2
Mineração de Textos: Fundamentos

O principal objetivo deste capítulo é fornecer uma visão do surgimento, estruturação e evolução dos procedimentos utilizados no processo de Descoberta de Conhecimento em Textos.

2.1.
Definição

Mineração de Textos (MT), Mineração de Dados Textuais ou Descoberta de Conhecimento em Textos surge, neste contexto, como uma abordagem ao processamento de grandes bases de dados textuais com o objetivo de extrair informação relevante e obter conhecimento implícito e útil a partir destas. Conhecimento útil é aquele que pode ser aplicado de forma a apoiar um processo de tomada de decisão, ou seja, é aquele que pode ser aplicado de forma a proporcionar benefícios.

De acordo com (ARANHA C. N., 2007), muitas são as definições encontradas na literatura:

- Pode-se então definir Descoberta de Conhecimento em Textos ou Mineração de Textos como sendo o processo de extrair padrões interessantes e não triviais, a partir de documentos textuais (TAN A.-H. , 1999).
- Mineração de Textos é a descoberta, através de meios computacionais, de informações desconhecidas ou novas, através da utilização de ferramentas de extração automática de informação, a partir de documentos de textos não estruturados (HEARST, 1999).
- Mineração de Textos é o estudo sobre a extração de informação de textos usando os princípios da linguística computacional (SULLIVAN, 2000).

Apesar de muitas definições, é fácil concluir que Mineração de Textos, de maneira análoga a Mineração de Dados, busca extrair informação útil de bases de dados através da identificação e exploração de padrões interessantes. Porém, é importante ressaltar a principal diferença entre a Mineração de Dados e a de Textos. Mineração de Textos é um processo de obtenção de conhecimento oriundo a partir de bases de dados textuais, ou seja, documentos em linguagem natural, e que, portanto, possuem pouca ou nenhuma estrutura de dados. Em Mineração de Dados, a obtenção de conhecimento ocorre em bases de dados fortemente estruturadas, geralmente armazenadas em Sistemas Gerenciadores de Bancos de Dados (SGBD).

Dados mantidos em Sistemas Gerenciadores de Bancos de Dados apresentam uma estrutura de representação ou esquema previamente definido. Um SGBD é uma coleção de programas que permitem ao usuário definir, construir e manipular bases de dados para as mais diversas finalidades (DATE, 2005). O principal objetivo de um SGBD é retirar da aplicação cliente a responsabilidade de gerenciar o acesso, manipulação e organização dos dados. Para isto, todo SGBD disponibiliza uma interface para que os seus clientes possam incluir, alterar ou consultar dados. O acesso e a manipulação destes esquemas são tarefas específicas do SGBD. Usuários ou aplicações realizam operações sobre estes dados com base neste esquema.

Muitas são as funções de um Sistema Gerenciador de Bancos de Dados, porém, uma delas provê integridade às bases de dados gerenciadas por este: integridade semântica. Integridade semântica é função de um SGBD responsável por garantir a armazenagem correta de dados em relação ao domínio (DATE, 2005). Por exemplo, na coluna de uma tabela destinada a armazenar informações sobre o saldo bancário de um cliente, somente valores numéricos podem ser inseridos. A Figura 2 demonstra o conceito de integridade semântica em um SGBD.

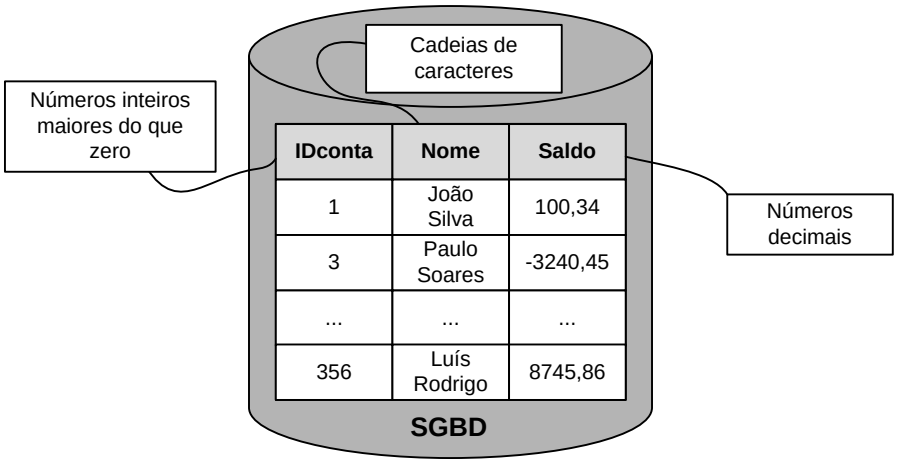

Figura 2 – Integridade semântica de um SGBD

Desta forma, os tipos de dados a serem retornados por um SGBD são conhecidos. Na coluna idade de uma pessoa física, sabe-se que somente valores inteiros e maiores do que zero serão retornados. Na coluna data de nascimento, sabe-se que somente dados no formato de data serão retornados. Portanto, não há o aspecto da imprevisibilidade de informação presente em dados textuais.

Já documentos em linguagem natural não possuem garantia alguma de integridade. Além disso, documentos deste tipo, ainda que dentro de um mesmo contexto, podem apresentar enormes diferenças estruturais entre si. Considerando dados referentes a um *curriculum vitae*, por exemplo, pode-se ter um pequeno texto informal descrevendo dados pessoais e experiência profissional ou, no extremo oposto, um documento organizado em seções e subseções, com *links* para dados das empresas e instituições onde a pessoa trabalhou.

A justificativa para tal fato decorre da própria natureza destes dados. Dados *Web*, por exemplo, apresentam uma organização bastante heterogênea (MARKOV & LAROSE, 2007), que pode variar de um texto sem nenhuma formatação até um conjunto de registros bem formatados. Em geral, dados textuais referem-se a massas de informação, quase sempre digitalizadas, e que não possuem rígida estrutura de dados. Exemplos deste tipo de dados são áudio, vídeo e texto livre, como por exemplo, o corpo de um e-mail.

Como Mineração de Dados lida com dados que já estão em formato estruturado, a maior parte das suas rotinas de pré-processamento concentra-se em duas tarefas: limpeza e integração de dados (ZHU & DAVIDSON, 2007). Sistemas de Mineração de Textos não submetem aos seus algoritmos de

descoberta de conhecimento coleções de textos despreparadas (GOMES, 2008). Em Mineração de Textos, operações de pré-processamento possuem como objetivo a identificação e a extração de características representativas de documentos para que estes possam ser representados adequadamente de maneira estruturada.

Contudo, embora a principal diferença entre Mineração de Dados e Mineração de Textos ocorra na apresentação da informação de trabalho, não significa que estes dois processos sejam completamente distintos. Ambos os processos são baseados em exemplos coletados em uma imensa base de dados e utilizam técnicas de **Aprendizado de Máquina**[6] semelhantes. Além disso, em grande parte dos casos o processo de Mineração de Textos ocorre por uma simples transformação de textos em dados estruturados, nos quais as técnicas já conhecidas de Mineração de Dados podem ser aplicadas sem qualquer restrição, não sendo necessário o entendimento de características específicas da língua em que se aplica o processo de Mineração de Textos.

Entretanto, em virtude da grande centralidade da linguagem natural nos processos de Mineração de Textos, a utilização da rica informação semântica presente em qualquer linguagem pode ser utilizada em proveito do processo de obtenção de conhecimento a partir de dados textuais. Para isto, fez-se necessário buscar avanços em áreas da Ciência que se relacionam com tratamento da linguagem como Ciência Cognitiva, Processamento de Linguagem Natural e Recuperação de Informação, dentre outras. Atualmente, muitas abordagens aos processos de Mineração de Textos tiram proveito da vasta informação linguística de suas coleções de documentos.

Independente da utilização, ou não, de dados linguísticos, da mesma forma que em Mineração de Dados, antes de ser iniciado um processo de Mineração de Textos, faz-se imprescindível aquisição de conhecimento básico sobre o domínio da aplicação no qual será realizada a mineração. Entender o domínio dos dados é naturalmente um pré-requisito para a obtenção de conhecimento útil (REZENDE, 2005).

Além disso, todo processo de Mineração de Textos visa um objetivo, ou seja, a realização de uma tarefa, e que precisa ser definido antes do início da

[6] Do termo em inglês, *Machine Learning –ML*.

mineração, pois, de acordo com este, todo o processo de Mineração de Textos será orientado. Em (REZENDE, 2005), algumas questões importantes sobre um processo de Mineração de Dados são sugeridas. Estas mesmas questões, citadas abaixo, também são aplicáveis ao processo de Mineração de Textos:

- *"Quais são os objetivos do processo?";*
- *"Quais critérios de desempenho são importantes?";*
- *"O conhecimento extraído deve ser compreensível a seres humanos ou um modelo tipo caixa-preta é apropriado?"*

Em (CARRILHO, 2007), as principais tarefas de Mineração de Textos são abordadas minuciosamente.

2.2. Principais Elementos

O principal elemento de um processo de Mineração de Textos é a coleção de documentos, pois constitui o conjunto de dados sob o qual este processo é realizado. Um documento, elemento básico de uma coleção, é definido como uma unidade discreta de dados em formato textual, como por exemplo, uma página *web* ou um *e-mail* (KONCHADY, 2006). Em geral, Mineração de Textos lida com enormes coleções de documentos, e é esta característica que, principalmente, torna impossível, em tempo hábil, a análise desta imensa base de dados por humanos (FELDMAN & SANGER, 2007).

Em alguns cenários, essa coleção de documentos pode ser estática, ou seja, o conjunto de documentos selecionado permanece inalterado tanto em elementos, como em conteúdo. Entretanto, em grande parte dos casos, essa coleção de documentos é dinâmica, podendo ter elementos incluídos, excluídos e até mesmo alterados, o que demanda desafios ainda maiores para um sistema de Mineração de Textos. Um bom exemplo deste segundo caso é a própria *Web* brasileira. A Figura 3 ilustra os dados obtidos em um estudo recente sobre a Internet no Brasil em que se constata que quase metade dos *sites* hospedados possui conteúdo dinâmico (MODESTO, PEREIRA, ZIVIANI, CASTILLHO, & BAEZA-YATES, 2005):

Figura 3 – Sites brasileiros quanto à frequência de modificação do conteúdo

É importante também ressaltar aspectos sobre a organização de uma coleção. Usualmente, os itens de uma coleção são documentos do mundo real, e são arranjados de acordo com as características que melhor os representem. Muitos são os critérios que podem ser levados em consideração na organização de uma coleção, como, por exemplo, a separação por tipo de documento (*e-mail*, memorando, currículo), mas a predominância do assunto abordado em cada documento como critério de organização é notória. Ao preparar uma coleção de documentos, até mesmo a simultaneidade de um mesmo documento pertencer a diversas coleções é possível. Na Figura 4, em um sistema fictício de Mineração de Textos para área médica, há duas coleções de documentos organizadas pelo assunto. Um documento que retrate o diagnóstico de um paciente com suspeitas de uma doença X e que foi tratada com um medicamento Y pode pertencer, simultaneamente, à coleção de documentos referentes à doença X, bem como à coleção de textos referentes ao medicamento Y.

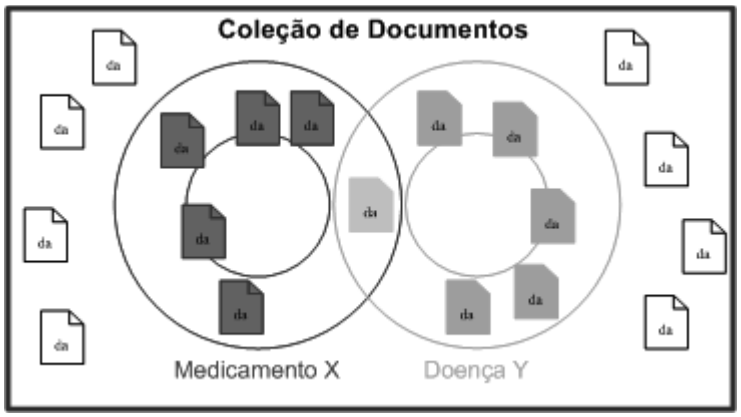

Figura 4 - Coleções de documentos com elementos em comum

2.3.
Documentos textuais são estruturados

Apesar de muitas vezes denominado não estruturado, um documento texto pode ser visto, sob muitas perspectivas, como um objeto estruturado. Especialmente, sob o enfoque da Linguística, mesmo um simples documento apresenta abundantes estruturas semânticas e sintáticas, ainda que estas estejam implícitas no texto. Elementos tipográficos, como pontuações, letras maiúsculas, números e outros caracteres especiais, ajudam a definir subcomponentes de um documento: parágrafos, títulos, datas, autores e outras informações. Até mesmo a sequencia das palavras pode definir características importantes de um documento. No exemplo da Figura 5, podemos visualizar importantes estruturas sintáticas presentes em um simples trecho de texto.

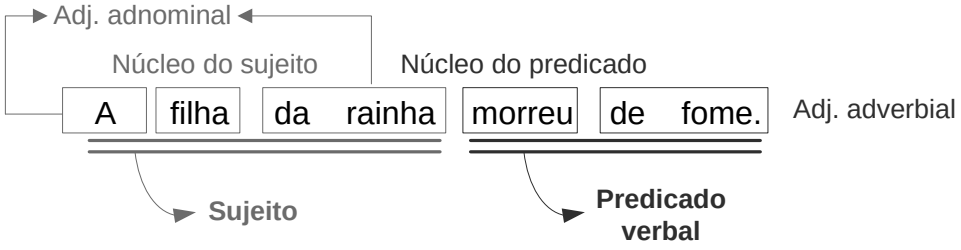

Figura 5 – Algumas estruturas sintáticas de um trecho de texto

Ainda que documentos textuais não possuam estruturas de dados definidas, é incorreto denominá-los não estruturados, pois, como visto, estes documentos possuem ricas estruturas sintáticas e semânticas. Atualmente, documentos textuais que apresentem poucos detalhes tipográficos são denominados fracamente estruturados.

Existem também documentos que fornecem mais informações sobre a sua estrutura do que aquela provida pelo texto. Documentos com extensivos e consistentes elementos de formatação (*tags*), estrutural ou visual, em que metadados podem ser facilmente inferidos, são denominados semiestruturados (SHOLOM, INDURKHYA, ZHANG, & DAMERAU, 2005) (FELDMAN & SANGER, 2007). Por exemplo, em uma mensagem de *e-mail*, informações sobre

remetente, destinatários e assuntos podem ser facilmente extraídas pela forma estrutural deste tipo de documento. Documentos em linguagem XML também podem ser considerados como semiestruturados (POWEL, 2007), quando apresentam diagramação estrutural que releva características adicionais sobre a sua estrutura.

Entretanto, não é o tipo de documento ou a linguagem de formatação do mesmo que o classifica em fracamente estruturado ou semiestruturado. Até mesmo documentos XML podem ser considerados como fracamente estruturados, pois, muitas vezes, não apresentam elementos que possam ajudar a inferir qualquer informação adicional sobre o texto que apresentam. Na Figura 6, há a representação de dois documentos na linguagem XML. Porém, somente um deles pode ser considerado semiestruturado, pois apresenta *tags* em suas definições que permitem a interpretação de informações adicionais sobre o conteúdo, o que não seria possível se o mesmo estivesse em formato livre.

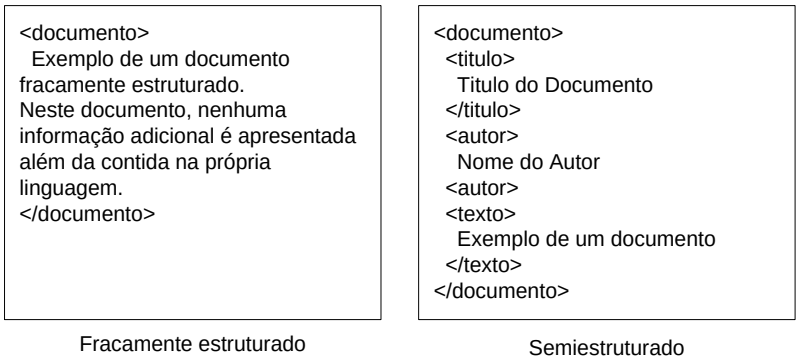

Figura 6 – Documentos em formatos fracamente estruturado e semiestruturado (respectivamente)

2.4. Características representativas de um documento

As operações de pré-processamento envolvidas em Mineração de Textos possuem como objetivo realçar os diferentes elementos presentes em um documento em linguagem natural para transformá-lo de uma representação estrutural irregular e implícita a uma representação explicitamente estruturada (FELDMAN & SANGER, 2007). Entretanto, dado o número potencialmente enorme de palavras, frases, sentenças, elementos tipográficos e *tags* de layout que,

até mesmo um simples e pequeno documento pode apresentar, uma tarefa essencial para qualquer processo de Mineração de Textos é a identificação do conjunto mais simples de características de um documento que pode representá-lo como um todo. Este conjunto de características recebe a denominação de modelo de representação, e cada documento em uma coleção é representado pelo conjunto de características que o seu modelo de representação possui. Os modelos de representação utilizados em Mineração de Textos serão explicados, com maiores detalhes, no capítulo 4.

Apesar de grandes esforços para desenvolver um modelo de representação eficiente, cada documento de uma coleção é, em geral, composto por um enorme número de características. Este enorme número de características afeta todo o processo de Mineração de Textos e influi principalmente na desempenho e design de um sistema de Mineração de Textos.

Problemas relacionados à alta dimensionalidade estão mais presentes em sistemas de Mineração de Textos do que em sistemas de Mineração de Dados. Isto está relacionado ao fato das inúmeras possibilidades existentes para a seleção de características de um documento com informações textuais (KONCHADY, 2006).

Outro fator relevante no aspecto de representação de um documento é a ausência de muitas características comuns a todos os documentos. Esta representação esparsa muitas vezes dificulta o descobrimento de padrões que são facilmente encontrados em tarefas de Mineração de Dados.

Ao selecionar quais características de cada documento serão utilizadas para construir o seu modelo de representação, há dois objetivos essenciais:

- O primeiro objetivo é determinar uma quantidade suficiente de características que permita representar os documentos sem que haja perda significativa de suas informações semânticas, o que, quase sempre, resulta em um grande número de características selecionadas.
- O segundo objetivo, por outro lado, é determinar o menor número possível de características para que os modelos de representação criados sejam computacionalmente eficientes.

Embora muitos critérios possam ser utilizados na seleção dos tipos de características que irão representar um documento, as três seguintes são as mais utilizadas:

- Caracteres: componentes individuais que são responsáveis pela formação de blocos com um nível semântico maior, como palavras e termos. Em geral, são utilizados junto com a posição em que ocorrem no texto. Abordagens que utilizam a combinação de um número predefinido de caracteres, como por exemplo, bigrama ou trigrama, são mais comuns do que a utilização de um único caractere. Embora a construção de um modelo de representação que utilize estas características seja considerada o mais próximo da realidade, a alta dimensionalidade, decorrente da escolha desta característica para representá-lo, torna impeditiva a utilização de diversas técnicas computacionais.
- Palavras: palavras retiradas de um documento constituem a menor unidade capaz de representar algum valor semântico. Por esta razão, é a característica mais utilizada para a construção de um modelo de representação de um documento. Entretanto, termos multipalavras como, por exemplo "casa da moeda", podem perder seus valores semânticos quando separados. Geralmente, para construir um modelo de representação de um documento baseado em palavras, seleciona-se somente aquelas que são mais representativas, eliminando-se *stopwords* (item "3.2.2"), caracteres simbólicos, dentre outros.
- Termos: termos podem ser compostos por uma única palavra ou por um conjunto de palavras que exprimem, por completo, a semântica que era desejada no texto. A extração de termos, na maioria das vezes, é auxiliada por um dicionário de palavras, o que permite identificar os termos que são compostos por mais de uma palavra.

No exemplo da Figura 7, as diferenças na utilização destes dois últimos critérios de seleção do tipo de características representativas de um documento podem ser visualizadas. No modelo baseado em palavras, houve perda semântica e um número maior de elementos. Porém, o modelo baseado em termos exigiu que o sistema tivesse conhecimento prévio dos termos que são compostos por mais de uma palavra.

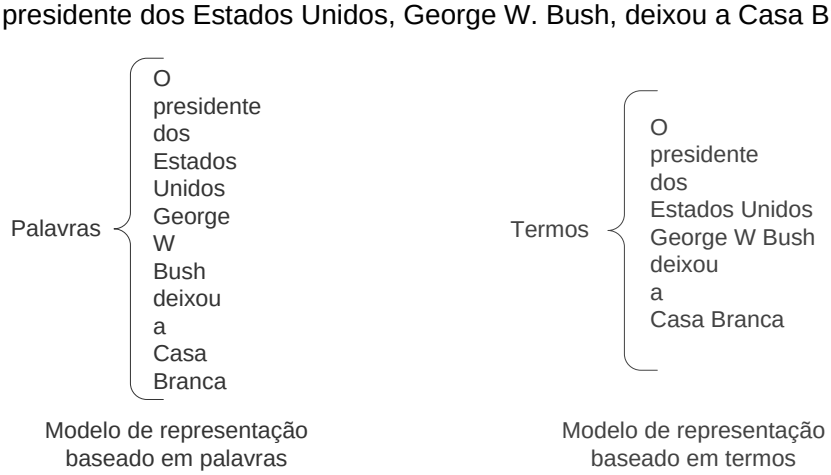

Figura 7 – Modelos de representação baseados em palavras e termos

2.5. Abordagens ao processo de Mineração de Textos

Bases de dados textuais fornecem dados semânticos e estatísticos em seu conteúdo. Diversas formas de abordagem aos dados textuais podem ser empregadas. As duas abordagens mais utilizadas são: a análise estatística, que é baseada na frequência de ocorrência dos termos nos textos, e a análise semântica, que é baseada na funcionalidade de cada termo no texto. Ambas as abordagens podem ser utilizadas sozinhas ou em conjunto.

2.5.1. Análise Estatística

Na Análise Estatística, a importância dos termos está diretamente ligada à frequência de ocorrência destes nos textos (SILVA A. A., 2007). Informações sobre contextualização, precedência ou sucessão de outros termos não são consideradas. Baseada no aprendizado estatístico, a principal vantagem desta abordagem é permitir a sua utilização em qualquer idioma.

2.5.2.
Análise Semântica

Na Análise Semântica, há a utilização da rica informação semântica, presente em qualquer linguagem, em proveito do processo de obtenção de conhecimento a partir de dados textuais. Mais do que considerar apenas aspectos estatísticos no tratamento de textos, a abordagem por Análise Semântica considera com grande centralidade a linguagem natural nos processos de Mineração de Textos.

Com o emprego de técnicas, estas baseadas no Processamento de Linguagem Natural, capazes de avaliar e identificar a funcionalidade correta de um determinado termo em uma sentença, é possível obter a verdadeira importância do mesmo em seu contexto, possibilitando aumento da qualidade dos resultados produzidos. Conforme (SILVA A. A., 2007), o emprego desse tipo de análise justifica-se pela melhoria em qualidade da Mineração de Textos quando incrementado de um processamento linguístico mais complexo.

A Tabela 3 resume as áreas de conhecimento mais envolvidas com os dois tipos de análise (CARRILHO, 2007). A Tabela 4 resume as principais características das duas abordagens. Na próxima seção, as áreas de conhecimento são explicadas de forma sucinta.

Tabela 3 - As duas abordagens para a Análise de Textos e suas principais Áreas de Conhecimento

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

Tabela 4 - As principais características de cada uma das abordagens para a Análise de Textos

Análise Estatística	**Análise Semântica**
Utilizável em qualquer idioma.	Necessita conhecimento específico do
Modelos com simples implementação e	idioma que será objeto de análise.

conhecidos na literatura. Descarta qualquer valor semântico presente nos textos.	Utiliza a informação semântica dos textos, tal como humanos.

2.6. Áreas correlatas a Mineração de Textos

Mineração de Textos é um campo multidisciplinar. Para o tratamento de textos e obtenção de conhecimento presente neles, fez-se necessário buscar e empregar avanços, técnicas e conceitos de diversas áreas como Ciência Cognitiva, Processamento de Linguagem Natural, Aprendizado de Máquina, Estatística, Recuperação de Informação e, principalmente, Mineração de Dados, da qual teve seu ponto de partida. A Figura 8 ilustra a demanda de ferramentas em outras áreas da Mineração de Textos.

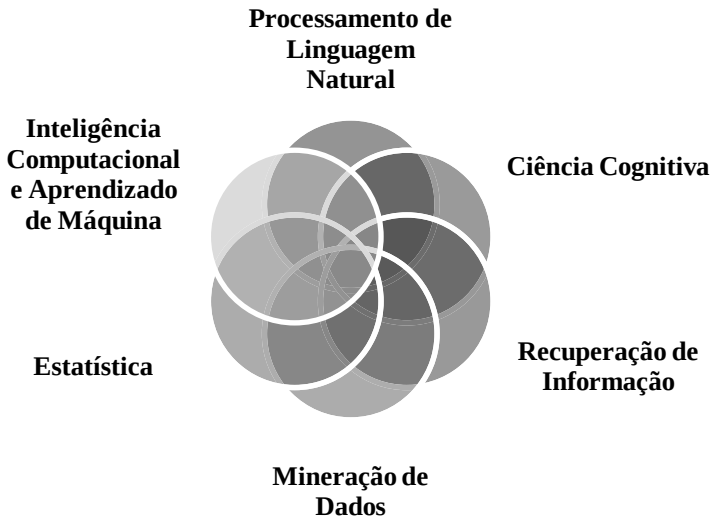

Figura 8 – Multidisciplinaridade da Mineração de Textos

2.6.1. Ciência Cognitiva

Ciência Cognitiva consiste em um conjunto de esforços interdisciplinares visando compreender, cientificamente, a mente e sua relação com o cérebro

humano. Entender a mente humana requer numerosos métodos e teorias, por isso, desta área fazem parte a Psicologia, a Filosofia, a Inteligência Artificial, a Neurociência e a Linguística (PINKER, 1998). As Neurociências colaboram na parte referente às estruturas cerebrais, a Psicologia, com as teorias de funcionamento da mente, a Filosofia, através da Lógica e da Epistemologia, a Linguística, com o exame da linguagem e a Inteligência Artificial, com os modelos de máquinas reais ou teóricas que poderiam simular o funcionamento do cérebro ou de suas partes.

O legado de contribuições da Ciência Cognitiva é enorme. Redes Neurais (RN) são exemplos de inteligência artificial conexionista, isto é, baseada na estrutura física e/ou biológica do cérebro; e que adquirem conhecimento através da experiência. Constituem um dos grandes exemplos dos avanços desta área.

Mineração de Textos busca, nesta área, principalmente, entender processo de formação da fala e da escrita.

2.6.2. Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é o conjunto de métodos formais utilizados para analisar textos e gerar frases escritas em um idioma humano (ARANHA C. N., 2007). Normalmente computadores estão aptos a compreender instruções escritas em linguagens de computação, que seguem uma forte estrutura sintática, mas possuem muita dificuldade em entender comandos escritos em uma linguagem humana. Isso se deve ao fato das linguagens de computação serem extremamente precisas, contendo regras fixas e estruturas lógicas bem definidas que permitem ao computador saber exatamente como proceder a cada comando. Já em um idioma humano uma simples frase normalmente contém ambiguidades, nuances e interpretações que dependem do contexto, do conhecimento do mundo, de regras gramaticais, culturais e de conceitos abstratos (INSITE, 2001).

O objetivo final do Processamento de Linguagem Natural é fornecer aos computadores a capacidade de entender e compor textos. Entender um texto significa reconhecer o contexto, fazer análise sintática, semântica, léxica e

morfológica, criar resumos, extrair informação, interpretar os sentidos e até aprender conceitos com os textos processados.

Em Mineração de Textos, técnicas de PLN são utilizadas principalmente na fase de pré-processamento. Tarefas como identificação de classes gramaticais de termos, reconhecimento de entidades e até mesmo redução da dimensionalidade de representação de documentos são auxiliadas por PLN.

2.6.3. Aprendizado de Máquina

Aprendizado de Máquina é uma área de Inteligência Artificial (IA) cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado, bem como a construção de sistemas capazes de adquirir conhecimento de forma automática (BISHOP, 2007).

Extrair conhecimento de bases de dados pode envolver, entre outras coisas, a utilização de algoritmos de Aprendizado de Máquina capazes de generalizar os exemplos encontrados em uma grande massa de dados na forma de regras de alto nível, isto é, compreensíveis ao ser humano.

Para que seja possível o aprendizado, um sistema de IA deve ser capaz de realizar três tarefas (RUSSELL & NORVIG, 2004):

- Armazenar conhecimento;
- Aplicar o conhecimento armazenado para resolver problemas;
- Adquirir novo conhecimento.

No modelo simples de aprendizagem de máquina representado pela Figura 9, o ambiente fornece alguma informação para um elemento de aprendizagem. O elemento de aprendizagem utiliza, então, esta informação para aperfeiçoar a base de conhecimento, e finalmente, o elemento de desempenho utiliza a base de conhecimento para executar a sua tarefa.

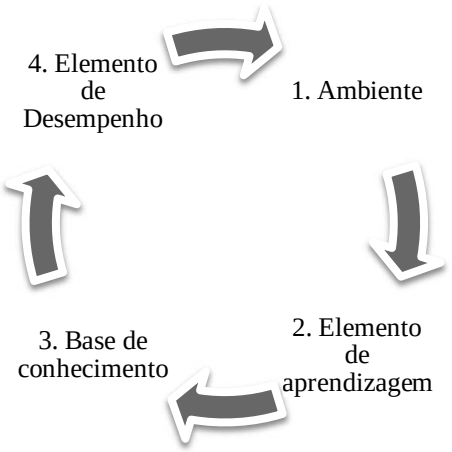

Figura 9 – Modelo simples de aprendizagem de máquina

Na área de Mineração de Textos, o Aprendizado de Máquina é utilizado para a realização de diversas tarefas como a classificação automática de documentos, na qual se destacam as **Máquinas de Vetor de Suporte**[7] (BURGES, 1998) e Classificadores Bayesianos (MCCALLUM & NIGAM, 1998) (KIM, HAN, RIM, & MYAENG, 2006), bem como, para a identificação de classes gramaticais, na qual **Cadeias de Markov Escondidas**[8] apresentam ótimos resultados (HARPER & THEDE, 1999) (SEYMORE, MCCALLUM, & ROSENFELD, 1999).

2.6.4. Estatística

Estatística é a ciência que, por meio de teorias probabilísticas, tem por objetivo a coleção, análise e interpretação de dados numéricos a respeito de fenômenos coletivos ou de massa, bem como a indução das leis a que tais fenômenos cabalmente obedecem e, ainda, a representação numérica e comparativa, em tabelas ou gráficos, dos resultados da análise desses fenômenos (SPIEGEL, 2003). Desta forma, a Estatística busca modelar a aleatoriedade e a incerteza de forma a estimar ou possibilitar a previsão de fenômenos futuros, conforme o caso.

[7] Do termo em inglês, *Support Vector Machines – SVM.*

[8] Do termo em inglês, *Hidden Markov Models – HMM.*

Há alguns anos a Estatística vem sendo utilizada no ramo da Computação. Muitos programas de e-mail modernos realizam a filtragem de *spams*[9] por meio do emprego de classificadores probabilísticos, como o classificador de Bayes. O classificador de Bayes é baseado na aplicação do Teorema de Bayes: a ideia principal é que a probabilidade de um evento A dado um evento B depende não apenas do relacionamento entre os eventos A e B, mas também da probabilidade simples da ocorrência de cada evento envolvido. Embora de simples formulação, como pode ser visto na Equação 1 , classificadores de Bayes apresentam ótimos resultados na categorização de documentos quanto ao assunto (MCCALLUM & NIGAM, 1998). Outras tarefas de Mineração de Textos são baseadas na Estatística, como por exemplo, a seleção de amostras de dados, o cálculo de aproximações, taxas de erro, médias e desvios, bem como as validações de hipóteses e conhecimentos adquiridos ao final do processo de mineração.

Teorema de Bayes

$$\mathrm{P}(A/B) = P(B/A) \times \frac{P(A)}{P(B)}$$

Equação 1 - Teorema de Bayes

2.6.5. Recuperação de Informação

Recuperação de Informação (RI) é a área da computação que lida com o armazenamento de documentos, geralmente textuais, e a recuperação automática de informação associada a eles (BAEZA-YATES & BERTIER, 1999) (MANNING, RAGHAVAN, & SCHÜTZE, 2007). De uma forma simplificada, Recuperação de Informação lida com documentos, termos de indexação e as expressões de buscas dos usuários.

[9] Mensagem eletrônica não-solicitada, geralmente, com fins publicitários. Abreviação em inglês de "*spiced ham*".

Sistemas de RI foram originalmente usados para gerenciar a explosão da informação na literatura científica na segunda metade do século XX. Muitas universidades e bibliotecas públicas usam estes sistemas para prover acesso a livros, jornais, periódicos e outros documentos.

Com a explosão demográfica da *Web*, técnicas de Recuperação de Informação passaram a ser utilizadas em máquinas de buscas**.** As máquinas de buscas surgiram logo após o surgimento da internet, com a intenção de prestar um serviço extremamente importante: a localização de qualquer informação na *Web*, apresentando os resultados de uma forma organizada**,** como um meio de prover a localização do conteúdo desejado. Atualmente, Google, Yahoo e MSN são as máquinas de buscas globais mais acessados e realizam milhões de consultas diárias em seus servidores.

Por tratar de coleções de documentos enormes, Mineração de Textos recorre à área de Recuperação de Informação para obter documentos relevantes ao tópico que será trabalhado, de forma rápida e eficiente. Este assunto será abordado em detalhes no capítulo 4.

2.6.6. Mineração de Dados

O processo de Descoberta de Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Databases,* KDD) consiste na exploração de grandes quantidades de dados à procura de padrões consistentes e potencialmente úteis, como regras de associação ou sequencias temporais, para a obtenção do conhecimento implícito presente nestes dados (GOLDSCHMIDT & PASSOS, 2005) (TAN, STEINBACH, & KUMAR, 2005). O termo processo implica que existem vários passos envolvendo preparação de dados, procura por modelos, avaliação de conhecimento e refinamento, todos estes repetidos em múltiplas iterações (FERRO & LEE, 2001).

Mineração de Dados (do inglês, *Data Mining*) é um dos principais passos no processo de *KDD* e utiliza muitas técnicas de análises estatísticas sofisticadas e algoritmos de Aprendizagem de Máquina, para descobrir padrões escondidos e relações em bases de dados.

É importante ressaltar a grande diferença existente entre processamento de dados e Mineração de Dados. O primeiro lida com operações comuns em uma base de dados: recuperação, exclusão, inserção e atualização de dados. O segundo encontra informação desconhecida nas bases de dados (KONCHADY, 2006).

Mineração de Textos buscou na área de Mineração de Dados os principais algoritmos e técnicas para a descoberta de conhecimento relevante em dados. O processo de *KDD* serviu como referência para a criação de uma metodologia de Mineração de Textos baseada em etapas bem definidas. Para uma referência completa e prática acerca do processo de *KDD* vide (GOLDSCHMIDT & PASSOS, 2005).