Extração de informação como base para descoberta de conhecimento em dados não estruturados

Rui Gureghian Scarinci*
José Palazzo Moreira de Oliveira**

Resumo

Métodos de Descoberta de Conhecimento em Texto ou *Knowledge Discovery in Text* - KDT tem sido aplicados a uma grande variedade de domínios, desde artigos para congressos, até receituários médicos. KDT é o processo de encontrar padrões e informações implícitas interessantes ou úteis em um corpo de informação textual não estruturado [LOH 97]. Este processo combina muitas das técnicas de Extração de Informação, Recuperação de Informação, Processamento da Linguagem Natural e Sumarização de Documentos com os métodos de *Data Mining* (DM).

Os dados estruturados, armazenados na maioria dos Sistemas de Gerência de Bancos de Dados, são mais fáceis de serem tratados por meios computacionais, porque existem linguagens formais, como SQL e QBE, que permitem sua manipulação e consulta de forma mais concisa e precisa [LOH 97]. Os dados não estruturados, por outro lado, necessitam de mecanismos computacionais diferentes dos tradicionalmente usados, para que possam ser coletados, armazenados, manipulados e consultados. Para aplicar métodos tradicionais de DM sobre textos, é necessário impor alguma estrutura para os dados [DIX 97]. Ou seja, alguém deve definir a estrutura destes dados, coletá-los e armazená-los num Banco de Dados convencional. Entretanto, tal processo necessita de apoio automatizado, pois é difícil, tedioso e sujeito a erros se feito por pessoas. Neste sentido, Descoberta de Conhecimento em Textos é uma área bastante relacionada com a área de Extração de Informação, bem como a de Recuperação de Informação, e realmente pode-se considerar que sistemas de KDT são construídos a partir de componentes que executam estas tarefas [FEL 99].

Extração de Informação

Recentemente, uma nova orientação nas pesquisas em Processamento da Linguagem Natural (*Natural Language Processing* - NLP) emergiu sob o nome de Extração de Informação (EI), sendo uma sub-área desta primeira [CO 97a]. Esta nova área é dedicada ao processamento de dados associados a grandes volumes de texto que contém informações de algum domínio de interesse [LEW 96].

* rgs@inf.ufrgs.br
** palazzo@inf.ufrgs.br

A meta de um sistema de extração de informações é extrair tipos específicos de informações a partir de textos [SCA 97]. A vantagem principal desta tarefa é o particionamento de um texto de entrada, permitindo que partes não pertinentes ao domínio possam ser efetivamente ignoradas. Extração de Informação é menos custosa computacionalmente que NLP tradicional, pois muitas frases ou até mesmo orações inteiras podem ser ignoradas se elas não são pertinentes ao domínio de interesse [RIL 94]. E, visto que o sistema está só preocupado com porções específicas de domínio do texto, alguns dos problemas mais difíceis em NLP são simplificados (por exemplo, resolução de ambigüidades). Extração de Informação é uma tecnologia mais prática e robusta que Processamento da Linguagem Natural tradicional e tem alcançado sucesso nos últimos anos [RIL 94a].

Segundo [COW 96], Extração de Informação é o nome dado a qualquer processo que seleciona estruturas e combina dados que são achados em um ou mais textos. A produção final do processo de extração varia de caso para caso; contudo, esta pode ser facilmente transformada em entradas para bancos de dados. Analistas de informação, que já trabalham a longo tempo em tarefas específicas de Extração de Informação manualmente, entendem EI como um processo com o claro objetivo de criação de bancos de dados.

Recuperação e Extração de Informação

Extração de Informação é diferente de Recuperação de Informação (RI). Técnicas de RI podem localizar os documentos pertinentes dentro de uma coleção, mas estas estão impossibilitadas de extrair informação dos documentos pertinentes de acordo com critérios específicos [WIL 97]. O poder de um sistema de extração de informações comparado a um sistema de recuperação de informações está na habilidade de extrair a informação relevante dos artigos de acordo com um critério específico e a representar em estruturas, as quais sistemas de recuperação são impossibilitados de produzir [COS 97a].

No entanto, Extração de Informação é uma tarefa profundamente relacionada com Recuperação de Informação de diferentes maneiras, nas quais as duas operações podem ser potencialmente combinadas [SME 97]. RI pode ser usada com EI, pré-processando uma grande coleção de documentos para um subconjunto menor gerenciável, no qual os custos computacionais das técnicas de EI possam ser aplicados. EI pode ser utilizada como um componente em RI, onde o processo de análise de documentos de EI é utilizado para identificar termos para índices a fim de representarem os documentos originais. Por exemplo, identificação de nomes próprios é uma importante habilidade de EI que certamente pode auxiliar em RI. Em um mesmo nível, através de um *browser* de pesquisa de informações, recursos de RI pesquisam e selecionam textos, enquanto EI é usado para sumarizar o resultado desta pesquisa em algo mais conciso e coerente.

Extração de Informações e *Data Mining*

Data Mining tem se tornado recentemente uma área de pesquisa muito popular. Esta nova área foca a exploração computadorizada de grandes volumes de dados e a descoberta de padrões relevantes e interessantes existentes nestes dados [FEL 99]. Enquanto muito do trabalho de DM está concentrado em bancos de dados estruturados, é claro que este paradigma é requerido para manipular grandes conjuntos de dados existentes somente na forma de textos não estruturados [SCA 97a]. Surpreendentemente, somente um pequeno conjunto de exemplos de KDT esta disponível. DM tende a ser muito matemática, e a maioria das pesquisas neste campo se preocupa somente com extração de conhecimento de bancos de dados [DIX 97]. Por isto, muitas das técnicas e teorias não se prestam livremente para KDT. A maioria dos pesquisadores de *Data Mining* não contempla entradas do tipo texto não estruturado em sua pesquisa. Por outro lado, considerando o volume de coleções de documentos existente, Descoberta de Conhecimento em Textos deve ser uma área de pesquisa muito útil.

KDT combina muitas das técnicas de Extração de Informação, Recuperação de Informação, Processamento da Linguagem Natural e Sumarização de Documentos com os métodos de *Data Mining*. O principal uso de KDT é para a extração de conhecimento previamente desconhecido armazenado em um volume de texto [FEL 95]. Para aplicar métodos tradicionais de DM em textos, é necessário impor alguma estrutura para os dados [DIX 97]. O primeiro passo neste processo é decidir qual estrutura impor aos dados. Para fazer isto, deve-se considerar muito cuidadosamente os processos posteriores de descoberta a serem usados. Dado as fortes limitações da tecnologia atual no processamento de textos, nós necessitamos definir normalmente estruturas simples que possam ser extraídas a partir dos textos de forma automática e com baixos custos [FEL 95]. Por outro lado, a estrutura deve ser boa o suficiente para permitir a execução das operações de *Data Mining* de forma interessante. [DIX 97] destacada a fase de pré-processamento como uma fase crucial para DM, efetivamente alterando a natureza da mineração de dados dependendo de como o texto foi processado inicialmente. Descoberta de Conhecimento em Textos é bastante relacionada com a área de Extração de Informação, bem como a de Recuperação de Informação, e realmente pode-se considerar que sistemas de MD são construídos a partir de componentes que executam estas tarefas [FEL 99].

Segundo [DIX 97], a melhor visão de um sistema de KDT seria como apresentado na sucessão de passos esboçados abaixo (Figura 3.1). Alguns pesquisadores apresentam de forma combinada a fase de Recuperação e a de Extração em uma única fase de pré-processamento. O primeiro passo (Recuperação de Informação) é localizar e recuperar documentos que possam ser considerados relevantes ao usuário. Tipicamente os usuários do sistema podem especificar conjuntos de documentos, mas ainda precisa-se de um sistema que filtre os documentos irrelevantes. A próxima fase (Extração de Informação) é extrair a informação dos documentos selecionados. Esta extração está em retirar dos documentos a informação que o usuário especificou através de modelos de informação. Uma vez que as informações foram estruturadas pela fase anterior para cada documento, nesta fase

(Mineração de Informação) nós temos um banco de dados que é compatível com técnicas de mineração de dados padrão. Assim, procura-se descobrir padrões dentro dos dados. O passo final (Interpretação) é encontrar uma interpretação para os padrões recuperados da fase de mineração. Idealmente, a interpretação deveria ser em formato de linguagem natural.

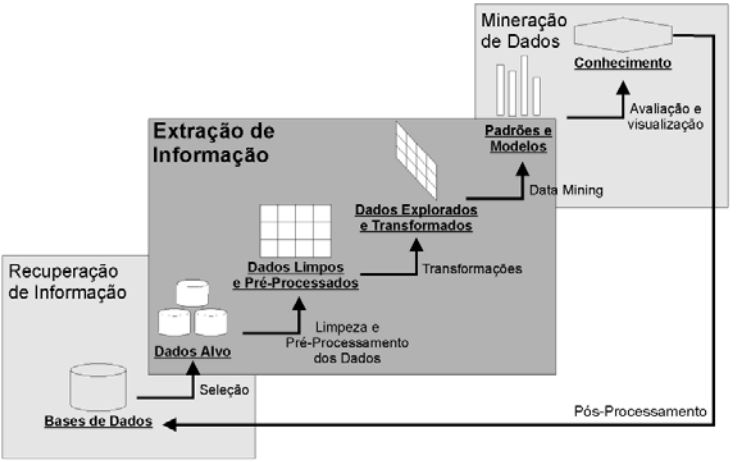

Figura 3.1 - Etapas de um Processo de KDT

Conclusão

O resultado estruturado gerado pela aplicação de um sistema de extração de informações sobre um conjunto de textos, como a representação da informação em *templates*, viabiliza o uso de técnicas de DM sobre a saída deste processamento [COS 97]. Além disso, o volume de informação a ser processado pelo módulo de DM é menor, podendo o mesmo se concentrar nos dados relevantes.

EI é indiscutivelmente o componente mais importante do processo de KDT [FEL 99]. Alguns artigos tratam EI e DM como a mesma tecnologia, entretanto elas não são equivalentes. Extração de Informação é um processo concebido para varrer um texto ou um conjunto de textos com a meta de extrair fatos apresentados nos mesmos. É provável que o sistema de EI seja voltado normalmente para um domínio específico, sendo pré-programado ou treinado para reconhecer cada campo no corpo do texto. *Data Mining* trata alguns problemas não contemplados por EI. Sistemas de mineração buscam deduzir um conjunto de regras ou um modelo de domínio com base no texto. Isto prevê um forte uso de técnicas de aprendizado de máquina, além dos componentes de NLP existentes em EI [COW 96]. A base de conhecimento que se espera extrair, normalmente é designada para um sistema especialista ou sistemas baseados em casos. DM é mais ambicioso quanto ao entendimento do texto que EI, cujo objetivo se restringe a extrair informações existentes no

texto, sem deduzir novas, encontrando padrões e informações implícitas em um corpo de informação textual não estruturado [LOH 97].

Referências

[COS 97] COSTANTINO, Marco. **Financial Information Extraction Using Pre-defined and User-definable Templates in the LOLITA.** Disponível por WWW em http://www.advanced-finance.co.uk/marco.html (1997).

[COS 97a] COSTANTINO, Marco et al. **Natural Language Processing and Information Extraction:** Qualitative Analysis of Financial News Articles. Disponível por WWW em http://www.advanced-finance.co.uk/marco.html (1997).

[COW 96] COWIE, Jim; LEHNERT, Wendy. Information Extraction. **Communications of the ACM**, New York, v. 39. n. 1, p. 80-91, Jan. 1996.

[DIX 97] DIXON, Mark. **An Overvioew of Document Mining Technology.** Disponível por WWW em http://www.geocities.com/ResearchTriangle/Thinktank/1997/mark/writings/main.html (out. 1997).

[FEL 95] FELDMAN, Ronen; DAGAN, Ido. **Knowledge Discovery in Textual Databases (KDT).** Disponível por WWW em www.cs.biu.ac.il:8080/~feldman/ (1995).

[FEL 99] FELDENS, Miguel A. **Knowledge Discovery in Databases.** Disponível por de WWW em http://www.inf.ufrgs.br/~feldens/ucpel.zip (nov. 1999).

[LEW 96] LEWIS, David D.; JONES, Karen S.. Natural Language Processing for Information Retrieval. **Communications of the ACM**, New York, v. 39. n. 1, p. 92-101. Jan. 1996.

[LOH 97] LOH, Stanley. **Descoberta de Conhecimento em Bases de Dados Textuais**. Disponível por WWW em http://atlas.ucpel.tche.br/~loh/dc-texto.htm (1997).

[RIL 94] RILOFF, Ellen; LEHNERT, Wendy. **Information Extraction as a Basis for High-Precision Text Classification.** Disponível por WWW em http://www.cora.justresearch.com (1994).

[RIL 94a] RILOFF, Ellen; LEHNERT, Wendy. **Information Extraction as a Basis for Portable Text Classification System.** Disponível por WWW em http://www.cora.justresearch.com (1994).

[SCA 97] SCARINCI, Rui G; PALAZZO, José M. SES – Sistema de Extração Semântica de Informações. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 1997, Fortaleza, Ceara. **Anais...** Fortaleza: Universidade Federal do Ceara, 1997. p. 65-79.

[SCA 97a] SCARINCI, Rui G. **SES – Sistema de Extração Semântica de Informações**. Porto Alegre: CPGCC da UFRGS, 1997. 170 p. Dissertação de Mestrado.

[SME 97] SMEATON, Allan F. **Information Retrieval:** Still Butting Heads with Natural Language Processing. Information Extraction - A Multidisciplinary Approach to an Emerging Information Technology. Lecture Notes in Artificial Inteligence. Edited by Pazienza Maria T., Springer-Verlag, Berlin Heldelberg, 1997. p. 115-138

[WIL 97] WILKS, Yorick. **Information Extraction as a Core Language Technology.** Disponível por WWW em http://www.dcs.shef.ac.uk/~yorick/papers.html (1997)