2
Mineração de Textos: Fundamentos e Aplicações

O principal objetivo deste capítulo é fornecer ao leitor uma visão global do surgimento e composição de procedimentos de investigação textual, com a adição de relatos sobre aplicações de sucesso.

As principais áreas de conhecimento que compõem e contribuem com a Mineração de Textos (MT) são: Aprendizado de Máquina, Processamento de Linguagem Natural (PLN), Estatística, Inteligência Computacional (IC), Recuperação de Informação (RI), Ciência Cognitiva, Mineração de Dados e Mineração na *Web* (do inglês, *Web Mining*).

Para o processamento da mineração, podemos ter duas abordagens diferentes: a **Análise Semântica**, na qual o foco é na funcionalidade dos termos dos textos e a **Análise Estatística**, que se preocupa com a freqüência de aparição de cada termo.

A primeira apóia-se no tratamento de textos conforme o ser humano faz, através do significado das palavras, de conhecimentos morfológicos, sintáticos, semânticos, pragmáticos e do contexto em geral.

Já na Análise Estatística, os termos são valorados, basicamente, pela sua freqüência de aparição na massa de dados, não importando a contextualização deste, como em que parágrafo está inserindo, que termos o antecedem ou que estão diretamente relacionados. Ambas as abordagens podem ser utilizadas sozinhas ou em conjunto. A Tabela 1 resume que áreas de conhecimento estão mais ligadas com os dois tipos de análise. Na próxima seção, as áreas de conhecimento são explicadas de forma sucinta.

Tabela 1 – As duas abordagens para a Análise de Textos e suas principais Áreas de Conhecimento

Análise Semântica	Análise Estatística
Ciência Cognitiva	Recuperação de Informação
Processamento de Linguagem Natural	Estatística
Mineração de Dados	Aprendizado de Máquina
Web Mining	Inteligência Computacional
	Mineração de Dados
	Web Mining

2.1. Áreas de Conhecimento em Mineração de Textos

2.1.1. Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) [6][7] é o conjunto de métodos formais para analisar textos e gerar frases escritas em um idioma humano. Muitos pesquisadores em todo mundo estão voltados para esta área, que está dividida em dois tipos de abordagens: **a baseada em texto** e a **baseada em diálogo**.

A abordagem baseada em diálogo ganhou impulso com os crescentes avanços na interface entre as máquinas e os seres humanos, pois estes estão cada vez mais sofisticados e caminhando aos poucos em direção às formas mais humanas de comunicação.

A abordagem baseada em texto trata de assuntos como: **busca de documentos** e **resumo e compreensão de textos**. Esta abordagem é a que tem sido objeto de interesse nas pesquisas relacionadas à Mineração de Textos, principalmente no campo da Lingüística Computacional (LC) [8] [9].

A teoria da LC é o ramo que trata dos aspectos computacionais da linguagem. Para tanto, são utilizados diversos algoritmos e estruturas de dados para examinar os seguintes tópicos: identificação de estruturas das frases, modelagem do conhecimento e raciocínio, e como usar a linguagem para realizar determinadas tarefas.

O Processamento de Linguagem Natural tem papel fundamental na Mineração de Textos, sendo utilizado no estágio inicial da etapa de **Pré-**

processamento, aonde sua principal função é fornecer um primeiro nível de estruturação da informação textual, como o reconhecimento de início e fim de sentenças e classificação de palavras quanto a sua função sintática. Entretanto, é necessário salientar que muitos dos problemas aonde são aplicadas técnicas de PLN não fazem parte de MT, como traduções automáticas de texto e corretores ortográficos.

2.1.2. Ciência Cognitiva

A Ciência Cognitiva é normalmente definida como o estudo científico da mente ou da inteligência. Quase toda a introdução à ciência cognitiva frisa a sua alta interdisciplinaridade; é normalmente caracterizada como tomando parte ou colaborando com as disciplinas de psicologia (especialmente através da psicologia cognitiva, lingüística, neurociência, inteligência artificial - em particular no ramo de redes neurais - e filosofia, especialmente a filosofia da mente e a filosofia da matemática, mas com aplicações na filosofia da ciência) [10] [11].

2.1.3. Recuperação de Informação

Recuperação de Informação (RI) é uma área da computação que lida com o armazenamento e recuperação automática de documentos, que são objetos de dados, geralmente textos.

RI é bastante relacionada com Mineração de Textos, principalmente na etapa de **Indexação** (seção 3.3), aonde são montadas estruturas de dados que permitem rápido acesso a documentos através de palavras-chave, reduzindo drasticamente o seu tempo de acesso.

Máquinas de Busca são os sistemas de RI mais conhecidos, como o **Google**, aonde são identificados quais documentos na *World Wide Web* são mais relevantes para um conjunto de palavras relacionadas pelos usuários.

Mineração de Textos tem como característica a execução de algoritmos de grande esforço computacional que são aplicados em grandes coleções de documentos. Sistemas de Recuperação de Informação podem acelerar esta etapa através da redução do número de documentos que são analisados. Por exemplo, se

o usuário tiver interesse em “minerar” apenas documentos que possuam as palavras “Pelé” e “futebol”, é possível restringir a coleção para que apenas os documentos que contenham ambas ou pelo menos uma das palavras-chave relacionadas sejam utilizados. Um sistema clássico de Recuperação de Informação pode ser estruturado conforme ilustrado na Figura 1.

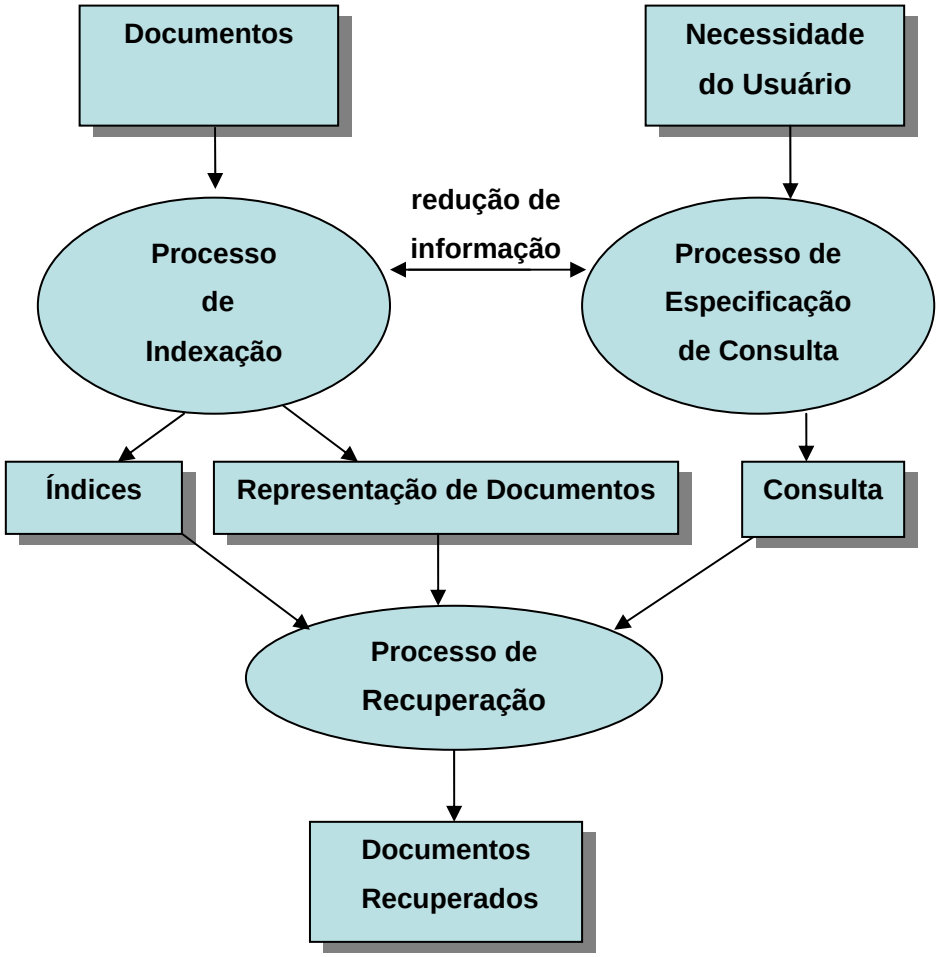

Figura 1 – Componentes de um sistema de Recuperação de Informação

O **Processo de Indexação** envolve a criação de estruturas de dados associadas à parte textual dos documentos, por exemplo, as estruturas de listas invertidas (seção 3.3.3).

O **Processo de Especificação da Consulta** geralmente é uma tarefa difícil. Há freqüentemente uma distância semântica entre a real necessidade do usuário e o que ele expressa na consulta formulada. Esta distância é gerada pelo limitado conhecimento do usuário sobre o universo de pesquisa e pelo formalismo da linguagem de consulta [12].

O **Processo de Recuperação** consiste na geração de uma lista de documentos recuperados para responder à consulta formulada pelo usuário. Os índices construídos para uma coleção de documentos são usados para acelerar esta tarefa. Além disso, a lista de documentos recuperados é classificada em ordem decrescente de um grau de similaridade entre o documento e a consulta.

2.1.4. Estatística

A estatística é uma ciência que utiliza teorias probabilísticas para a explicação de eventos, estudos e experimentos. Tem por objetivo obter, organizar e analisar dados, determinar as correlações que apresentem, tirando delas suas conseqüências para descrição e explicação do que passou e previsão e organização do futuro [13].

Em Mineração de Textos utiliza-se estatística em vários momentos, como na aplicação de modelos probabilísticos para, por exemplo, classificar um texto de acordo com seu assunto. Dentre os classificadores probabilísticos de maior sucesso está o ***Naive Bayes*** (seção 4.1.3)**,** apresentado nesta dissertação.

2.1.5. Aprendizado de Máquina

Aprendizado de Máquina é uma subárea da Inteligência Artificial concentrada em desenvolver modelos que possam ”aprender” através da experiência. O aprendizado se dá através de algoritmos dedutivos que baseados em estatística, extraem regras e padrões em grandes massas de dados.

ML (do inglês, *Machine Learning*) tem sido bastante utilizado em tarefas de pré-processamento de textos como etiquetagem morfossintática [14] e no processo de classificação automática de textos [15].

Dentre as técnicas de aprendizado de máquina mais utilizadas podemos citar as **Cadeias de Markov Escondidas** (do inglês, *Hidden Markov Models* ou simplesmente HMM), as **Máquinas de Vetor de Suporte** (do inglês, *Support Vector Machines* ou apenas SVM) e o já citado ***Naive Bayes***.

2.1.6.
Inteligência Computacional

A Inteligência Computacional (IC) busca, através de técnicas inspiradas na natureza, o desenvolvimento de sistemas inteligentes que imitem aspectos do comportamento humano, tais como: aprendizado, percepção, raciocínio, evolução e adaptação [16].

É também chamada de *Soft Computing* [17], distinguindo da computação convencional (*Hard Computing*) que é baseada na lógica binária. O termo "*soft*" vem da característica que permite a construção de sistemas tolerantes à imprecisão, incerteza e verdades parciais.

A IC Engloba algumas outras áreas como: Redes Neurais Artificiais, Lógica Nebulosa, Computação Evolucionária e Inteligência Coletiva. Em Mineração de Textos, podemos citar alguns exemplos de aplicação desta classe de soluções, como os modelos de ***Hopfield*** [18] e ***Backpropagation*** [19], ambos exemplos de Redes Neurais; e o uso de **Algoritmos Genéticos** [17] como exemplificação do uso da **Computação Evolucionária**.

2.1.7.
Mineração de Dados

Mineração de Dados é o processo de descoberta de conhecimento em grandes volumes de dados. Tem como principal objetivo revelar padrões, descobrir fatos e associações não percebidas e prover métodos que apóiem a tomada de decisão. É um processo composto por diversas etapas e que faz uso de dados estruturados, normalmente provenientes de dados operacionais de grandes corporações, como registros de vendas de produtos, cadastro de clientes, registro de pedidos a fornecedores, etc.

A área de Mineração de Dados contribuiu bastante para a de Textos, em especial no que se refere à separação e estruturação do processo em etapas, assim como dos algoritmos e soluções utilizadas, principalmente na etapa de **Mineração**, que será apresentada em maiores detalhes no capítulo que segue. Para uma referência completa e prática acerca de Mineração de Dados vide [1].

2.1.8. *Web Mining*

A Internet, sem nenhuma dúvida, revolucionou o conceito de Sistemas de Informação. Devido à sua grande magnitude, a *World Wide Web* pode ser considerada o maior repositório de dados já observado. Entretanto, como citado no terceiro parágrafo do primeiro capítulo, grande parte dos dados na Internet – cerca de 80% - está na forma não-estruturada, ou seja, textos, imagens, gráficos, vídeos ou sons. A Internet também possui características únicas em sua estrutura, como a formação natural de clusters de páginas HTML através de ligação entre *hiperlinks*. A própria estrutura dos documentos HTML, que possuem *tags* de marcação indicando estruturas do texto, é algo bastante interessante para a extração de padrões, conhecimento e até mesmo comportamento, através da observação da navegação de usuários entre sites e portais.

Web Mining agrupa em três diferentes abordagens um conjunto de ferramentas importantes que além de descobrir as fontes de informação relevantes, pretende mapear e analisar o padrão de acesso e armazenamento de informações na web. As três abordagens, mostradas na Figura 2, são: *mining* no conteúdo da web, *mining* na estrutura da web e *mining* do uso na web [20].

***Web Mining* de conteúdo** consiste em analisar textos, imagens e outros componentes presentes nos documentos HTML. Esta técnica é essencialmente utilizada como forma de facilitar o acesso ao conteúdo predominantemente desestruturado encontrado nestes tipos de documento.

***Web Mining* de estrutura** estuda o relacionamento entre as páginas da web através de seus *hiperlinks*, com o objetivo de identificar páginas pertinentes a uma determinada área de conhecimento.

***Web Mining* de uso** é a descoberta de conhecimento através do registro de visitação e de busca de usuários entre os diferentes sites na Internet. Desta forma, é possível identificar padrões de acesso, requisito essencial para, por exemplo, implementar o processo de personalização de uso que permite a utilização de um contexto próprio na busca de documentos na Internet, gerando resultados também personalizados.

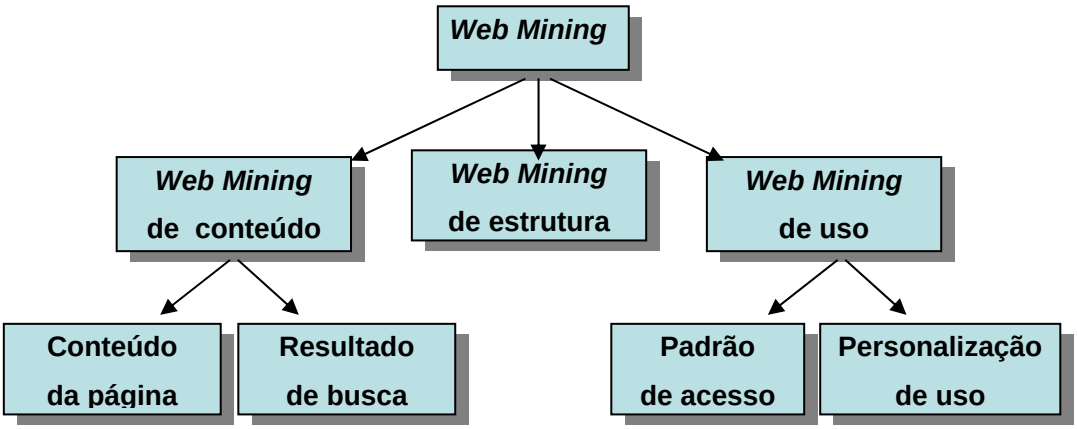

Figura 2 – Abordagens de *Web Mining*

2.2. Aplicações

Inúmeras são as aplicações da Mineração de Textos em diversas áreas e segmentos da sociedade. O objetivo desta seção é identificar aonde e como utilizar o potencial da análise de dados não-estruturados, bem como relatar alguns casos reais já implementados em áreas como Medicina e Negócios.

2.2.1. Negócios

2.2.1.1. Análise de Sentimento em Pesquisas de Opinião

Muitas organizações utilizam sistemas de pesquisa de opinião para que possam melhor avaliar a satisfação de seus clientes quanto aos produtos e serviços oferecidos. Estes sistemas podem ser utilizados como poderosas ferramentas de apoio à tomada de decisão, indicando, por exemplo, que determinado produto ou serviço está enfrentando uma péssima aceitação em um segmento de mercado.

A utilização de questionários aonde a resposta do entrevistado é do tipo pré-formatada, isto é, cada pergunta só pode ser respondida com uma e apenas uma opção dentre todas oferecidas, podem não expressar de forma completa uma opinião, receber informações imprecisas ou até mesmo ignorar uma sugestão valiosa.

Em contrapartida, questionários que permitem respostas em linguagem natural admitem maior nível de detalhe, capturando de forma mais precisa o real

sentimento do entrevistado em face ao que lhe foi indagado. Entretanto, em pesquisas aonde o número de entrevistados é grande, essa modalidade por muitas vezes é inviabilizada, devido à dificuldade de processamento, necessidade de tratamento individualizado e a demanda por grande esforço humano.

A área de Mineração de Textos que tem como objetivo classificar automaticamente opiniões quanto à satisfação de quem a escreveu é conhecida como *Sentiment Analysis* [21]. O principal desafio é identificar como sentimentos são expressos em textos e se tais sentimentos indicam uma opinião positiva (favorável) ou negativa (desfavorável) com relação a um tópico. Para maiores referencias de trabalhos já realizados consultar [22] [23] .

2.2.1.2. Inteligência Competitiva

Inteligência Competitiva é o processo de monitorar os "rastros" deixados por concorrentes ou outros agentes que integram determinado setor do mercado, a fim de se obter informações que possam agregar valor ao planejamento tático e estratégico de uma empresa.

Um exemplo de aplicação com essa finalidade é o monitoramento de páginas na *Web* realizado com a tecnologia de *crawlers* (seção 3.1) – robôs que visitam páginas na Internet e coletam textos e documentos. A Tabela 2 exibe os principais focos de monitoramento e seus respectivos propósitos.

Tabela 2 – Principais focos no monitoramento de páginas na *web*

Foco de Monitoramento	Propósito
Concorrentes	Identificar e realizar antecipação frente a movimentação dos concorrentes
Novos Negócios	Identificar novas oportunidades
Produtos	Melhorar o processo de desenvolvimento de produtos
Tecnologia	Identificar o desenvolvimento de novas tecnologias impactantes ao negócio
Macroeconomia	Antecipar mudanças de impacto
Fusões e Aquisições	Identificar compras e vendas de concorrentes
Vendas	Melhorar o processo de vendas na organização
Leis de Regulamentação	Monitorar e Antecipar mudanças que ocorram em normas e regulamentações

2.2.1.3. Suporte e Atendimento ao Usuário

Em geral, quanto maior a organização, maior a dificuldade de se transmitir informação a quem realmente precisa receber. Isso pode ser percebido, por exemplo, em empresas que implementam sistemas de solicitação de manutenção, aonde funcionários registram comportamentos anormais em *softwares*, *hardware*s e equipamentos. É esperado que esta requisição chegue até o especialista no funcionamento do objeto defeituoso e providencie o conserto o mais breve possível.

Entretanto, o caminho percorrido pela requisição pode ser demorado, necessitando que alguns outros funcionários precisem ler o conteúdo da informação e analisar para quem ou para qual setor direcioná-la. Esta triagem pode passar por vários níveis até que finalmente chegue à caixa de e-mail do especialista no assunto.

Através da análise automática do texto, é possível programar uma aplicação que classifique o conteúdo da requisição e a encaminhe diretamente para o destinatário final [24]. Por ser realizado de forma automática, evita-se a necessidade de intervenção humana, aumentando de forma significativa a qualidade de entrega.

Neste mesmo contexto, outras soluções podem ser implementadas, como a identificação de mais de um pedido de manutenção para o mesmo problema e a busca por soluções de casos passados. Basta que seja consultada a base histórica de atendimentos e que se recupere os documentos de maior relevância, tal como funciona o sistema de buscas **Google**.

2.2.1.4. Análise e Extração de Informações em Contratos

O gerenciamento de contratos em organizações é uma tarefa que pode ser tediosa e que pode custar caro caso haja algum equívoco. É necessário que seja observado prazos de validade, valores envolvidos, entidades citadas etc.

A implementação de um sistema que agrupa os contratos por setor ou assunto, melhora o gerenciamento e evita a redundância, quando da existência de contratos que se sobrepõe.

É possível também a aplicação de técnicas de **Extração de Informação** (com o objetivo de extrair do texto blocos com informações específicas, como datas e entidades) e montar uma base estruturada. Por exemplo, o seguinte *template* poderia ser gerado a partir de uma coleção de contratos:

- Início da vigência;
- Fim da vigência;
- Empresa contratante/contratada;
- Valores acordados.

2.2.2. Direito

O armazenamento e recuperação de textos jurídicos é um grande desafio devido à grande quantidade de documentos que diariamente são gerados. Exemplos de documentos dessa categoria são: jurisprudências, leis, decretos, normas, medidas provisórias, constituição e ofícios. O público interessado é constituído de advogados, juízes, professores e estudantes de Direito, integrantes do Ministério Público, empresários e profissionais liberais.

Muitas são as aplicações de Mineração de Textos na área jurídica, como a aquisição de conhecimento jurisprudencial [25]. Diversas outras ferramentas podem sem implementadas, como a geração automática de resumos a partir de depoimentos e de relatórios policiais, ambos geralmente escritos em texto livre.

2.2.3. Medicina

Participantes da Medicina, em geral, produzem e mantêm grandes volumes de informação, normalmente provenientes de registros de hospitais, prontuários, fichas de pacientes, pesquisas com doentes, etc.

Sistemas de Mineração de Texto podem auxiliar médicos a diagnosticar doenças, tendo com base relatos de sintomas de pacientes, normalmente adquiridos a partir de prontuários e fichas de acompanhamento [26][27][28].

Outra aplicação de bastante interesse de cientistas e pesquisadores é a mineração de documentos médicos científicos gerados a partir de pesquisas, aonde

o elevado número de termos e combinações não são tratadas facilmente pelos tradicionais sistemas de Recuperação de Informação. Existem ferramentas especializadas em capturar informação a partir de bases de dados públicas, como a base *Medline*[1].

[1]Disponível no endereço: http://medline.cos.com