TEXT-MINING: TÉCNICAS DE EXTRAÇÃO DE CONHECIMENTO EM TEXTO BASEADAS EM APRENDIZADO DE MÁQUINA

André Felipe Borba*

RESUMO: *A evolução tecnológica, juntamente com a popularização do acesso aos chamados microcomputadores, trouxe consigo um novo formato de produzir, comunicar, armazenar e disseminar informações: o documento eletrônico. Entretanto a recuperação e extração de conhecimento, armazenado em meio magnético não é uma tarefa trivial, o que incentiva o desenvolvimento dos chamados Sistemas de Recuperação de Informações (SRI), cujo objetivo é recuperar informações tidas como relevantes para o usuário e armazenadas em dispositivos mediados por computadores. A natureza destes sistemas baseia-se normalmente em técnicas de Text-Mining (mineração de textos). Assim, o presente artigo pretende apresentar uma proposta de criação de um SRI para extrair informações textuais armazenadas em documentos eletrônicos, usando conceitos oriundos de área de Inteligência Artificial, a qual tem dotado a área de informática com uma série de metodologias para desenvolvimento de sistemas inteligentes capazes de conferir a equipamentos microprocessados a capacidade de aprender conceitos de alto nível relativos ao domínio de conhecimento humano (Aprendizado de máquina).*

Palavras-chave: Text-Mining; Aprendizado de máquina; Documento eletrônico.

INTRODUÇÃO

No passado, com o surgimento das bibliotecas, nascia o profissional responsável por organizar e armazenar os documentos (livros, manuscritos, etc.). Este profissional, conhecido como bibliotecário, passou a preocupar-se também com o aprimoramento dos meios de armazenamento destas informações afim de que pudéssemos ter uma maior conservação das mesmas e uma maior facilidade no momento em que precisássemos recuperá-las.

Para facilitar a recuperação, o que se fazia era uma organização dos documentos de acordo com regras pré-estabelecidas que, por exemplo, poderiam ser por assunto. Podemos dizer então que se fazia uma indexação dos documentos, uma indexação, segundo Rocha (2002) é o "ajuste de algo de acordo com um índice cuja variação pode ser determinada".

Com o sucessivo aprimoramento das técnicas computacionais de gerenciamento de documentos, questões relacionadas à localização e acesso às informações passaram a ser estudadas pela área de Recuperação de Informações (Information Retrieval) que, de acordo com Yates (1999), trata do armazenamento, da organização e do acesso às informações que sejam de interesse do usuário.

Em nossos dias, a produção de documentos eletrônicos apresenta-se dinâmica e sofisticada, promovendo o surgimento de vários formatos de documentos, como dados, imagens, textos, vídeos, dentre outras. Desta forma, várias técnicas surgiram com o intuito de estudar e promover a interação com estes novos formatos da informação. Podemos citar, dentre eles, a técnica de Mineração de Dados (Data Mining), para tratar dados, e a Mineração de Textos (Text Mining), para tratar textos. Esta é a abordagem considerada adequada para implementação do

* Acadêmico do Curso de Bacharelado em Informática da Universidade Católica do Salvador – UCSal. E-mail: borba.andre@terra.com.br. Orientador: Professor Arnaldo Bispo de Jesus. E-mail: arnaldo@nedael.org.

modelo para o qual se propõe este trabalho, onde se pretende aliar as técnicas de mineração de textos com a metodologia de aprendizado de máquina proposta pela Inteligência Artificial (IA). Os fundamentos que norteiam cada técnica bem como a integração entre elas é o foco do estudo teórico empreendido, o qual será descrito mais detalhadamente ao longo deste trabalho, de acordo com seguinte estrutura: Seção 1 – Recuperação de informações; Seção 2 – Descoberta do conhecimento em textos; Seção 3 – Aprendizado de máquina; Seção 4 – Arquitetura proposta.

RECUPERAÇÃO DE INFORMAÇÕES

O entendimento dos princípios da recuperação de informações podem ser entendidas de acordo com Wives (2002):

> Calvin Moores[1] define Recuperação de Informações como sendo uma atividade que envolve os aspectos de descrição de informação (indexação, padronização) e sua especificação para busca, além de qualquer técnica, sistema ou máquina empregada para realizar ou auxiliar estas tarefas.

Com o constante desenvolvimento da tecnologia, novos tipos de documentos como, por exemplo, páginas de internet e imagem, foram surgindo. Este fato é responsável pelo surgimento de novas técnicas de RI. Os Sistemas de Recuperação de Informações (SRI) têm como principal objetivo ser a interface entre um usuário e os documentos de uma base, visando o auxílio na recuperação da informação requisitada. Estas precisam ser relevantes à solicitação do usuário para que o mesmo não precise efetuar a leitura de todos os documentos ou informações disponíveis para encontrar a informação que deseja.

A relevância da informação, que consiste na relação entre o que foi solicitado e o que foi apresentado, é de fundamental importância. Um elemento que dificulta o processo é o fato de que o usuário muitas vezes não consegue expressar com clareza a sua necessidade. Desta forma o SRI pode apresentar informações que são relevantes para a descrição feita pelo usuário, mas não para a necessidade do mesmo.

A figura abaixo descreve o processo de um SRI:

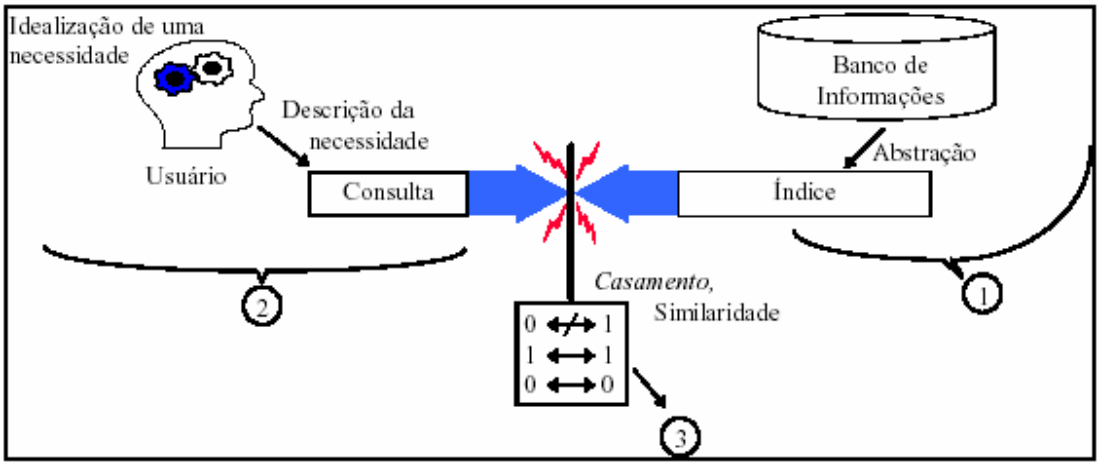


Figura 1 - Paradigma de Recuperação de Informações. Fonte: (WIVES, 1997-A, P.12)

[1] Empresário que trabalhava na área de RI, por volta de 1950, sendo um dos pioneiros neste tipo de estudo.

> "Na figura 1, o primeiro passo é o processo de abstração das informações, determinada pela modelagem do sistema. O segundo passo é decorrente da abstração que o usuário faz ao descrever a informação do que necessita. E, o último passo é o processo de casamento (Matching) que o sistema faz entre a consulta do usuário e as informações do sistema, a fim de determinar quais informações são relevantes". (WIVES, 1997-A, P.12)

Os problemas ocorrem pela complexidade em se trabalhar com Linguagem Natural[2], a qual é utilizada nos textos armazenados e na descrição das consultas. Na abstração das informações, o SRI deve identificar o que é mais relevante, ou seja, o que caracteriza aquela informação. Após esta identificação, torna-se necessário idealizar algum formalismo que possa descrevê-la e armazená-la.

É uma tarefa difícil armazenar o conteúdo da informação, mas extremamente importante para o funcionamento do SRI. Uma das técnicas que pode ser utilizada é a indexação automática, pois, através dos índices que serão gerados, as informações poderão ser acessadas.

Não só o sistema precisa estar preparado para as abstrações, o usuário também deve ter a capacidade de descrever suas necessidades de forma clara e precisa a fim de recuperar as informações relevantes às suas necessidades. Torna-se fundamental que se façam alguns esclarecimentos para o usuário de como o sistema trata as informações, visando o aperfeiçoamento da descrição da necessidade do mesmo.

Um dos problemas que ocorrem neste momento é o conhecido problema do vocabulário (Vocabulary Problem) descrito em Wives (1997-A), no qual várias pessoas podem descrever uma necessidade de forma diferente. Para solucionar este problema, pode ser utilizado o dicionário de sinônimos na avaliação da solicitação.

Outro auxílio para a descrição de uma solicitação pode ser dado também pelo Thesaurus, que segundo Wives (1997-B, p.44) "é uma estrutura hierárquica de palavras, permitindo que o usuário descubra os relacionamentos entre as palavras". Os sistemas precisam estar preparados para auxiliar os usuários a efetuarem suas consultas, visando uma maior eficiência do resultado.

A última etapa do processo é o mecanismo que efetua a identificação das informações que são relevantes para a consulta especificada. Um problema nesta etapa pode ocorrer devido ao problema tanto da abstração das informações armazenadas, quanto da abstração do usuário ao especificar uma consulta, o que pode resultar na perda de informações importantes.

Neste momento, podemos, então, medir a eficiência do sistema, analisando os itens de abrangência e precisão no resultado apresentado.

Na avaliação da abrangência, estaremos verificando a proporção de itens relevantes recuperados na consulta. Wives (1997-A) representa abrangência pela fórmula:

$$\textbf{Recall} = \frac{\textbf{Total de Informações relevantes encontradas}}{\textbf{Total de informações relevantes no sistema}}$$

A precisão medirá a proporção de itens recuperados que são realmente relevantes para o usuário. Utilizaremos a fórmula, proposta por Wives (1997-A):

$$\textbf{Precision} = \frac{\textbf{Total de informações relevantes encontradas}}{\textbf{Total de informações encontradas}}$$

[2] Linguagem Natural deve-se ao fato de ser a linguagem normalmente utilizada pelo homem para comunicar-se (exemplo: Português, Inglês, Alemão...).

Na figura 2, teremos uma idéia das combinações possíveis para a eficiência do software.

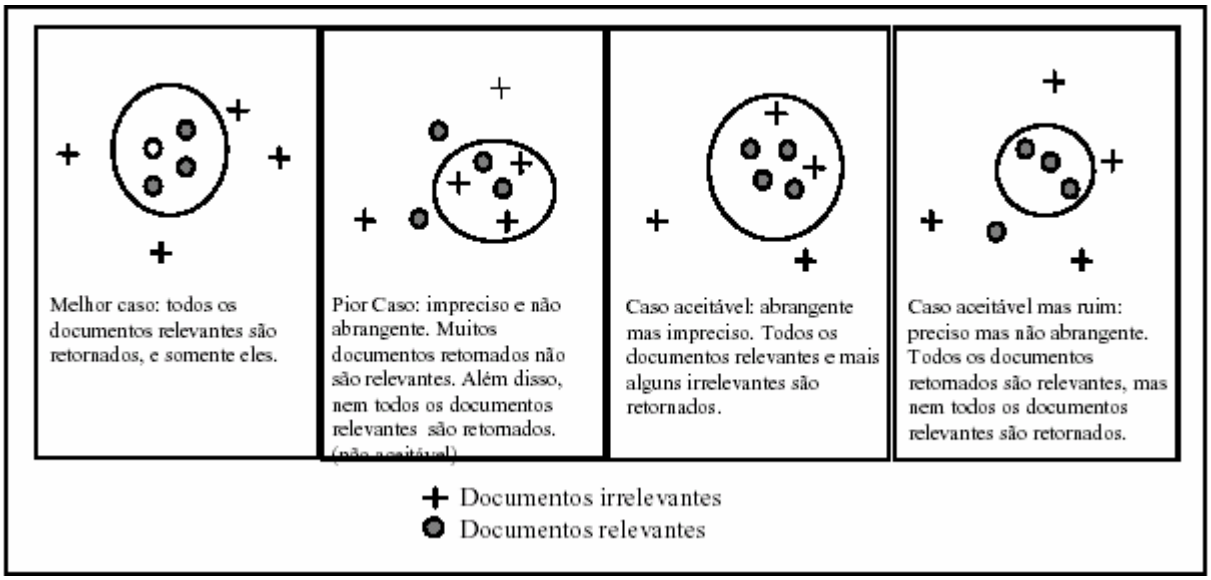


Figura 2 - Relações entre Recall e Precision. (WIVES, 1997-A, p.17)

DESCOBERTA DE CONHECIMENTO EM TEXTOS

Na grande maioria dos documentos codificados em forma textual, existem conhecimentos implícitos, ou seja, que não estão expressos claramente. Com ferramentas adequadas, todo este conhecimento pode ser inferido, mas isto não é uma tarefa simples, pois envolve um conjunto de técnicas diferentes até a obtenção do resultado final.

Como tudo o que ocorre na área de informática surge a partir da evolução de alguma técnica ou estudo já existente, com a descoberta de conhecimento em textos (Knowledge Discovery from Text - KDT) não foi diferente. O surgimento ocorreu devido ao desenvolvimento da área de Recuperação de Informações.

De acordo com Loh (1999), "pode-se definir Descoberta de Conhecimento em Textos (KDT) ou Text Mining como sendo o processo de extrair padrões ou conhecimento, interessantes e não-triviais, a partir de documentos textuais".

O processo de mineração de dados ou de textos tem como foco principal a descoberta de padrões e relacionamentos entre os dados e os textos envolvidos no processo.

Como falado anteriormente, temos atualmente uma sobrecarga de informações, ou seja, uma grande quantidade de informações que, mesmo com uma ferramenta de Recuperação de Informações, apresentando documentos relevantes em uma busca, fica difícil para o ser humano tratá-las por completo. Por este motivo, é de grande importância o desenvolvimento de ferramentas de KDT, visando os benefícios destas para a evolução do conhecimento humano. Podemos citar, por exemplo, a ferramenta de apoio à decisão utilizada em empresas para o auxílio da tomada de decisões.

APRENDIZADO DE MÁQUINA

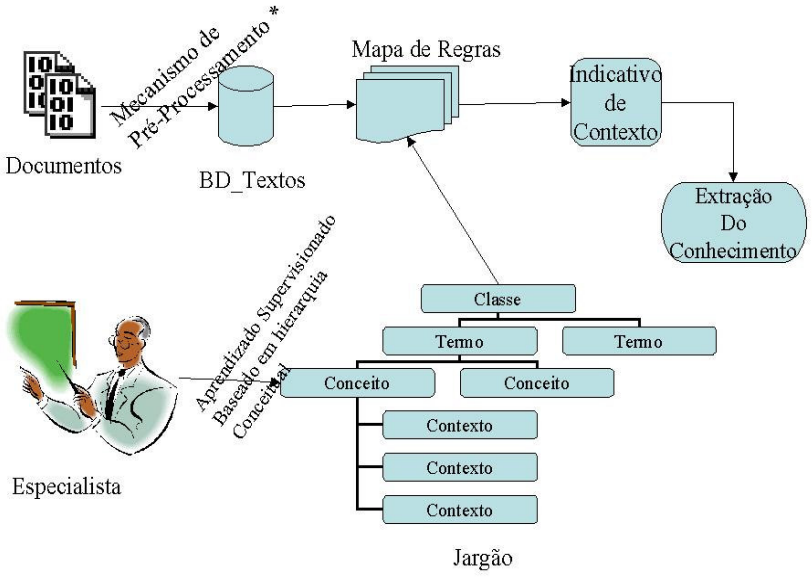
Os chamados sistemas inteligentes marcam uma nova era no desenvolvimento de software. Existe uma real demanda da sociedade mundial por melhores sistemas que apoiem o

processo de tomada de decisão e permitam um acesso mais rápido e eficiente a informações armazenadas em meio magnético (POMERO 1997). Neste sentido, podemos ainda afirmar que existe uma real e premente demanda por desenvolvimento de Sistemas Especialistas (SE). Sistemas especialistas, conforme Booty (2001), são sistemas projetados e desenvolvidos para atender a uma aplicação determinada e limitada do conhecimento humano. "São capazes de emitir uma decisão, apoiado em conhecimento justificado, a partir de uma base de informações, tal qual um especialista de determinada área do conhecimento humano".

Como tal, espera-se que estes sistemas evoluam com o tempo, que se adaptem aos diversos contextos onde são inseridos, ou seja, que sejam capazes de aprender, sobretudo com suas limitações (POLIKAR 2001). A solução que se propõe neste trabalho pretende fundamentar-se neste conceito de aprendizado de máquina, permitindo então ao especialista a construção de campos de domínio que atendam aos requisitos esperados.

ARQUITETURA PROPOSTA

O modelo proposto atende ao conceito de jargão modelado por Rocha (1992), atuando como uma camada entre uma rede neural e um sistema especialista. O jargão é o mecanismo responsável por fazer as associações entre o mapa de regras e o modelo conceitual introduzido pelo especialista. A figura 3 ilustra a arquitetura do modelo.



•O Mecanismo de Pré-Processamento consiste na Identificação das Palavras, Remoção de StopWords e Verificação de Sinônimos

Figura 3 - Arquitetura do modelo proposto

Este modelo pode ser descrito da seguinte forma:

- os documentos a serem investigados constituem a fonte ou base de dados de entrada;
- o mecanismo de pré-processamento é responsável por extrair do texto os elementos para análise;
- os elementos filtrados são armazenadas no repositório chamado base de textos;
- o mapa de regras associa os elementos entre si, gerando o indicativo de contexto;

- as regras aplicadas pelo jargão fornecem os elementos para a saída do sistema, ou seja, o **conhecimento** extraído dos textos de entrada.

CONCLUSÕES

O volume de produção de documentos e arquivos magnéticos de conteúdo textual, gerados por computador, nos dias atuais, é expressivo. Sabemos que, cada vez mais, documentos eletrônicos são gerados e disponibilizados em ambientes computacionais heterogêneos dentro das organizações públicas e privadas, independente do caráter de sua função social.

A produção desordenada destes documentos, aliada à necessidade atual sempre crescente de informatização de processos documentais, podem levar estas organizações a produzirem documentos em duplicidade, oneram os custos de processamento e armazenamento de arquivos, dificultam o processo de organização, manutenção, compartilhamento, recuperação e consulta das informações eletronicamente armazenadas.

Este artigo propõe um modelo de sistema computacional, cujo objetivo é interagir com a massa de documentos eletrônicos de domínio do usuário, a fim de recuperar neles as informações julgadas como relevantes, a partir de um esquema de recuperação previamente definido. O desdobramento desta pesquisa, bem como a implementação do sistema proposto, é fruto do projeto final de graduação do primeiro autor deste trabalho, ao final do qual se espera contribuir com as comunidades que pesquisam sobre o tema e com aqueles que precisam interagir com sistemas de produção de documentos eletrônicos.

REFERÊNCIAS

BOOTY, W. G. LAM.: Design And Implementation of An Environmental Decision Support System. Environmental Modelling Software - 2001.

POLIKAR, R. et al.: Artificial intelligence methods for selection of an optimized sensor array for identification of volatile organic compounds, - 2001.

POMERO, J. C. Artificial Intelligence And Human decision Making. European Journal of Operational Research - 1997

ROCHA A, F. Neural Nets – A theory for brains and machines - Lecture Notes in Artificial Inteligence. Springer-Verlag 1992.

ROCHA, Ruth. Minidicionário, 10 ed, São Paulo, Scipicione, 2002.

WIVES, Leandro Krug. Indexação de Documentos Textuais, Porto Alegre, 1997-B.

WIVES, Leandro Krug. Um Estudo sobre Técnicas de Recuperação de Informações com ênfase em Informações Textuais, Porto Alegre, 1997-A.

YATES, R. Baeza; NETO, B. Ribeiro. Modern Information Retrieval, New York, ACM Press Series/Addison Wesley, 1999.