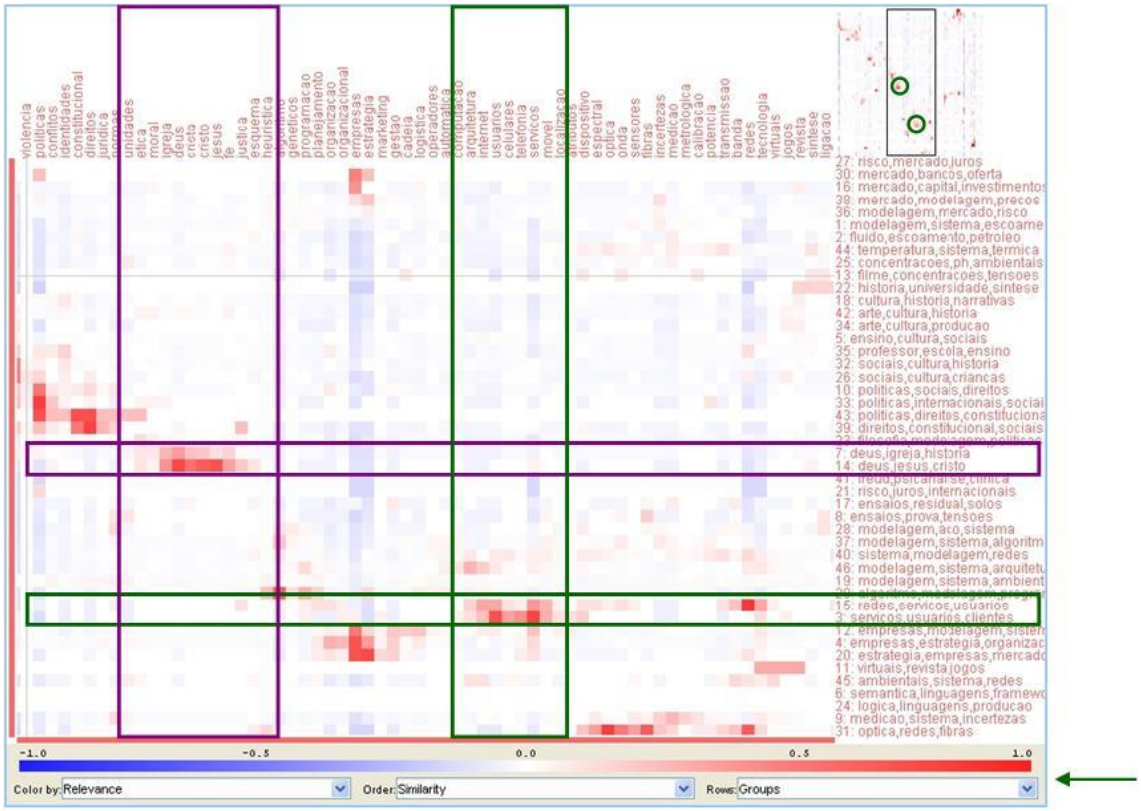

Figura 12: *Matrix* – Clusters e Palavras

Na parte inferior existem alguns critérios para alterações da visualização dos clusters. Nesta visualização, os clusters estão numerados, etiquetados por suas palavras-chave e representados pelas linhas. Estão ordenados segundo a similaridade entre os clusters e, o colorido é apresentado segundo o grau de relevância das palavras da coleção, representadas pelas colunas, nos respectivos clusters. Quanto mais vermelha a célula representante de uma palavra num determinado cluster, maior a sua relevância ou freqüência. Na parte superior direita, está disponível uma miniatura da figura inteira, que possui um marcador de posicionamento no espaço, representada pelo retângulo preto.

Assim, a análise do resultado é facilitada, visto que estes recursos de visualização permitem diversas comparações viabilizando conclusões mais expressivas.

Percebe-se que os pontos em vermelho estão sempre próximos, em se tratando de clusters similares. Os clusters 15 e 3, marcados pelos retângulos, vertical e horizontal, verdes, são propostos para esta análise. As linhas 15 e 3 apresentam pontos vermelhos, com graduações diferentes, relativamente nas mesmas colunas ao longo de toda a figura. As palavras-chave do cluster 15 são **redes**, **serviços** e **usuários**, enquanto que as do cluster 3 são **serviços**, **usuários** e **clientes**. Outras

palavras como, internet, arquitetura, computação, tecnologia, banda entre outras, também estão presentes nos dois grupos e reunidas próximas ao posicionamento atual. Na miniatura superior, nota-se que para o lado esquerdo não há palavras relevantes para estes grupos, já que a coloração está de branco para azul.

Outros dois clusters bastante similares são os de números 7 e 14, também delimitados por retângulos, vertical e horizontal, na cor uva. Além da visualização acima, as Tabelas 22 e 23, logo abaixo apresentam comparações entre os 15 termos mais freqüentes nos dois clusters. O cluster 7 é composto por 40 referências, enquanto o cluster 14 possui 19 referências. É importante destacar novamente, que a freqüência das palavras está relacionada ao número de referências em que aparecem e não ao número de vezes que aparecem no cluster.

Tabela 22: *Major Topics*

***Major Topics* - 15 mais freqüentes**			
Cluster 14	**F**	**Cluster 7**	**F**
deus	15	deus	17
jesus	14	igreja	12
cristo	13	historia	11
crista	11	cultura	11
igreja	8	crista	11
fe	7	cristo	9
historia	6	sociais	8
justica	4	fe	8
servicos	2	politicas	6
opcoes	2	modelagem	6
narrativas	2	jesus	5
mulheres	2	etica	5
fundo	2	unidades	4
etica	2	narrativas	4
esquema	2	mulheres	4

Tabela 23: *Minor Topics*

***Minor Topics* - 15 mais freqüentes**			
Cluster 14	**F**	**Cluster 7**	**F**
humana	14	humana	17
vida	11	dissertacao	17
presente	10	mundial	16
religioso	9	experiencia	16
experiencia	9	vida	15
teologia	8	religioso	14
pessoa	7	homem	12
mundial	7	dialogo	12
atuais	7	busca	12
humanidade	6	teologia	11
espirito	6	atuais	11
acao	6	sentido	10
tema	5	presente	10
superacao	5	pensamento	10
pensamento	5	obra	10

A *Matrix* possibilita uma visão geral da coleção de documentos, fornecendo um panorama dos conceitos existentes. Possibilita ainda, uma melhor compreensão das sobreposições e associações existentes entre os clusters e entre os conceitos. Assim, como já era sabido, constatou-se que a coleção é composta por textos de diversas áreas do conhecimento. Porém, o software apresentou maior granularidade na coleção, com um número elevado de clusters, conseqüentemente menores e mais específicos. As teses em questão fazem parte de 24 programas de pós-graduação, provavelmente bastantes abrangentes. Sabe-se que os programas de pós-graduação

também possuem subdivisões em áreas de concentração, mas esses dados não foram utilizados no *clustering*.

Hoje em dia, como as áreas são inter e multidisciplinares, propôs-se analisar a alocação de textos de diferentes áreas de concentração e programas de pós-graduação num mesmo cluster. Ainda no cluster 7, verificou-se que existem referências de alguns programas de pós-graduação como Teologia, História, Filosofia e Letras.

6.2.2 RefViz sem Stoplist

O processamento sem a redução das s*top words* foi realizado, mas como o *RefViz* está baseado apenas em análise estatística, tornou-se inviável o *clustering* da base com resultados confiáveis.

De acordo com esta abordagem, analisa apenas palavras e não termos ou lexemas. Por exemplo, o lexema “Rio de Janeiro” foi dividido em 3 palavras: “Rio”, “de”, “Janeiro”. Sem a redução das s*top words*, além de não reconhecer este lexema, o que já é prejudicial para a análise, a ferramenta contabiliza a preposição “de” para realizar os agrupamentos. Ou seja, preposições, artigos, pronomes entre outros se tornam as palavras mais freqüentes em toda a base.

Percebe-se que apesar destes problemas, o *RefViz* não inseriu tais palavras no *Major Topics,* grupo que possui os conceitos mais importantes, e sim nos outros dois grupos. Este é um ponto positivo, pois seu algoritmo consegue perceber que quando a freqüência de uma palavra é muito alta, provavelmente esta não é uma palavra representativa para o processo de mineração. Dependendo da palavra e da sua freqüência, dificilmente não estará presente em todos os documentos. Ou seja, deve ser desconsiderada. Na tabela abaixo pode-se observar a incidência das dez palavras mais freqüentes no *Minor Topics* e no *Other Topics* após este processamento.

Tabela 24: *RefViz* - Processamento sem *Stoplist*

Sem Stoplist			
Minor Topics	**F**	**Other Topics**	**F**
na	1655	de	2139
com	1642	e	2129
os	1638	a	2127
como	1575	o	2079
as	1565	da	1989
dos	1511	do	1986
por	1368	em	1912
das	1294	que	1900
trabalho	1132	para	1860
se	1101	uma	1782

É importante ressaltar que a freqüência apresentada na coluna F é relacionada ao total de documentos em que a palavra aparece e não ao total de vezes em que ela aparece nos documentos. Desta forma, a palavra “de” do exemplo acima, deixou de aparecer em apenas um documento da base, visto que o universo é composto por 2140 textos.

6.2.3 RefViz com Cortex

Após a realização dos processamentos descritos nos itens anteriores, estudou-se a possibilidade de *clustering* no *RefViz*, utilizando arquivos pré-processados pelo Cortex para posterior comparação dos resultados. Para isso, alguns passos tiveram que ser realizados e refeitos.

O pré-processamento foi novamente realizado, visto que as entidades identificadas e nomeadas pelo Cortex deveriam ser expressas em apenas uma palavra, visto que o *RefViz* não analisa lexemas. A maneira adotada para este processo foi implementar um algoritmo que unisse os lexemas, quando identificados, através do caractere *underline* **(_)** e nomeá-los com símbolos que representam a sua classe gramatical (verbo, adjetivo, substantivo) ou a semântica (pessoa, lugar, valor monetário) dentro dos contextos dos documentos. O exemplo abaixo esclarecerá o procedimento adotado.

O lexema “rio de janeiro”, quando processado pelo *RefViz* é separado em 3 palavras: “rio”, “de” e “janeiro”. Após a realização deste novo pré-processamento, o *output* deste lexema passou para “rio_de_janeiro_**ng**”. Ou seja, agora o lexema será reconhecido pelo *RefViz* como uma palavra, conforme seu processamento. O símbolo “ng” possui a função de nomear a entidade reconhecida “rio_de_janeiro”. Na tabela abaixo, pode-se observar a lista dos símbolos mais freqüentes com seus respectivos significados.

Tabela 25: Símbolos das Entidades Nomeadas pelo Cortex

CORTEX	
Símbolos	**Significados**
a	agente
aj	cargo
currency	dinheiro
date	data
n	nome
na	nome abstrato
ng	nome geográfico
no	nome de organização
np	nome de pessoa
numeric	números
nw	nome de obra, publicação
p	particípio
petroleum	petróleo
q	qualificador
qg	qualificador de graduação
s	substantivo
v	verbo

É importante ressaltar que as s*top words* não foram eliminadas antes e nem após a identificação das entidades. As s*top words* que não compuseram as entidades identificadas pelo Cortex, não foram nomeadas conforme todas as outras palavras, podendo ser excluídas no momento do processamento, de acordo com o processo do *RefViz*. Desta forma, não foi necessário a construção de nova *stoplist*.

Este processo pode melhorar os índices de acertos, pois o RefViz não diferencia maiúsculas de minúsculas e não reconhece acentos, uma palavra como “Pará” seria confundida com a *stop word* “para”, sendo eliminada no processamento. Já com o procedimento realizado pelo Cortex, a palavra “Pará” será transformada em “para_ng” enquanto que a *stop word* “para” não receberá qualificador.

A construção do *Thesaurus* foi outro processo refeito. Desta vez, com o léxico ampliado, visto que uma palavra pôde ter recebido símbolos diferentes, pequenas

variações nos lexemas foram identificadas e eliminadas através da lista dos "sinônimos". O mesmo "rio_de_janeiro_**ng**" também apareceu como "rio__de_janeiro_**ng**". Pode-se perceber a existência do espaço duplo entre o "rio" e o "de" que fez o *RefViz* identificar duas palavras ao invés de uma. Associações como número e gênero também foram realizadas.

Após alguns processamentos de teste, verificou-se uma grande quantidade de verbos entre outras palavras não representativas, que pela alta freqüência foram alocadas no *Major Topics* e no *Minor Topics*. Assim, realizou-se novo pré-processamento, excluindo do arquivo de output, as palavras nomeadas como verbos, datas, números, dentre outras. Este procedimento facilitou a análise dos termos alocados nos dois grupos citados e a realização das alterações necessárias para obtenção de melhores resultados no *clustering*. Os resultados são apresentados logo abaixo.

O número de clusters formados a partir deste processamento é 46 no total. Desses 46, nenhum cluster possui apenas 1 documento e 10 são formados por até 10 documentos. A partir da Figura 13, pode-se observar a distribuição de todos os clusters para realização das análises, em seguida.

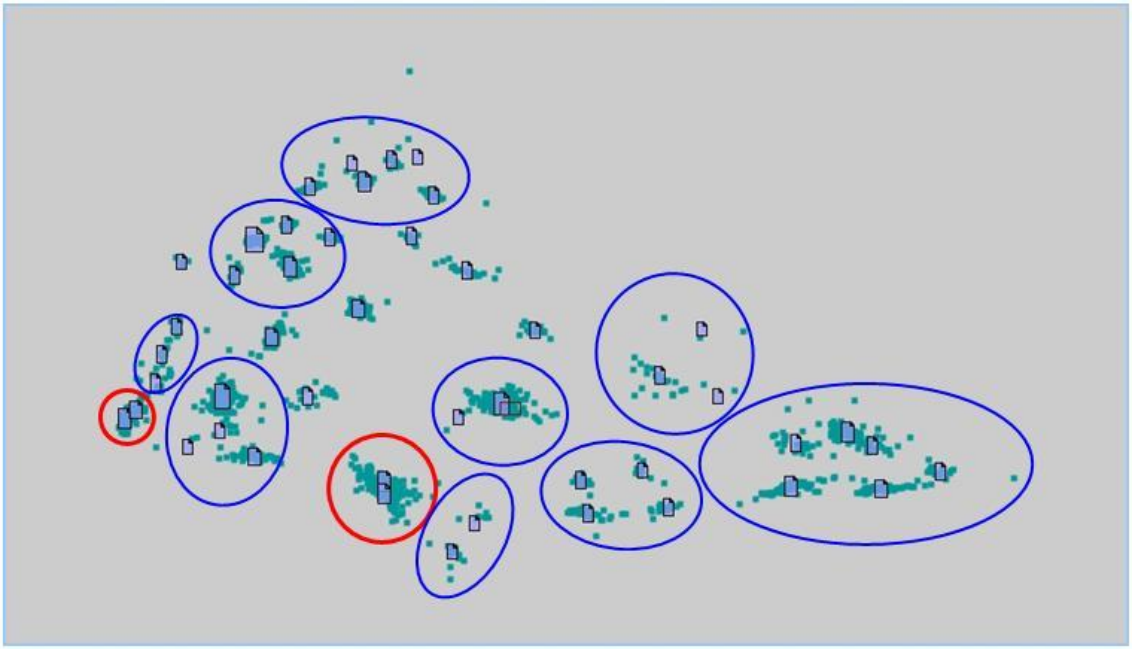

Figura 13: *Galaxy* – 46 clusters e 2140 teses

A distribuição está um pouco esparsa, porém apresenta grupos de clusters, que foram demarcados pelos envoltórios na cor azul para melhor percepção.

Apresenta, também, clusters mais isolados e alguns sobrepostos, apresentando alta similaridade. Na figura, os clusters sobrepostos estão envolvidos por círculos em vermelho.

A visualização da Figura 14 auxilia a interpretação dos resultados e da distribuição acima. Nela, pode-se observar a similaridade entre os clusters, além das palavras-chave representantes dos mesmos. Ressalta-se que as palavras da coleção de documentos estão identificadas segundo o reconhecimento de entidades no pré-processamento e que as siglas e seus significados podem ser consultados na Tabela 25, conforme mencionado anteriormente.

Como resultados deste processamento, os clusters dispostos relativamente no centro da distribuição serão apresentados para análise. São os clusters 3, 11 e 23, que podem ser identificados na Figura 14. Na Tabela 26 está disponível o comparativo do *Major Topics* mais freqüentes em cada cluster**.**

Tabela 26: *Major Topics* – Com Cortex

Major Topics	
Cluster 3	**F**
arquitetura_s	12
internet_s	9
sistema_s	7

Major Topics	
Cluster 11	**F**
virtual_q	3
jogo_s	2
arquitetura_s	2

Major Topics	
Cluster 23	**F**
sistema_s	33
politicas_q	31
producao_s	23

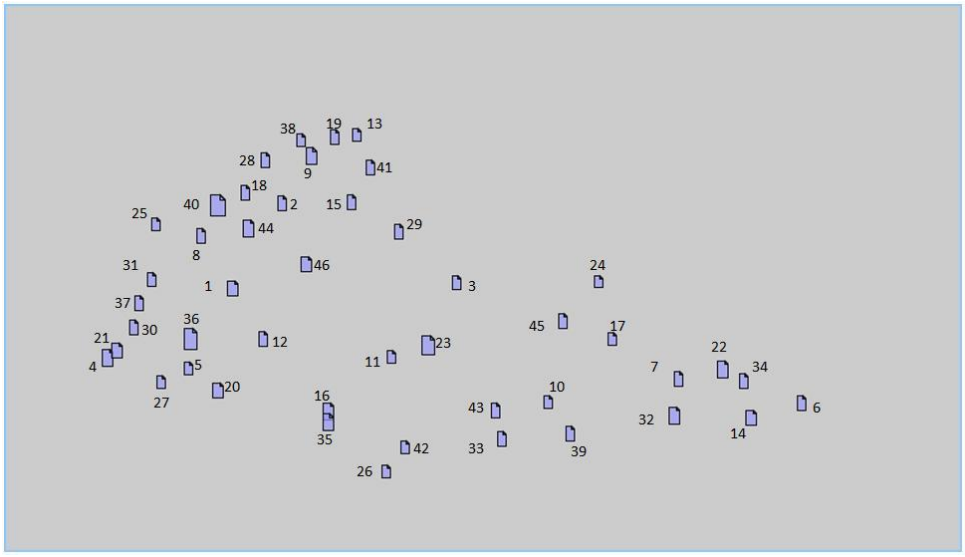

Figura 14: *Galaxy* – Clusters Identificados

Através da tabela e da figura acima, pode-se verificar que os 3 clusters estão próximos uns dos outros representando que são ligeiramente similares entre si. Dentre os 3, o cluster 23 é o maior e foi o que obteve mais erros em relação à classificação

dos documentos. Ao compará-los com as reais áreas dos documentos, verifica-se que é composto por 7 delas. O cluster 11 é o menor, com 3 documentos e todos relacionados à Informática, sendo 2 deles específicos de jogos.

Na tabela abaixo são apresentadas as palavras mais freqüentes dentre as palavras-chave atribuídas pelos autores. Analisando a Tabela 27, a seguir, nota-se que o cluster 3 também é formado por documentos de Informática, porém, trata de assuntos diferentes dentro desta grande área.

Segundo esta análise, pode-se dizer que o *RefViz* obteve um bom desempenho neste processamento, visto que não reuniu o cluster 3 e 11, posicionando-os relativamente próximos, já que são similares, mas com atenção em tratar os documentos de acordo com suas especificidades.

Tabela 27: RefViz Com Cortex – Clusters Centrais

Palavras-chave	
Cluster 3	**F**
software	10
web	9
semantica	6

Palavras-chave	
Cluster 11	**F**
virtuais	4
ambientes	3
rede	2

Palavras-chave	
Cluster 23	**F**
design	22
cultura	17
etica	14

Finalizada a etapa dos processamentos e das análises realizadas a respeito de cada um ao longo da pesquisa, partiu-se para uma análise geral dos resultados obtidos.

Após analisar todos os processamentos, observou-se que os melhores resultados foram obtidos a partir dos processamentos realizados pelo *RefViz*. Dentre estes, a melhor performance foi quando houve a utilização do Cortex na fase do pré-processamento. Os resultados foram mais relevantes, além do baixo tempo de processamento.

Com as análises realizadas, verificou-se um ponto bastante interessante em relação às comparações realizadas. Pode-se considerar que a análise realizada é bidirecional. Ao mesmo tempo em que o desempenho dos softwares foram analisados, comparando os resultados com a realidade, que são as palavras atribuídas pelos próprios autores ou pelo sistema de informação; a realidade também é colocada em *check,* visto que nem sempre os autores utilizam palavras ou termos que representem o seu documento adequadamente. Muitas vezes foram encontradas palavras muito genéricas ou muito específicas, descaracterizando o real teor do documento. Isto acarreta ao usuário pesquisador, um pouco de frustração na realização de uma busca

ou até mesmo prejuízos na pesquisa. Se um documento não for bem indexado na entrada de um sistema, por exemplo, com certeza não terá uma boa saída. Isso quer dizer, que pode ser recuperado em buscas indevidas e que pode não ser localizado em uma busca específica sobre o assunto tratado. A partir desta análise, nem sempre as palavras atribuídas pelos autores podem ser tomadas como verdade dentro de seu próprio documento. A Figura 15 ilustra o tratamento de documentos, tanto na entrada como na saída de um sistema de informação.

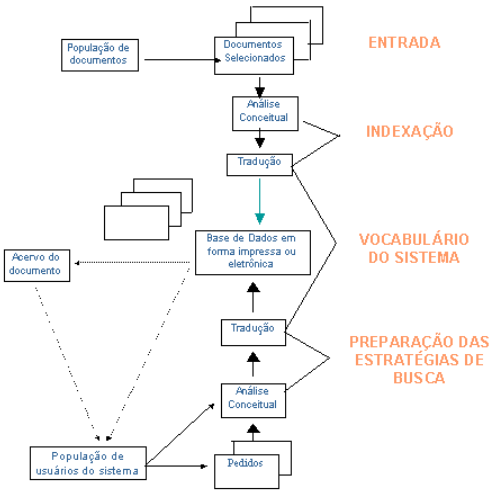

FIGURA 15 – Sistema de Recuperação da Informação

7 Conclusões

O trabalho consistiu no desenvolvimento de uma metodologia de Mineração de Textos em língua portuguesa. Assim, algumas definições da Mineração de Textos foram apresentadas com intuito de uma descrição geral das técnicas disponíveis.

As ferramentas e ambientes de aplicação da metodologia em desenvolvimento foram detalhados, incluindo a descrição da base de textos utilizada para todos os testes.

Durante o desenvolvimento do trabalho, pode-se constatar a necessidade de cuidadosa atenção à fase do pré-processamento e as técnicas e processos relacionados à esta etapa. Pequenos detalhes foram tratados por conta de impactos que poderiam ser gerados em fases posteriores. Estes detalhes obtiveram maior peso, devido à realização da análise semântica. Por conta dos diferentes formatos de entrada dos softwares utilizados, algumas adaptações foram realizadas.

Houve a criação de uma *stoplist* que foi utilizada na etapa do processamento da base em cada ferramenta e a contrução do *Thesaurus,* funcionalidade adaptada da ferramenta *RefViz*. A *stoplist* e o *Thesaurus* foram criados a partir da base de dados para que fossem menos abrangentes e mais eficientes.

Após a verificação da possibilidade de realização de processamentos no *RefViz,* com a base de dados pré-processada através do Córtex, foi identificada a necessidade da criação de uma nova *stoplist* e um novo *Thesaurus.* A utilização destes dois recursos no novo processamento foi considerada de grande relevância, por isso decidiu-se realizar as adaptações e adequações, mesmo que trabalhosas.

Notou-se que ao realizar-se um bom pré-processamento, torna-se mais simples a aplicação das técnicas de Mineração de Textos.

De acordo com a metodologia desenvolvida, percebe-se a importância da análise semântica, quando se trata de Mineração de Textos. Percebe-se, também, que os melhores resultados foram alcançados a partir da combinação das análises estatística e semântica. Este fato pode ser comprovado, a partir dos resultados encontrados. Nos processamentos realizados em todas as ferramentas, os que apresentaram melhores resultados foram os que na etapa do pré-processamento, o Cortex foi utilizado. Tratando-se apenas de análise estatística, a redução das *stop words* tornou-se imprescindível. E quando realizada juntamente com a análise semântica continuou agregando valor ao resultado final.

Além disso, constatou-se que a clusterização como outras técnicas de Mineração de Textos podem ser utilizadas em diferentes universos, dependendo do objetivo de suas aplicações.

Deve-se ressaltar que na análise dos resultados, também, foi realizada a análise da classificação real do documento, que não é atribuída pelo Sistema Maxwell, mas pelo próprio autor do texto. A não existência de um vocabulário controlado prejudica o sistema, pois apesar de disponibilizar diferentes buscas, os documentos não estão sendo igualmente tratados. Desta forma, sugere-se o desenvolvimento ou adaptação de algum vocabulário para a utilização no sistema.

Outro ponto relevante foi constatar que a utilização de mais de uma técnica de Mineração de Textos num mesmo contexto pode apresentar resultados mais relevantes. No caso desta pesquisa, indica-se além da aplicação da clusterização, a aplicação da técnica de classificação, como sugestão para resultados ainda mais interessantes. A criação de classes de documentos é bem satisfatória, quando se trata de textos que integram coleções de um sistema de informação como uma biblioteca. Considera-se que a aplicação dessas duas técnicas, trará resultados de grande relevância nesse universo de informações.

No entanto, o estudo possibilitou efetuar a aplicação de apenas uma técnica. Segue como sugestão para trabalhos futuros o desenvolvimento de outras pesquisas desta natureza que possam contribuir para o aprimoramento e validação desta metodologia, incluindo sua aplicação em novas bases de textos científicos ou literários, como as bases da Biblioteca Nacional.

Referências

ARANHA, C. N., **Uma abordagem de pré-processamento automático para mineração de textos em Português: sob o enfoque da Inteligência Computacional.** 2007. 144 f. Tese (Doutorado em Engenharia Elétrica) - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica, Rio de Janeiro, 2007.

BATISTA, G. E. A. P. A. **Pré-processamento de dados em aprendizado de máquina supervisionado**. 2003. Tese (Doutorado em Ciências da Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2003.

BASTOS, Valeria M. **Ambiente de Descoberta de Conhecimento na Web para a língua portuguesa.** 2006. Tese (Doutorado em Engenharia Civil) - Instituto Alberto Luiz Coimbra de Pós Graduação e Pesquisa de Engenharia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2006.

BIGGS, M. Resurgent text-mining technology can greatly increase your firm's 'intelligence' factor**. InfoWorld**. v. 11, n.2, 52, 2000.

BRASIL. Ministério da Cultura. Biblioteca Nacional. **Fundação Biblioteca Nacional.** Disponível em: <http://www.bn.br/>. Acesso em: 01 dez. 2006.

CHEN, H. **Knowledge management systems: a text mining perspective**. Tucson, Arizona: University of Arizona Press, 2001.

CORTEX INTELLIGENCE (Brasil). **Cortex Intelligence.** Disponível em: <www.cortex-intelligence.com>. Acesso em: 20 jan. 2008.

CUTTING, D. R. *et al.* Scatter/gather: a cluster-based approach to browsing large document collections. *In:* ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL*, 15.*, 1992, Copenhagen. **Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.** Copenhagen, 1992. p. 318–329.

DONGEN, Stijn van. **Graph Clustering by Flow Simulation**. 2000. 175 f. PhD thesis, University of Utrecht: 2000. Disponível em: <http://micans.org/mcl/lit> Acesso em: 15 jul. 2008.

EBECKEN, N. F. F.; LOPES, M.C.S.; COSTA, M.C.A. Mineração de Textos. *In*: REZENDE, S.O (Coord.). **Sistemas Inteligentes**: Fundamentos e Aplicações. Barueri, SP: Manole, 2005.

GOLDSCHMIDT, R., PASSOS, E. **Data Mining**: Um Guia Prático. Rio de Janeiro: Elsevier. 2005.

GRIFFITHS, A.; ROBINSON, L. A.; WILLETT, P. Hierarchical agglomerative clustering methods for automatic document classification. **Journal of Documentation**, v. 40, n.3, p. 175–205, Sept. 1984.

HEARST, M. A.. Untangling Text Data Mining. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 37., 1999, College Park, Maryland. **Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics.** College Park, Maryland: University Of Maryland, 1999. p. 3 – 10.

LAMBDA/PUC-Rio. **Sistema Maxwell.** Disponível em: <http://www.maxwell.lambda.ele.puc-rio.br/>. Acesso em: 03 dez. 2006.

LANCASTER, F. W. **Indexação e Resumos: teoria e prática**. Brasília: Briquet de Lemos, 1993.

_____. **Vocabulary Control for Information Retrieval**. 2nd. ed. Arlington, VA, Information Resources Press, 1986.

LOPES, M. C. S. **Mineração de Dados Textuais utilizando Técnicas de Clustering para o idioma Português.** 2004. 191 f. Tese (Doutorado em Engenharia Civil) - Instituto Alberto Luiz Coimbra de Pós Graduação e Pesquisa de Engenharia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2004.

LUCAS, M. Mining in textual mountains, an Interview with Marti Hearst. **Mappa Mundi Magazine**, Trip-M 005, 2000. Disponível em: <http://www.mundi.net/trip-m/hearst/ > Acesso em: 05 jun. 2007.

MARTINS, C. A. **Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado**. 2003. Tese (Doutorado em Ciências da Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2003.

MCL System. Versão 1.0. Rio de Janeiro, 2008.

REFVIZ. Stamford, Conn.: Thomson Research Soft, 2005. Disponível em : <www.refviz.com>. Acesso em: 08 ago. 2007.

SESC. Projeto: **Atualização em Indexação e Controle de Vocabulário à Distância**. Unidade II: Instrumentos de Indexação / Recuperação da Informação. Módulo 1: Linguagens Documentárias. Rio de Janeiro: SESC, 1998.

SULLIVAN, D. The need for text mining in business intelligence. **DM Review**, 2000. Disponível em: <http://www.dmreview.com/master.cfm> . Acesso em: 05 jun. 2007.

TAN, Ah-hwee. Text mining: The state of the art and the challenges. In: PAKDD 1999 WORKSHOP ON KNOWLEDGE DISOCOVERY FROM ADVANCED DATABASES, 1999, Beijing. **Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases.** Beijing, 1999. p. 65 - 70.

THURAISINGHAM, B. **Data mining:** technologies, techniques, tools, and trends. Florida: CRC Press, 1999.

THOMSOM ResearchSoft. RefViz. Version 2.0. Stamford, Conn., 2005. Disponível em: <www.refviz.com>. Acesso em: 08 ago. 2007.

WEISS, S. M. et al. **Text Mining**: Predictive Methods for Analyzing Unstructured Information. New York: Springer, 2004.

ZAMIR, O. et al. Fast and intuitive clustering of Web documents. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 3., 1997, Newport Beach, CA. **Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining.** Newport Beach, CA: AAAI Press, 1997. p. 287 - 290.

APÊNDICE A - Stoplist

a
à
acerca
adeus
agora
aí
ainda
além
algmas
algo
algumas
alguns
ali
ambos
ano
anos
antes
ao
aos
apenas
apoio
apontar
após
aquela
aquelas
aquele
aqueles
aqui
aquilo
área
as
às
assim
até
atrás
através
baixo
bastante
bem
bom
breve
cá
cada
caminho
catorze
cedo
cento
certamente
certeza
cima
cinco
coisa
com
como
como
comprido
conhecido
conselho
contra
corrente
custa
da
dá
dão
daquela
daquele
dar
das
de
debaixo
demais
dentre
dentro
depois
desde
desligado
dessa
desse
desta
deste
deve
devem
deverá
dez
dezenove
dezesseis
dezessete
dezoito
dia
diante
direita
diz
dizem
dizer
do
dois
dos
dos
doze
duas
dúvida
e
é
ela
elas
ele
eles
em
embora
enquanto
então
entre
era
és
essa
essas
esse
esses
esta
está
estão
estar
estará
estas
estás
estava
este
este
estes
esteve
estive
estivemos
estiveram
estiveste
estivestes
estou
eu
exemplo
faço
falta
fará
favor
faz
fazeis
fazem
fazemos
fazer
fazes
fazia
fez
fim
final
foi
fomos
for
fora
foram
forma
foste
fostes
fui
geral
grande
grandes
grupo
há
hoje
horas
iniciar
inicio
ir
irá
isso
ista
iste
isto
já
lá
lado
ligado
local
logo
longe
lugar
maior
maioria
maiorias
mais
mal
mas
máximo
me
meio
menor
menos
mês
meses
mesmo
meu
meus
mil
minha
minhas
momento
muito
muitos
na
nada
não
naquele
nas
nem
nenhuma
nessa
nesse
nesta
neste
nível
no
noite
nome
nos
nós
nossa
nossas
nosso
nossos
nova

nove
novo
novos
num
numa
número
nunca
o
obrigada
obrigado
oitava
oitavo
oito
onde
ontem
onze
os
ou
outra
outras
outro
outros
para
parece
parte
partir
pegar
pela
pelas
pelo
pelos
perto
pode
pôde
podem
poder
poderá
podia
põe
põem
ponto
pontos
por
porque
porquê
posição
possível
possivelmente
posso
pouca
pouco
povo
primeira
primeiro
promeiro
próprio
próximo
puderam
qual
qualquer
quando
quanto
quão
quarta
quarto
quatro
que
quê
quem
quer
quero
questão
quieto
quinta
quinto
quinze
relação
sabe
saber
são
se
segunda
segundo
sei
seis
sem
sempre
ser
seria
sete
sétima
sétimo
seu
seus
sexta
sexto
sim
sob
sobre
sois
somente
somos
sou
sua
suas
tal
talvez
também
tanto
tão
tarde
te
tem
têm
temos
tempo
tendes
tenho
tens
tentar
tentaram
tente
tentei
ter
terceira
terceiro
teu
teus
teve
tipo
tive
tivemos
tiveram
tiveste
tivestes
toda
todas
todo
todos
trabalhar
trabalho
três
treze
tu
tua
tuas
tudo
último
um
uma
umas
uns
usa
usar
vai
vais
valor
vão
vários
veja
vem
vêm
vens
ver
verdade
vez
vezes
viagem
vindo
vinte
você
vocês
vos
vós
vossa
vossas
vosso
vossos
zero

APÊNDICE B – Amostra do Thesaurus construído para o RefViz

THESAURUS DO REFVIZ - Lista de Sinônimos												
Termo 1	**Termo 2**	**Termo 3**	**Termo 4**	**Termo 5**	**Termo 6**	**Termo 7**	**Termo 8**	**Termo 9**	**Termo 10**	**Termo 11**	**Termo 12**	**Termo 13**
abastecimento	abastecimentos											
aberto	abertura	abertos	aberto	abertas	aberta	aberturas						
aborda	abordagem	abordagens	abordar									
abstracao	abstracoes											
abuso	abusos											
bancos	banco											
banda	bandas											
baseada	baseadas	baseado	baseados	baseia								
basica	basicas	basico	basicos	basicamente								
brasil	brasileira	brasileiras	brasileiro	brasileiros								
cadeia	cadeias											
calculados	calculo											
calibracao	calibracoes											
calor	calores											
camada	camadas											
dado	dados											
decada	decadas											
decisao	decisoes											
definicao	definido	definidos	definir									
demanda	demandas											
economica	economicas	economico	economicos									
educacao	educacoes											
efeito	efeitos											
eficacia	eficaz											
eficiencia	eficiente	eficientes										
faixa	faixas											

familia	familias											
familiar	familiares											
fase	fases											
fato	fatos											
ganho	ganhos											
genero	generos											
geneticos	genetico											
gerada	geradas	gerado	gerados									
gerencia	gerenciamento											
heranca	herancas											
heuristica	heuristicas											
hibrido	hibridos											
hidrocarboneto	hidrocarbonetos											
hipotese	hipoteses											
idade	idades											
ideal	ideais											
ideia	ideias											
identidades	identidade											
identifica	identificacao											
jogos	jogo											
jovens	jovem											
juridica	juridicas											
justica	justicas											
laboratorio	laboratorios											
lacuna	lacunas											
legislacao	lei											
leitor	leitores											
leitura	leituras											
manifestacao	manifestacoes											
mapas	mapa	mapeamento	mapeamentos	mapeados	mapeadas	mapear	mapeia	maps	maps-	map		
mapeamento	mapeamentos											
maquina	maquinas											

marca	marcas											
nacao	nacoes											
nacionais	nacional											
narrativas	narrativa											
naturais	natural											
natureza	naturezas											
objetiva	objetivo	objetivos										
objeto	objetos											
obra	obras											
observacao	observacoes	observada	observadas	observado	observados	observar	observou-se					
obtencao	obter	obtida	obtidas	obtido	obtidos							
padrao	padroes											
palavra	palavras											
panorama	panoramas											
papeis	papel											
paradigma	paradigmas											
quadro	quadros											
questionario	questionarios											
quimica	quimicas	quimicamente	quimico	quimicos								
raio	raios											
raio-x	raios-x											
razao	razoes											
reacao	reacoes											
reais	real											
seculo	seculos											
segmento	segmentos											
seguranca	segurancas											
semantica	semanticas	semanticamente	semantico	semantico-	semanticos	semantic						
semelhante	semelhantes											
tarefa	tarefas											
tecnica	tecnicas											

tecnologia	tecnologias											
telefonia	telefone	telefones	telefonica	telefonico	telefonicos							
televisao	tv	televisivas	televisivos	televisiva								
ultima	ultimas	ultimos										
unica	unico											
unidades	unidade											
universidade	universidades											
usada	usadas	usado										
validacao	validar											
valorizacao	valorizado	valorizados	valorizadas	valores	valoriza	valoriza-se	valorizam	valorizando	valorizar	valorizaram	valorizem	valorize
vantagem	vantagens											
variacao	variacoes											
variaveis	variavel											
zona	zonas											