Mineração de Textos

E. A. M. Morais *A. P. L. Ambrósio*

Technical Report - INF_005/07 - Relatório Técnico
December - 2007 - Dezembro

Instituto de Informática
Universidade Federal de Goiás
www.inf.ufg.br

Mineração de Textos

Edison Andrade Martins Morais * Ana Paula L. Ambrósio †
edison@inf.ufg.br apaula@inf.ufg.br

Abstract. *The objective of this technical report is describe the state of the art in Text Mining, through description of its concepts and definitions, of the detailing of each stage of its process and the identification of several techniques of knowledge discovery in texts.*

Keywords: Text Mining, Knowledge Discovery.

Resumo. *O objetivo deste relatório técnico é descrever o estado da arte em Mineração de Textos, através de descrição de seus conceitos e definições, do detalhamento das etapas de seu processo e da identificação de várias técnicas de descoberta de conhecimento em textos.*

Palavras-Chave: Mineração de Textos, Descoberta de Conhecimento.

1 Introdução

Considerada uma evolução da área de Recuperação de Informações (RI) [20], *Mineração de textos* (*Text Mining*) é um *Processo de Descoberta de Conhecimento*, que utiliza técnicas de análise e extração de dados a partir de textos, frases ou apenas palavras. Envolve a aplicação de algoritmos computacionais que processam textos e identificam informações úteis e implícitas, que normalmente não poderiam ser recuperadas utilizando métodos tradicionais de consulta, pois a informação contida nestes textos não pode ser obtida de forma direta, uma vez que, em geral, estão armazenadas em formato não estruturado[1].

Os benefícios da mineração de textos pode se estender a qualquer domínio que utilize textos [12], sendo que suas principais contribuições estão relacionadas à busca de informações específicas em documentos, à análise qualitativa e quantitativa de grandes volumes de textos, e a melhor compreensão do conteúdo disponível em documentos textuais.

O objetivo deste relatório técnico é descrever o estado da arte em mineração de textos. Neste sentido este relatório está estruturado da seguinte forma: a Seção 2 descreve o processo de descoberta de conhecimento, dividindo-o em duas formas, descoberta de conhecimento em dados estruturados (Seção 3) e descoberta de conhecimento em dados não estruturados (Seção 4).

*Mestrando em Ciência da Computação – INF/UFG.
†Orientadora– INF/UFG.

[1]Formato não estruturado está relacionado ao fato de um texto ser livre de formato ou padrão de armazenamento [28].

2 O Processo de Descoberta de Conhecimento

De acordo com Wives [27], descobrir conhecimento significa identificar, receber informações relevantes, e poder processá-las e agregá-las ao conhecimento prévio de seu usuário, mudando o estado de seu conhecimento atual, a fim de que determinada situação ou problema possa ser resolvido. Neste sentido, observa-se que a o processo de descoberta de conhecimento está fortemente relacionado à forma pela qual a informação é processada.

Sabe-se que o volume de informações disponíveis é muito grande, neste sentido, mecanismos automáticos de processamento tendem a tornar o processo de descoberta de conhecimento mais eficiente. Logo, faz-se necessário automatizar este processo, principalmente através da utilização de softwares e computadores.

Neste contexto surge a *Descoberta de Conhecimento Apoiada por Computador* (*Knowledge Discovery - KD*), que é um processo de análise de dados ou informações, cujo principal objetivo é fazer com a que as pessoas possam adquirir novos conhecimentos a partir da manipulação de grandes quantidades de dados.

Basicamente, existem duas abordagens utilizadas nesta área, a *Descoberta de Conhecimento em Dados Estruturados* e a *Descoberta de Conhecimento em Dados não Estruturados* (Figura 1).

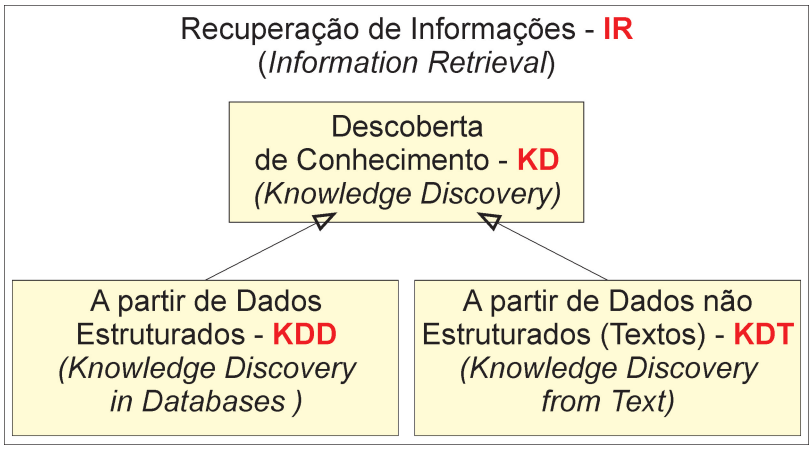

Figura 1: Tipos de Descoberta de Conhecimento

3 Descoberta de Conhecimento em Dados Estruturados

As primeiras aplicações nesta área foram realizadas em Bancos de Dados corporativos. Os métodos e ferramentas utilizados para realizar este processo foram desenvolvidos com base em métodos estatísticos, métodos de Inteligência Artificial e métodos provenientes da área de Recuperação de Informações (*Information Retrieval*) [27].

A partir destas primeiras aplicações surgiu a área conhecida como *Descoberta de Conhecimento em Bancos de Dados* (*Knowledge Discovery in Databases - KDD*). O principal objetivo desta área está relacionado à descoberta de co-relacionamentos e dados implícitos em registros de bancos de dados, através do estudo e desenvolvimento de processos de extração de conhecimento. Seu principal objetivo é encontrar conhecimento a partir de um conjunto de dados para ser utilizado em algum processo decisório.

Desta forma, é importante que o resultado do processo de *KDD* seja compreensível a humanos, além de útil e interessante para usuários finais do processo, que geralmente são to-

madores de decisão. Os processos de *KDD* devem ser vistos como práticas para melhorar os resultados das explorações feitas utilizando ferramentas tradicionais de exploração de dados, como os *Sistemas de Gerenciamento de Bancos de Dados (SGBD)* [21].

Normalmente, o processo de *KDD* deve ser feito seguindo alguma metodologia, que por sua vez envolve um conjunto de fases (Figura 2). Estas fases podem ser específicas para uma determinada aplicação do processo, ou podem ser genéricas. Geralmente as etapas do processo de *KDD* são (Figura 2): *identificação do problema*; *pré-processamento ou preparação dos dados*; *mineração de dados* (*data-mining*); *pós-processamento*.

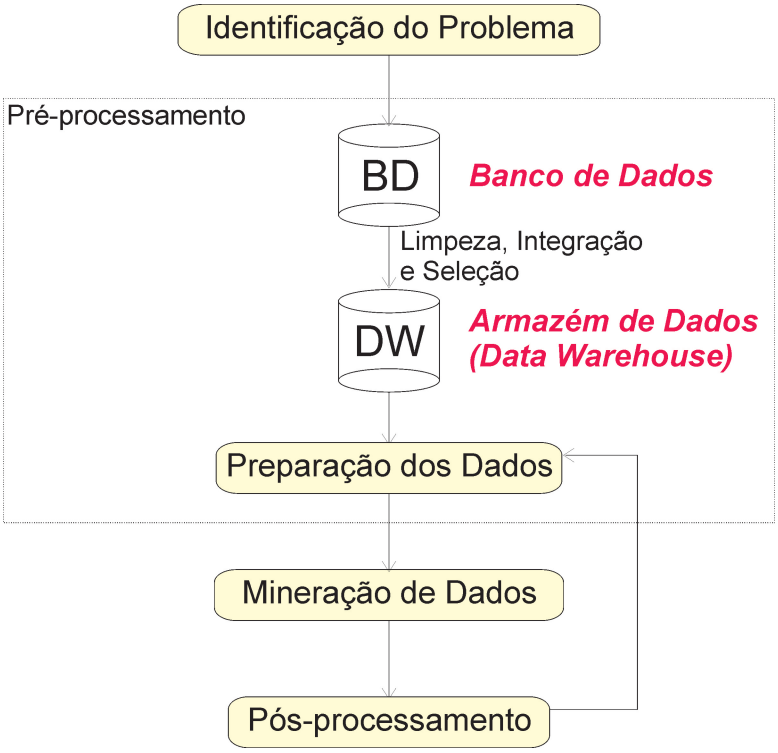

Figura 2: Etapas do Processo de KDD

- *Identificação do problema*

  Nesta fase, um estudo do domínio da aplicação, e a definição dos objetivos e metas a serem alcançados no processo de *KDD* são identificados. O sucesso de todo o processo depende, neste momento, do envolvimento e participação de especialistas do domínio, no sentido de fornecerem conhecimento sobre a área.

  O conhecimento adquirido nesta fase servirá como insumo para todas as outras etapas.

  A etapa de pré-processamento, poderá auxiliar na escolha do melhor conjunto de dados a serem utilizados para extração de padrões.

  A etapa de mineração de dados, poderá auxiliar na escolha de um critério de preferência entre os modelos gerados, ou mesmo na geração de um conhecimento inicial a ser fornecido como entrada para o algoritmo de mineração.

  Já na etapa de pós-processamento, este conhecimento servirá como critério de avaliação das saídas produzidas, no sentido de verificar se o conhecimento extraído pode ser útil para o usuário final.

- *Pré-processamento*

Uma vez que dados armazenados em bases de dados normalmente não estão em formato adequado para extração de conhecimento, faz-se necessária a aplicação de métodos para *extração e integração*, *transformação*, *limpeza*, *seleção e redução* de volume destes dados, antes da etapa de mineração.

Extração e integração envolvem a obtenção dos dados nas várias bases de dados disponíveis, e sua posterior unificação, formando uma única fonte de dados.

Transformação é a adequação dos dados para serem utilizados em algoritmos de extração de padrões. Estas transformações variam de acordo com o domínio da aplicação, por exemplo, aplicações financeiras que necessitam trabalhar com vários tipos de valores monetários.

Uma vez que o processo de coleta de dados pode apresentar problemas, como erros de digitação, por exemplo, técnicas de limpeza destes dados se fazem necessárias no sentido de garantir a qualidade dos mesmos. Este processo também pode ser utilizado para outros fins, como a remoção de valores inválidos para determinados atributos, por exemplo.

A seleção e redução dos dados, normalmente, é necessária em virtude de restrições de espaço em memória ou tempo de processamento. Neste caso, o número de exemplos e de atributos disponíveis para análise pode inviabilizar a utilização de algoritmos de extração de padrões. A redução dos dados pode ser feita diminuindo o número de exemplos, diminuindo o número de atributos ou diminuindo o número de valores de determinados atributos [26]. Segundo Glymour et al. [9], é importante que esta redução mantenha as características do conjunto original de dados, por meio de amostras representativas dos mesmos.

- *Mineração de dados* (*Data Mining*)

  Mineração de Dados é uma área de pesquisa multidisciplinar, incluindo tecnologia de bancos de dados, inteligência artificial, aprendizado de máquina, redes neurais, estatística, reconhecimento de padrões, sistemas baseados em conhecimento, recuperação da informação, computação de alto desempenho e visualização de dados. Em seu sentido relacionado a banco de dados, trata-se do processo de extração ou mineração de conhecimento a partir de grandes volumes de dados [1].

  De acordo com Fayyad et al. [6], mineração de dados é um *processo* de identificação de *padrões válidos*, *novos*, potencialmente *úteis* e *compreensíveis* disponíveis nos dados.

  Processo quer dizer que mineração de dados envolve diversas etapas, por exemplo, preparação dos dados, busca por padrões e avaliação do conhecimento.

  Padrão significa alguma abstração de um subconjunto de dados em alguma linguagem descritiva de conceitos.

  Válidos denota que os padrões descobertos devem possuir algum grau de certeza, ou seja, devem garantir que os casos relacionados ao padrão encontrado sejam aceitáveis.

  Um padrão encontrado deve fornecer informações novas, úteis e compreensíveis sobre os dados. O grau de novidade serve para determinar o quão novo ou inédito é um padrão, além disso estes devem ser descritos em alguma linguagem que possa ser compreendida por seus usuários.

  Esta etapa deve ser direcionada para o cumprimento dos objetivos definidos na etapa de identificação do problema. Na prática, envolve a escolha, configuração e execução de um ou mais algoritmos para extração de conhecimento. Estes algoritmos poderão ser

executados diversas vezes (processo iterativo), até que resultados mais adequados aos objetivos possam ser alcançados.

- *Pós-processamento*.

  O conhecimento extraído na fase de mineração de dados pode gerar uma grande quantidade de padrões. Muitos destes padrões podem não ser importantes, relevantes ou interessantes para seu usuário. Portanto, é necessário fornecer a estes usuários apenas os padrões que possam lhes interessar.

  Neste sentido, diversas medidas para avaliação de padrões [18] têm sido pesquisadas com a finalidade de auxiliar seu usuário no entendimento e utilização do conhecimento representado por estes padrões. Estas medidas podem ser divididas em medidas de desempenho e medidas de qualidade. Alguns exemplos de medidas de desempenho são precisão, erro, confiança negativa, sensitividade, especificidade, cobertura, suporte, satisfação, velocidade e tempo de aprendizado.

  Medidas subjetivas também podem ser utilizadas para avaliação da qualidade de padrões. Estas medidas levam em consideração que fatores específicos de conhecimento de domínio e de interesse de usuários devem ser considerados. Exemplos destas medidas são inesperabilidade e utilidade [22].

  Finalmente, caso o resultado final do pós-processamento não seja satisfatório para seu usuário final, todo o processo pode ser repetido até que este objetivo seja alcançado.

O processo de *KDD* é voltado para análise de dados armazenados em formato estruturado. A próxima sub-seção descreve técnicas, em especial a *Mineração de Textos*, de descoberta de conhecimento em dados não estruturados.

4 Descoberta de Conhecimento em Dados não Estruturados

Análise de dados armazenados em formato não estruturado pode ser considerada uma atividade mais complexa, se comparada à análise de dados estruturados, justamente pelo fato dos dados possuírem a característica da não estruturação. Logo, são necessárias técnicas e ferramentas específicas para tratamento deste tipo de dados. Este conjunto de técnicas e ferramentas também fazem parte da área de *Recuperação de Informações*, mais especificamente da área conhecida como *Descoberta de Conhecimento em Textos (Knowledge Discovery from Text - KDT)* [28] [17].

De acordo com Beppler et al. [3], KDT engloba técnicas e ferramentas inteligentes e automáticas que auxiliam na análise de grandes volumes de dados com o intuito de "garimpar" conhecimento útil, beneficiando não somente usuários de documentos eletrônicos da Internet, mas qualquer domínio que utiliza textos não estruturados.

Logo, como a forma mais comum de armazenamento de informação é através de texto, KDT, teoricamente, tem um potencial maior de utilização do que KDD, pois cerca de 80% das informações contidas nas organizações estão armazenadas em documentos textuais [3].

Recuperação de informação, KDT, e mineração de textos possuem alto grau de dependência no que diz respeiro a *processamento de linguagem natural*, especialmente utilizando processos de *lingüística computacional*. O processamento de linguagem natural corresponde ao uso de computador para interpretar e manipular palavras como parte da linguagem. A lingüística computacional é o ramo que lida com a gramática e a lingüística, onde é desenvolvido o

ferramental necessário para investigar textos e extrair informação sintática e gramaticalmente classificada dos mesmos [15].

Na prática, o processo de KDT é centrado no processo de Mineração de Textos, que é um campo multidisciplinar, que envolve recuperação de informação, análises textuais, extração de informação, *clusterização*, categorização, visualização, tecnologias de base de dados, e mineração de dados.

4.1 Mineração de Textos - Conceitos e Definições

A mineração de textos tem sua origem relacionada a área de *Descoberta de Conhecimento em Textos* (*Knowledge Discovery from Text - KDT*), tendo seus processos sido descritos pela primeira vez em [7], descrevendo uma forma de extrair informações a partir de coleções de texto dos mais variados tipos.

As principais contribuições desta área estão relacionadas à busca de informações específicas em documentos, à análise qualitativa e quantitativa de grandes volumes de textos, e à melhor compreensão de textos disponíveis em documentos. Textos estes que podem estar representados das mais diversas formas, dentre elas: *e-mails*; arquivos em diferentes formatos (pdf, doc, txt, por exemplo); páginas *Web*; campos textuais em bancos de dados; textos eletrônicos digitalizados a partir de papéis. Moura [15], por exemplo, descreve uma proposta de utilização de mineração de textos para seleção, classificação e qualificação de documentos.

Atualmente, mineração de textos pode ser considerado sinônimo de descoberta de conhecimento em textos. Nomenclaturas como *Mineração de Dados em Textos* (*Text Data Mining*) ou *Descoberta de Conhecimento a partir de Bancos de Dados Textuais* (*Knowledge Discovery from Textual Databases*), também podem ser encontrados na literatura [14], uma vez que o processo de mineração de textos também pode ser realizado a partir de técnicas de *Descoberta de Conhecimento em Bancos de Dados [21]* (*Knowledge Discovery in Databases - KDD*) aplicadas sobre dados extraídos a partir de textos [25].

Em [12] podem ser encontrados outros termos que já foram utilizados como sinônimo de mineração de textos: Busca de Informação (*Information Seeking*); Conhecimento Público não Descoberto (*Undiscovered Public Knowledge*); Recuperação de Conhecimento (*Knowledge Retrieval*).

Existem várias definições para mineração de textos. Segundo Lopes [14], o termo se refere ao processo de extração de padrões interessantes e não triviais, ou conhecimento a partir de documentos em textos não-estruturados. Moura [15] descreve a mineração de textos, como sendo uma área de pesquisa tecnológica cujo objetivo é a busca por padrões, tendências e regularidades em textos escritos em linguagem natural.

Já Wives [27], afirma que a mineração de textos pode ser entendida como a aplicação de técnicas de KDD sobre dados extraídos de textos. Entretanto, KDT não inclui somente a aplicação das técnicas tradicionais de KDD, mas também qualquer técnica nova ou antiga que possa ser aplicada no sentido de encontrar conhecimento em qualquer tipo de texto. Com isso, muitos métodos foram adaptados ou criados para suportar esse tipo de informação semi-estruturada ou sem estrutura, que é o texto.

Ao utilizar os recursos de mineração de textos, um usuário não solicita exatamente uma busca, mas sim uma análise de um documento. Entretanto, este não recupera o conhecimento em si. É importante que o resultado da consulta seja analisado e contextualizado para posterior descoberta de conhecimento.

Na prática, a mineração de textos define um processo que auxilia na descoberta de conhecimento inovador a partir de documentos textuais, que pode ser utilizado em diversas áreas do

conhecimento.

4.2 O processo de Mineração de Textos

De forma geral, as etapas do processo de mineração de textos são as seguintes: seleção de documentos, definição do tipo de abordagem dos dados (análise semântica ou estatística), preparação dos dados, indexação e normalização, cálculo da relevância dos termos, seleção dos termos e pós-processamento (análise de resultados).

Estas etapas podem ser visualizadas através do diagrama de atividades representado na figura 3.

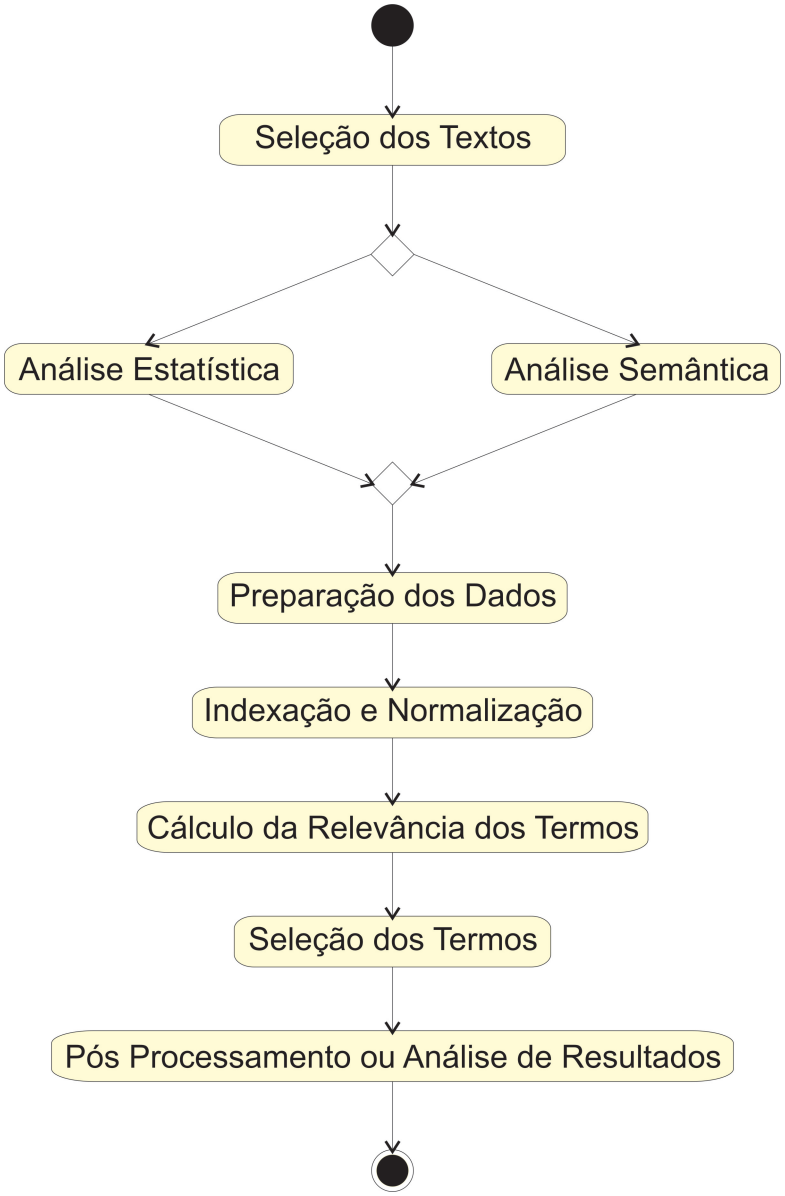

Figura 3: Etapas do Processo de Mineração de Textos

4.2.1 Tipos de Abordagens de Dados

De acordo com Ebecken et al. [5] existem dois tipos de abordagens para análise de dados textuais na área de mineração de textos: a *Análise Semântica*, baseada na funcionalidade

dos termos encontrados nos textos, e a *Análise Estatística*, baseada na freqüência dos termos encontrados nos textos. Estas abordagens podem ser utilizadas separadamente ou em conjunto.

- *Análise Semântica*

  Este tipo de análise emprega técnicas que avaliam a seqüência dos termos no contexto dos textos, no sentido de identificar qual a sua função. Ela é fundamentada em técnicas de *Processamento de Linguagem Natural (PNL)* [19]. Sua utilização justifica-se principalmente pela melhoria da qualidade dos resultados do processo de mineração de textos, especialmente se for incrementado por *Processamento Lingüístico* [19].

  Para processamento de linguagem natural é preciso ter, pelo menos, conhecimento morfológico, sintático, semântico, pragmático, do discurso e do mundo.

  1. *Conhecimento Morfológico*
     É o conhecimento da estrutura, da forma e das inflexões das palavras.
  2. *Conhecimento Sintático*
     É o conhecimento estrutural das listas de palavras e como elas podem ser combinadas para produzir sentenças.
  3. *Conhecimento Semântico*
     É o conhecimento do significado das palavras, independente do contexto. Também designa outros significados mais complexos, podem ser obtidos pela combinação destas palavras.
  4. *Conhecimento Pragmático*
     É o conhecimento do uso da língua em diferentes contextos e como estes afetam seu significado e a interpretação.
  5. *Conhecimento do Discurso*
     É o conhecimento de como as sentenças imediatamente precedentes afetam a interpretação da próxima sentença.
  6. *Conhecimento do Mundo*
     É o conhecimento geral do domínio ou o mundo com o qual a comunicação da linguagem natural se relaciona.

  Segundo Cordeiro [4], técnicas de análise semântica de textos procuram identificar a importância das palavras dentro da estrutura de suas orações. Porém, quando se utiliza um único texto algumas funções podem ser identificadas com um grau de importância. Entretanto, para algumas tarefas isso não é suficiente. Como exemplo podem ser citadas as categorizações, onde é interessante analisar um documento comparando-o com bases de conhecimento de diferentes assuntos para descobrir a que categoria ele pertence.

- *Análise Estatística*

  Neste tipo de análise, a importância de um termo é dada pelo número de vezes que este aparece no texto. Basicamente, seu processo envolve *aprendizado estatístico a partir de dados*, que normalmente inclui as etapas de *codificação dos dados*, *estimativa dos dados* e *modelos de representação de documentos*.

1. *Codificação dos Dados*

   Uma codificação inicial dos dados é escolhida com base em indicações de especialistas. Também pode ser feita de acordo com critérios que representem propriedades interessantes dos dados em relação aos objetivos da seleção dos mesmos.

   Se informações relevantes forem descartadas nesta etapa, não poderão ser recuperadas depois. Entretanto, se a codificação inicial dos dados contém muita informação irrelevante ou ruídos, a busca por uma seleção adequada pode se tornar difícil ou consumir muito tempo. Além disso, propriedades importantes destes dados podem ser perdidas em meio ao ruído.

2. *Estimativa dos Dados*

   Esta etapa envolve a procura por um modelo adequado a partir de um conjunto de modelos (espaço de modelos). Um modelo pode ser obtido a partir da aplicação de um algoritmo de aprendizado ou de um método de estimativa.

3. *Modelos de Representação de Documentos*

   Documentos podem ser vistos como *"containers"* de palavras. Esta abordagem, também conhecida como *bag of words*, ignora a ordem que as palavras aparecem nos textos, assim como qualquer informação de pontuação ou de estrutura, mas retém o número de vezes que a palavra aparece.

   Esta técnica é considerada uma simplificação de toda a abundância de informações que um texto pode expressar, não fornecendo portanto uma descrição fiel de seu conteúdo. O desenvolvimento de modelos mais ricos, que sejam computacionalmente viáveis e possíveis de serem estimados, continua sendo um problema desafiador para a computação.

   Entretanto, apesar desta técnica não ser suficiente para interpretação completa a respeito do textos, ela provê uma quantidade considerável de informações sobre associações entre palavras e documentos que tem se apresentado suficiente para *clustering* e para recuperação de informações a partir de grandes coleções de textos.

Uma vez definido o tipo de abordagem de dados utilizada no processo de mineração de textos, a próxima etapa é a *Preparação dos Dados*.

4.2.2 Preparação dos Dados

A preparação dos dados é a primeira etapa do processo de descoberta de conhecimento em textos. Envolve a seleção dos dados que constituem a base de textos de interesse e o trabalho inicial para tentar selecionar o núcleo que melhor expressa o conteúdo destes textos. Além de prover uma redução dimensional, o objetivo desta etapa é tentar identificar similaridades em função da morfologia ou do significado dos termos nos textos [5].

Neste sentido, a área de Recuperação de Informação (RI) desenvolveu modelos para a representação de grandes coleções de textos que identificam documentos sobre tópicos específicos. Segundo Ebecken et al. [5], pode-se considerar que a RI seja o primeiro passo de um processo de preparação dos dados.

Já um *Sistema de Recuperação de Informações textuais (SRI Textual)* é um sistema desenvolvido para indexar e recuperar documentos do tipo textual. Nesse tipo de sistema, as consultas são descritas através de palavras (termo). Os usuários devem escolher os termos mais adequados para caracterizar sua necessidade de informação. Os documentos relevantes a essa consulta

são selecionados de acordo com a quantidade de palavras semelhantes que eles possuem com a consulta [27].

Este sistema possui um mecanismo de *Análise de Relevância* (filtro) sobre um conjunto de documentos, retornando ao seu usuário o resultado de um problema particular. Quem faz essa análise de relevância é uma função denominada *Função de Similaridade*. Essa função busca identificar uma relação entre os termos da consulta e os termos dos documentos. Teoricamente pode ser feita uma comparação direta entre esses termos, mas, devido a problemas de sinonímia[2], polissemia[3] e outros relacionados ao vocabulário, essa simples comparação nem sempre oferece resultados satisfatórios e os documentos recuperados podem estar relacionados a assuntos variados.

Existem vários métodos de cálculo de similaridade. Além disso, diversos métodos de RI foram criados, formando uma taxonomia de modelos [27], que incluem o *booleano*, o *espaço-vetorial*, o *probabilístico*, o *difuso (fuzzy)*, o da *busca direta*, o de *aglomerados (clusters)*, o *lógico* e, o *contextual* ou *conceitual*. Mais detalhes sobre estes modelos podem ser obtidos em [27] e [5].

- *Modelo Booleano*

  Considera os documentos como sendo conjuntos de palavras. Possui esse nome justamente por manipular e descrever esses conjuntos através de conectivos de *boole* (*and*, *or* e *not*). As expressões *booleanas* são capazes de unir conjuntos, descrever intersecções e retirar partes de um conjunto.

  Em uma busca, por exemplo, o usuário indica quais são as palavras (elementos) que o documento (conjunto) resultante deve ter para que seja retornado. Assim, os documentos que possuem interseção com a consulta (mesmas palavras) são retornados. Os documentos podem ser ordenados pelo grau de interseção, onde o mais alto é aquele que contém todas as palavras especificadas na consulta do usuário, e o mais baixo o que contém somente uma.

- *Modelo Espaço-Vetorial*

  Cada documento é representado por um vetor de termos e cada termo possui um valor associado que indica o grau de importância (denominado peso) desse no documento. Portanto, cada documento possui um vetor associado que é constituído por pares de elementos na forma (*palavra 1*, *peso 1*), (*palavra 2*, *peso 2*)...(*palavra n*, *peso n*).

  Nesse vetor são representadas todas as palavras da coleção e não somente aquelas presentes no documento. Os termos que o documento não contém recebem grau de importância zero e os outros são calculados através de uma fórmula de identificação de importância. Isso faz com que os pesos próximos de um (1) indiquem termos extremamente importantes e pesos próximos de zero (0) caracterizem termos completamente irrelevantes (em alguns casos a faixa pode variar entre -1 e 1).

  O peso de um termo em um documento pode ser calculado de diversas formas. Esses métodos de cálculo de peso geralmente se baseiam na contagem do número de ocorrências dos seus termos (freqüência).

[2] O problema de sinonímia (*synonymy*) ocorre porque o significado de uma palavra pode existir em uma variedade de formas [27].

[3] Polissemia ou homografia (*homography*) é o nome dado a característica de linguagem que um termo (a mesma forma ortográfica) possui de poder pode conter vários significados [27].

- *Modelo Probabilístico*

  Trabalha com conceitos provenientes da área de probabilidade e estatística. Nesse modelo, busca-se saber a probabilidade de um documento *x* ser relevante a uma consulta *y*, caso os termos especificados por esta apareçam naquele.

  Existem diversas formas de se obter estatisticamente essa informação. Porém, a base matemática comumente adotada para esse modelo é o *Método Bayesiano*. Devido a isso, esse modelo também é chamado de *Modelo Bayesiano*

- *Modelo Difuso (Fuzzy)*

  Os documentos também são representados por vetores de palavras com seus respectivos graus de relevância. A diferença está no conceito relacionado à relevância.

  Na teoria de conjuntos difusos, todas as características de determinado universo estão presentes em todos os conjuntos. A diferença é que a presença pode ser medida e pode não ser exata, ou seja, pode haver incerteza. Logo, não há conjunto vazio, mas sim um conjunto cujos elementos possuem uma relevância (importância) muito baixa (próxima de zero).

  A teoria difusa permite trabalhar com esses valores intermediários que indicam o quanto determinado objeto pertence ou não ao conjunto, pois esta foi construída com a finalidade de tratar incertezas e imprecisões.

- *Modelo Busca Direta*

  O modelo de busca direta também é denominado de *Modelo de Busca de Padrões* (*pattern search*), e utiliza métodos de busca de *strings* para localizar documentos relevantes.

  Na prática, esse modelo é utilizado na localização da *strings* em documento. As buscas são realizadas diretamente nos textos originais, em tempo de execução. O resultado da busca é a localização de todas as ocorrências do padrão de consulta em um documento ou conjunto de documentos.

  Sua utilização é aconselhada em casos onde a quantidade de documentos é pequena, sendo muito utilizada em softwares de edição de documentos para que o usuário possa localizar palavras ou expressões no texto que está editando.

- *Modelo Aglomerados (Clusters)*

  Também conhecida como *Clustering Model*, utiliza técnicas de Agrupamento (ou *Clustering*) de documentos.

  Seu funcionamento consiste em identificar documentos de conteúdo similar (que tratem de assuntos parecidos) e armazená-los ou indexá-los em um mesmo grupo ou aglomerado (*cluster*). A identificação de documentos similares em conteúdo dá-se pela quantidade de palavras similares e freqüentes que eles contêm.

  Quando o usuário especifica sua consulta, o sistema identifica um documento relevante e retorna para o usuário todos os documentos pertencentes ao mesmo grupo.

- *Modelo Lógico*

  Baseia-se em métodos e teorias provenientes da lógica matemática para modelar o processo de recuperação de documentos. Não se tem conhecimento de SRI comerciais e práticos que utilizem esse modelo. As aplicações existentes são aparentemente de âmbito acadêmico e teórico.

Para que o modelo lógico funcione torna-se necessário modelar os documentos através de lógica predicativa, o que exige um grande esforço no trabalho de modelagem, incorporando semântica ao processo de recuperação. Com isso o sistema passa a "ter uma noção" do conteúdo dos documentos, podendo julgar melhor a relevância desses para seu usuário.

- *Modelo Contextual ou Conceitual*

  Os modelos anteriores consideram a presença de termos em documentos. Eles realizam o "casamento" entre um documento e uma consulta somente se as palavras contidas no documento tiverem a mesma morfologia que as palavras especificadas na consulta. Logo, os documentos que não possuem as palavras identificadas são considerados irrelevantes (mesmo que os termos tenham o mesmo sentido) por possuírem uma morfologia diferente.

  Devido à ambigüidade [10] e incerteza inerentes à linguagem natural, esta abordagem torna-se muito restritiva, causando problemas de sinonímia e polissemia.

  Neste sentido, o modelo contextual (ou conceitual) é desenvolvido a partir do princípio de que todo documento possui um contexto, pois a pessoa que escreve um texto o faz desenvolvendo um assunto específico, e utiliza frases interconectadas ou encadeadas que fazem sentido dentro do assunto (o contexto).

  A consulta do usuário também possui um contexto que é definido por sua necessidade de informação. Uma vez identificado o contexto dessa necessidade de informação e os contextos dos documentos de uma coleção (base de documentos), o processo de recuperação e de identificação de informações relevantes pode ser feito ao nível de contextos e não mais ao nível de palavras isoladas. Espera-se com isso que os resultados em relação à relevância dos documentos retornados sejam melhores.

  Entretanto, esta não é uma tarefa simples. Os processos de cognição humana ainda não são completamente compreendidos e, portanto, não é possível saber que elementos são necessários para modelar um contexto. Atualmente, isso é feito selecionando algumas palavras que, em conjunto (e estando correlacionadas), podem definir (dar uma idéia de) esse contexto. Também podem ser utilizadas ontologias para modelá-lo.

  Como cada palavra pode estar presente em mais de um contexto, deve haver um grau (peso) indicando quanto uma palavra é relevante (importante) para aquele assunto. Esse conjunto de palavras é então utilizado para representar o contexto.

  Esse modelo não elimina o problema do vocabulário, mas pode minimizá-lo se o conjunto de palavras utilizado na descrição dos contextos for bem escolhido. Vários termos podem ser utilizados nessa descrição. Porém, muitos deles certamente são encontrados em vários contextos. Devem ser escolhidos apenas aqueles que caracterizam bem cada contexto sem que indiquem ou apareçam em outros (ou muitos outros) contextos. Ou seja, as palavras devem ter um alto grau de discriminação.

  Um dos principais problemas desse modelo está no fato dos descritores de contextos poderem ser elaborados incorretamente, o que ocasionaria uma busca errada onde os documentos retornados provavelmente seriam irrelevantes para a necessidade do usuário. Ou seja, a descrição dos contextos deve ser elaborada cuidadosamente para que a recuperação contextual funcione de forma a oferecer resultados relevantes e coerentes.

Após a escolha do modelo mais apropriado, o próximo passo é a *Indexação e Normalização* dos textos.

4.2.3 Indexação e Normalização

O objetivo principal da indexação e normalização dos textos é facilitar a identificação de similaridade de significado entre suas palavras, considerando as variações morfológicas e problemas de sinonímia [5]. Nesta fase as características dos documentos são identificadas e adicionadas ao SRI.

Este processo tem como resultado a geração de um índice. Esse índice é construído através de um processo de indexação. Indexar, portanto, significa identificar as características de um documento e colocá-las em uma estrutura denominada índice.

Um documento pode ser indexado por termos diferentes que são correspondentes ao vocabulário utilizado em sua área. Nesse caso, geralmente, há um conjunto de termos predefinidos e específicos para cada assunto da área em questão.

Essa técnica facilita muito a localização de informações, pois usuários de áreas específicas estão acostumados a utilizar os termos comuns. Por outro lado, se o SRI for utilizado em uma área diferente da área para a qual foi indexado ele não será tão eficiente porque os problemas relacionados à diferença de vocabulário serão mais freqüentes.

Quando a indexação é realizada manualmente, a pessoa encarregada de fazê-la deve analisar o conteúdo de cada documento e identificar palavras-chave que o caracterizem. Essas palavras, quando adicionadas ao índice, passam a ser chamadas de *termos de índice*. A geração automática de índices deve produzir o mesmo resultado, isto é, produzir os termos de índice.

Em mineração de textos, indexação é um processo automático (Figura 4). Suas principais fases são: *identificação de termos* (simples ou compostos); a remoção de *stopwords* (palavras irrelevantes); e *normalização morfológica (stemming)*.

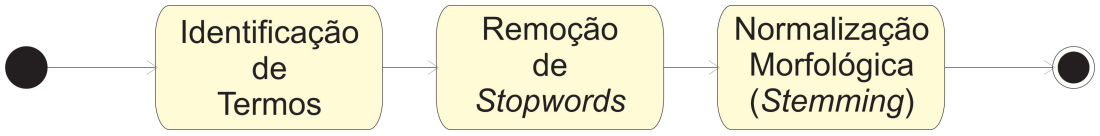

Figura 4: Etapas do Processo de Indexação Automática [27]

- *Identificação de Termos*

  Esta fase tem como objetivo principal a identificação dos termos contidos no texto, sejam eles *simples* ou *compostos*.

  - *Identificação de Termos Simples*

    Corresponde a aplicação de um *parser* (analisador léxico[4]) que identifique as palavras (termos simples) presentes nos documentos, e elimine os símbolos e caracteres de controle de arquivo ou de formatação [27].

    Caso seja necessário verificar a ocorrência de erros ortográficos, pode ser utilizado um dicionário de termos.

    Também pode ser utilizado um dicionário de sinônimos para normalização do vocabulário, caso seja necessário trabalhar com *vocabulário controlado*[5].

[4] Análise Léxica: consiste na conversão de uma cadeia de caracteres de entrada em uma cadeia de palavras ou *tokens* [27].

[5] Vocabulário Controlado: é um conjunto de termos predefinidos e específicos para cada assunto de uma determinada área. Esta abordagem facilita muito a localização de informações, pois normalmente seus usuários estão acostumados aos termos comumente utilizados em sua área de interesse [27].

Outras alterações no texto podem ser feitas nesta fase: os caracteres podem ser todos convertidos para maiúsculo ou minúsculo; múltiplos espaços e tabulações podem ser substituídos por espaços simples; números e datas podem ser padronizados; hífens podem ser eliminados.

Entretanto nem todas as alterações realizadas nos textos durante esta fase são benéficas. A conversão de seus caracteres para maiúsculo, por exemplo, pode fazer com que não seja possível diferenciar substantivos próprios de comuns, ocorrendo, portanto, perda de semântica em posterior análise do texto.

– *Identificação de Termos Compostos*

Muitas palavras têm significado diferente quando utilizadas em conjunto. Isso pode ocorrer porque existem conceitos que somente são descritos através da utilização de duas ou mais palavras adjacentes. Como, por exemplo, "Processo Cível" ou "Processo Criminal".

Esta fase, também conhecida como *Word-phrase formation*, busca identificar essas expressões compostas de dois ou mais termos [27].

Esta identificação pode ser feita de duas formas. A primeira envolve a identificação de termos que co-ocorrem com muita freqüência em uma coleção de documentos. Neste caso, o sistema pode apresentar a seu usuário expressões identificadas e deixar que ele decida quais são válidas ou não. A segunda consiste na utilização de um dicionário de expressões.

Uma vez que as expressões identificadas são armazenadas no índice na forma composta, o usuário somente poderá localizar este texto utilizando a forma composta da expressão. Para resolver este problema, palavras que formam as expressões compostas também devem ser armazenados em separado no índice.

- *Remoção de Stopwords*

Esta fase envolve a eliminação de algumas palavras que não devem ser consideradas no documento, conhecidas como *stopwords*. *Stopwords* são palavras consideradas não relevantes na análise de textos, justamente por não traduzirem sua essência. Normalmente fazem parte desta lista as preposições, pronomes, artigos, advérbios, e outras classes de palavras auxiliares.

Além dessas, existem também palavras cuja freqüência na coleção de documentos é muito alta. Palavras que aparecem em praticamente todos os documentos de uma coleção não são capazes de discriminar documentos e também não devem fazer parte do índice.

As *stopwords* formam uma lista de palavras conhecida como *stoplist*. As palavras que compoem a *stoplist* dificilmente são utilizadas em uma consulta, além disso sua indexação somente tornaria o índice maior do que o necessário.

- *Normalização Morfológica (Stemming)*

Segundo Wives [27], durante o processo de indexação, dependendo do caso, torna-se interessante eliminar as variações morfológicas de uma palavra. Elas são eliminadas através da identificação do radical de uma palavra. Os prefixos e os sufixos são retirados e os radicais resultantes são adicionados ao índice. Essa técnica de identificação de radicais é denominada lematização ou *stemming*, que em inglês significa reduzir uma palavra ao seu radical (ou raiz).

Além da eliminação dos prefixos e sufixos, características de gênero, número e grau das palavras são eliminados. Isso significa que várias palavras acabam sendo reduzidas para um único termo, o que pode reduzir o tamanho de um índice em até 50%.

Entretanto, a aplicação de técnicas de *stemming* ocasionam uma diminuição na precisão das buscas, já que o usuário não consegue mais procurar por uma palavra específica. Na classificação de documentos, por exemplo, variações morfológicas são importantes, pois aumentam o poder de discriminação entre documentos.

Além disso, ao realizar *stemming* deve-se ter cuidado com *overstemming* e *understemming*. *Overstemming* ocorre quando a cadeia de caracteres extraída não é um sufixo, mas sim parte do radical. Por exemplo, a palavra "gramática", após o processamento reduz para "grama", o que não representa o seu radical, que é "gramat". *Understemming* ocorre quando o sufixo não é removido totalmente. Por exemplo, a palavra "referência", após o processamento reduz para "referênc", ao invés de "refer", que é o radical correto [2].

Ebecken [5] descreve três métodos de *stemming*: *método do stemmer S*; *método de Porter*; *método de Lovins*.

- *método do stemmer S*

  Este é considerado o método mais simples. Consiste na eliminação de apenas alguns finais de palavras, geralmente sufixos que formam o plural. Em palavras da língua inglesa, são removidos apenas os sufixos *ies*, *es*, s.

- *método de Porter*

  Consiste na identificação de diferentes inflexões referentes à mesma palavra e sua substituição por um radical comum.

  Seu algoritmo remove cerca de 60 sufixos diferentes para palavras da língua inglesa e é baseado nas seguintes etapas [2]: redução do plural, troca de sufixos, retirada de sufixos, remoção de sufixos padrões e remoção da vogal "e" ao final da palavra.

- *método de Lovins*

  Este método remove cerca de 250 sufixos diferentes para palavras da língua inglesa. Seu algoritmo remove apenas um sufixo por palavra, retirando o sufixo mais longo conectado à mesma.

Todos os métodos acima consideram palavras da língua inglesa. Um algoritmo de *stemming* para língua portuguesa pode ser encontrado em [16]. Este algoritmo foi implementado em linguagem C e é composto por oito etapas (Figura 5). Cada etapa, por sua vez, tem um conjunto de regras, que são examinadas em seqüência, e somente uma regra na etapa pode ser aplicada.

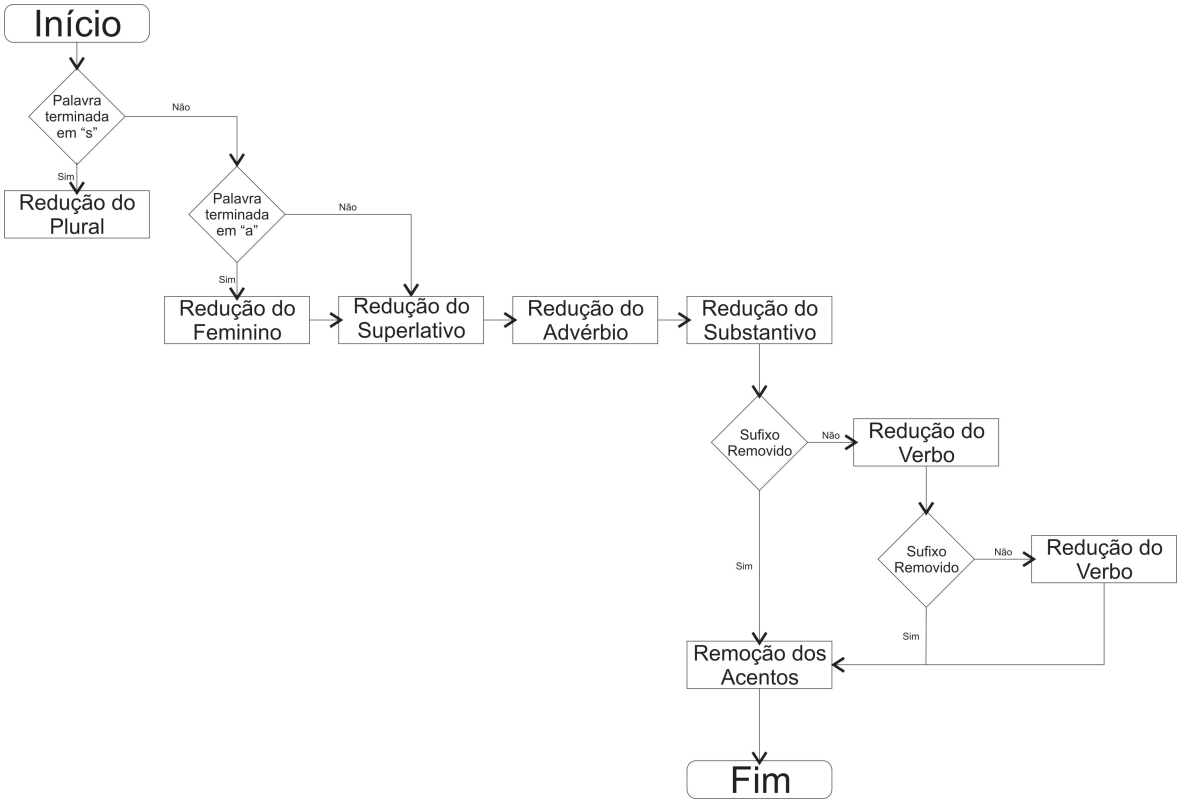

Figura 5: *Portuguese Stemming*

As etapas são descritas a seguir [2]:

1. *Remoção do plural*

   Consiste basicamente em remover o “s” do final das palavras. Há uma lista de exceções como a palavra “lápis” por exemplo.

2. *Remoção do feminino*

   Nesta etapa as formas femininas são transformadas na correspondente masculina. Ex.: “chinesa” → “chinês”.

3. *Remoção de advérbio*

   Esta é a etapa mais simples, uma vez que o único sufixo que denota um advérbio é “mente”. Neste caso também há uma lista de exceções.

4. *Remoção de aumentativo e diminutivo*

   Remove os sufixos dos substantivos e adjetivos que podem ter aumentativo e diminutivo. Por exemplo, “gatinha” ou “menininha”.

5. *Remoção de sufixos em nomes*

   Esta etapa testa as palavras contra uma lista de 61 sufixos para substantivos e adjetivos. Se o sufixo for removido, as etapas seis e sete não são executadas.

6. *Remoção de sufixos em verbos*

   Os verbos da língua portuguesa possuem mais de 50 formas diferentes de conjugação (na língua inglesa existem apenas quatro). Cada uma delas possui seu conjunto de sufixos específico. Os verbos podem variar de acordo com o tempo, a pessoa, o número e o modo. A estrutura das formas verbais pode ser representada por: radical

+ vogal temática[6] + tempo + pessoa. Por exemplo: “andaram” = “and + a + ra + m”. As formas verbais são reduzidas ao seu radical correspondente.

7. *Remoção de vogais*

   Esta etapa consiste em remover a última vogal (“a”, “e” ou “o”) das palavras que não foram examinadas pelas etapas cinco e seis. Ex.: “menino” → “menin”.

8. *Remoção de acentos*

   Esta atividade é necessária porque existem vários casos onde algumas variantes são acentuadas e outras não, como em “psicólogo” e “psicologia”, por exemplo.

   A execução deste passo por último é importante, porque a presença de acentos é significativa em algumas regras, por exemplo: “óis” para “ol” transformando “sóis” em “sol”, por exemplo. Se a regra fosse “ois” para “ol”, poderia causar erros no caso de “dois” para “dol”.

Das etapas de *identificação de termos*, remoção de *stopwords* e *normalização morfológica (stemming)* citadas anteriormente, as duas últimas têm como objetivo reduzir a dimensionalidade do problema, dando foco ao tratamento das palavras que concentram a carga semântica do texto.

A seguir são descritas técnicas de *cálculo de relevância* e *seleção de termos*, que são as etapas finais do processo de indexação e normalização.

4.2.4 Cálculo da Relevância

Nem todas as palavras presentes em um documento possuem a mesma importância. O termos mais freqüentemente utilizados (com exceção das *stopwords*) costumam ter significado mais importante, assim como as palavras constantes em títulos ou em outras estruturas, uma vez que provavelmente foram colocadas lá por serem consideradas relevantes ou descritivas para a idéia do documento. Substantivos e complementos também podem ser considerados mais relevantes que os demais termos [27].

Logo, o cálculo de relevância de uma palavra em relação ao texto em que está inserido pode basear-se na freqüência da mesma, na análise estrutural do documento ou na sua posição sintática de uma palavra. As análises baseadas em freqüência costumam ser as mais utilizadas por serem mais simples, pois os outros tipos de análise muitas vezes necessitam de outras técnicas mais complexas, como processamento de linguagem natural, por exemplo.

A este grau de relacionamento de uma palavra com um texto dá-se o nome de *peso*. Logo, é o peso que indica a importância da palavra em relação a um texto.

Existem várias fórmulas para cálculo do peso. As mais comuns são baseadas em cálculos simples de freqüência: *freqüência absoluta*, *freqüência relativa*, *freqüência inversa de documentos*.

- *freqüência absoluta*

  Também conhecida por freqüência do termo ou *term frequency* (TF), representa a medida da quantidade de vezes que um termo aparece em um documento. Essa é a medida de peso mais simples que existe, mas não é aconselhada em alguns casos, porque, em análise de coleções de documentos, não é capaz de fazer distinção entre os termos que aparecem em poucos ou em muitos documentos. Este tipo de análise também não leva em conta

[6] Existem três classes de verbos na língua portuguesa, de acordo com a terminação da forma infinitiva: “ar”, “er” e “ir”. A vogal temática é a letra (“a”, “e” e “i”) que agrupa verbos e categorias.

a quantidade de palavras existentes em um documento. Com isso, uma palavra pouco freqüente em um documento pequeno pode ter a mesma importância de uma palavra muito freqüente de um documento grande.

- *freqüência relativa*

  Este tipo de análise leva em conta o tamanho do documento (quantidade de palavras que ele possui) e normaliza os pesos de acordo com essa informação. A freqüência relativa (F_{rel}) de uma palavra x em um documento qualquer é calculada dividindo-se sua freqüência absoluta (F_{abs}) pelo número total de palavras no mesmo documento (N):

$$F_{rel}(x) = \frac{F_{abs}(x)}{N}$$

- *freqüência inversa de documentos*

  A fórmula de freqüência relativa considera apenas a quantidade de documentos em que um termo aparece (freqüência de documentos - *DocFreq*). Com base na informação da freqüência absoluta e da freqüência de documentos é possível calcular a freqüência inversa de documentos (*inverse document frequency* - IDF), que por sua vez é capaz de aumentar a importância de termos que aparecem em poucos documentos e diminuir a importância de termos que aparecem em muitos, justamente pelo fato dos termos de baixa freqüência de serem, em geral, mais discriminantes.

  A fórmula mais comum utilizada para cálculo do peso de um termo utilizando a freqüência inversa é:

$$Peso_{td} = \frac{Freq_{td}}{DocFreq_{td}}, \text{ onde:}$$

  $Peso_{td}$: é o grau de relação entre o termo t e o documento d;

  $Freq_{td}$: número de vezes que o termo t aparece no documento d;

  $DocFreq_{td}$: número de documentos que o termo t aparece.

Segundo Wives [27], não existe estudo que indique a superioridade de uma técnica sobre outra de forma significativa. Porém, algumas são mais adequadas do que outras para certas aplicações ou modelos conceituais.

4.2.5 Seleção de Termos

Seleção de termos corresponde à etapa de seleção das palavras retiradas do texto, após o processos de pré-processamento e cálculo da relevância. Esta técnica pode ser baseada no peso dos termos ou na sua posição sintática em relação ao texto.

As principais técnicas de seleção de termos são: *filtragem baseada no peso do termo*, *seleção baseada no peso do termo*, *seleção por análise de co-ocorrência*, *seleção por Latent Semantic Indexing* e *Seleção por análise de linguagem natural*.

- *Filtragem baseada no peso do termo*

  Como a determinação da importância de um termo geralmente é dada pelo seu peso, esta técnica consiste em eliminar todos os termos abaixo de um limiar (*threshold*) estabelecido pelo usuário ou pela aplicação.

- *Seleção baseada no peso do termo*

  Mesmo depois de filtrados, o número de termos resultantes ainda pode ser alto. Esse número pode ser reduzido pela seleção dos n termos mais relevantes. Essa técnica de seleção, denominada *truncagem* [27], estabelece um número máximo de características a serem utilizadas para um documento e todas as outras são eliminadas.

  Para tanto, é necessário que as características estejam ordenadas de acordo com seu grau de relevância (ou importância). Assim, somente as primeiras x características são utilizadas.

  Um dos maiores problemas dessa técnica consiste em definir a quantidade mínima de palavras necessárias para uma descrição adequada dos documentos, sem que suas características mais relevantes sejam perdidas no processo. Alguns experimentos indicam que, na grande maioria dos casos, um número de 50 características é suficiente [27], mas esse valor pode variar dependendo da coleção de documentos.

- *Seleção por Análise de Co-ocorrência*

  Em alguns casos pode ser necessário fazer a análise dos pesos dos termos contidos em um conjunto de documentos. Nestes casos, essa análise pode ser feita levando-se em consideração os termos que ocorrem em vários documentos ao mesmo tempo (co-ocorrência).

  Esta técnica também pode ser utilizada para verificar qual o grau de relacionamento entre textos, isto é, considerando dois ou mais textos, quanto maior o número de termos iguais entre ambos, maior tende a ser seu relacionamento semântico.

  Loh et al. [13] propõe a realização desta análise utilizando as seguintes fórmulas:

  1. Fórmula que analisa o grau de relação entre a palavra e o documento que a contém.

     $d_{ij} = tf_{ij} \times \log\left(\frac{N}{df_j}\right)$, onde:

     d_{ij}: representa o valor combinado da palavra j no documento i;

     tf_{ij}: representa a freqüência da palavra j no documento i;

     N: representa o número total de documentos considerados;

     df_j: representa a freqüência inversa de documentos (número de documentos em que a palavra j aparece).

  2. Fórmula que analisa co-ocorrência das palavras nos textos (baseado nos resultados gerados pela fórmula anterior).

     Co-ocorrência $= \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}}$, para $d_{ijk} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}}\right)$, onde:

     tf_{ijk}: representa o número de ocorrências de ambas as palavras j e k no documento i (o menor número de ocorrências entre as palavras deve ser escolhido);

     df_{jk}: representa o número de documentos (de uma coleção N) no qual as palavras j e k ocorrem ao mesmo tempo.

  Duas palavras podem estar relacionadas entre si em mais de um contexto, portanto, em cada contexto deve existir um grau diferente de relação.

- *Seleção por Latent Semantic Indexing [11]*

Esta técnica foi desenvolvida com o intuito de reduzir o número de dimensões utilizadas pelo modelo espaço-vetorial. Seu principal objetivo é transformar os vetores de documentos originais em um espaço dimensional pequeno e significativo, fazendo uma análise da estrutura co-relacional de termos na coleção de documentos.

- *Seleção por análise de linguagem natural*;

  É possível aplicar técnicas de análise *sintática* e *semântica* para identificar as palavras mais importantes de um documento.

  A análise sintática pode ser realizada a partir de um dicionário ou gramática bem definida para um domínio específico (um *lexicon*, por exemplo). Neste tipo de análise é possível influenciar o peso dos termos encontrados em posições sintáticas específicas do texto (substantivo, por exemplo) a fim de torná-los mais ou menos relevantes.

  De acordo com Wives [27], atualmente, principalmente para a língua portuguesa, esse tipo de análise não funciona em 100% dos casos, sendo que a maioria dos SRI não costuma utilizá-la. A análise lingüística é complexa em termos de implementação e não existem estudos que indiquem que sua utilização oferece resultados estatisticamente melhores do que os obtidos pelas técnicas que são baseadas em freqüência.

  Já a análise semântica baseia-se no princípio de que as partes mais relevantes de um documento já estão de alguma forma demarcadas por estruturas de formatação específicas para isso. A utilização deste tipo de abordagem vem crescendo, especialmente pela possibilidade da utilização de novos formatos (HTML e XML, por exemplo) de armazenamento de textos.

  A maioria destes novos formatos permitem incluir algum tipo de marca (*tag*), que serve para estruturar o documento, facilitando a identificação das estruturas mais relevantes. Cada uma dessas marcas possui uma definição, um significado semântico.

  Infelizmente, para que essa técnica funcione os documentos precisam ter essas marcas. Isso exige que a pessoa responsável pela elaboração do documento as coloque, o que nem sempre é feito de forma correta.

4.2.6 Análise de Resultados

Esta fase envolve a aplicação de técnicas de análise dos resultados de um sistema de recuperação de informações, particularmente os resultados do processo de mineração de textos.

De acordo com Wives [27], esta análise pode ser realizada com base em técnicas de uma área conhecida como bibliometria [29], que é uma sub-área da biblioteconomia encarregada de estudar e aplicar métodos matemáticos e estatísticos em documentos e outras formas de comunicação.

Na prática, estas métricas podem ser utilizadas nos SRI como forma de avaliação do mesmo, isto é, para saber se o mecanismo funcionou ou não como deveria. Nesse caso, as métricas poderiam informar para o usuário quantos e quais documentos lhe são relevantes, além de quanto cada um deles tem importância no contexto.

Porém, para que essas métricas funcionem corretamente, é necessário que a coleção de documentos a ser analisada pelo sistema seja muito bem conhecida. Ou seja, para cada documento é necessário saber, a *priori*, para quais consultas (ou assuntos) eles são relevantes.

Tendo-se uma coleção de documentos conhecida, pode-se adotar a seguinte estratégia: quando o usuário de um SRI faz uma busca, o conjunto de documentos é dividido em quatro segmentos lógicos: *documentos relevantes à consulta que são recuperados*, *documentos*

relevantes que não foram recuperados, *documentos irrelevantes e recuperados* e *documentos irrelevantes que não foram recuperados*.

A eficiência e a eficácia de um SRI é então avaliada de acordo com a sua capacidade em recuperar o máximo possível de itens relevantes ao mesmo tempo em que filtra o maior número de itens irrelevantes. Finalmente, são aplicadas métricas sobre os resultados dessa estratégia.

As principais métricas de análise de SRI são [27]: *recall*, *precision*, *fall-out* e *effort*.

- *Recall*

  O *recall* (abrangência ou revocação) mede a habilidade do sistema em recuperar os documentos mais relevantes para seu usuário, com base no termo ou expressão utilizado na formulação de sua busca.

  Fórmula:

  $$recall = \frac{n\text{-}recuperados\text{-}relevantes}{n\text{-}possíveis\text{-}relevantes}, \text{ onde:}$$

  n-recuperados-relevantes: é o número de documentos relevantes recuperados.

  n-possíveis-relevantes: é o número total de documentos relevantes do sistema. Essa informação geralmente não é conhecida e só pode ser estimada estatisticamente.

- *Precision*

  A *precision* (precisão) mede a habilidade do sistema em manter os documentos irrelevantes fora do resultado de uma consulta.

  Fórmula:

  $$precision = \frac{n\text{-}recuperados\text{-}relevantes}{n\text{-}total\text{-}recuperados}, \text{ onde:}$$

  n-recuperados-relevantes: é o número de documentos relevantes recuperados.

  n-total-recuperados: é o número total de documentos do sistema.

  A precisão é capaz de indicar o *overhead* (esforço) que o usuário teria para analisar uma determinada busca. Isso significa que, se 60% dos itens retornados fossem relevantes, o usuário teria, teoricamente, desperdiçado 40% de seu esforço analisando itens irrelevantes.

  Logo, quanto maior a precisão, menor o esforço do usuário em analisar itens. Portanto, a precisão pode ser utilizada nas interações do usuário com o sistema para indicar o quanto ele ainda necessita interagir para conseguir filtrar os itens irrelevantes e retornar itens mais relevantes.

- *Fall-out*

  Esta técnica considera que a quantidade de documentos irrelevantes poder ser modificada pelo crescimento ou decrescimento da base de dados. Neste sentido, ela mede a quantidade de documentos irrelevantes, permitindo que se identifique se a quantidade de documentos relevantes permanece a mesma quando o número de documentos varia.

  Fórmula:

  $$fall\text{-}out = \frac{n\text{-}recuperados\text{-}irrelevantes}{n\text{-}possíveis\text{-}irrelevantes}, \text{ onde:}$$

n-recuperados-irrelevantes: é o número de documentos irrelevantes recuperados.

n-possíveis-irrelevantes: é o número total de documentos irrelevantes do sistema. Essa informação, geralmente, também não é conhecida e só pode ser estimada estatisticamente.

- *Effort*

  Mede o esforço gasto pelo usuário durante o processo de busca, desde a preparação da consulta, passando pelo processo de análise de resultados até a reformulação da consulta quando necessário, ou seja, toda a interação do usuário com o sistema.

  A taxa de precisão é uma medida do esforço que o usuário deve realizar para obter determinada precisão. Consegue-se saber se o usuário está "caminhando" na direção certa se, a cada iteração com o sistema, seu grau de precisão aumenta. Esse grau de imprecisão (o complemento da precisão) pode ser considerado como o esforço que ele ainda tem que realizar para localizar todos os itens relevantes.

  A medida de esforço deve levar em conta os documentos relevantes existentes na base de dados e que ainda não foram recuperados (ou seja, não basta analisar somente o complemento da precisão). Além disso, o grau de precisão também está relacionado com o tamanho da base de dados e não indica exatamente quantos documentos ainda devem ser recuperados.

Uma das estapas mais importantes do processo de mineração de textos é o cálculo da relevância dos termos. Ultimamente, um método que vem ganhando destaque neste processo é a Análise Semântica Latente - LSA (*Latent Semantic Analysis - LSI*) [11].

4.3 Análise Semântica Latente

Análise Semântica Latente (LSA) é o método para extração e representação do significado semântico de palavras em um contexto, obtidos através de cálculos estatísticos aplicados a um conjunto numeroso de textos [11].

Seu modelo de indexação semântica é baseado na co-ocorrência de palavras em textos. A suposição é que palavras que tendem a ocorrer juntas dentro de um mesmo documento (parágrafos, frases, textos, etc) são consideradas como representativas de similaridade semântica.

Na prática este método é utilizado para a construção de um *espaço semântico*, onde não só palavras, mas, sentenças, parágrafos, textos ou qualquer outro conjunto de palavras, podem ser representados por vetores [24].

A primeira etapa deste processo é a selação de uma coleção de textos sobre determinados assuntos. Esta coleção serve como base para indexação, i.e., geração de um vetor semântico, como os termos que representam semanticamente estes documentos.

A este conjunto de textos são aplicadas técnicas de pré-processamento tradicionais (descritas na sub-seção 4.2.3). Em seguida, é construída uma matriz de representação desta coleção, com as linhas correspondendo às palavras selecionadas e as colunas correspondendo aos textos da coleção. Inicialmente, à cada entrada desta matriz é atribuido o valor da frequência absoluta de cada palavra em cada texto.

A frequência absoluta de cada palavra, em cada uma de suas entradas na matriz, é transformada em seu logaritmo. Isto é feito baseando-se no fato de que um documento com, por exemplo, três ocorrências de uma mesma palavra, tende a ser mais importante do que um documento com apenas uma ocorrência, porém não três vezes mais importante.

Em seguida, cada um dos novos valores de entrada é dividido pelo somatório do produto destes valores pelo logarítimo dos mesmos, para salientar a sua importância.

O próximo passo é a aplicação da técnica conhecida como *Decomposição de Valor Singular* (SVD) [8]. Através desta técnica são geradas outras três matrizes a partir da matriz original, e uma nova matriz é obtida a partir do produto destas três matrizes.

$$M = T \text{ x } S \text{ x } D, \text{ onde:}$$

T = matriz de vetores singulares à esquerda;
S = matriz diagonal de valores singulares em ordem decrescente;
D = matriz de vetores singulares à direita.

A dimensão destas matrizes é reduzida, eliminando as linhas e colunas correspondentes aos menores valores singulares da matriz S assim como as colunas da matriz T e linhas da matriz D correspondentes.

A Decomposição de Valor Singular é normalmente utilizada para localizar a informação semântica essencial em uma matriz de co-ocorrência de palavras. Com isto, partir desta decomposição, é possível, com a redução de dimensão das matrizes T, S e D (mantendo somente os maiores valores singulares), descartar as informações acidentais que geralmente estão presentes.

O objetivo com o produto destas três novas matrizes reduzidas, é obter um um espaço semântico condensado que representa as melhores relações entre as palavras e documentos. A proximidade entre duas palavras é obtida calculando-se o cosseno do ângulo entre seus vetores (linhas da matriz) correspondentes. Quanto maior o cosseno do ângulo entre os vetores de duas palavras, maior a proximidade entre elas.

Finalmente, o vetor de representação de um dado conjunto de palavras, como parágrafos ou textos, no espaço, pode ser obtido através do centróide (média) de todos os vetores das palavras deste conjunto. Permitindo, assim a obtenção da proximidade entre uma palavra e um texto, e até mesmo entre dois textos.

4.4 Outras Técnicas de Descoberta de Conhecimento em Textos

- *Descoberta por Extração de Passagens*

  Este tipo de descoberta tem por objetivo encontrar informações específicas, de forma um pouco mais independente de domínio, auxiliando usuários a encontrar detalhes de informação, sem que este precise ler todo texto. Entretanto, ainda assim, é necessário que o usuário leia e interprete as partes do texto que forem recuperadas para extrair a informação desejada.

- *Descoberta por Análise Lingüística*

  Nesta abordagem, informações e regras podem ser descobertas através de análises lingüísticas em nível léxico, morfológico, sintático e semântico.

- *Descoberta por Análise de Conteúdo*

  Semelhante à descoberta por extração de passagens e à descoberta por análise lingüística, este tipo de descoberta investiga lingüisticamente os textos e apresenta ao seu usuário informações sobre o conteúdo dos textos. Entretanto, a sua diferença está relacionada à forma de análise de conteúdo, onde há maior esforço no tratamento semântico dos textos, extrapolando o limite léxico-sintático. Em relação à extração de passagens, a diferença é que, neste caso, o objetivo é encontrar o significado do texto pretendido pelo autor ao invés de partes ou informações específicas.

- *Descoberta por Sumarização*

  Este tipo de descoberta utiliza as técnicas de descoberta por extração de passagens, descoberta por análise de conteúdo, e descoberta por análise lingüística, com ênfase na produção de resumos ou sumários (abstração das partes mais importantes do conteúdo do texto) a partir de textos.

- *Descoberta por Associação entre Passagens*

  Este tipo de descoberta tem por objetivo encontrar automaticamente conhecimento e informações relacionadas no mesmo texto ou em textos diferentes. Esta abordagem combina a recuperação de informações por passagens com a recuperação contextual. Sua principal aplicação está relacionada à definição automática de *links* em sistemas de hipertexto, sendo que principal vantagem é apresentar ao usuário partes de textos que tratam do mesmo assunto específico.

- *Descoberta por Listas de Conceitos-Chave*

  O objetivo deste tipo de descoberta é apresentar uma lista com os conceitos principais de um único texto, utilizando técnicas semelhantes à geração de centróides de classes [30], que permitem, por exemplo, extrair os termos mais freqüentes dos textos.

- *Descoberta de Estruturas de Textos*

  Determinar a estrutura de um texto ajuda a entender seu significado. Neste sentido, esta técnica analisa as coesões léxicas de um texto, tendo como resultado cadeias de termos relacionados que contribuem para a continuidade do seu significado léxico. Estas cadeias léxicas delimitam partes do texto que têm forte unidade de significado e ajudam também na resolução de ambigüidades, além da identificação da estrutura do discurso. Além da estrutura léxica, também são analisadas as relações de coesão entre as partes e elementos do texto, e as relações de coerência entre sentenças.

- *Descoberta por Recuperação de Informações (RI)*

  RI é parte de um processo maior de exploração, correlação e síntese de informação. Suas técnicas podem ajudar apresentando documentos com visão geral das informações ou assuntos (RI tradicional) ou apresentando partes de documentos com detalhes de informações (recuperação por passagens). Existem ferramentas de RI por filtragem que contribuem garimpando documentos interessantes para seus usuários, sem que este precise formular consultas.

- *Descoberta Tradicional após Extração*

  É o tipo de descoberta mais simples. Nesta abordagem os dados são extraídos dos textos e formatados em bases de dados estruturadas, com o auxílio de técnicas de Extração de Informações (EI). Depois, são aplicadas técnicas e algoritmos de Mineração de Dados Estruturados (KDD) [18], no sentido de descobrir conhecimento útil para seus usuários.

  Basicamente, este processo segue os seguintes passos:

  1. Tratar o problema de erros de digitação nos textos do universo considerado;
  2. Recuperar documentos textuais que contenham as informações a serem estruturadas;
  3. Extrair as partes que interessam dos documentos recuperados;
  4. Extrair as informações destas partes com técnicas de EI;

5. Estruturar as informações coletadas para um formato próprio;
6. Extrair padrões nos dados coletados, com técnicas de descoberta de conhecimento;
7. Formatar a saída para o usuário (por exemplo, em linguagem natural).

- *Descoberta por Clusterização*

  A clusterização auxilia o processo de descoberta de conhecimento, facilitando a identificação de padrões (características comuns dos elementos) nas classes. Esta técnica pode ser utilizada para estruturar e sintetizar o conhecimento, quando este é incompleto ou quando há muitos atributos a serem considerados. Também pode ser utilizada para facilitar o entendimento e identificação de classes potenciais para descoberta de algum conhecimento útil. Geralmente, esta técnica vem associada com alguma técnica de descrição de conceitos, para identificar os atributos de cada classe.

- *Descoberta por Descrição de Classes de Textos*

  Dada uma classe de documentos textuais (já previamente agrupados) e uma categoria associada a esta classe (por exemplo, tema ou assunto dos textos), este tipo de descoberta busca encontrar as características principais desta classe, as quais possam identificá-la para os usuários e distingui-las das demais classes. Esta abordagem geralmente também segue as técnicas para construção do centróide de classes e pode ser utilizada em conjunto com a clusterização. Ela é diferente da abordagem por listas de conceitos-chave, porque descobre características comuns em vários textos e não em um único texto.

- *Descoberta por Associação entre Textos*

  Esta técnica procura relacionar descobertas presentes em vários textos diferentes. As descobertas estão presentes no conteúdo ou significado dos textos. Esta abordagem é diferente do que acontece na descoberta por associação entre passagens, cujo objetivo é somente relacionar partes de textos sobre o mesmo assunto. Na associação entre textos, a interpretação semântica é fundamental.

- *Descoberta por Associação entre Características*

  Esta abordagem procura relacionar tipos de informação (atributos) presentes em textos, aplicando a técnica de correlação ou associação tradicional em Mineração de Dados diretamente sobre partes do texto. Uma das diferenças é que os valores para os atributos são partes do texto e não necessariamente dados extraídos por técnicas de extração de informações.

- *Descoberta por Hipertextos*

  Um caso especial de descoberta utilizando técnicas de recuperação de informações (RI) é a descoberta com uso de hipertextos. Nesta abordagem, a descoberta é exploratória e experimental, feita através de mecanismos de navegação (*browsing*). Com tais ferramentas, é possível expandir e comparar o conhecimento através dos *links* que relacionam as informações, funcionando de modo análogo à mente humana (memória associativa). A aprendizagem pode ocorrer acidentalmente e de forma cumulativa, não exigindo estratégias cognitivas. A criatividade e a curiosidade guiam tal processo. Tal abordagem é útil quando os problemas de falta de informação são mal definidos e quando se quer explorar novos domínios.

- *Descoberta por Manipulação de Formalismos*

  Uma vez que é possível representar o conteúdo dos textos em formalismos, mecanismos de manipulação simbólica podem inferir novos conhecimentos, simplesmente por transformações na forma. As representações resultantes podem ser depois transformadas para estruturas na linguagem natural, facilitando a compreensão de usuários leigos no formalismo.

- *Descoberta por Combinação de Representações*

  Um caso especial da descoberta por associação entre textos é a descoberta por combinação de representações. A diferença é que os textos, antes de serem combinados, passam por um processo de representação interna. Então, na verdade, não são os textos que são combinados, mas sim seus conteúdos, conforme o formalismo e as regras internas. A combinação de representações diferentes, permite que pontos de vista diferentes possam ser usadas para criar novas representações e conseqüentemente novo conhecimento. Os formalismos internos podem ser modelos conceituais ou tradicionais (por exemplo, o modelo relacional) ou ontologias, linguagens baseadas em lógica, etc. A saída do processo de combinação deve estar representada em linguagem natural, podendo ser usadas técnicas de processamento de linguagem natural como as citadas anteriormente.

- *Descoberta por Comparação de Modelos Mentais*

  Esta abordagem procura representar documentos textuais e o estado de conhecimento do usuário (modelo mental das informações) em um formalismo padrão, para após compará-los. Se for possível verificar o que há nos documentos que falta no estado mental do usuário, então um conhecimento novo foi descoberto. O problema maior desta abordagem está na aquisição ou elicitação do conhecimento ou estado mental do usuário para poder representá-lo.

- *Descoberta por Análise de Seqüências Temporais*

  Segundo Palazzo et al. [17], esta técnica permite descobrir dependências entre conceitos que aparecem em textos dentro de uma mesma janela de tempo. O objetivo é saber se um conceito condiciona a aparição de outro no futuro. Os textos a serem analisados neste processo devem obrigatoriamente seguir uma ordem cronológica, formando uma seqüência temporal. Essas seqüências podem ser independentes, não havendo relação explícita entre textos de uma seqüência e de outra.

5 Conclusão

Mineração de textos faz parte da área de processo de descoberta de conhecimento, provendo técnicas efetivas de descoberta de conhecimento em bases de dados não estruturados.

Uma vez que a maioria dos dados disponibilizados, não somente na Internet, mas nas empresas em geral, é armazenado neste formato, este tipo de técnica possui um vasto campo de aplicação. Entretanto, alguns estudos [23] indicam que as tecnologias disponibilizadas por esta área ainda são muito pouco aplicadas na prática, i.e., fora do meio acadêmico.

Além disso, um dos principais problemas da área é a falta de técnicas efetivas de análise semântica de textos. Observa-se vários trabalhos que implementam análises estatísticas, entretanto, pouco se evoluiu em termos de semântica (existem diversas propostas teóricas, mas muito poucas foram implementadas na prática).

Isso ocorre principalmente pelo fato de que análise semântica de textos é muito difícil de ser realizada de fato, justamente pelas características destes textos. Logo, o futuro deste tipo de análise possivelmente estará relacionado ao fato de que textos não estruturados irão precisar, necessariamente, de vir acompanhados de alguma tipo de marca semântica, que auxilie na sua análise, seja através de arquivos de metadados ou *tags* explicativas.

Neste sentido, os formatos popularizados após o surgimento da Internet (HTML, XML, por exemplo) são mais adequados que os formatos tradicionais de armazenamento (DOC, PDF, TXT, por exemplo).

Entretanto, gerar conteúdos semânticos, que explicam conteúdos textuais, gera maior carga de trabalho para os usuários que produzem estes textos. Isso pode fazer com que nem sempre estas marcas semânticas sejam geradas ou a qualidade das mesmos não seja 100% confiável.

6 Agradecimento

Ao Prof. Dr. Cedric Luiz de Carvalho, pela avaliação do presente texto e pelas sugestões feitas, as quais muito contribuiram para a melhoria do texto original.

Referências

[1] AMO, S. **Técnicas de mineração de dados**. Universidade Federal de Uberlândia, Faculdade de Computação, disponível em `http://www.deamo.prof.ufu.br/`, 2003.

[2] BASTOS, V. M. **Ambiente de Descoberta de Conhecimento na Web para a Língua Portuguesa**. PhD thesis, Universidade Federal do Rio de Janeiro, COPPE, 2006.

[3] BEPPLER, M; FERNANDES, A. **Aplicação de text mining para a extração de conhecimento jurisprudencial**. In: PRIMEIRO CONGRESSO SUL CATARINENSE DE EDUCAçãO, 2005.

[4] CORDEIRO, A. D. **Gerador Inteligente de Sistemas com Auto-aprendizagem para Gestão de Informações e Conhecimento**. PhD thesis, Universidade Federal de Santa Catarina, Departamento de Engenharia da Produção, 2005.

[5] EBECKEN, N; LOPES, M; COSTA, M. **Mineração de Textos**, chapter 13, p. 337–370. Manole, 2003.

[6] FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. **From data mining to knowledge discovery in databases**. Ai Magazine, 17:37–54, 1996.

[7] FELDMAN, R; DAGAN, I. **Knowledge discovery in textual databases (KDT)**. In: KNOWLEDGE DISCOVERY AND DATA MINING, p. 112–117, 1995.

[8] FORSYTHE, G. E; MALCOM, M; MOLER, C. **Computer Methods for Mathematical Computations.** Prentice Hall, New Jersey, 1977.

[9] GLYMOUR, C; MADIGAN, D; PREGIBON, D; SMYTH, P. **Statistical themes and lessons for data mining**. Data Mining and Knowledge Discovery, 1(1):11–28, 1997.

[10] GROSS, G. **Eliminating semantic ambiguity by means of a lexicon-gramar**. Laboratoire de Linguistique Informatique. URA 1576, Universite de Paris 13, CNRS-INALF, 2005.

[11] LANDAUER, T; K., F. P. W; LAHAM, D. **Introduction to latent semantic analysis**. In: DISCOURSE PROCESSES, volume 25, p. 259–284, 1998.

[12] LOH, S. **Abordagem Baseada em Conceitos para Descoberta de Conhecimento em Textos**. PhD thesis, Universidade Federal do Rio Grande do Sul, Instituto de Informática, 2001.

[13] LOH, S; WIVES, L; FRANIER, A. **Recuperação semântica de documentos textuais na internet**. Programa de Pós Graduação em Computação - Univerisdade Federal do Rio Grande do Sul, 2004.

[14] LOPES, M. C. S. **Mineração de Dados Textuais Utilizando Técnicas de Clustering para o Idioma Português**. PhD thesis, Universidade Federal do Rio de Janeiro, 2004.

[15] MOURA, M. F. **Proposta de utilização de mineração de textos para seleção, classificação e qualificação de documentos**. Embrapa Informática Agropecuária, 2004, ISSN 1677-9274, 2004.

[16] ORENGO, V. M; HUYCK, C. **A stemming algorithm for the portuguese language**. In: EIGHTH INTERNATIONAL SYMPOSIUM ON STRING PROCESSING AND INFORMATION RETRIEVAL (SPIRE 2001), p. 186–193, 2001.

[17] PALAZZO, M. D. O; LOH, S; AMARAL, L. A; WIVES, L. K. **Descoberta de conhecimento em textos através da análise de seqüências temporais**. In: WORKSHOP EM ALGORITMOS E APLICAçõES DE MINERAçãO DE DADOS - WAAMD, SBBD, ISBN 85-7669-088-8, FLORIANóPOLIS: SOCIEDADE BRASILEIRA DE COMPUTAçãO, volume II, p. 49–56, 10 2006.

[18] REZENDE, S; PUGLIESE, J; MELANDA, E; PAULA, M. **Mineração de Dados**, chapter 13, p. 307–335. Manole, 2003.

[19] ROSA, J. **O significado da palavra para o processamento de linguagem natural**. In: ESTUDOS LINGüíSTICOS XXVII (ANAIS DOS SEMINáRIOS DO GEL), p. 807–812. Trabalho apresentado no XLV Seminário do GEL na UNICAMP, UNESP-IBILCE, 1998.

[20] SALTON, G; MCGILL, M. J. **Introduction to Modern Information Retrieval**. John Wiley and Sons, New York, 1983.

[21] SILBERCHATZ, A; KORTH, H; SUDARSHAN, S. **Sistema de Banco de Dados**. ELSEVIER, 5 edition, 2006.

[22] SILBERSCHATZ, A; TUZHILIN, A. **On subjective measures of interestingness in knowledge discovery**. In: KNOWLEDGE DISCOVERY AND DATA MINING, p. 275–281, 1995.

[23] SILVA, E. M. **Descoberta de conhecimento com o uso de *Text Mining*: Cruzando o abismo de moore**. Master's thesis, Universidade Católica de Brasília, Mestrado em Gestão do Conhecimento e da Tecnologia da Informação, 2002.

[24] SILVA, R. A. V; MARTINEZ, A. S; RUIZ, E. E. S. **Categorização e análise de informações médicas**. Technical report, Departamento de Física e Matemática Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto - FFCLRP, Universidade de São Paulo - USP, Brasil, 2006.

[25] TAN, A. **Text mining: The state of the art and the challenges**. Disponível em `http://citeseer.ist.psu.edu/tan99text.html`, acessado em Jul/06, 1999.

[26] WEISS, S. M; INDURKHYA, N. **Predictive Datamining: a pratical guide**. Morgan Kaufman, San francisco, CA, 1998.

[27] WIVES, L. **Tecnologias de descoberta de conhecimento em textos aaplicadas à inteligência competitiva**. Exame de Qualificação EQ-069, PPGC-UFRGS, 2002.

[28] WIVES, L. **Recursos de text mining**. Disponível em `http://www.inf.ufrgs.br/~wives/portugues/textmining.html`, acessado em Ago/2006, Dez 2005.

[29] ZANASI, A. **Competitive intelligence though datamining public sources**. In: SCIP, editor, COMPETITIVE INTELLIGENCE REVIEW, volume 9 de **1**. Alexandria, Virginia, 1998.

[30] ZENG, H; HE, Q; CHEN, Z; MA, W; MA, J. **Learning to cluster web search results**. SIGIR 04, p. 25–29, Mai 2004.