

Descoberta Proativa de Conhecimento em Coleções Textuais: Iniciando sem Hipóteses

Stanley Loh
(UCPEL, ULBRA,
PPGC/II/UFRGS)
loh@inf.ufrgs.br

Leandro Krug Wives
(PPGC/II/UFRGS)
wives@inf.ufrgs.br

José Palazzo M. de Oliveira
(II/UFRGS)
palazzo@inf.ufrgs.br

Endereço:

Programa de Pós-Graduação em Computação
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Avenida Bento Gonçalves, 9500
Bloco IV, Prédio 43412 - Campus do Vale
Porto Alegre - RS - 91501-970
BRASIL

Resumo

Este artigo discute o processo de descoberta de conhecimento em textos (KDT) segundo a abordagem proativa, isto é, segundo uma abordagem que inicia sem hipóteses predefinidas e é baseada em uma ação efetiva do pesquisador. A abordagem proativa difere da reativa, porque no primeiro caso o usuário não tem uma necessidade ou problema consciente, enquanto que no segundo o usuário sabe o que está procurando ou do que precisa. Na descoberta proativa, o usuário tem por objetivo encontrar conhecimento novo e útil, mas não sabe o que está à procura, muito menos por onde começar. Neste artigo, exemplos de descoberta proativa são apresentados e discutidos para mostrar como as técnicas de KDT podem ser usadas e que tipos de resultados podem ser obtidos segundo este paradigma. O trabalho apresenta também estratégias para descoberta de conhecimento no modo proativo (sem hipóteses iniciais) e discute a necessidade de intervenção humana no processo e os aspectos que podem influenciar negativamente os resultados (ruídos).

Palavras-chave: *descoberta de conhecimento, text mining, análise de textos*

1 Introdução

Com o crescente uso de computadores e principalmente da Internet, cada vez mais documentos eletrônicos estão sendo armazenados e colocados à disposição das pessoas. Davies [DAV89] afirma que muito conhecimento pode ser inferido a partir destes documentos. Entretanto, encontrar tal conhecimento é uma tarefa árdua.

Existem técnicas e ferramentas para Recuperação de Informação (RI), as quais auxiliam as pessoas a encontrar documentos que contenham informações relevantes [SPA97]. Entretanto, é necessário examinar os documentos resultantes para encontrar a informação desejada, o que não é uma tarefa fácil. Esta dificuldade é causada pelo fato de que documentos são insatisfatórios como respostas, por serem grandes e difusos em geral [WIL94]. Um exemplo prático são os serviços de busca (*search engines*) da Web, que encontram para o usuário um volume enorme de documentos, mas o usuário tem de examiná-los para encontrar o que deseja. Este problema é chamado de “sobrecarga de informações” (*information overload*).

Alguns trabalhos estão sendo desenvolvidos para tentar resolver ou minimizar tal problema. A área de Extração de Informações (EI – *information extraction*) estuda

metodologias, técnicas e sistemas que possam encontrar dados específicos dentro de textos. Tais sistemas extraem automaticamente valores de atributos (como campos de um banco de dados). Infelizmente, em geral, tais sistemas são muito dependentes do domínio, isto é, só funcionam com certos tipos de documentos [CRO95]. Além disto, para criar tais sistemas é necessário muita engenharia de conhecimento, examinando amostras de textos para saber como a informação é codificada em frases da língua natural [CHI93].

Estes trabalhos partem de uma necessidade específica do usuário ou de um objetivo previamente definido em uma aplicação e, portanto, são classificados sob o paradigma **reativo**. Ou seja, o usuário deve definir sua necessidade ou problema e fornecer caminhos para a solução (por exemplo, as técnicas e parâmetros a serem utilizados). Em geral, esta é uma premissa falsa que distorce o processo de busca, uma vez que as pessoas não estão aptas a especificar precisamente o que é necessário para resolver seu problema. Pedir ao usuário para formular o que precisa é uma premissa irreal, se é isto justamente o que falta [BEL97]. Belkin e outros [BEL97] chamam a necessidade de informação de um estado anômalo de conhecimento (*ASK - Anomalous State of Knowledge*), portanto (por ser anômalo) difícil de representar.

Contrária a esta abordagem, surge um novo tipo de paradigma (**proativo**) que procura automaticamente informações novas e úteis em uma coleção de documentos, sem que seja necessário que o usuário estabeleça inicialmente uma necessidade. Na área de Banco de Dados, esta abordagem é conhecida como Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases – KDD*) [FAY96]. Em geral, os sistemas de KDD utilizam técnicas estatísticas conhecidas como técnicas de *Data Mining* para encontrar automaticamente padrões nas distribuições de valores em atributos ou campos de um banco de dados

Tais técnicas têm tido sucesso mas trabalham somente sobre dados estruturados. Em se tratando de coleções de textos (dados não-estruturados), o problema é recente e ainda necessita mais estudos. Descoberta de Conhecimento em Textos (*Knowledge Discovery in Texts*) é o termo utilizado para designar a aplicação das mesmas técnicas de KDD só que sobre características ou atributos extraídos de textos [FEL95]. Estas características podem ser valores de atributos/campos extraídos dos textos por algum tipo de inferência ou até algo mais simples como as próprias palavras do texto. Lin e outros [LIN98], por exemplo, descobrem associações entre palavras extraídas automaticamente dos textos. Os termos mais frequentes são utilizados como atributos do texto. Já Feldman e Dagan [FEL98] aplicam as técnicas de KDD sobre palavras-chave associadas a textos. Tal associação já deve ter sido feita antes, seja por pessoas (atividades manuais e intelectuais) ou automaticamente por ferramentas de software. Estas técnicas usam análises estatísticas sobre as distribuições das características para descobrir padrões no formato de regras associativas.

Já Swanson e Smalheiser [SWA97] sugerem algumas estratégias mais complexas para descoberta de conhecimento em textos. Uma delas procura identificar analogias em diferentes textos através da análise de termos comuns. Sua sugestão é omitir termos relacionados à área para reunir documentos de diversas áreas que possam estar relacionados. Tais autores têm relatado sucesso com descobertas de novas e úteis alternativas na área médica. Entretanto, a relação entre os textos não é simples de ser feita, exigindo a intervenção de especialistas humanos no assunto.

Chen [CHE93], por sua vez, sugere a construção automática de resumos combinando partes de distintos textos, usando para isto estruturas internas (redes semânticas) e termos comuns aos textos. Davies [DAV89] sugere examinar as correlações escondidas em textos, isto é, combinações de conceitos através de relações estatísticas. Para tanto, Davies sugere que

sejam analisadas as distribuições de termos numa coleção. Assim, por exemplo, foi possível identificar uma hipótese de relação entre um certo tipo de falha num sistema e alguns itens mais frequentes (possíveis causas das falhas). Davies afirma que o todo é mais que a mera soma das partes, o que permite que conhecimentos novos não explicitamente presentes nos textos possam ser descobertos analisando relações semânticas entre os textos. No caso, as relações são identificadas através da análise dos termos presentes nos textos.

Feigenbaum (*apud* [DAV89]) compara as bibliotecas de hoje com as do futuro. As primeiras são como um armazém de objetos passivos. Já as bibliotecas do futuro serão uma coleção de documentos ativos que ajudarão às pessoas fornecendo conexões desconhecidas, associações e analogias, entendimento de conceitos novos, descoberta de novos métodos e teorias, sem que as pessoas precisem definir claramente quais são suas necessidades de informação. Minsky e Feigenbaum falam que os documentos devem ser capazes de “conversarem entre si” [DAV89] [CHE93].

Este artigo discute o uso de técnicas de KDT sob o paradigma **proativo**, ou seja, é analisado o processo de descoberta sem que seja necessário estabelecer hipóteses iniciais. Para tanto, a seção 2 contém uma breve revisão sobre o assunto, apresentando as principais técnicas de KDT e as diferenças entre os paradigmas **reativo** e **proativo**. Já na seção 3, são apresentados resultados de experimentos sob este paradigma. Na seção 4, são discutidos os aspectos envolvidos em processos deste tipo e que o influenciam (estratégias de descoberta, intervenção humana e ruídos). Por fim, a conclusão discute vantagens e limitações deste tipo de abordagem.

2 Revisão sobre o assunto

Nesta seção, é feita uma breve revisão dos assuntos envolvidos neste artigo. Primeiro, são apresentadas as principais técnicas para KDT, as quais são utilizadas nos experimentos apresentados mais adiante. Depois, na segunda subseção, são discutidas as diferenças entre as abordagens reativa e proativa.

2.1 Técnicas para KDT

Existem muitas técnicas e ferramentas de software para realizar KDT. Nesta subseção, são apresentadas as principais técnicas.

A técnica mais básica é a recuperação de informações (RI), cujo objetivo é encontrar textos que podem conter determinada informação. Métodos para RI são discutidos em [SPA97] e [SAL83]. Uma técnica similar é a recuperação de passagens, que aplica as mesmas técnicas de RI só que sobre partes do texto [CAL94] [KAS97]. Já a técnica de extração de informação (EI) procura valores de atributos dentro dos textos [COW96]. A técnica de sumarização (*summarization*) tem por objetivo extrair resumos de um texto ou de uma coleção, podendo ser uma visão geral ou as partes mais importantes ou mais interessantes [SPA97] [SAL97] [MCK95] [CHE93].

A técnica de listagem de conceitos-chave (*key-concept listing*), por sua vez, analisa uma coleção de textos em busca de características comuns (palavras, palavras-chave, temas, etc). A ferramenta de Moscarola e outros [MOS98], por exemplo, encontra e apresenta ao usuário uma lista de termos relacionados por proximidade. Já Maarek [MAA92] extrai afinidades léxicas, definidas como relações entre unidades da linguagem, por exemplo sujeito-verbo, substantivo-adjetivo. A partir desta técnica pode-se fazer o inverso, isto é, descobrir diferenças comparando textos ou coleções, o que é chamado de técnica da diferença.

Já a técnica de agrupamento (*clustering*) é um pouco mais complexa. Ela é utilizada para identificar automaticamente, sem intervenção humana, grupos de textos similares [WIL88]. Sua principal utilidade é permitir encontrar características comuns em subgrupos quando não há nada em comum na coleção toda. Já a técnica de classificação ou categorização procura encontrar temas ou assuntos no conteúdo dos textos (do que os textos estão tratando). A já comentada técnica de associação (ou correlação) descobre relações de dependência entre textos ou características dos textos.

Existem também técnicas para visualização de resultados, que ajudam o usuário a entender melhor o conhecimento descoberto [VEE97] [BAE98] [SCH96b].

Apesar de apresentadas em separado, nada impede que elas sejam usadas de modo integrado, uma após a outra, de forma que a saída de uma seja a entrada da seguinte. Por exemplo, Moens e Uyttendaele [MOE97] usam a técnica de sumarização associada com EI, para criar resumos de casos jurídicos. Moscarola e outros [MOS98] [MOS98b] apresentam uma ferramenta que integra diversas técnicas para KDT.

2.2 Descoberta Reativa X Proativa

A maioria dos pesquisadores concorda que o processo de descoberta é cíclico, tendo como passos principais: [PAR89] [AGR93] [ING96]

- a) a formulação de hipóteses;
- b) o teste das hipóteses;
- c) a observação dos resultados (para refutar ou confirmá-las);
- d) a revisão das hipóteses e a sua modificação (reiniciando o processo), até que o usuário se dê por satisfeito.

Entretanto, esta estratégia só pode ser aplicada quando o usuário consegue formular hipóteses iniciais, ou seja, quando ele tem idéia de qual é o seu objetivo ou necessidade e sabe do que precisa.

De acordo com Choudhury e Sampler [CHO97], existem dois modos para aquisição de informação: o modo reativo e o modo proativo. No primeiro caso, a informação é adquirida para resolver um problema específico do usuário (uma necessidade resultante de um estado anômalo de conhecimento). Nestes casos, o usuário sabe o que quer e poderá identificar a solução para o problema quando há encontrar.

Por outro lado, no modo proativo, o propósito de adquirir informação é exploratório, para detectar problemas potenciais ou oportunidades. Neste segundo caso, o usuário não tem um objetivo específico.

Oard e Marchionini [OAR96] classificam as necessidades de informação em estáveis ou dinâmicas e em específicas ou abrangentes (gerais). Taylor (citado em [OAR96]) define 4 tipos de necessidades, os quais formam uma escala crescente para a solução do problema:

- necessidades viscerais: quando existe uma necessidade ou interesse, mas esta não é percebida de forma consciente;
- necessidades conscientes: quando o usuário percebe sua necessidade e sabe do que precisa;
- necessidades formalizadas: quando o usuário expressa sua necessidade de alguma forma;
- necessidades comprometidas: quando a necessidade é representada no sistema.

As necessidades tratadas pela abordagem de descoberta reativa poderiam ser classificadas como estáveis e específicas, segundo a classificação de Oard e Marchionini, e

como conscientes (no mínimo), segundo Taylor. Isto porque o usuário sabe o que quer, mesmo que não consiga formalizar.

Exemplos de objetivos que caracterizam um processo reativo são:

- encontrar atributos comuns nos produtos mais vendidos;
- encontrar motivos que levam à evasão ou a reclamações de clientes;
- achar perfis de grupos de clientes;
- encontrar clientes potenciais para propaganda seletiva;
- encontrar concorrentes no mercado.

No modo reativo, o usuário tem uma idéia, mesmo que vaga, do que pode ser a solução ou, pelo menos, de onde se pode encontrá-la. Pode-se dizer então que o usuário possui algumas hipóteses iniciais, que ajudarão a direcionar o processo de descoberta. Neste caso, é necessário algum tipo de pré-processamento, por exemplo para selecionar atributos (colunas em uma tabela) ou valores de atributos (células). Isto exige entender o interesse ou objetivo do usuário para limitar o espaço de busca (na entrada) ou filtrar os resultados (na saída). É o caso típico de quando se deseja encontrar uma informação específica, por exemplo, um valor para um atributo ou um processo (conjunto de passos) para resolver um problema.

Já as necessidades da abordagem proativa poderiam ser classificadas como dinâmicas e abrangentes, segundo a classificação de Oard e Marchioninni. São dinâmicas porque podem mudar durante o processo, já que o objetivo não está bem claro, e são abrangentes porque o usuário não sabe exatamente o que está procurando. Pela taxonomia de Taylor, as necessidades do modo proativo são viscerais. Isto quer dizer que há uma necessidade ou objetivo, mas o usuário não consegue definir o que precisa para resolver o problema. A necessidade típica do modo proativo poderia ser representada pela expressão: “*diga-me o que há de interessante nesta coleção*”. Neste caso, o usuário não tem de forma definida o que lhe seja de interesse (o que precisa), podendo tal interesse mudar durante o processo. Pode-se dizer que é um processo exploratório, sendo, em geral, iterativo (com retroalimentação) e interativo (com ativa participação e intervenção do usuário).

Na abordagem proativa, não há hipóteses iniciais ou elas são muito vagas. O usuário deverá descobrir hipóteses para a solução do seu problema e explorá-las, investigá-las e testá-las durante o processo. Em geral, acontece porque o usuário não sabe exatamente o que está procurando. É o caso típico de quando se quer monitorar alguma situação ou encontrar algo de interessante que possa levar a investigações posteriores. Depois que hipóteses são levantadas, o processo pode seguir como no paradigma reativo, talvez sendo necessário avaliar as hipóteses, para verificar se são verdadeiras ou não.

3 Experimentos

Foi implementado um conjunto de ferramentas de software para KDT. Há uma ferramenta diferente para cada técnica descrita na seção anterior. As ferramentas estão integradas de forma que os resultados de uma podem ser usados como entrada de outra. Assim, processos complexos de descoberta podem ser realizados.

Nesta seção, são apresentados experimentos sob a abordagem proativa, usando as ferramentas implementadas. Para os experimentos foram usadas 3 coleções de textos, a saber:

- a) coleção política: formada por textos extraídos de um jornal publicado na Web falando sobre um prefeito; os textos foram extraídos usando a ferramenta local de recuperação de informação do *site* e tendo como consulta o nome do prefeito; esta coleção está dividida

em dois segmentos (sub-coleções), uma com 180 textos publicados em 1997 e outra com 178 textos publicados em 1999;

- b) coleção médica: composta por 1040 prontuários médicos escritos por médicos sobre pacientes de uma clínica psiquiátrica (26 prontuários eram referentes à internação do paciente);
- c) coleção sobre guerra: 18 textos versando sobre guerras na história mundial, extraídos de uma enciclopédia (em inglês).

A seguir, são apresentados exemplos de processos de descoberta sobre estas coleções, iniciando com diferentes técnicas de KDT e sem hipóteses ou interesse inicial (abordagem proativa). Estes exemplos ajudam a entender como se pode utilizar a abordagem proativa e que tipo de resultados podem ser alcançados. As técnicas utilizadas durante cada processo aparecem em **negrito**.

- **1º Experimento: coleção política**

Considerando que há dois segmentos nesta coleção, a técnica de **listagem de conceitos-chave** foi utilizada para comparar as palavras que apareciam nas duas sub-coleções (1997 x 1999). Analisando os termos mais frequentes nas duas coleções, chegou-se a uma descoberta interessante: o nome da esposa do prefeito aparecia mais vezes na sub-coleção referente a 1999 (35 textos) do que na de 1997 (somente 6 textos). Isto suscitou a hipótese de que a esposa do prefeito era citada em situações diferentes. Passou-se então a analisar estes dois subgrupos (35 x 6 textos) em separado, usando a mesma técnica de **listagem**. Notou-se que, no grupo referente a 1997, um termo aparecia em todos os textos. Com um pouco de conhecimento prévio sobre domínio (*background knowledge*), sabia-se que tal termo designava um escândalo no qual o prefeito era acusado de corrupção e no qual sua esposa fora envolvida também. A conclusão final é de que a esposa do prefeito só aparece em 1997, nesta mídia, envolvida neste escândalo. Já no segmento de 1999, não foi possível identificar termos significativos comuns em todos os 35 textos (com a mesma técnica de **listagem**). Então usou-se a técnica de **agrupamento** sobre este pequeno segmento. Analisando a **listagem** de termos comuns a cada *cluster* resultante, notou-se que os termos eram muito genéricos sobre o assunto. Então estes termos foram eliminados dos textos da coleção e a **agrupamento** foi refeita. Como resultado, foram encontrados dois *clusters* com forte coesão, isto é, cujos documentos tinham alto grau de similaridade entre si. Usando a técnica de **sumarização**, verificou-se que o primeiro *cluster* continha textos que tratavam de uma reunião na casa de uma certa pessoa. No segundo *cluster*, notou-se que os termos “*new*” e “*york*” apareciam em todos os textos. Examinando estes textos com a técnica de **sumarização**, foi possível saber que a esposa do prefeito viveu por uns tempos na cidade de Nova York. A conclusão final é que, em 1999, a esposa do prefeito é citada em diferentes situações.

- **2º Experimento: coleção política**

Utilizando-se a técnica da **diferença** para comparar os termos mais frequentes das duas sub-coleções (1997 x 1999), notou-se que o termo “separação” aparecia somente no segundo segmento. A primeira hipótese levantada era de que o prefeito e sua esposa estavam terminando seu casamento. Para verificar tal hipótese, foram extraídos, com a técnica de **sumarização**, resumos com as frases onde o tal termo aparecia. Os resultados confirmaram a hipótese levantada, o que leva à conclusão de que a separação do casal somente aconteceu após 1997.

- **3º Experimento: coleção médica**

Nesta coleção em particular, havia 26 textos sobre a internação de pacientes diferentes. Realizando um processo de **agrupamento**, descobriu-se um *cluster* forte. Analisando em separado este grupo com a técnica de **listagem**, notou-se a predominância de termos relacionados a “familiares” e “suicídio”. A interpretação inicial para estes resultados é de que “a maioria das pessoas com tendências suicidas possuem família”. Tal hipótese está sendo avaliada por especialistas médicos da área, os quais acharam a princípio tais descobertas interessantes mas merecedoras de estudos mais profundos.

- **4º Experimento: coleção de guerra**

Utilizando a técnica de **associação** sobre as palavras dos textos desta coleção, encontrou-se que em, 100% dos casos (grau de confiança),

- quando o termo “doença” aparecia, então o termo “verão” também aparecia.
- quando o termo “ditadura” aparecia, então o termo “invasão” também aparecia; e
- quando o termo “ditadura” aparecia, então o termo “assassinato” (ou um de seus correlatos, por exemplo, “assassinar”, “matar”) também aparecia.

Tais resultados são interessantes mas não permitem grandes conclusões, já que a coleção não era representativa, como será discutido na próxima seção.

4 Observações e Discussão

Observando os processos de descoberta realizados e seus resultados, pode-se chegar a algumas conclusões e também são levantadas algumas dúvidas. O interesse deste trabalho foi o de analisar três aspectos principais:

- a) que tipos de estratégias podem ou devem ser usadas na abordagem proativa (por exemplo, por onde começar e que passos seguir depois);
- b) qual a importância da intervenção humana no processo e como o conhecimento prévio sobre o assunto pode influenciar tais tipos de abordagem;
- c) que aspectos influenciam tal processo, podendo levar a interpretações erradas.

A seguir, cada um destes aspectos é analisado com base nos experimentos realizados.

4.1 Estratégias para Descoberta Proativa

Um dos problemas do paradigma proativo é definir um plano de uso das técnicas ou de como a coleção textual deverá ser investigada de forma automática pelas ferramentas, a fim de serem descobertas hipóteses.

Kuhlthau [KUH91] determinou seis fases em processo de descoberta de informação: iniciação, seleção, exploração, formulação, coleção e apresentação. Cada fase é caracterizada por atitudes diferentes do usuário (por exemplo, em relação a sentimentos, pensamento, ações e tarefas). Uma das descobertas mais interessantes desta pesquisadora é que o usuário inicia procurando algum tipo de conhecimento mais geral, depois ele procura informação relevante em grupos mais restritos e termina procurando informações mais focadas ou específicas. Durante este processo, o usuário reconhece, identifica, investiga, formula, reúne e complementa o conhecimento.

Watts e Porter [WAT97] propõem um esboço de metodologia (*framework*) sobre algumas ferramentas de descoberta. Entretanto, a estratégia somente foi testada envolvendo problemas da área de Inteligência Competitiva. Neste caso, pode-se dizer que a estratégia

proposta está ainda no paradigma reativo, pois é apropriada para encontrar nomes de pessoas e companhias e para examinar o vocabulário técnico, necessitando bastante conhecimento específico do domínio.

Seguindo as sugestões destes trabalhos e com base nas observações feitas durante os experimentos, foi possível identificar alguns passos comuns. Sugere-se então uma estratégia para descoberta proativa de conhecimento em textos. Não se pode considerar esta estratégia uma metodologia, mas sim um esboço (*framework*), que poderá conduzir os usuários no processo, indicando os passos principais (técnicas ou ferramentas a serem usadas). Os passos são resumidamente descritos a seguir:

- 1) seleção de textos: o primeiro passo é selecionar uma coleção de textos sobre os quais serão aplicadas as técnicas; as técnicas automáticas mais indicadas são a recuperação de informação (que encontra textos procurando por palavras-chave ou termos presentes nos textos) e a classificação (que separa textos por assunto); outra possibilidade, é o usuário mesmo encontrar ou selecionar os textos, o que demanda mais trabalho manual;
- 2) análise da coleção toda ou de partes: neste ponto, o usuário deve decidir se irá aplicar as técnicas de descoberta sobre todos os textos ou sobre partes; a sugestão é que se comece analisando toda a coleção e depois se examine sub-coleções; em alguns casos, nada de interessante é encontrado na coleção toda, o que leva o usuário necessariamente a investigar pequenos grupos; a separação em grupos pode ser feita de forma automática, com a técnica de agrupamento, ou sob algum critério estabelecido pelo usuário, como por exemplo selecionando partes de interesse com as técnicas de recuperação ou classificação;
- 3) análise de grupos de textos (toda a coleção ou partes): uma boa maneira de começar a análise é extraíndo uma lista de termos comuns a todos os textos ou que aparecem em mais de um (técnica de listagem de conceitos-chave); a técnica de diferença pode ser usada depois para levantar novas hipóteses; por fim, a técnica de associação, mesmo que demorada, pode ajudar a descobrir algo interessante;
- 4) comparação de sub-coleções entre si ou em relação à coleção toda: os resultados conseguidos com as técnicas de listagem de conceitos-chave, diferença e associação aplicadas a cada grupo particular podem ser comparados entre si ou com os resultados obtidos com a coleção toda;
- 5) validação de hipóteses: em geral, a técnica de resumos traz bons resultados, pois possibilita ao usuário ler as frases mais significativas e interpretar os resultados;
- 6) retroalimentação: como o processo é cíclico, os passos ou o processo todo podem ser refeitos.

4.2 Necessidade de Intervenção Humana e Conhecimentos Prévios

Uma discussão que surge é se ferramentas de software poderão extrair automaticamente conhecimento a partir de coleções textuais. Os experimentos realizados mostram que é possível automatizar partes do processo de descoberta, minimizando a dependência ao usuário. Entretanto, fica claro que algum tipo de intervenção humana é necessária e útil. Por exemplo, o primeiro passo do processo obrigatoriamente precisa da intervenção do usuário, para selecionar os textos da coleção, seja de forma manual ou fornecendo parâmetros para as ferramentas de recuperação. Também será necessário que o usuário interprete os resultados no contexto da realidade, para que as descobertas sejam úteis. Segundo Aamodt and Nygard [AAM95], o conhecimento é imprescindível para que os dados possam ser interpretados e se

tornem informação. O conhecimento é subjetivo e depende das pessoas. Por isto, Moscarola and Bolden [MOS98b] sugerem o modelo construtivista ao invés do positivista para os processos de descoberta, ou seja, o processo deve ser guiado pelo usuário.

Por outro lado, o conhecimento prévio (*background knowledge*) de que dispõe o usuário ajuda no processo de descoberta, limitando o espaço de pesquisa ou análise, sem que o usuário precise ainda definir hipóteses. Por exemplo, Feldman e Hirsh [FEL97] aplicam as mesmas técnicas descritas em [FEL98] mas permitem que o usuário intervenha no processo, fazendo uso de seus conhecimentos prévios sobre o domínio ou assunto (*background knowledge*). Isto acelera o processo e filtra os resultados de acordo com o interesse do usuário.

Um exemplo de uso do conhecimento prévio nos experimentos, aparece no 2º experimento, quando o usuário interpreta o termo “separação” como algo ligado ao casal citado. Deste caso, conclui-se que não é suficiente encontrar termos comuns ou diferentes. É necessário algum tipo de conhecimento prévio, mesmo que mínimo e limitado à linguagem. Um exemplo de uso de conhecimento sobre o domínio, aparece no 1º experimento, quando o usuário pode verificar que um termo chama mais à atenção que os outros (no caso, o termo que designava o escândalo de corrupção). No 3º experimento, pode-se notar que a falta de conhecimento especializado sobre o domínio pode resultar em descobertas que não podem ser aproveitadas, conseqüentemente também em desperdício de esforços. Pelos experimentos, descobriu-se que uma boa maneira de obter um pouco mais de conhecimento sobre o domínio é examinando os termos mais freqüentes, com a técnica de listagem. Isto permite ao usuário conhecer o estilo dos textos ou o escopo do conteúdo (do que se fala e do que não se fala nos textos). Para maiores discussões teóricas, Choudhury and Sampler [CHO97] discutem tipos de conhecimento prévio em processos de aquisição de conhecimento.

4.3 Ruídos no Processo de KDT

Durante processos de descoberta, alguns aspectos podem influenciar o resultado final. Estes são chamados de ruídos e diminuem a qualidade das descobertas. O primeiro problema notado nos experimentos é o grau de representatividade da coleção. Por exemplo, analisando os resultados do 4º experimento, não se pode afirmar ou concluir que doenças só acontecem em guerras durante o verão. Isto porque a coleção pode não descrever todas as guerras ou os textos podem não conter todas as informações sobre a guerra que descrevem.

Mesmo que a coleção seja representativa, outros tipos de ruído podem aparecer, como o caso dos sinônimos. Por exemplo, no 4º experimento, após utilizar a técnica de diferença, notou-se que somente um texto continha o termo “fuga”, levantando assim a hipótese de que somente uma das guerras teve fuga. Para verificar tal hipótese, foi usada a técnica de sumarização, procurando frases que tivessem termos relativos a fuga (como “fugir”, “escapar”, etc). Os resultados provaram que a hipótese inicial estava errada.

Outro ruído relativo ao vocabulário pode acontecer quando são usados termos polisêmicos, com mais de um significado. Por exemplo, nos experimentos com a coleção política, o termo “família” aparecia referenciando “parentes” ou como parte da “Secretaria da Família e do Bem-Estar”. Isto pode levar a hipóteses erradas quando usando a técnica de listagem, por exemplo. Problemas com o vocabulário, como sinônimos e termos polisêmicos, são discutidos em [FUR87].

Outro cuidado que se deve ter é com o contexto em que os termos aparecem. Por exemplo, o termo “melhora” aparecia freqüentemente nos prontuários médicos em frases negativas (por exemplo, “o paciente não apresentou melhora”). Outros problemas ainda podem surgir por erros ortográficos.

Além disto tudo há ainda o problema da confiabilidade das fontes de informação. No caso da Web, tal problema é ainda mais preocupante, já que os documentos mudam rapidamente [CHE93]. Alguns trabalhos sugerem estratégias para avaliação da qualidade da informação disponível na Web [SCH96] [OWE97] [SMI97]. Hersh [HER95] discute a falta de qualidade em informações textuais e [PAR96] sugere que problemas podem ocorrer devido a 5 tipos de imperfeições:

- informação incompleta: quando faltam detalhes de informação (por exemplo, atributos sem valores);
- informação imprecisa: devido à diferença de granularidade (por exemplo, datas sem o dia);
- informação incerta: quando não pode ser provada;
- informação vaga: devido a imprecisões do vocabulário;
- informação inconsistente: por exemplo, quando há valores contraditórios.

5 Conclusão

Este artigo discutiu o processo de descoberta de conhecimento em textos segundo a abordagem proativa, que é aquela que inicia sem que o usuário tenha hipóteses. Os experimentos realizados com as ferramentas de software implementadas permitiram concluir que tal abordagem é viável, ou seja, é possível realizar descoberta e obter resultados interessantes sem ter algum tipo de hipótese ou interesse inicial.

Exemplos de descoberta foram apresentados para mostrar que técnicas podem ser usadas, como elas podem ser usadas e a que tipos de resultados elas podem levar. O trabalho também apresentou estratégias para descoberta de conhecimento no modo proativo (sem hipóteses iniciais). Também foi discutida a necessidade de intervenção humana no processo e como os conhecimentos prévios sobre o domínio ou sobre a linguagem podem ajudar no processo. As contribuições ainda incluem uma análise dos possíveis problemas, chamados de ruídos, que podem interferir no processo, levando a interpretações errôneas.

6 Agradecimentos

Este trabalho é parcialmente apoiado por CNPq e CAPES.

7 Referências Bibliográficas

- [AAM95] AAMODT, Agnar; NYGARD, Mads. Different roles and mutual dependencies of data, information and knowledge - an AI perspective on their integration. **Data & Knowledge Engineering**, v.16, n.3, Setembro de 1995.
- [AGR93] AGRAWAL, Rakesh; IMIELINSKI, Tomasz. Database mining: a performance perspective. **IEEE Transactions on Knowledge and Data Engineering**, v.5, n.6, Dezembro de 1993.
- [BAE98] BAEZA-YATES, Ricardo e alli. A model and a visual query language for structured text. In: String Processing and Information Retrieval: A South American Symposium - SPIRE'98. **Proceedings...** 1998.
- [BEL97] BELKIN, N. J.; ODDY, R. N.; BROOKS, H. M. ASK for information retrieval: part I. background and theory. In: [SPA97]
- [CAL94] CALLAN, James P. Passage-level evidence in document retrieval. In: VII International ACM-

- SIGIR Conference on Research and Development in Information Retrieval. **Proceedings...** London: Springer-Verlag. 1994.
- [CHE93] CHEN, Z. Let documents talk to each other: a computer model for connection of short documents. **Journal of Documentation**, v.49, n.1, Março de 1993.
- [CHI93] CHINCHOR, Nancy; HIRSCHMAN, Lynette; LEWIS, David D. Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3). **Computational Linguistics**, v.19, n.3, Setembro de 1993.
- [CHO97] CHOUDHURY, Vivek; SAMPLER, Jeffrey L. Information specificity and environmental scanning: an economic perspective. **MIS Quarterly**, Março de 1997.
- [COW96] COWIE, Jim; LEHNERT, Wendy. Information extraction. **Communications of the ACM**, v.39, n.1, Janeiro de 1996.
- [CRO95] CROFT, W. Bruce. Machine learning and information retrieval. In: COLT Conference. **Proceedings...** July 1995. (invited talk). Online in <http://www.ee.umd.edu/medlab/filter/>
- [DAV89] DAVIES, Roy. The creation of new knowledge by information retrieval and classification. **Journal of Documentation**, v.45, n.4, Dezembro de 1989.
- [FAY96] FAYYAD, Usama M. et alli (ed) **Advances in Knowledge Discovery and Data Mining**. Menlo Park, The MIT Press, 1996.
- [FEL95] FELDMAN, Ronen and DAGAN, Ido. Knowledge discovery in textual databases (KDT). In: 1st International Conference on Knowledge Discovery (KDD-95). **Proceedings...** Montreal, Agosto de 1995.
- [FEL97] FELDMAN, Ronen and HIRSH, Haym. Exploiting background information in knowledge discovery from text. **Journal of Intelligent Information Systems**, v.9, n.1, Julho/Agosto de 1997.
- [FEL98] FELDMAN, Ronen; DAGAN, Ido. Mining text using keyword distributions. **Journal of Intelligent Information Systems**, v.10, n.3, 1998.
- [FUR87] FURNAS, G. W. et al. The vocabulary problem in human-system communication. **Communications of the ACM**, v.30, n.11, Novembro de 1987.
- [HER95] HERSH, William R. et alli. Towards new measures of information retrieval evaluation. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval. **Proceedings...** 1995.
- [ING96] INGWERSEN, Peter. Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. **Journal of Documentation**, v.52, n.1, Março de 1996.
- [KAS97] KASZKIEL, Marcin; ZOBEL, Justin. Passage retrieval revisited. In: XX International ACM SIGIR Conference on Research and Development in Information Retrieval. **Proceedings...** Philadelphia: ACM Press, 1997.
- [KUH91] KUHLTHAU, Carol C. Inside the search process: information seeking from the user's perspective. **Journal of the American Society for Information Science**, v.42, n.5, Junho de 1991.
- [LIN98] 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). **Proceedings...** 1998.
- [MAA92] MAAREK, Yoëlle S. Automatically constructing simple help systems from natural language documentation. IN: JACOBS, Paul S. (ed) **Text-based intelligent systems: current research and practice in information extraction and retrieval**. New Jersey: Lawrence Erlbaum, 1992.
- [MCK95] McKEOWN, Kathleen; RADEV, Dragomir R. Generating summaries of multiple news articles. IN: International ACM-SIGIR Conference on Research and Development in Information Retrieval. **Proceedings...** Seattle, 1995.
- [MOE97] MOENS, Marie-Francine; UYTENDAELE, Caroline. Automatic text structuring and categorization as a first step in summarizing legal cases. **Information Processing & Management**, v.33, n.6, Novembro de 1997.
- [MOS98] MOSCAROLA, Jean; BAULAC, Yves; BOLDEN, Richard. **Technology watch via textual data analysis**. Note de Recherche n° 98-14, Université de Savoie. Julho de 1998.
- [MOS98b] MOSCAROLA, Jean; BOLDEN, Richard. **From the data mine to the knowledge mill: applying the principles of lexical analysis to the data mining and knowledge discovery process**. Note de Recherche n° 98-15, Université de Savoie. Setembro de 1998.
- [OAR96] OARD, Douglas W.; MARCHIONINI, Gary. **A conceptual framework for text filtering**. Technical Report, University of Maryland. Maio de 1996. Online at <http://www.ee.umd.edu/medlab/filter/>

- [OWE97] OWENS, Janet; RAGAINS, Patrick. **Evaluating Information Sources**. Janeiro de 1997. Online at <http://www.library.unr.edu/~ragains/eval.html>
- [PAR89] PARSAYE, Kamran et alli. **Intelligent databases: object-oriented, deductive hypermedia technologies**. New York: John Wiley & Sons, 1989.
- [PAR96] PARSONS, Simon. Current approaches to handling imperfect information in data and knowledge bases. **IEEE Transactions on Knowledge and Data Engineering**, v.8, n.3, Junho de 1996.
- [SAL83] SALTON, Gerard; MCGILL, M. J. **Introduction to Modern Information Retrieval**. McGraw-Hill, 1983.
- [SAL97] SALTON, Gerard et alli. Automatic text structuring and summarization. **Information Processing & Management**, v.33, n.2, Março de 1997.
- [SCH96b] SCHAFFER, Doug et alli. Navigating hierarchically clustered networks through fisheye and full-zoom methods. **ACM Transactions on Computer-Human Interaction**, v.3, n.2, Junho de 1996.
- [SCH96] SCHOLZ, Ann. **Evaluating World Wide Web Information**. Fevereiro de 1996. Online at <http://thorplus.lib.purdue.edu/research/classes/gsl75/3gs175/evaluation.html>
- [SMI97] SMITH, Alastair. **Criteria for evaluation of Internet Information Resources**. Março de 1997. Online at <http://www.vuw.ac.nz/~agsmith/evaln/index.htm>
- [SPA97] SPARCK-JONES, Karen; WILLET, Peter (eds). **Readings in Information Retrieval**. San Francisco: Morgan Kaufmann, 1997.
- [SWA97] SWANSON, D. R.; SMALHEISER, N. R., An interactive system for finding complementary literatures: a stimulus to scientific discovery. **Artificial Intelligence**, 91 (1997) 183-203.
- [VEE97] VEERASAMY, Aravindan; HEIKES, Russell. Effectiveness of a graphical display of retrieval results. In: XX International ACM SIGIR Conference on Research and Development in Information Retrieval. **Proceedings...** 1997.
- [WAT97] WATTS, Robert J.; PORTER, Alan L. Innovation forecasting. **Technological Forecasting and Social Change**, 56. 1997.
- [WIL94] WILKINSON, Ross. Effective retrieval of structured documents. In: VII International ACM-SIGIR Conference on Research and Development in Information Retrieval. **Proceedings...** London: Springer-Verlag. 1994.
- [WIL88] WILLET, Peter. Recent trends in hierarchic document clustering: a critical review. **Information Processing & Management**, v.24, n.5, 1988.