1
Introdução

Métodos de recuperação de textos sempre foram utilizados para organizar documentos, porém, com o aumento do volume de textos que vem ocorrendo, principalmente, pela digitalização do conteúdo e pela Internet, técnicas de tratamento automático de textos começaram a se tornar cada vez mais importantes para se encontrar e trabalhar a informação. Para solucionar esses problemas surge uma nova linha de pesquisa, a mineração de textos.

Quando uma nova área surge, precisa-se de algum tempo e muita discussão acadêmica antes que seus termos e conceitos sejam padronizados (Kroeze, H. J. et al, 2003). No caso da mineração de textos não é diferente, abaixo encontram-se algumas definições encontradas nos textos estudados.

> Text Mining realiza várias funções de busca, análise lingüística e categorização. Mecanismos de busca se restringem à Internet. (Chen, H., 2001):5,9
>
> Text Mining é o estudo e a prática de extrair informação de textos usando os princípios da lingüística computacional. (Sullivan, D., 2000)
>
> Text Mining é ideal para inspecionar mudanças no mercado, ou para identificar idéias. (Biggs, M., 2005)
>
> Text Mining é uma forma de examinar uma coleção de documentos e descobrir informação não contida em nenhum dos documentos. (Lucas, M., 2000):1
>
> 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. (Thuraisingham, B., 1999):167
>
> 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. (Hearst, M. A., 1999):6-7
>
> Pode-se então definir 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. (Tan, A.-H., 1999)

A partir dessas definições/conceitos podemos ver que a área de mineração de textos tem uma origem forte na área de mineração de dados e KDD (*Knowledge Discovery in Databases*), sendo, por isso, chamada também de *Text Data Mining* (Hearst, M. A., 1999) e KDT (*Knowledge Discovery in Texts*) (Dörre, J. et al, 1999). Existe, também, uma interseção com a área de busca de informação na Internet, assim como influências de áreas correlatas como Processamento da Linguagem Natural (PLN), de Recuperação da Informação (RI), Inteligência Artificial (IA) e Ciência Cognitiva. A conjunção do conhecimento dessas áreas fizeram dela uma área própria, chamada apenas de Mineração de Textos (*Text Mining*).

Um processo de mineração de textos que vem sendo bastante utilizado envolve três etapas principais: a seleção, a indexação e a análise de textos. A seleção de textos tem por objetivo montar a base a ser analisada; a indexação tem por objetivo viabilizar uma busca rápida por um documento específico (algumas vezes opcional); e, finalmente, a análise de textos tem por objetivo extração de informação por algoritmos inteligentes e interpretação do conhecimento contido no texto.

O modelo de mineração de textos, proposto pelo autor, propõe a inserção de um módulo de pré-processamento antes da fase de indexação do processo tradicional descrito. Segundo o modelo proposto, a fase de indexação de textos passa a ser realizada em duas etapas: uma de pré-processamento de textos e outra de indexação. Com a nova etapa, o processo passa a conter quatro etapas no total.

O presente trabalho apresenta, então, uma pesquisa onde é proposto um novo modelo, automático, de mineração de textos, de ênfase na língua portuguesa e inspirado em técnicas de inteligência computacional. O trabalho é complementado com o desenvolvimento e implementação do sistema de pré-processamento.

A hipótese do autor é de que o pré-processamento dos textos, tanto na codificação como no enriquecimento dos textos a serem minerados, é uma parte importante para uma análise eficaz. Sua importância é sugerida pelo maior aproveitamento dos aspectos linguísticos como morfologia, sintaxe e semântica. Os resultados mostram que um pré-processamento simplista pode deixar de lado parte da informação relevante. Essa tarefa, no entanto, apresenta dificuldades com relação a sua complexidade e à própria natureza dos dados, que via de regra, são

não-estruturados e dinâmicos. É necessário, por isso, saber conviver com exceções.

Apesar do objeto principal desta tese ser a etapa de pré-processamento, passaremos por todas as etapas do processo de mineração de textos com o intuito de fornecer a teoria básica completa para o entendimento do processo como um todo. Além de apresentar a teoria de cada etapa individualmente, será descrito um processamento completo (com seleção, indexação, pré-processamento, mineração e pós-processamento) utilizando nas outras etapas modelos já consagrados na literatura e implementados durante esse trabalho. A implementação visa mostrar funcionalidades e algumas aplicações como: classificação de documentos, extração de informações e interface de linguagem natural (ILN).

Outro produto dessa tese é a implementação do sistema de pré-processamento, que terá seus módulos, e funcionalidades, detalhados clareando toda sua complexidade. São especificados, também, o ambiente de desenvolvimento, plataformas utilizadas, formas de integração com outros sistemas de mineração de textos e descrição do funcionamento do sistema em termos de formato das entradas e saídas de cada sub-módulo.

1.1. Motivação

O primeiro motivo é o fato de um processo clássico de mineração de dados, onde não envolve um pré-processamento tão complexo, esta fase consome 60% do tempo total (Goldschmidt, R. e Passos, E., 2005). Estimamos, então, que o pré-processamento automático dos textos para análise seja um procedimento importante e essencial, tanto para a economia de tempo como para o bom funcionamento das etapas seguintes.

O formato textual parece ser a forma mais natural de armazenar informações já que cresce a informação produzida e disponível para os sistemas computacionais armazenada em forma de textos como livros, revistas, artigos científicos, manuais, memorandos, e-mails, relatórios, projetos e outros tipos de formalização de conhecimento. Isso ocorre porque o meio mais intuitivo de externalização (transformação do conhecimento tácito em explícito) é registrar, em textos livres, pensamentos, idéias, sentimentos e opiniões de pessoas. Além de

dados empíricos, duas pesquisas podem comprovar esse pensamento. A primeira mostrou que 80% do conteúdo da Internet está em formato textual (Chen, H., 2001). A segunda mostrou que, além da Internet, nas organizações há também muito conhecimento deste tipo disponível. Verificou-se que 80% das informações armazenadas por uma empresa são também dados não-estruturados (Tan, A.-H., 1999).

Essas pesquisas mostram também que inúmeras novas páginas contendo textos são lançados diariamente na Internet, assim como outros tipos de documentos (como relatórios de acompanhamento, atas de reuniões, históricos pessoais, etc.) são periodicamente gerados, atualizados e armazenados nas empresas. Por esses motivos, a importância da análise automática de textos é reconhecida em todos os segmentos que lidam com informação e conhecimento.

Adicionalmente, grande parte das atividades de tomada de decisões, hoje, envolve a análise de grandes volumes de texto. O processo decisório, que era orientado a análise (automática ou não) de séries temporais e fluxo de dados (*data-driven*) desde os anos 70, está cada vez mais, principalmente das áreas estratégicas das empresas, orientado pelas informações (*information-driven*) (Koenig, M. E. D., 2000).

Entretanto, o grande volume dessas informações faz com que as organizações e as pessoas tenham dificuldade para gerenciar adequadamente estas informações, principalmente as não-estruturadas. Durante muito tempo as técnicas de mineração de dados (Goldschmidt, R. e Passos, E., 2005) cresceram para elaborar soluções para as informações estruturadas da empresa. Seguindo esse mesmo caminho, a área de mineração de textos surge para minimizar o problema de tratar dados não-estruturados, ajudando a explorar conhecimento armazenado em meios textuais e assim gerar algum tipo de vantagem competitiva.

A tarefa de gerar inteligência a partir da análise das informações capturadas e documentadas em textos livres já é realizada atualmente e demanda cada vez mais tempo dos participantes envolvidos devido ao volume cada vez maior a ser tratado. É exatamente nesse ponto que a mineração de textos pode contribuir. A Figura 1 ilustra a forma como o autor entende esse processo e como pretende contribuir na eficiência do processo de análise e decisão.

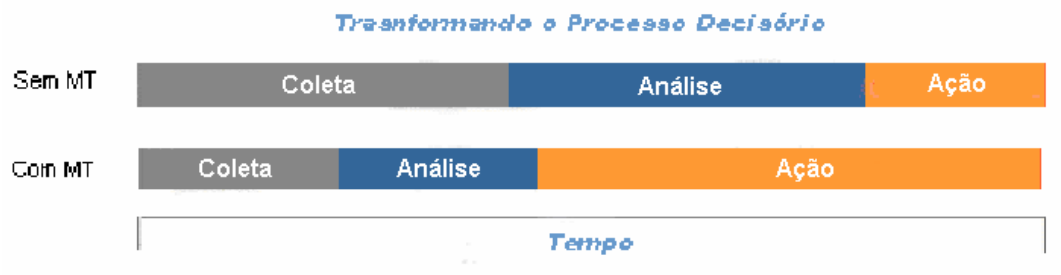

Figura 1 – Valor agregado pela Mineração de Textos (MT) no processo de análise de informações textuais.

Os processos de mineração de textos podem representar uma nova visão das informações disponíveis nas empresas. As aplicações são inúmeras, exemplos como o acompanhamento da gerência de projetos a partir de relatórios, documentação de projeto, comunicação com o cliente, o desenvolvimento do planejamento de marketing baseado em detalhes de planos passados, opções de anúncios e pesquisas de marketing. Aplicações não tão pretensiosas já se encontram implantadas atualmente, como a categorização automática de mensagens de correio eletrônico.

Quanto ao idioma desenvolvido (língua portuguesa), de acordo com projeto CLIC 2004 (CNPq pequenos grupos), o português é uma língua falada por uma parcela significativa da população mundial (aproximadamente 3%, em 1999, de acordo com The Ethnologue, em www.ethnologue.com), sendo o sexto idioma mais usado no mundo.

Na Internet, o Brasil ocupa a oitava posição em número de *hosts*, segundo os dados de 2004 do Comitê Gestor da Internet no Brasil[1] da Tabela 1

Tabela 1 – Ranking de quantidade de Hosts por país

1º	Estados Unidos*	162.195.368
2º	Japão (.jp)	12.962.065
3º	Itália (.it)	5.469.578
4º	Reino Unido (.uk)	3.715.752
5º	Alemanha (.de)	3.421.455
6º	Holanda (.nl)	3.419.182
7º	Canadá (.ca)	3.210.081
8º	**Brasil (.br)**	**3.163.349**

[1] www.cg.org.br/indicadores/index.htm

Ainda de acordo com a tabela fornecida pelo comitê, o português brasileiro está entre as seis línguas mais publicadas na Internet. No entanto, observou-se que a disponibilidade de ferramentas automáticas de processamento de textos em português, em termos de recuperação de informação textual, não atende às necessidades decorrentes desta participação significativa. Os recursos computacionais desenvolvidos para outras línguas, particularmente para o inglês, vêm sendo adaptados para o português, sem que as peculiaridades de nossa língua sejam levadas em consideração.

Em (Inoki, S., 1992) é enfatizada a necessidade de uma especialização a língua. Mostra como a língua portuguesa do Brasil é rica em vocábulos e flexibilidade gramatical, ressaltando as dificuldades do idioma português. As dificuldades surgem desde a diversidade de verbos, formas verbais, problemas de concordância, regências verbais, sem contar ainda com as de flexões de verbos irregulares.

Além da Internet, um grande grupo de instituições nacionais lida prioritariamente com grandes massas de conhecimento documental tais como legislações, notícias e patentes, além de relatórios e outros tipos de documentos produzidos internamente. Este panorama justifica as pesquisas em processamento de conteúdos digitais em português.

1.2. Objetivo da Tese

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 textual. Assim, para que esses dados textuais possam ser submetidos a algoritmos de mineração, é necessário um tratamento diferenciado na etapa de pré-processamento de dados.

O objetivo dessa tese de doutorado é, portanto, inovar na fase de pré-processamento da mineração de textos, propondo um modelo automático de enriquecimento dos dados para uma análise mais eficiente. O modelo é uma extensão da tradicional abordagem por conjunto de palavras (*bag-of-words*, em inglês), um dos procedimentos mais usados atualmente em mineração de textos (Bekkerman, R. et al, 2003). Foi desenvolvido e implementado um modelo computacional automático de pré-processamento valorizando o conteúdo do texto,

isto é, transformamos e o modelo baseado em palavras em um modelo baseado em lexemas, *bag-of-words* em *bag-of-lexems*. O conteúdo de um texto, no entanto, é dependente da língua em que está escrito, sendo assim, o modelo de pré-processamento ora proposto utiliza conhecimentos das áreas de PLN e Lingüística Computacional para formatar soluções com mais ênfase no conteúdo.

Apostou-se nesse caminho, embora alguns resultados experimentais mostrem que representações mais sofisticadas, às vezes, perdem em desempenho com relação à representação *words* usando palavras simples (Apté, C. et al, 1994); (Dumais, S. et al, 1998); (Lewis, D. D., 1992).

De acordo com (Lewis, D. D., 1992), a razão mais provável para explicar esses resultados é que, embora o uso de representações mais sofisticadas tenham qualidade semântica superior, a qualidade estatística é inferior em relação a representações baseadas em palavras simples. Assim, de acordo com (Joachims, T., 2002), a abordagem *bag-of-words* é uma boa relação entre expressividade e complexidade do modelo. Enquanto representações mais expressivas capturam melhor o significado do documento, sua complexidade é maior e degrada a qualidade de modelos estatísticos.

A Figura 2 apresenta um modelo do processo de mineração de textos tal como é proposto nessa tese. A figura apresenta o processo do início ao fim, passando por todas as etapas. A seqüência como é mostrado na figura é uma tendência encontrada nos recentes trabalhos da literatura como (Mathiak, B. e Eckstein, S., 2004) (Batista, G. E. A. P. A., 2003); (Ferneda, E. e Smit, J., 2003); (Kao, A. e Poteet, S., 2004). Essa figura será usada como modelo didático para a descrição tecnológica, passando as etapas uma a uma nessa tese.

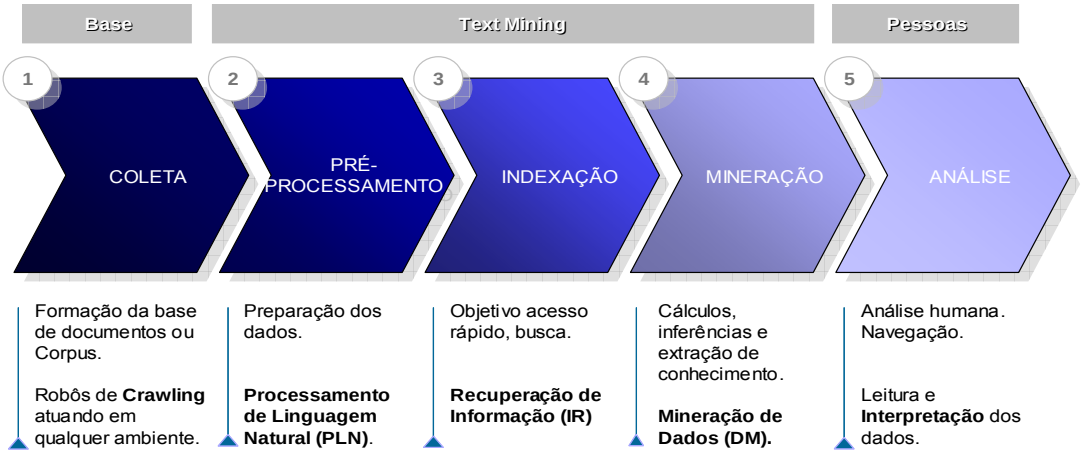

Figura 2 – Diagrama de camadas abstratas de um sistema computacional

Seguindo a Figura 2, o objetivo dessa tese é então concatenar um novo 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.

Serão analisados também os impactos dessa construção, apontando o que muda nas outras etapas para atender a uma boa solução de pré-processamento. Será mostrado nessa tese como a mineração de textos pode concatenar as diversas áreas do conhecimento em um processo multidisciplinar.

Apesar desse passeio multidisciplinar, o objetivo principal da mineração de textos é a resolução de problemas. Não pertence a esse trabalho explorar a fundo as áreas de Recuperação de Informação e Lingüística Computacional. Os conhecimentos dessas áreas são utilizados na medida em que oferecem idéias interessantes para a abordagem dos problemas. Esta perspectiva é compatível, por exemplo, com o pensamento de (Santos, D., 2001):

> "(...) é ao tentar resolver um dado problema (isto é, ao tentar construir um programa que manipula a língua) que surge o momento de nos debruçarmos quer sobre algumas características do léxico ou da gramática, quer sobre as teorias que pretendam dar respostas a esse problema" (:229)

1.3. Estrutura da Tese

Capítulo 2 – São apresentadas diversas abordagens de trabalhos recentes da área de Text Mining mostrando a origem das soluções e sua correlação com a área de Data Mining.

Capítulo 3 – É descrita a teoria de mineração de textos, assimcomo a teoria envolvida em cada uma das etapas do diagrama de mineração da **figura 2**.

Capítulo 4 – É exposto teoricamente o modelo de pré-processamento, núcleo dessa tese. Nesse mesmo capítulo é feita uma revisão de alguns pontos da área do Processamento da Linguagem Natural.

Capítulo 5 – Nesse capítulo é especificado o modelo do sistema desenvolvido apresentando e aprofundando um pouco mais cada um dos seus módulos que viabilizam os resultados.

Capítulo 6 – Os resultados do sistema de pré-processamento avaliado segundo medidas estatísticas de precisão (*precision*), abrangência (*recall*) e medida-F.

Capítulo 7 – Três exemplos de aplicação do trabalho dessa tese. Será mostrado como utilizar TM para classificar documentos automaticamente, extrair informação e responder perguntas, assim como a importância dessa tese para a área emergente de Web Semântica. Na apresentação da solução de classificação será mostrada a diferença dos resultados com e sem o pré-processamento proposto.

Capítulo 8 – Resumo dos pontos principais dessa tese, as contribuições à pesquisa acadêmica, aponta os desafios que ainda permanecem e sugestões de trabalhos futuros.