1
Mineração de textos

"A mente que se abre a uma nova idéia jamais retorna ao seu tamanho original."

Albert Einstein

Informações podem ser armazenadas das mais variadas maneiras, sendo as mais comuns sob formatos que utilizam linguagem natural. Primordialmente, dentre esses formatos tinha-se apenas o impresso e o acesso à informação estocada dessa forma é lento, difícil, e de pouco rendimento. Prescinde de interação humana, o que traz grandes limitações à capacidade de aquisição de conhecimentos em se tratando de grandes volumes de dados.

O advento dos computadores propiciou o surgimento de meios mais eficientes de armazenamento da informação. Os discos magnéticos são, no momento, os meios mais utilizados para esse armazenamento e combinam grande capacidade com alta velocidade de acesso. Já é viável para uma empresa média ter, em um escritório, uma capacidade de estocar o equivalente a uma biblioteca de porte razoável. A redução dos custos desse tipo de memória tem contribuído para uma verdadeira explosão de armazenamento de informações. Porém mais uma vez nos deparamos com problemas para processar essa imensa quantidade de informação disponível nos dias atuais. O simples aumento da velocidade de acesso, trazido pelas novas tecnologias de armazenamento magnético, não é suficiente para resolver essa questão. Nesse contexto se insere a pergunta: como extrair conhecimento útil oriundo desse enorme volume de informações, armazenado de forma tão distribuída e desorganizada? A solução para este problema passa pelas mais variadas formas de conhecimento envolvendo, por exemplo, técnicas de inteligência computacional, aprendizado de máquina, recuperação de informação, mineração de dados dentre outras. No caso específico das informações em forma de texto, ou seja, em formato desestruturado, uma nova área está ganhando cada vez mais força, mineração de textos.

A *Wikpedia*[1] define **mineração de textos**[2], também conhecida como **análise inteligente de textos**[3] ou ainda **descoberta de conhecimento em textos**[4] como o processo de se extrair conhecimento interessante e não-trivial de dados desestruturados. Para tal se aproveita, de forma interdisciplinar, de conhecimentos já citados e também pode utilizar como forte aliado a área de lingüística. Isso pode modificar a abordagem de utilização dos dados de entrada, atribuindo a estes, mais significado, sob a tutela de técnicas de **processamento de linguagem natural** (PLN). Em compensação, a utilização dessas técnicas tende a tornar os procedimentos dependentes da língua utilizada nos textos, pois cada língua possui propriedades e características lingüísticas próprias.

1.1
O que a Mineração de Textos é capaz de fazer

O grande crescimento de publicações em formato textual tem dificultado, e porque não dizer impossibilitado, que, até os leitores mais ávidos, se mantenham atualizados. Mineração de texto oferece um auxílio para esse problema através do fornecimento de sistemas automáticos para os leitores humanos desencorajados pela explosão de textos (REDFEARN, 2006). Isto envolve a análise de uma grande coleção de documentos para descobrir informação previamente desconhecida. A informação pode ser um relacionamento ou padrão que está escondido em uma coleção de documentos e que de outra forma seria extremamente difícil, se não impossível, de ser encontrada (REDFERAN, 2006). Conforme já citado, mineração de textos utiliza técnicas de várias áreas como **recuperação de informações**, onde um sistema identifica documentos, em uma coleção, enquadrados em uma consulta submetida pelo usuário; **processamento de linguagem natural,** onde o sistema analisa a linguagem utilizada pelos humanos de maneira a fornecer dados lingüísticos à próxima etapa do processo; **extração de informações**, onde o sistema obtém dados estruturados a partir de documentos desestruturados em linguagem natural; e **mineração de dados**, onde

[1] Enciclopédia livre que todos podem editar disponível na *web* em http://www.wikipedia.org.
[2] Do inglês, *text mining*.
[3] Do inglês, *intelligent text analysis*.
[4] Do inglês, *knowledge discovery in text – KDT*.

o sistema identifica padrões previamente desconhecidos, que serão úteis para a extração de algum tipo de conhecimento. Esses vários estágios do processo de mineração de textos podem ser combinados em um único **fluxo de tarefas**[5], e uma proposta para esse fluxo será apresentada mais adiante.

1.2
Elementos básicos e estruturação

O processo de mineração de textos é todo escorado em uma **coleção de documentos** ou ***corpus***. Essa coleção pode ser estática, ou seja, seu conteúdo permanece imutável, ou dinâmica, que é o termo aplicado a coleções de documentos caracterizadas pela inclusão de documentos novos ou atualizações com o decorrer do tempo (FELDMAN, 2007).

Outro elemento básico é o **documento**, propriamente dito. Para propósitos práticos um documento pode ser conceituado como uma unidade finita de dados textuais, que normalmente, mas não necessariamente, estão organizados dentro de uma coleção que os correlacionam segundo algum critério de agrupamento do mundo real como, por exemplo, relatórios, *emails*, artigos científicos, manuscritos e etc. Este exemplo de agrupamento do mundo real estabelece separações dos textos segundo categorias mais relacionadas ao tipo de documento, porém é importante destacar que a forma de agrupamento depende das características a serem consideradas. As coleções podem ser organizadas de acordo com os tipos de documento, assunto do texto ou segundo outro critério proposto, podendo ainda utilizar mais de um critério estabelecendo uma organização hierárquica, como no exemplo da figura 1:

[5] Do inglês, *workflow*.

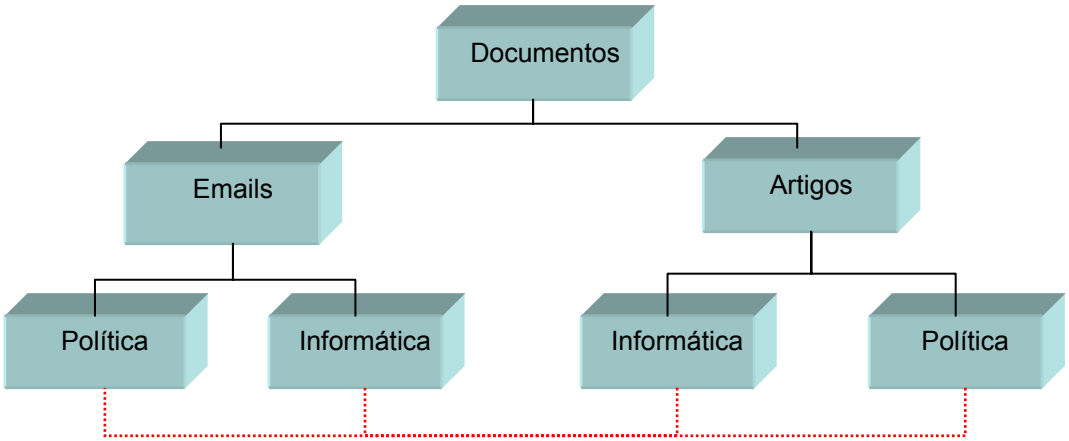

Figura 1 – Organização hierárquica com o tipo em nível superior ao assunto.

É interessante que a estrutura tenha flexibilidade para suportar união e interseção de grupos ou outras operações interessantes. Por exemplo, caso haja interesse na mineração de textos sobre política sem se importar com o tipo de documento, uma organização baseada na inversão do segundo e terceiro níveis seria mais apropriada ou a simples união de todos os documentos referentes ao ramo política. Dessa forma pode-se se pensar na reorganização hierárquica, conforme a figura 2, como sendo uma simples operação de união.

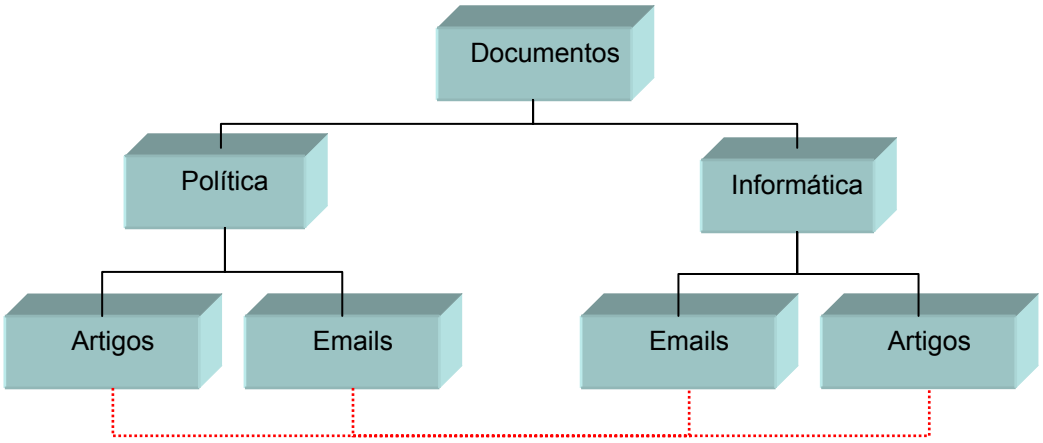

Figura 2 – Organização hierárquica com o assunto em nível superior ao tipo.

Um documento pode ainda pertencer a mais de uma coleção e de forma análoga pode haver o interesse na mineração de um subconjunto específico de

textos contido nas referidas coleções. Dessa maneira o sistema selecionaria o subconjunto através de uma operação de interseção. Outras operações referenciadas em teoria dos conjuntos podem também ser interessantes como complemento, por exemplo. Um exemplo seria a mineração de textos a um determinado assunto e que não fazem parte da coleção de outro determinado assunto. De qualquer maneira uma organização em grupos, seja ela hierárquica ou não, é bastante interessante para algumas das tarefas envolvidas na mineração de textos. Esses agrupamentos podem ser feitos manualmente ou automaticamente por intermédio de técnicas de classificação ou *clusterização*, temas que serão apresentados mais à frente. Podem ainda, tirar proveito de alguma estruturação prévia que o documento possa ter, como por exemplo, no caso de *emails*, que possuem campos específicos como emissor, assunto, etc. O assunto de um *email* costuma trazer características que o descrevam muito bem, assim como os textos das âncoras (estrutura interna de documentos HTML) em uma página *web*, que muitas vezes servem como um indício melhor do que o próprio título da página no que diz respeito a caracterização da própria página (BOYAN, 1996). Quanto às questões de estruturação cabe destacar alguns termos comumente utilizados. Documentos textuais sem nenhuma estrutura são ditos **desestruturados** ou em **formato livre**[6]. Porém esse conceito guarda alguns aspectos relativos. Por exemplo, um documento tipo XML que possua rótulos como `<documento>` e `</documento>` respectivamente no início e no fim do documento será possuidor de um tipo de estruturação. Mas que diferença, na prática, isso pode ter em relação a um documento composto apenas pelo texto entre esses rótulos? Pode-se dizer que quase nenhuma. Outro aspecto diz respeito à estrutura implícita que um documento pode conter. A simples organização em parágrafos, utilização de letras maiúsculas, pontuação, títulos, pode ser considerada uma forma de estruturação do ponto de vista lingüístico. De uma perspectiva lingüística, mesmo o mais inócuo documento demonstra uma rica quantidade de estrutura sintática e semântica, embora essa estrutura esteja implícita no conteúdo textual (FELDMAN, 2007). Alguns autores utilizam uma classificação diferenciada, denominando documentos textuais com algum tipo de rotulação na forma de

[6] Do inglês, *free format*.

metadados ou padronização rígida de formatação como sendo **semi-estruturados**[7] (BADR, 2001), (JIANWU, 2002) e (KUDO, 2004). Documentos do tipo XML podem ser considerados bons exemplos dessa categoria. Porém é preciso cuidado ao se fazer essa categorização de forma absoluta, pois os aspectos utilizados são bastante relativos. Da mesma forma que a etiquetagem de documentos XML, com rótulos apenas no início e no fim, não agrega quase nenhuma estruturação na prática, um documento XML pode ser etiquetado de forma a representar integralmente uma base de dados. Um documento XML é capaz de representar dados de várias tabelas, e até mesmo de várias bases, o que pode ser considerado o ápice de estruturação, pensando-se nos dados dispostos em um banco de dados como sendo o topo de uma organização estruturada (SHOLOM, 2005).

Um documento XML é uma base de dados, afinal (POWEL, 2007):

1. os elementos e atributos XML descrevem as propriedades dos dados, o que é equivalente a tabelas e campos em uma base de dados relacional;
2. um documento XML propriamente estruturado descreve relacionamentos entre diferentes tipos de dados em um conjunto de dados. Isso está mais próximo dos relacionamentos entre tabelas em uma base de dados relacional do que associações estabelecidas entre classes em um modelo de dados baseado em objetos.

É perfeitamente possível representar qualquer base de dados no formato XML utilizando como rótulos o nome da base, das tabelas e respectivas colunas. Um pequeno exemplo pode elucidar melhor essa questão. Supondo que as tabelas a seguir, Clientes e Produtos, representem uma base de dados chamada Loja.

Clientes	
Código	Nome
001	João
002	Pedro

Produtos		
Código	Preço	Quantidade
001	10	5
002	20	3

Figura 3 – Tabelas da base de dados Loja.

[7] (CHARKABARTI, 2003) classifica documentos tipo hipertexto como sendo desestruturados ou semi-estruturados, por não possuírem uma descrição precisa dos dados. Essa descrição é chamada de *schema* e é um requisito mandatório nas bases de dados relacionais.

Podemos ter um documento XML que represente integralmente essa base de dados escrito da seguinte forma, por exemplo:

```
<loja>
   <clientes>
      <registro>
         <codigo>001</codigo>
         <nome>João</nome>
      </registro>
      <registro>
         <codigo>002</codigo>
         <nome>Pedro</nome>
      </registro>
   </clientes>
   <produtos>
      <registro>
         <codigo>001</codigo>
         <preco>10</preco>
         <quantidade>5</quantidade>
      </registro>
      <registro>
         <codigo>002</codigo>
         <preco>20</preco>
         <quantidade>3</quantidade>
      </registro>
   </produtos>
</loja>
```

Figura 4 – Documento XML representando a base de dados Loja.

É importante enfatizar que o documento da figura 4 é apenas uma das formas. Poder-se-ia ainda utilizar os nomes das colunas de uma tabela como sendo elementos do rótulo registro da seguinte forma:

```
<registro codigo="001" preco="10" quantidade="5">
```

Tendo em vista tais questões, os arquivos XML podem ter uma estruturação variante dentro de um espectro bastante grande, e o termo semi-estruturado até lhe cai bem considerando o seu papel intermediário, mesmo ele podendo alcançar os dois limites extremos de estruturação.

Qualquer informação que contribua com a estruturação é bem vinda no auxílio às tarefas referentes à mineração de textos. No caso mais extremo, onde os dados estejam completamente estruturados, mineração de textos se torna puramente mineração de dados, que é relativamente mais fácil e de certa forma está mais consolidada. Arquivos XML também contribuem sobremaneira com o trabalho de mineração de textos devido à estruturação inerente dada pelas suas etiquetas.

Chegando ao nível mais atômico existem os **termos** ou *tokens* (veja seção 2.2.1 – *Tokenização*), que de forma mais simplória poderiam ser representados pelas palavras, menores unidades sintáticas de uma língua, que não poderiam ser quebradas em segmentos menores. Foi dito de forma simplória porque na seção 2.2 – Processamento de Linguagem Natural - serão apresentadas algumas técnicas para construção dos termos que prevêem tanto a "quebra" das palavras, reduzindo-as a sua forma raiz (morfema principal) quanto à reunião delas para a formação de um único termo. O conjunto de todos os termos de um determinado universo composto pelo *corpus*[8] usados é conhecido como dicionários ou léxico (por conter todos os lexemas).

1.3
Ferramentas para Mineração de Textos

Dentro do objetivo mais amplo de descoberta de conhecimento em textos podemos citar algumas aplicações que contribuem para essa meta final, como classificação, agrupamento e organização de documentos; predição e avaliação de novas respostas a partir de novos exemplos; recuperação e extração de informações.

1.3.1
Classificação de Documentos

Documentos são organizados em tópicos pré-rotulados. A cada novo documento apresentado um elemento categorizador do sistema decide sobre o

[8] O plural de *corpus* é *corpora*.

tópico mais propício a abrigar o documento. Em geral classificação é o processo de aprendizado de uma função que mapeia os dados de entrada em uma ou diversas classes de saída.

Em mineração de dados a metodologia aplicada em diversas aplicações do mundo real, como soluções poderosas para o problema de classificação, é baseada em árvores de decisão ou regras de decisão (KANDARTZIC, 2003). Em sua forma mais simples a classificação é binária, ou seja, um documento é classificado como pertencente ou não a determinado tópico, porém algumas máquinas de aprendizado podem ser capazes de categorizar documentos entre diversas classes distintas. Uma abordagem possível para classificação dentre diversas categorias é a utilização de um classificador binário específico para cada categoria. Nesse caso a escolha da categoria mais adequada é dada através de um sistema de pontuação obtido por cada classificador. Outra abordagem é a utilização de comitês de classificadores, em que saídas de diferentes classificadores individuais são combinadas de várias formas, podendo produzir resultados mais acurados do que os obtidos pelos elementos participantes do comitê de forma individual. Os comitês mais comumente utilizados são os baseados em redes neurais (ELIS, 1997) e *Support Vector Machine* (JEREBKO, 2005).

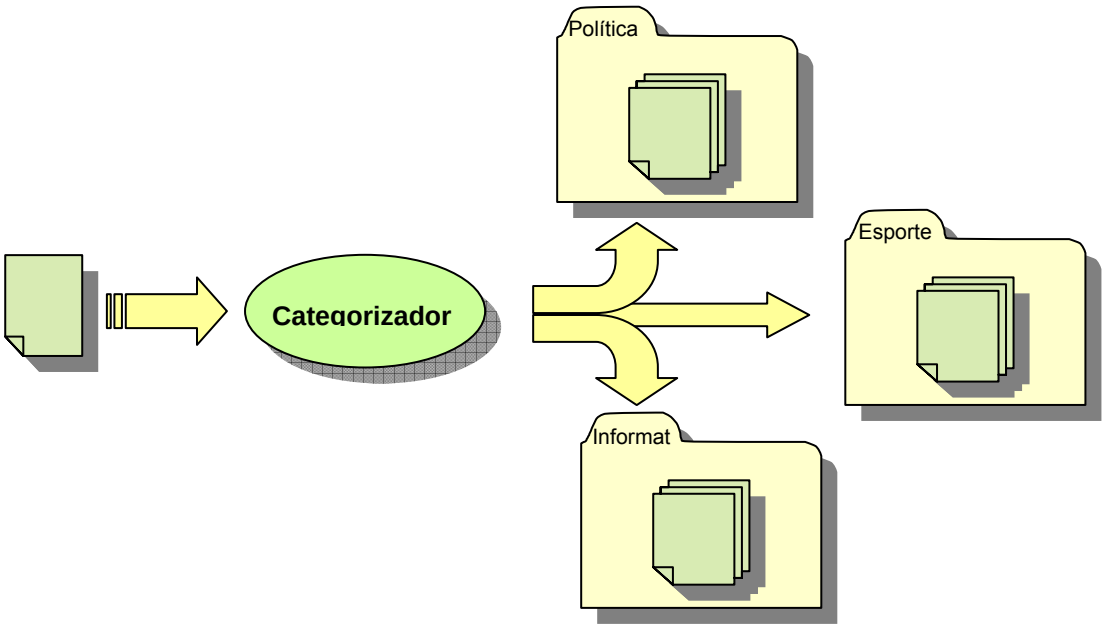

Figura 5 – Classificação ternária de documentos.

1.3.2
Agrupamento de Documentos

Dada uma coleção de documentos, um elemento organizador estabelece uma organização em grupos de forma que esses grupos contenham documentos similares (figura 1). A designação dos itens nos grupos é baseada no cálculo do grau de associação entre eles de maneira que os grupos formados tenham um alto grau de associação entre seus membros e baixo grau entre os membros de grupos diferentes. Vários modelos (veja capítulo 3 – Recuperação de Informações Textuais) e métricas de similaridade (veja seção 4.3 – Medidas de Similaridade), que determinam o grau de associação entre documentos podem ser utilizados (DUDA, 2000) e (RIJSBERGEN, 1999). Dentre os modelos destaca-se o **vetorial** (veja seção 3.3 – Modelo Vetorial) com medidas de similaridade dadas pelo co-seno do ângulo formado pelos vetores de características que representam os documentos. Existem diversas outras formas de se medir a similaridade de dois documentos (DUDA, 2000) e (GIUDICI, 2003).

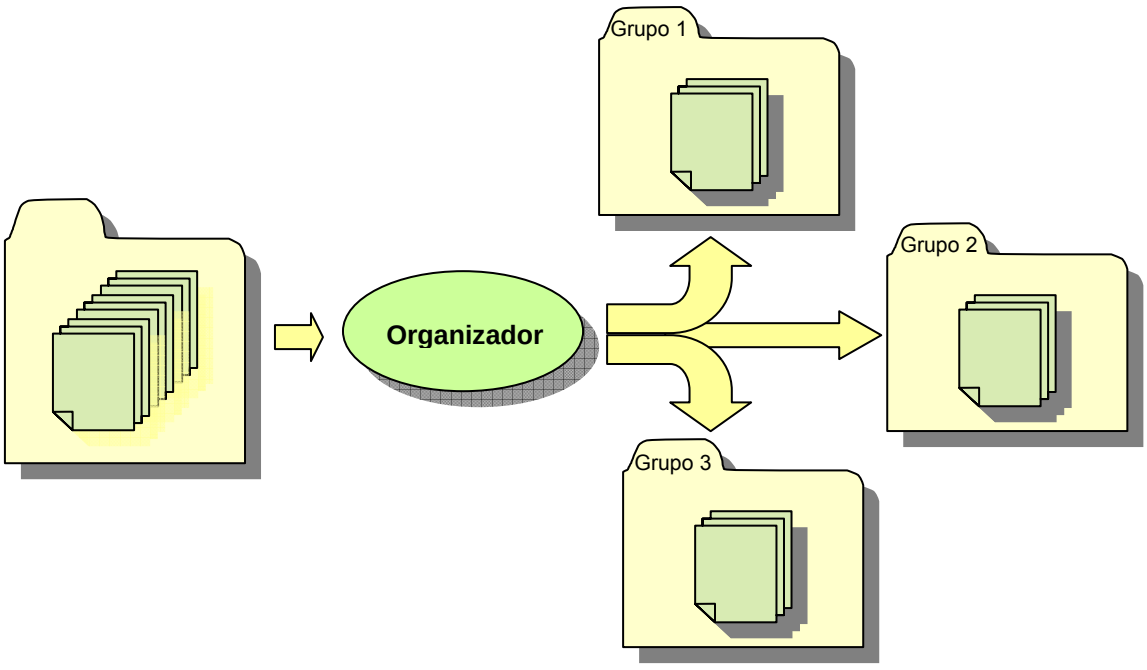

Figura 6 – Agrupamento não-hierárquico de documentos em 3 grupos.

Os métodos de agrupamento costumam ser diferenciados de acordo com o tipo de estrutura gerada, podendo ser hierárquicos ou não-hierárquicos (DUDA,

2000). Os **não-hierárquicos** são mais simples, apenas dividindo os ***N*** documentos em ***M*** grupos sem sobreposição. Cada item pertence ao grupo que melhor represente suas características. Esses métodos são heurísticos por natureza, pois certas decisões precisam ser tomadas *a priori* como número de grupos, tamanho do grupo e critério de agrupamento. São mais utilizados quando os recursos computacionais são limitados e os mais conhecidos e utilizados são *K-means* (HARTINGAN, 1979), *Fuzzy C-means* (DUNN, 1973), dentre outros.

Já os **hierárquicos** são mais complexos, produzindo um conjunto de dados aninhados, onde pares de itens são sucessivamente ligados (ou separados) até que todos os itens do conjunto estejam conectados (ou desconectados). Os métodos hierárquicos podem ser ou **aglomerativos**, onde aos ***N*** documentos são aplicadas ***N-1*** junções de pares de documentos ou grupos anteriormente não agrupados; ou **divisivos**, começando com todos os objetos em um mesmo grupo e ***N-1*** divisões progressivas em grupos menores. A representação mais natural para agrupamentos hierárquicos corresponde a uma árvore conhecida como **dendograma**, que mostra como os documentos são agrupados (DUDA, 2000). A figura 7 apresenta, como exemplo, um dendograma para agrupamento de 6 documentos (***D1***, ***D2***, ..., ***D6***).

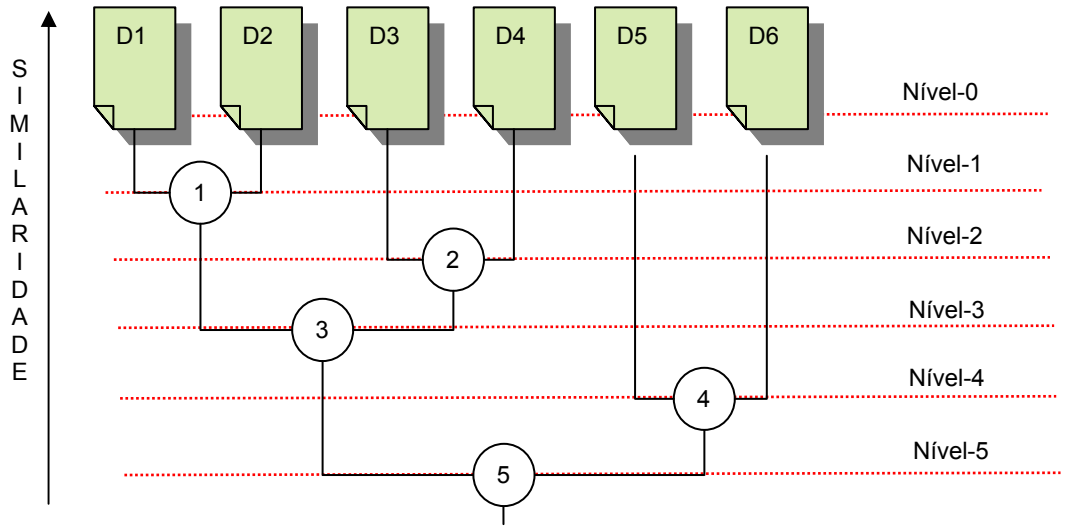

Figura 7 – Agrupamento de não-hierárquico aglomerativo de documentos (dendrograma).

O nível-0 mostra os grupos compostos por cada documento separado, conhecidos como *singleton*. Já no nível-1 se formaram dois grupos, ambos com

um par de documentos cujos graus de similaridade entre eles mais se destacava, um agrupando os documentos ***D1*** e ***D2***, e outro com os documentos ***D3*** e ***D4***. Procede-se dessa forma até que todos os documentos pertençam a um mesmo grupo (método aglomerativo). É possível medir a similaridade dos agrupamentos em cada nível, que vai caindo na medida em que grupos com mais documentos vão sendo formados. Pode-se dizer que no nível-0 a escala de similaridade seja máxima, ou 100 considerando-se termos percentuais, pois todo documento é completamente similar a si mesmo. A origem do uso desse tipo de agrupamento é atribuída a (JOHNSON, 1969). A cada iteração, quando dois documentos ou grupos são reagrupados, um novo valor de similaridade é calculado para esse agrupamento. Existem várias formas para se fazer esse cálculo e algumas delas serão apresentadas no capítulo 4.

1.3.3
Predição e Avaliação

Tem como objetivo projetar, de uma amostra de exemplos conhecidos *a priori*, novos exemplos ainda não vistos ou concluir algo sobre um novo exemplo apresentado. Utiliza-se de aprendizado de máquinas que estuda os documentos e encontra algumas regras generalizadas capazes de dar respostas corretas sobre novos exemplos. Nesse ramo de pesquisa destacam-se as mais variadas técnicas de **aprendizado de máquina**[9] (MITCHELL, 1997) e técnicas de inteligência computacional. Essas máquinas aprendem e são capazes de realizar inferência e tirar conclusões a respeito de determinada situação, baseadas em alguma forma de aprendizado sobre dados representativos. Como exemplos de aplicações em mineração de textos destacam-se ***Support Vector Machine* - SVM** (JOACHIMS, 1998) e (BURGES, 1998), ***Tranformation Based Learning* - TBL** (BRILL, 1995) e (FLORIAN, 2001), ***Hidden Markov Model* - HMM** (RABINER, 1989), (SEYMORE, 1999) e (CAPPÉ, 2005), **Redes Neurais Artificiais - RNA** (HAIKIN, 1999), dentre outras.

[9] Do inglês *machine learning*.

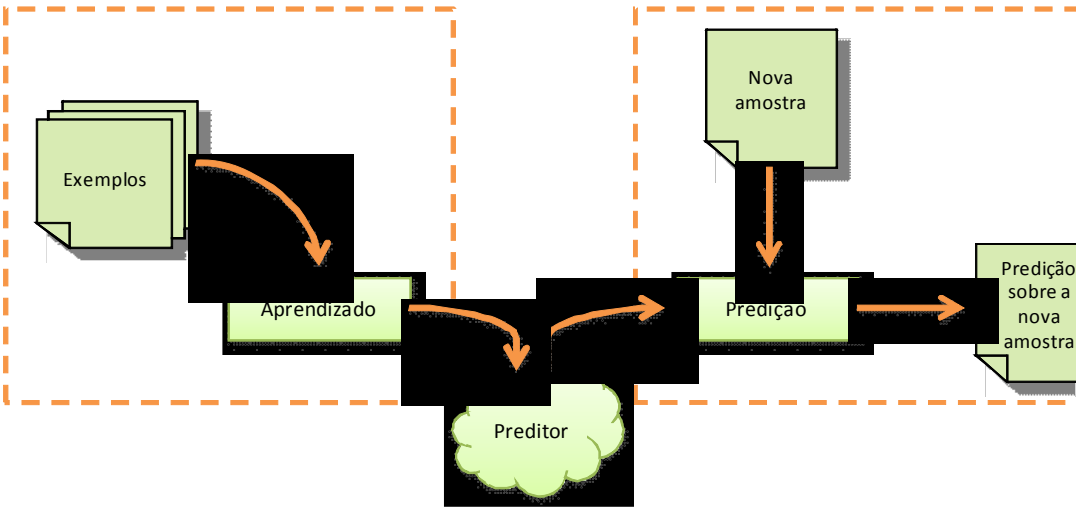

Figura 8 – Processo de predição pelo aprendizado de máquina.

1.3.4
Recuperação de Informações

Com base em uma coleção de documentos obtida e armazenada de forma conveniente recuperam-se documentos relevantes a partir de termos fornecidos como entrada ao sistema. O sistema realiza um casamento entre as palavras passadas, seja diretamente de um usuário ou coletadas de outro documento, e identifica outros documentos relevantes segundo algum critério de similaridade (DUDA, 2000). É importante destacar que em uma fase anterior a indexação desses documentos deve ser provida para compor um sistema de recuperação de informações. Os Principais exemplos de aplicações desse tipo são as máquinas de busca na *Web*. Mais uma vez, assim como nas aplicações de agrupamento, medidas de similaridade são utilizadas, só que agora para efetuar os casamentos para escolha dos documentos a serem recuperados. O usuário de um sistema de recuperação de informações tem que traduzir a informação procurada em uma linguagem de consulta provida pelo sistema. Normalmente, existem duas tarefas que podem ser executadas por um usuário que utilize um sistema de recuperação de informações: **busca**, quando os interesses do usuário não estão claramente definidos e podem ir mudando na medida em que ele interage com o sistema, algo muito comum em sistemas baseados em *hiperlinks*; **recuperação**, quando o usuário submete uma expressão de consulta (algumas vezes incluindo expressões

regulares[10]) utilizada para comunicar as restrições a serem satisfeitas para que o objeto a ser procurado, pertença ao conjunto de resposta. Sistemas clássicos de recuperação de informação normalmente estão associados apenas à tarefa de recuperação. Sistemas hipertexto são normalmente ajustados a prover busca rápida. Bibliotecas digitais modernas e interfaces *web* podem tentar combinar essas tarefas para trazer incremento à capacidade de recuperação por parte dos sistemas mais modernos. Esse novo paradigma de combinação entre as tarefas de busca e recuperação vem ganhando força com as pesquisas para aumento da inteligência dos motores de busca na *web*. De maneira geral as tarefas associadas ao processo de recuperação de informação mediante consulta fornecida por usuário podem ser descrito pela figura 9 (BAEZA-YATES, 1999).

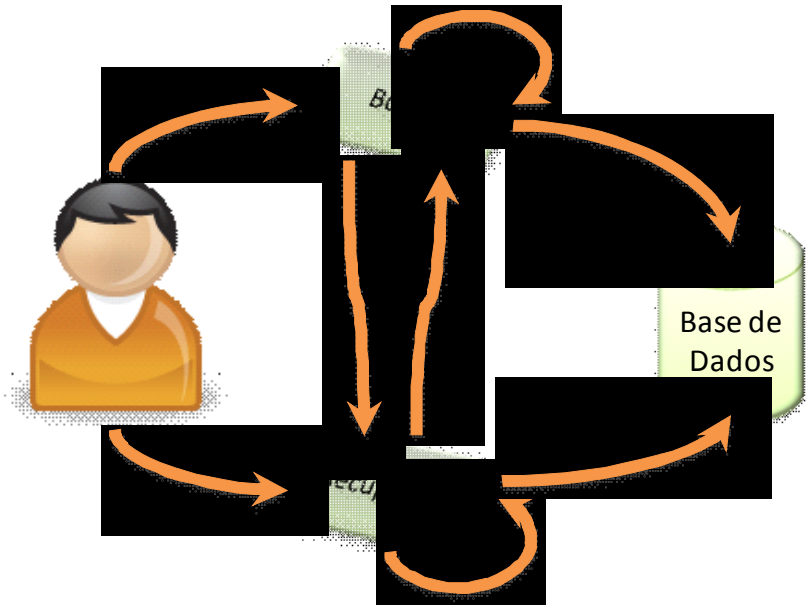

Figura 9 – Processo de recuperação da informação.

[10] Expressões expressas por meio da teoria de linguagens formais, que consistem de operadores indicativos de conjuntos de cadeias de caracteres e operações e operações sobre esses conjuntos, provendo uma forma concisa e flexível de identificar padrões nesses conjuntos de caracteres.

1.3.5
Extração de Informações

Tem como objetivo obter automaticamente, a partir de informações desestruturadas, um modelo estruturado capaz de passar por procedimentos de mineração de dados. Em outras palavras visa a extrair dados de documentos de forma a preencher as células de uma tabela, atribuindo, dessa maneira, estrutura às informações, conforme apresentado na figura 10. Com isso tenta-se migrar de mineração de textos para mineração de dados, ou seja, trabalhar num universo estruturado. É muito comum em mineração de texto a utilização de técnicas de extração de informação para obtenção de entidades dentro de um texto como, por exemplo, nomes de pessoas, organizações, datas e etc.

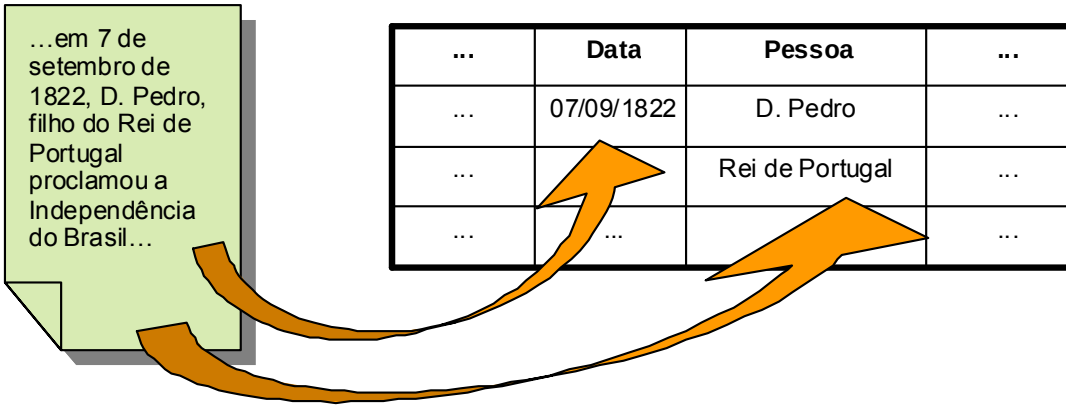

Figura 10 – Extração de informações de um documento.

O Instituto Americano de padronizações (NIST) realizou, em 1987, uma pesquisa em forma de competição entre grupos visando desenvolver e avaliar técnicas de extração de informações – MUCs (*Message Understanding Conference*). Como resultado apresentou algumas padronizações e conceitos para que os pesquisadores da área se comunicassem em uma mesma linguagem (DRAEGER, 2007). Dentre os conceitos básicos, de forma resumida se destacam (NIST, 2003):

a) *Corpus* – coleção de textos que guardam afinidades entre si por possuírem características que exemplifiquem de maneira mais ampla possível aquele conjunto lingüístico que se deseja investigar.

b) Entidade – todo objeto que seja de interesse na tarefa de extração, estando portanto, intimamente ligada ao domínio da tarefa de extração de informação. No exemplo da figura 10 temos duas entidades, data e pessoa.

c) Fato – relacionamento existente entre duas ou mais entidades dentro de um contexto. No exemplo da figura 10 o fato poderia ser D. Pedro filho do Rei de Portugal, por estabelecer uma relação entre duas entidades, através do termo filho.

d) Evento – possui uma semântica temporal e pode ser entendido como uma ação que ocorreu em algum ponto da linha do tempo. No exemplo da figura 10 o fato poderia ser Proclamação a República.

e) *Template* – está diretamente relacionado ao formato estruturado dos dados de saída do processo de extração da informação e intuitivamente poderia ser encarado como uma ficha com lacunas a serem preenchidas pelas informações extraídas. A estruturação pode ser feita através o uso de uma tabela conforme apresentado na figura 10.