# Descoberta de conhecimento em textos baseada em conceitos

**Stanley Loh***
**José Palazzo M. de Oliveira****

## Resumo

A proposta deste trabalho é aplicar técnicas de Descoberta de Conhecimento sobre características extraídas de textos. Ao invés de aplicar as técnicas sobre termos (como faz [LIN 98]) ou sobre palavras-chave associadas aos textos (como faz [FEL 98]), a proposta é identificar conceitos presentes nos textos e depois aplicar as técnicas de descoberta sobre estes conceitos. Assim, seria possível diminuir o problema do vocabulário e permitir descobertas a nível de conceitos e não de palavras ou valores de atributos.

A idéia de trabalhar com conceitos é semelhante ao trabalho de [BOW 96], que extrai de textos conceitos tais como exemplos e definições, e ao de [LIM 97], que identifica nos textos categorias extraídas de um *thesaurus* médico. Entretanto, estes trabalhos não aplicam técnicas posteriores de descoberta. A proposta também tem certa semelhança com o trabalho de Swanson (citado em [DAV 89]) que utiliza um sistema de indexação para reconhecer relações lógicas entre textos. Entretanto, o referido trabalho não aplica técnicas estatísticas, mas somente avalia possíveis relações entre conteúdos para formar conhecimento novo.

A abordagem proposta combina um processo de categorização, para identificar conceitos presentes nos textos, com a posterior aplicação de técnicas estatísticas (processo de mineração), para descobrir padrões através da análise das distribuições dos conceitos em uma coleção de documentos textuais. Conforme [CHE 94b], os conceitos podem ser representados por conjuntos de termos. Estes permitem identificar a presença do conceito em um texto. De acordo com [FEL 95], a categorização é um tipo extração de informação. Só que, ao invés de usar métodos complicados de processamento de língua natural (PLN), a proposta é utilizar técnicas simples baseando-se em que os conceitos podem ser identificados por sinais (no caso de textos, sinais são termos). Como cada conceito é definido por um conjunto de termos, o processo de categorização busca encontrar a presença destes termos (sinais) nos textos. Depois, usando um processo de raciocínio *fuzzy*, sugerido por [NAK 93], os sinais (termos) encontrados são computados para avaliar a probabilidade de presença do conceito no texto.

Para criar as definições dos conceitos, é necessário um processo de classificação, ou seja, escolher os conceitos que serão definidos e descrever cada um com um conjunto de palavras. Este é um processo de aprendizado e pode ser feito manualmente ou de forma automática (supervisionada ou não) com ajuda de ferramentas de software.

* loh@inf.ufrgs.br
** palazzo@inf.ufrgs.br

O processo de mineração aplica técnicas estatísticas sobre os conceitos descobertos na etapa de categorização. De acordo com a analogia proposta por [LIN 98] e [GAR 99], os textos (ou documentos) são tratados como transações e os conceitos como os itens do banco de dados. Uma das técnicas que pode ser aplicada é a listagem de conceitos-chave, que avalia o quanto um conceito está presente numa coleção. Outra técnica possível é associação ou correlação, que permite avaliar a implicação entre conceitos ou de forma mista (conceitos e palavras). As associações são apresentadas na forma X ➔ Y, com a interpretação de que "se X aparece num texto, então Y também aparece com certo grau de confiança e suporte".

A vantagem do uso de conceitos é que estes representam melhor que palavras os objetos, eventos, sentimentos, ações, etc do mundo real. Em geral, são usados em áreas como análise de discurso para identificar idéias e ideologias presentes em textos. [CHE 94], por exemplo, usou com sucesso a identificação de conceitos para organizar idéias discutidas num processo de *brainstorming* eletrônico. Abordagens baseadas em conceitos (*concept-based approaches*) já são usadas com sucesso na área de Recuperação de Informação (IR). [LIN 93] comenta que a vantagem deste tipo de abordagem em relação à busca por palavras-chave é poder minimizar o problema do vocabulário. Mas o principal objetivo deste trabalho é poder realizar um novo tipo de descoberta, de mais alto nível, baseada em conceitos e não em palavras ou valores de atributos. Assim, as descobertas permitem análises de idéias, ideologias, tendências e intenções presentes em textos. Também, é possível realizar descobertas sobre novos conceitos, sem ter que ficar preso aos previamente definidos em *thesauri* ou ontologias. Neste caso, basta definir cada novo conceito na etapa de classificação.

Experimentos iniciais foram realizados sobre 4 coleções textuais: uma contendo notícias extraídas de um jornal *online* falando sobre o prefeito de uma grande cidade; uma referente a prontuários médicos de internação de pacientes numa clínica psiquiátrica; uma com textos extraídos da Web sobre ferramentas de Descoberta de Conhecimento em Bancos de Dados e outra coleção também extraída da Web sobre ferramentas de *Text Mining*. Os experimentos permitiram verificar que padrões interessantes e úteis podem ser extraídos rápido e facilmente, auxiliando o usuário a entender as idéias e as tendências presentes nas coleções. Comparando subcoleções, foi possível determinar a variação nos tópicos dominantes. Os experimentos também permitiram avaliar diferentes maneiras de realizar a classificação (definição de conceitos), de acordo com o objetivo do usuário (conceitos interessantes, rapidez na definição, análise de termos mais comuns, etc). Já que a qualidade da categorização pode influenciar o conhecimento descoberto, avaliações inicias foram conduzidas e os resultados mostraram-se promissores.

## Referências


[BOW 96] BOWDEN, Paul R.; HALSTEAD, Peter; ROSE, Tony G. Extracting conceptual knowledge from text using explicit relation markers. In: SHADBOLT, Nigel et alli (eds). IX European Knowledge Acquisition Workshop. **Proceedings...** Lecture Notes in Artificial Intelligence, 1076. Maio de 1996.

[CHE 94] CHEN, Hsinchum. The vocabulary problem in collaboration. *IEEE Computer, special issue on CSCW*, v. 27 , n. 5, May 1994. Online at http://ai.bpa.arizona.edu/papers/cscw94/cscw94.html

[CHE 94b] CHEN, Hsinchum e al. Automatic concept classification of text from electronic meetings. *Communications of the ACM*, 1994. Online at http://ai.bpa.arizona.edu/papers/ebs92/ebs92.html

[DAV 89] DAVIES, Roy. The creation of new knowledge by information retrieval and classification. **Journal of Documentation**, v.45, n.4, Dezembro de 1989.

[FEL 95] FELDMAN, Ronen; DAGAN, Ido. Knowledge discovery in textual databases (KDT). In: 1st International Conference on Knowledge Discovery (KDD-95). Montreal, August 1995.

[FEL 98] FELDMAN, Ronen; DAGAN, Ido. Mining text using keyword distributions. **Journal of Intelligent Information Systems**, v.10, n.3, 1998.

[GAR 99] GAROFALAKIS, Minos N. et al. Data mining and the web: past, present and future. In: ACM Workshop on Information and Data Management, Kansas City, 1999.

[LIM 97] LIMA, Luciano R. S.; LAENDER, Alberto H. F.; RIBEIRO NETO, Berthier A. Um modelo para recuperação de informação especializada aplicado a bases de dados médicas semi-estruturadas. In: Simpósio Brasileiro de Banco de Dados. Outubro de 1997.

[LIN 93] LIN, Chung-hsin; CHEN, Hsinchun. An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) documents. *IEEE Transactions on Systems, Man and* Cybernetics, v. 26, n.1, February 1996. Online at http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html

[LIN 98] LIN, Shian-Hua et al. Extracting classification knowledge of Internet documents with mining term associations: a semantic approach. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98). 1998.

[NAK 93] NAKANISHI, H.; TURKSEN, I. B.; SUGENO, M. *A review and comparison of six reasoning methods*. **Fuzzy Sets and Systems**, 57, 1993.