Uma Abordagem Completa para a Construção de Taxonomias de Tópicos em um Domínio

Maria Fernanda Moura[1,2], **Ricardo Marcondes Marcacini**[1], **Bruno Magalhães Nogueira**[1], **Merley da Silva Conrado**[1], **Solange Oliveira Rezende**[1]

[1]Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação.
Caixa Postal 668, São Carlos, SP, Brasil - 13560-970

[brunomn,merleyc,solange]@icmc.usp.br

[2]Embrapa Informática Agropecuária.
Caixa Postal 6041, Campinas, SP, Brasil - 13083-970

fernanda@cnptia.embrapa.br

Resumo. Neste trabalho é descrita uma proposta metodológica para a construção de taxonomias de tópicos em um domínio de conhecimento. A proposta vem sendo desenvolvida de forma completamente automatizada, porém permite a intervenção dos especialistas do domínio em algumas de suas etapas, de modo que algumas necessidades de compreensibilidade possam ser satisfeitas. A base metodológica é um processo de mineração de textos, que auxilia a identificação do vocabulário do domínio presente na coleção de textos, utiliza agrupamento hierárquico de documentos e gera automaticamente uma versão da taxonomia; que pode ser editada pelo usuário final com o auxílio de medidas estatísticas de validação das edições realizadas.
Palavras-chave: *taxonomia de tópicos, mineração de textos, rotulação de agrupamentos hierárquicos, seleção de atributos em textos*

1. Introdução

Encontrar tópicos em coleções de textos tem sido uma prática utilizada em aplicações voltadas para a recuperação de informações textuais, como a geração de indexadores para máquinas de busca, ou mesmo a própria apresentação de resultados de busca organizados em grupos mais significativos. Ainda, na maioria dos casos, os tópicos são encontrados sob agrupamentos disjuntos não hierárquicos. A organização hierárquica de tópicos, em geral, é realizada manualmente, por meio de um intenso trabalho humano, como para o site Yahoo, ou completada, como na construção de taxonomias ou ontologias auxiliadas por processos semi-automáticos ((Maedche et al., 2002), (Bloehdorn et al., 2005)), ou na construção de mapas de tópicos ((Librelotto et al., 2004) e (Evangelista et al., 2003)).

Considerando que uma coleção de textos de um domínio de conhecimento deva ser organizada de forma hierárquica para melhor ser compreendida, neste trabalho é proposta uma metodologia para tal fim. Deve-se notar que a metodologia se propõe a auxiliar o processo de organização da coleção de textos, fornecendo um mapa de tópicos que auxilia a análise e compreensão da coleção de textos e que pode ser utilizado junto a ferramentas de recuperação de informação sobre a coleção, fornecendo bons indexadores de busca. O objetivo da metodologia é obter uma organização que permita construir mapas de tópicos que reflitam exatamente as publicações existentes e recuperados em um domínio.

A proposta em desenvolvimento é completamente automatizada, porém permite a intervenção dos especialistas do domínio em algumas de suas etapas, de modo que algumas necessidades de compreensibilidade possam ser satisfeitas, se os especialistas assim o desejarem. Caso contrário, o ferramental pode ser utilizado sem intervenções. A base metodológica é um processo de mineração de textos, que auxilia a identificação do vocabulário do domínio presente na coleção de textos, utiliza um algoritmo de agrupamento hierárquico e gera automaticamente uma versão da taxonomia, que pode ser editada pelo usuário final com o auxílio de medidas estatísticas de validação das edições realizadas. Além disso, todo o suporte computacional vem sendo desenvolvido de modo a poder acoplar novos procedimentos, sempre que se mostrem necessários ou alternativos aos já incorporados, orientando os especialistas do domínio na escolha dos mesmos, ou escolhendo-os automaticamente.

Ainda, a metodologia aqui proposta trata-se de uma compilação das iniciativas já amplamente discutidas na literatura; porém, procurando cobrir cada passo do processo com soluções simples e o mais automatizáveis possíveis, bem como deixando espaço para novos trabalhos de pesquisa e desenvolvimento em cada uma das etapas. Não se pretende, com as soluções automatizáveis, retirar o papel do especialista do domínio na construção dessas taxonomias, mas sim, deixá-lo livre pra escolher em que etapas intervir no processo, ou mesmo, não intervir.

No próximo item são discutidos alguns trabalhos relacionados a este. Então, a metodologia é apresentada, seguida de um dos exemplos utilizados na validação, com a utilização de algumas ferramentas desenvolvidas, e a discussão dos resultados obtidos, bem como algumas necessidades futuras.

2. Trabalhos Relacionados

Há muitos trabalhos na literatura sobre hierarquia de tópicos visando a melhoria de processos de busca web, isto é, encontrar esses agrupamentos acaba por tornar mais eficientes as máquinas de busca e, por muitas vezes, a perda de detalhes de informação não é tão importante, isto é, não altera muito os resultados de busca; pois os resultados serão julgados *a posteriore*, quando um usuário os recupera. No entanto, o objetivo deste trabalho é obter uma taxonomia de alta qualidade, pois se quer organizar a informação de um domínio restrito e o procedimento todo é realizado com a intervenção de um especialista e conseqüentemente no seu ritmo de trabalho, com o único objetivo de facilitá-lo. Assim, nesta revisão encontram-se apenas os trabalhos mais diretamente relacionados aos objetivos deste projeto de pesquisa, que utilizam mais abordagens estatísticas e mais contemporâneos.

Na linha de encontrar uma sumarização para os tópicos de uma clusterização de documentos, no trabalho de Neto et al. (2000) foi usado o software AutoClass para obter uma boa clusterização de um conjunto de documentos e então cada tópico foi sumarizado. Foi realizado um pré-processamento convencional dos textos: *case-folding*, para diferenciar maiúsculas de minúsculas; *stemming*; eliminação de *stopwords*; e *n-gram words*, utilizando apenas trigramas. A medida utilizada para eliminar os casos com peso de ocorrência zero foi a tf-idf. Obtidos os clusters, as palavras mais descritivas de cada cluster foram consideradas suas palavras-chaves e para o documento centróide elaborou-se um sumário dos documentos mais relacionados a ele. No trabalho é proposto que se con-

siderem as palavras-chaves como um ultra-resumo, que ajuda a identificar os tópicos de cada cluster. Então, quando o usuário julga algum cluster mais interessante pode resumir os seus documentos para ter uma melhor idéia do que foi encontrado. Como a abordagem usada é *bag of words*, não se conhece a ordem das palavras, logo para resumir (sumarizar) foi utilizado um método extrativo para obter as sentenças mais significativas dos textos, usando um peso semelhante a tf-idf e estabelecendo um *threshold* para cortes. A clusterização obtida não é hierárquica e a avaliação do método é muito subjetiva, uma vez que o objetivo é fornecer uma ferramenta de análise exploratória para o especialista mais voltada à obtenção de resumos e palavras-chaves dos clusters.

No trabalho de descoberta e comparação de hierarquia de tópicos de Lawrie e Croft (2000) o interesse é gerar taxonomias de domínios específicos, clusterizar um conjunto grande de documentos e então obter hierarquias de cada cluster, isto é, hierarquias sob cada tópico. A idéia é que grupos homogêneos de documentos propiciam melhores hierarquias, pois o conjunto de atributos/termos tem um uso mais coeso e melhor delimitado, ou seja, com a restrição do domínio reduz-se o problema de polissemia. Assim, foram obtidas hierarquias pré-assumidas com características de dependência semântica e hierarquias léxicas; e propuseram duas diferentes formas de avaliar as diferenças entre hierarquias, com duas diferentes métricas. A primeira mede a velocidade com que se chega ao conjunto de documentos relevantes, correspondendo a quantas arestas anda-se no grafo, e outra quantifica a similaridade entre duas hierarquias. A questão da similaridade entre duas hierarquias é razoavelmente complicada de ser calculada e a proposta colocada é bastante interessante. O resultado final é um *k-means* modificado e cluster hierárquico num domínio restrito, bastante parecido com o problema de pesquisa deste trabalho.

Procurando-se por atributos mais significativos nos clusters obtidos, tem-se métodos para clusterizar e atribuir pesos simultaneamente como proposto por Frigui e Nasraoui (2004). Cada cluster é caracterizado por um conjunto de palavras-chave, então encontra os pesos das palavras-chave não só para particionar os documentos em grupos semânticos mais significativos, mas também para indicar a relevância dos termos nos clusters. o trabalho também mostra como levar essas relevâncias para uma proposta nebulosa (*fuzzy*). A medida utilizada para o termo no documento, casela da matriz atributo-valor, precisa ser contínua para ser mais rica em informação, dado que o algoritmo proposto é utilizado para resolver, simultaneamente, os clusters ótimos e os pesos dos atributos. O problema é que é necessário estabelecer o número de clusters a priori e não é hierárquico. Assim, embora essa fosse uma alternativa aos métodos que apenas encontram centróides, as medidas de relevância dos atributos são atualizadas junto à clusterização e dependem completamente do método; logo, para utilizar as mesmas idéias é necessário adaptá-las a outros métodos de interesse.

Para a melhoria de taxonomias de documentos, alguns autores trabalham com uma classificação pré-existente, como no caso de Punera et al. (2005). Nesse trabalho, a partir de um corpus de documentos rotulados extrai-se uma estrutura hierárquica viável; construída a partir do conhecimento prévio das classes e, para essas classes, então usam um classificador SVM refinar a classificação dos documentos; ou seja, o objetivo é melhorar a taxonomia pré-existente. Já em trabalhos em que o objetivo é extrair a taxonomia a partir dos documentos, como em Dupret e Piwowarski (2005), se o domínio não é res-

trito a representação *bag of words* não facilita a enfrentar problemas de polissemias e sinônimos. Nesse trabalho, por exemplo, procura-se explorar a indução de taxonomias de termos, a partir de uma decomposição espectral do tipo *singular value decomposition* (SVD) e identificar conceitos. Os conceitos devem ser ortogonais, logo não correlacionados e então menos sensíveis ao vocabulário particular utilizado nos documentos. Esse autores definem conceito a partir dos *ranks* dos termos na decomposição obtida; ou seja, a base matemática é uma decomposição espectral da matriz de similaridades obtida. Obtêm a taxonomia, mas não mostram a relação de documentos carregados juntamente, pois não existe essa preocupação no trabalho; é obtido um grafo direto e acíclico de termos e conceitos.

Na linha de obter taxonomias a partir dos documentos, o trabalho de Kashyap et al. (2004a) tem por objetivo reduzir o esforço humano para construir ontologias. O ponto de partida é uma clusterização hierárquica de documentos, obtendo uma hierarquia de tópicos, ou seja, uma taxonomia de conceitos; para chegar a isso, clusterizam os documentos, extraem uma hierarquia de tópicos desse agrupamento, e, assinalam rótulos aos tópicos. A seguir, no trabalho, é definida uma taxonomia como um sistema de organização do conhecimento que representa relações entre tópicos de tal forma que eles arranjem os conceitos dos mais genéricos para os mais específicos - conceitos mais formais foram deixados para trabalhos futuros. Eles também utilizaram técnicas de processamento de língua natural na geração da taxonomia, porém restritas ao pré-processamento, para retirar sintagmas verbais da análise. A técnica de cluster foi o *bissecting k-means* e o PDDP (*Principal Direction Divisive Partitioning*); extraindo-se a taxonomia a partir da hierarquia e avaliando a qualidade da mesma. Após a clusterização, são calculados parâmetros que refletem a coesão dos grupos e que são utilizados pelo algoritmo de extração da taxonomia; o que, de fato, resulta em pontos de poda na hierarquia inicialmente obtida, ou seja, ficam-se com os nós mais relevantes apenas - uma árvore mais enxuta e mais coesa. Assim, estabeleceram um procedimento para rotular a taxonomia e também para avaliar. As técnicas de avaliação exigem a comparação com uma taxonomia padrão ou então apenas a avaliação estrutural. No trabalho do TaxaMiner as hierarquias são obtidas de uma coleção de textos controlada, recuperada a partir de um critério de busca que garante a relação de pertinência em uma hierarquia pré-conhecida. Os autores mostram que, de acordo com a presença de termos mais amplos (*broader*) nos nós antecessores e mais específicos (*narrower*) nos nós descendentes de uma taxonomia pré-definida, conseguem avaliar o quanto o método se aproxima da taxonomia existente. Em um segundo trabalho (Kashyap et al., 2004b), os autores usaram a técnica *Latent Semantic Indexing* (LSI) para reduzir a dimensão do problema e auxiliar a identificação da desambigüidade de termos; propondo algumas adaptações aos critérios de avaliação da taxonomia, com base na nova forma de indexar (encontrar as medidas da matriz atributo-valor) os termos nos documentos - usando os *scores* da decomposição da matriz de termos na LSI.

Trabalhos que partem da existência de uma taxonomia e procuram melhorá-la contém idéias interessantes para o crescimento dos agrupamentos. Em Tang et al. (2006) dada uma hierarquia/taxonomia e um conjunto de treinamento, ajustam-na de acordo com o aprendizado sobre o conjunto de treinamento, considerando-o como novos exemplos a serem classificados e mostram que vão obtendo melhores hierarquias em relação à original. A abordagem de Zhao e Karypis (2005) parte de uma hierarquia de tópicos pré-definidos, por um especialista do domínio; dado que essa hierarquia é incompleta,

procuram montar os clusters a partir dela e para completá-la.

Algumas abordagens para a obtenção de taxonomias são muito dependentes dos conceitos de ontologias e processamento de língua natural, especialmente trabalhos como os de Hotho et al. (2002) e Jiang e Tan (2005), que usam analisadores léxicos e sintáticos bastante complexos, que consultam ontologias pré-definidas, para encontrar as similaridades termo a termo e expandir ou retrair as informações encontradas. Nessa linha, encontra-se a ferramenta Text-to-Onto, que está bem longe do escopo original deste trabalho.

Neste trabalho a idéia é usar uma abordagem estatística para o problema, mais próxima a do ambiente TaxaMiner, porém sem uma taxonomia pré-fixada para validar os resultados. Em lugar de taxonomias pré-fixadas pretende-se utilizar a intervenção do especialista do domínio, mediante a análise de parâmetros fornecidos. Obtida a primeira taxonomia, poder-se-á adaptar métodos de ajustes de taxonomia ou outros para crescer os clusters, mas sempre se permitindo que o especialista interfira no processo. Assim, nenhuma das abordagens vistas resolve o problema proposto, como um todo; pode-se utilizar várias das idéias colocadas, mediante adaptações, para resolvê-lo, que é a abordagem deste trabalho.

3. Metodologia Proposta

Considera-se um domínio específico de conhecimento por dois motivos maiores: evitar polissemias, dado que o tratamento de dados é puramente estatístico; e, evitar a obtenção de sub-tópicos sem nexo. Assim, organiza-se hierarquicamente a coleção e tenta-se identificar os tópicos e subtópicos nos quais os documentos se inserem. Embora a proposta seja bastante automatizada, permite-se a intervenção dos especialistas do domínio em algumas de suas etapas, fornecendo-se critérios estatísticos para guiar as escolhas.

A base metodológica é um processo de mineração de textos, cujas etapas são ilustradas na Fig. 1. Em um primeiro passo, faz-se uma **Identificação do Problema**, na qual delimita-se o problema a ser trabalhado, bem como as fontes de dados e as ferramentas utilizadas. Logo após, realiza-se um **Pré-processamento** da coleção de documentos, padronizando-a e delimitando o vocabulário que a representa. Passa-se, então, para a etapa de **Extração de Padrões**, por meio da utilização métodos de agrupamento hierárquico e rotulação de hierarquia. Com a taxonomia extraída, efetua-se o **Pós-processamento** da mesma, procurando validá-la e refiná-la. Por fim, avança-se à exploração e **Uso do Conhecimento**. Todas estas etapas estão detalhadas nas seções a seguir.

3.1. Identificação do Problema

Nessa etapa definem-se os objetivos da aplicação da metodologia, que poderia ser identificar tendências ou simplesmente auxiliar a organização de uma coleção de documentos. A identificação do problema é uma etapa muito importante, dado que não existe descoberta de conhecimento sem demanda pelo mesmo. Nesta etapa o especialista do domínio identifica e delimita o problema, o sub-domínio do problema, a coleção de textos a ser analisada ou sua fonte de busca, se existe algum pré-conhecimento de domínio que possa ser utilizado na análise, o que se espera obter e como os resultados poderão ser utilizados. É uma etapa que demanda muito esforço tanto do especialista do domínio quanto do

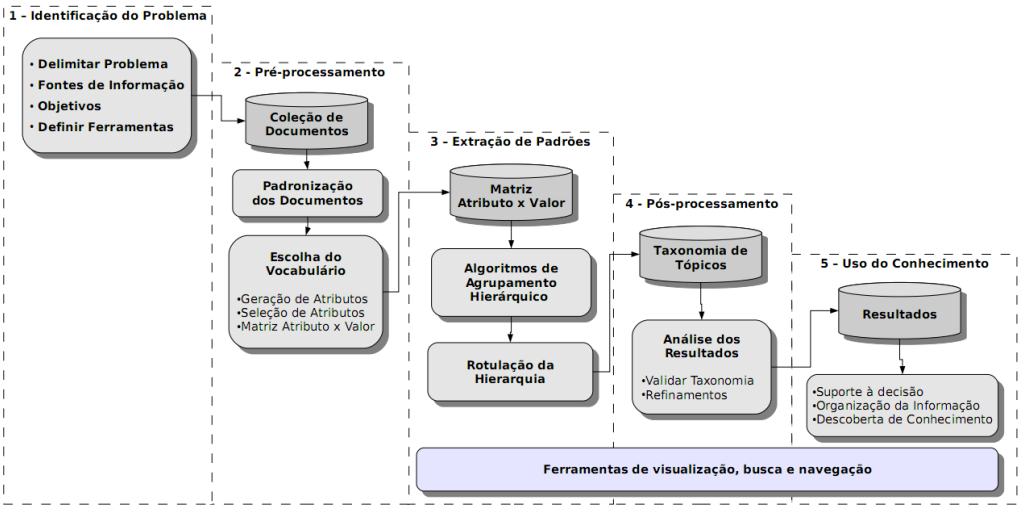

Figura 1. Metodologia pra obtenção de taxonomia de tópicos baseada em mineração de textos

especialista em mineração de dados, pois esta etapa fornece subsídios a todo o processo, permitindo identificar requisitos e possíveis ferramentas para cada passo.

Identificado o problema, a coleção de textos, que efetivamente corresponde aos dados a serem analisados, deve ser delimitada com o auxílio do especialista do domínio do problema. Neste ponto é conveniente que sejam observadas as seguintes restrições e requisitos:

- a coleção de textos a ser tratada deve estar em meio digital que possa ser convertido para arquivos ASCII comuns; por exemplo: documentos gerados por algum editor de textos ou apresentações e resultados de pesquisa em uma máquina de busca para *web*. Se os documentos a serem analisados não se encontrarem em meio digital, por exemplo, se forem relatórios ou publicações antigas, deve-se também avaliar o custo do processo de digitalização;
- deve-se conhecer a língua ou línguas em que os textos estejam escritos, de modo que se possa pelo menos classificá-los por esse quesito. Boa parte das ferramentas automáticas de mineração de textos são dependentes da língua, dado que no mínimo realizam uma análise léxica dos textos (Martins, 2003). Assim, se os textos em análise estiverem em mais de uma língua é necessário avaliar se essa informação pode ser desprezada ou é objeto da análise;
- dados textuais provenientes de alguma base de dados pré-organizada ou publicações catalogadas em alguma biblioteca possuem metadados - **metadados** são dados que descrevem os dados originais em um padrão pré-estabelecido e complementam as informações sobre os mesmos (Duval et al., 2002; Souza et al., 2005). Nesses casos, os metadados devem ser considerados, pois existe uma pré-classificação (rotulação) dos dados, que pode ser explícita ou implícita;
- outra questão a ser considerada é o número de figuras nos documentos da coleção e a sua importância entre os mesmos. Quando os documentos são convertidos para texto simples sem formatação, as figuras são perdidas, restando muitas vezes, apenas as legendas. Logo, se essas figuras pertencerem a alguma base de dados e houver metadados a respeito delas, há que se considerar trocá-las nos textos

convertidos por seus metadados; se a operação for factível.

Determinar o que será feito em termos de mineração dos dados é um processo que se estende desde a delimitação do problema (Ebecken et al., 2003); depende, além da identificação, das necessidades, dos objetivos estabelecidos e do quanto se dispõe de dados e informações sobre eles, bem como do quanto se dispõe de ferramentas adequadas à solução do problema.

Assim, nesta etapa, delimita-se o escopo de tratamento de dados e a fonte de informação, bem como as ferramentas que devem ser utilizadas no processo e, conseqüentemente, a forma de representação dos dados e resultados.

3.2. Pré-processamento

No pré-processamento faz-se uma padronização dos documentos, obtém-se a relação de atributos de interesse e reduz-se a dimensão dos atributos obtidos.

Durante a **padronização** analisa-se subjetivamente o número de documentos disponíveis que possam ser utilizados no processo. Nem sempre os documentos recuperados para serem utilizados estão em formatos que permitem seu uso direto e, muitas vezes, após as conversões, um grande volume de dados é descartado devido a problemas como trocas de caracteres ou perda de partes de palavras. Esta sub-etapa, muitas vezes, é repetida após novas buscas por documentos e ou outras fontes dos mesmos até que se obtenha uma coleção de textos confiável. À coleção final aplicam-se algumas padronizações, como transformar todas as letras em minúsculas, eliminar caracteres especiais, eliminar *stopwords* e inserir tags XML para delimitar os textos, bem como delimitar metadados, quando existentes.

Ainda no pré-processamento, tem-se como meta identificar e selecionar o método que melhor se adapte às necessidades desta metodologia para a **geração de atributos**, que são os termos de interesse. **Termo**, neste trabalho, é usado no mesmo contexto que para recuperação de informação, pois se considera uma palavra ou composição de palavras, removidas as inflexões ou não. Tratam-se os termos com uma única palavra como *unigramas*, com duas palavras como *bigramas* e com três como *trigramas*; e as coleções sempre como *bag-of-words*. Para esta geração, busca-se por técnicas que simplifiquem as diversas formas de apresentação de termos mantendo o mesmo significado. Entre as técnicas mais utilizadas, encontram-se a radicalização, lematização e substantivação.

A técnica de *stemming* ou "stemmização", aqui tratada como radicalização, reduz as palavras às suas formas inflexionáveis e, às vezes, às suas derivações, ou seja, eliminação de prefixos e sufixos das palavras ou a colocação de um verbo em sua forma infinitiva (Manning et al., 2008), sendo cada palavra analisada isoladamente. No trabalho de (da Silva Conrado e Rezende, 2008), é avaliada a geração de termos utilizando a técnica de radicalização a partir de coleções textuais de domínios específicos.

O processo de *stemming* pode depender da linguagem, pois existem algoritmos de radicalização que necessitam de conhecimento lingüístico (Silla Jr e Kaestner, 2002). Porém, os mais atuais não utilizam informações do contexto para determinar o sentido correto de cada palavra (Ebecken et al., 2003), devido à maioria das palavras poder ser considerada por um único significado, sendo que nestes casos, o contexto não ajudará no processo de *stemming*. No entanto, deve-se atentar aos possíveis erros resultantes de

análise incorreta do sentido das palavras, já que tais algoritmos ignoram o significado dos termos levando, às vezes, a erros. Como algoritmos de radicalização para a língua inglesa, pode-se citar o de Lovins (1968), Stemmer S (Harman, 1991) e de Porter (*Porter Stemming Algorith*) (Porter, 1980); para o português e espanhol surgiram algumas adaptações a partir desses algoritmos, como a ferramenta PreTexT (Martins et al., 2003) e Stemmer (Orengo e Huyck, 2001).

A **lematização** reduz as palavras a seus lemas. Para as línguas francesa e inglesa, tem-se o software proprietário Sphinx, versão 4 (Sphinx Brasil, 2008). Já para o português, tem-se o Lematizador de Nunes (Nunes, 1996). Há também ferramentas que etiquetam morfologicamente as palavras, como MXPOST (*Maximum entropy pos tagger*) (Ratnaparkhi, 1996), TreeTagger (Schmid, 1994) e o etiquetador de Brill (Brill, 1995), sendo necessário, em seguida, a utilização de ferramenta que lematize tais palavras.

A **substantivação** ou "Nominalização" faz com que as palavras exibam um comportamento sintático/semântico semelhante àquele próprio de um nome[1]. Para a língua portuguesa, cita-se as ferramentas FORMA e CHAMA (Gonzalez et al., 2006).

Com a utilização destas técnicas, é possível identificar e selecionar a que obtém termos simplificados que melhor se adapte às necessidades desta metodologia para a geração de termos.

Para os *bigramas* e *trigramas* são realizados alguns testes de escolha, considerando-se a ordem de ocorrência de cada um de seus componentes em toda a coleção. Essas escolhas tentam identificar potenciais colocações do domínio na coleção, que são posteriormente tratadas como termos únicos; além disso, são eliminadas combinações estatisticamente não significantes (Banerjee e Pedersen, 2003).

Por exemplo, eliminadas as *stopwords* e caracteres especiais na frase: "tecnicas inteligencia artificial aplicacoes tem sido utilizadas sistemas apoio decisao"; "tecnicas-inteligencia-artificial", "inteligencia-artificial", "sistemas-apoio-decisao" e "sistemas-decisao" são potenciais colocações e termos do domínio, mas "tecnicas-inteligencia", "artificial-aplicacoes", "aplicacoes-tem-sido" e "aplicacoes-tem" não acrescentam termos relevantes ao domínio; logo, se os testes assim o indicarem essas combinações são eliminadas. Como os testes estatísticos, em geral, resultam em *ranking* de atributos, pode-se optar por um corte pré-definido ou deixar que o especialista do domínio decida o corte.

O número de atributos, mesmo após um cuidadoso processo de limpeza da coleção e geração dos mesmos, é exageradamente grande e, não possui representatividade em cada documento da coleção, levando a representações esparsas de suas ocorrências. Desta forma, outra tarefa do pré-processamento é a **seleção de atributos**. A escolha de um método de seleção de atributos adequado nesta sub-etapa contribuirá muito com a performance dos algoritmos de aprendizado utilizados nas demais etapas. Como a metodologia faz uso de aprendizado não-supervisionado, sem rótulos de classes, existe uma dificuldade a mais ao se delimitar o que é relevante. O método mais comumente utilizado corresponde aos cortes de Luhn (Luhn, 1958). Para realizar esses cortes ordenam-se as freqüências de ocorrência dos termos na coleção e, após obter o gráfico dessa curva, escolhem-se, subjetivamente, os pontos de corte, próximos aos pontos de inflexão, considerando-se que palavras com freqüência de ocorrência muito baixa ou muito alta sejam irrelevantes. Porém,

[1] http://www.dacex.ct.utfpr.edu.br/paulo3.htm

a eliminação de termos de baixa ocorrência não é consenso, sendo inclusive privilegiada pelo indexador tf-idf que é muito utilizado em recuperação de informação (Salton et al., 1975).

Outro método bastante difundido é o de Salton et al. (1975), o qual usa a medida de DF (*document frequency*: número de documentos nos quais um determinado termo aparece) para a seleção dos termos. Nele é sugerido considerar termos que possuam DF entre 1% e 10% do número total de documentos, sendo considerado um corte bastante agressivo, reduzindo muito o número de termos.

Nogueira et al. (2008) propõem um método denominado "Luhn-DF" aproveitando as idéias dos cortes de Luhn e Salton. No método proposto, geram-se os histogramas das DF dos termos de forma descendente, efetuando os cortes nos pontos de inflexão da curva de tendência, tal qual o primeiro método supra-citado. Este método seleciona, assim como o de Salton, termos cuja DF não é tão grande, nem tão pequena, sendo, porém, menos agressivo.

Além destes três métodos, outros métodos de seleção de atributos baseados em variância dos termos e similaridade de documentos são explorados em trabalhos da área, apontando bons resultados. Por exemplo, *Term Contribution* (TC) (Liu et al., 2003) é uma medida que representa o quanto um termo contribui para a similaridade entre documentos na coleção. Outra medida é *Term Variance* (TV) (Liu et al., 2005), a qual calcula a variância de todos os termos da coleção, atribuindo os maiores *scores* àqueles termos que não possuem baixa freqüência em documentos e possuem uma distribuição não-uniforme ao longo da coleção. Por fim, *Term Variance Quality* (TVQ) (Dhillon et al., 2003) é uma outra medida bastante similar à TV, usando a variância no cálculo da qualidade dos termos.

Vários experimentos comparando a eficiência destes métodos podem ser encontrados em (Nogueira et al., 2008). Ao final dessa etapa, gera-se a **matriz atributo-valor** para representar os dados (documentos). Cada documento é representado por um vetor comum de atributos, no qual os valores considerados para cada atributo representam suas freqüências de ocorrência nos documentos. A matriz atributo-valor é utilizada na etapa de **extração de padrões**, na qual realiza-se um agrupamento hierárquico dos documentos e rotula-se automaticamente os agrupamentos obtidos.

3.3. Extração de Padrões

Esta metodologia se propõe a obter uma taxonomia de tópicos a partir de um agrupamento hierárquico de documentos de um único domínio. Os tópicos são inferidos a partir de listas de termos mais discriminativos de cada agrupamento, esse processo é conhecido como rotulação de agrupamentos (*cluster labelling*). Tem-se como meta separar o algoritmo de agrupamento utilizado do processo de seleção de termos mais discriminativos em cada grupo para a rotulação dos grupos; lembrando que os termos correspondem aos atributos escolhidos. Dessa forma pode-se utilizar diferentes algoritmos de agrupamento e, conseqüentemente, trabalhar diferentes critérios de corte dos agrupamentos, sem prejuízo das demais escolhas. Assim, dada a matriz atributoxvalor gerada a partir dos critérios de escolha do vocabulário, geração e seleção de atributos, constrói-se um agrupamento, aplicam-se critérios de corte aos agrupamentos e então geram-se seus rótulos.

Dada uma hierarquia qualquer de documentos, para a qual sejam conhecidas as

freqüências de ocorrência de cada termo em cada nível da hierarquia (grupo), aplica-se um processo próprio de **rotulação da hierarquia** (Moura e Rezende, 2007), que depende exclusivamente das freqüências observadas de cada termo em cada grupo de documentos. Essa rotulação permite gerar uma versão da taxonomia na qual não se encontram repetições de termos ao longo do mesmo ramo; dado a suposição que rótulos de nós pais aplicam-se também aos seus filhos. A taxonomia é gerada a partir dos termos mais discriminativos de cada nó da hierarquia, funcionando como mais um processo de seleção de atributos, porém supervisionado. O método de rotulação, além da seleção de termos discriminativos, faz uma análise objetiva dos termos não discriminativos, eliminando os que estatisticamente não contribuem para discriminar ramos da coleção de textos. Além disso, pode ser feita uma análise subjetiva sobre os termos que se encontram em altos níveis da hierarquia visando identificar *stopwords da coleção*. Essas *stopwords da coleção* são aqueles termos que o método automático não pode eliminar com segurança, dentro da margem estatística imposta ao teste de aceitação ou rejeição do termo, mas que os especialistas do domínio conseguem identificar visualmente como desnecessários à compreensão dos grupos. Sempre que termos são retirados da análise é necessário reprocessar as etapas anteriores.

Para teste dos métodos de rotulação da hierarquia, vêm sendo utilizados **algoritmos de agrupamento hierárquico** aglomerativos, com a medida de similaridade de cosseno, para obter a relação hierárquica entre os textos da coleção. Os algoritmos utilizados têm sido variados entre o "*single linkage*" para identificar documentos que fogem aos padrões na coleção (discussão sobre "*outliers*" em (Mardia et al., 1979)) e, o "*average*" ou o "*complete linkage*" para obter a hierarquia final. Paralelamente, corre um trabalho de pesquisa para encontrar candidatos a melhores podas do cluster, bem como testes de outros algoritmos. Deve-se observar, no entanto, que a taxonomia pode ser gerada sobre os resultados de qualquer algoritmo de agrupamento hierárquico.

3.4. Pós-processamento

No **pós-processamento** realiza-se uma avaliação objetiva, com base na acurácia demonstrada na recuperação de informações da coleção, utilizando-se os conjuntos de termos discriminativos como expressões de busca. A taxonomia é, também, subjetivamente avaliada por especialistas do domínio, apoiados por ferramentas de visualização. Caso a taxonomia não seja satisfatória, escolhem-se etapas do processo às quais se pode voltar, para que alguma mudança satisfaça critérios aí identificados. Quando a taxonomia é aceita, ela pode ser diretamente utilizada ou, ainda, passar por um processo de refino.

No **refinamento** da taxonomia obtida, os especialistas podem editá-la, auxiliados por medidas estatísticas de validação das edições realizadas. Os especialistas são orientados pelos critérios de corte de ramos dos agrupamentos e de corte do vocabulário utilizado, mas detêm a decisão final, mesmo se contrária às indicações estatísticas. A ferramenta de edição permite que os especialistas criem novos termos, que podem ser resumos de conjuntos de termos; por exemplo, determinar que um ramo corresponde à "família" e que pode ser "filho", "pai" e "mãe". Ou seja, a edição permite que se defina um termo geral, para abrigar os termos identificados discriminativos no ramo, de forma completamente subjetiva.

3.5. Uso do Conhecimento

Na etapa de **uso do conhecimento**, a taxonomia pode ser utilizada para facilitar processos de organização e recuperação de informação, bem como a própria compreensão da coleção de textos organizada; ou mesmo servir de suporte a sistemas de apoio à decisão.

Para apoiar esta etapa, é importante que sejam desenvolvidas técnicas de auxílio ao especialista do domínio com a finalidade de facilitar o entendimento e a utilização do conhecimento adquirido (Silberschatz e Tuzhilin, 1995). As técnicas e ferramentas para visualização de informação cumprem este objetivo, pois facilitam o reconhecimento de relacionamentos, tendências e padrões do conjunto de dados analisado, potencializando a exploração do conhecimento (Rezende et al., 1998).

Nesta metologia, é utilizada uma ferramenta específica para exploração da taxonomia (Marcacini, 2008), que permite navegar na taxonomia através de diferentes representações gráficas, sumarizar grandes quantidades de dados, realizar buscas, visualizar e editar informações pertencentes a cada tópico, comparar duas ou mais taxonomias obtidas por métodos distintos e, também, fornecer uma interface de avaliação para que seja possível validar, de forma subjetiva, a taxonomia resultante de um processo de mineração de textos.

4. Validação da Metodologia em Desenvolvimento

Vários conjuntos de dados têm sido utilizados em experimentos com objetivos específicos de produção, como no caso do estudo de tendências em produção científica em gado de corte e leite junto à Embrapa Pecuária Sudeste (Mazzari, 2007) e organização de uma biblioteca digital (Marcacini et al., 2007), bem como em estudos de prospecção e validação de métodos e tecnologias utilizados no projeto de construção do ambiente computacional para a aplicação da metodologia. Para os estudos, têm sido utilizadas bases de dados do domínio de inteligência artificial e afins, devido à disponibilidade de especialistas do domínio em grupos de pesquisa próximos que colaboram nas avaliações.

De acordo com essa suposição, neste trabalho, o uso da metodologia é exemplificado com uma coleção de textos do domínio de lingüística computacional, formada por 51 artigos completos, em português, apresentados nos últimos *workshops* anuais de "Tecnologia da Informação e da Linguagem Humana", TIL, de 2005 a 2007. Ao processo de padronização, aplicado à base de textos do TIL, foi acrescentada a retirada de informações às autorias dos artigos e às referências bibliográficas[2], para que os termos utilizados nesses itens, bem como os nomes próprios, não interferissem nos resultados.

A metodologia foi aplicada aos artigos padronizados, primeiro obtendo-se os atributos na forma de *unigramas* com o uso da ferramenta PreText (Matsubara et al., 2003), que elimina as *stopwords* e inflexões. Para os 51 artigos foram obtidos 7300 *unigramas*. A seguir, obteve-se o gráfico de ocorrência dos *unigramas* na coleção e foram definidos os cortes de Luhn, com valor mínimo 12 e máximo 450; resultando em 1154 *unigramas* (atributos). O agrupamento hierárquico dos documentos foi obtido com o MatLab, escolhendo-se similaridade de cosseno e o algoritmo "*average*". À hierarquia aplicou-se o algoritmo próprio de obtenção dos rótulos dos agrupamentos (Moura e Rezende, 2007).

[2]http://labic.icmc.usp.br/projects

Então, na primeira avaliação subjetiva visual, realizada pela equipe de mineração de textos, verificou-se que o resultado ainda era muito complexo e não permitia reconhecer eficientemente os grupos.

As observações dos resultados resultaram na identificação de problemas, como o grande valor estatisticamente discriminativo dos verbos e de algumas palavras muito comuns a artigos científicos, como "trabalho", "artigo", "figura" e "tabelas". Assim, realizaram-se escolhas subjetivas de *unigramas* a serem eliminados do processo, que passaram a ser tratados como *stopwords* específicas da coleção de documentos. Ao todo foram identificados 290 *unigramas* a serem eliminados. Após isso, repetiram-se os procedimentos do pré-processamento, aplicando-se os cortes de Luhn, coincidentemente, novamente com um mínimo de 12 e máximo de 450, restando 870 *unigramas* a serem considerados como atributos. E, novamente, foi aplicado o algoritmo de cluster *"average"* com medida de similaridade de cosseno e, a seguir, o método próprio de rotulação de cluster hierárquico, ordenando-se os rótulos por freqüência e mantendo um máximo de 30 rótulos por grupo, isto é, os 30 *unigramas* mais discriminativos de cada grupo.

Para exemplificar o resultado, na Fig. 2, pode-se observar um ramo da taxonomia, cujo comportamento é muito semelhante antes e depois do corte dos termos subjetivamente considerados *stopwords*. Alguns dos *unigramas* cortados estão hachurados na figura, em sua parte 1. Deve-se notar que a sua retirada não alterou a semântica dos ramos, apenas melhorou a sua interpretação. Na parte 2 da figura, nota-se claramente que os artigos sob esse ramo tratam de modelos de extração de conceitos e termos de domínio, que manipulam ontologias, por meio de várias ferramentas de construção de ontologias (owl, protegé, etc).

Em alguns outros experimentos, nos quais as coleções correspondiam a resumos dos textos, e não a textos completos, o peso dos verbos e termos específicos, como tabela e figura, não tiveram tanta importância. Dessa experiência, incorporou-se aos requisitos do ambiente computacional de suporte à metodologia, também a necessidade de uma ferramenta para eliminar verbos na obtenção dos atributos. Para eliminar verbos pode-se utilizar alguma ferramenta de identificação de partes do discurso no texto ou que, pelo menos, identifique sintagmas verbais. Uma ferramenta desse tipo deve ser incorporada ao pré-processamento, imediatamente antes da obtenção de atributos.

5. Considerações Finais

Neste trabalho é apresentada, em linhas gerais, uma proposta de metodologia para a construção de taxonomias de tópicos, com base em um processo de mineração de textos, utilizando agrupamento de documentos e rotulação automática dos mesmos. A metodologia vem sendo utilizada em processos de organização de coleções de documentos, bem como o uso dos resultados produzidos como facilitadores de processo de recuperação de informação; e, ainda na identificação de tendências de tópicos e sub-tópicos nas coleções.

A configuração atual do ambiente de suporte computacional à metodologia permite utilizá-la e produzir bons resultados. Adicionalmente, prevê-se a avaliação e validação de outras técnicas em cada parte do processo, com o objetivo de avaliar técnicas que melhor se adaptem a cada tipo de coleção de textos. Com isso, espera-se melhor auxiliar os especialistas do domínio, interessados em construir as taxonomias, a tomarem suas decisões, disponibilizando ferramental para executar diferentes tarefas sob diferen-

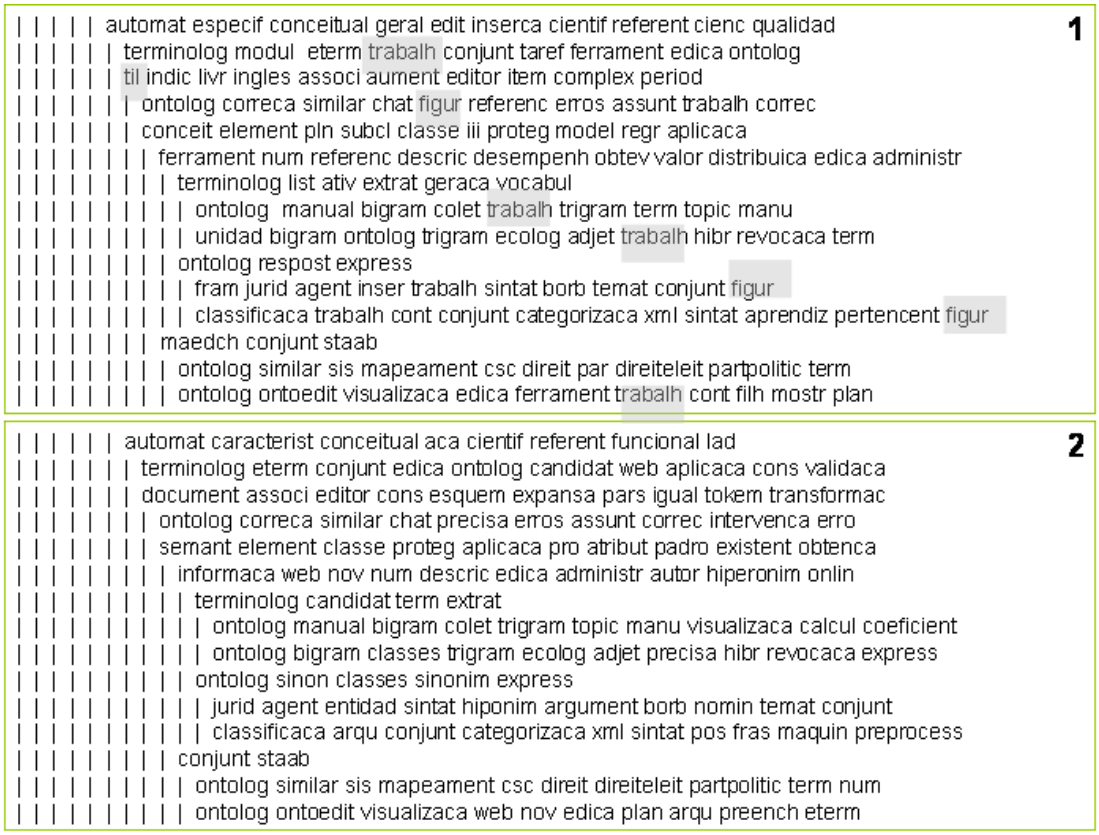

Figura 2. Exemplo de visualização, antes (1) e depois (2) de retiradas *stopwords* específicas da coleção

tes condições. Por esses motivos, o ambiente em desenvolvimento permite a incorporação de novas ferramentas, sempre que se mostrem necessárias, como visto na avaliação aqui discutida.

Como trabalho futuro, pretende-se disponibilizar ferramentas que viabilizem a realização de todas as etapas da metodologia apresentada, bem como formas de tratamento do crescimento da taxonomia de tópicos, tanto no que diz respeito ao simples aumento dos documentos, mantendo os grupos originais, quanto aos desmembramentos ou acoplamentos dos tópicos existentes.

Agradecimentos.

Os autores agradecem a CAPES, ao CNPq e ao Instituto Fábrica do Milênio pelo apoio técnico-financeiro.

Referências

Banerjee, S. e T. Pedersen (2003). The design, implementation, and use of the ngram statistics package. In A. F. Gelbukh (Ed.), *CICLing*, Volume 2588 of *Lecture Notes in Computer Science*, pp. 370–381. Springer.

Bloehdorn, S., P. Cimiano, A. Hotho, e S. Staab (2005). An ontology-based framework for text mining. *LDV Forum 20*(1), 87–112.

Brill, E. (1995). Transformation-based error-driven learning of natural language: A case study in part of speech tagging. *Computational Linguistics 21*, 543–565.

da Silva Conrado, M. e S. O. Rezende (2008, Salvador/BA: Outubro). Avaliando a geração de termos a partir de coleções textuais. *IV Workshop de Teses e Dissertações em Inteligência Artificial (WTDIA)*. A ser publicado.

Dhillon, I., J. Kogan, e C. Nicholas (2003). Feature selection and document clustering. In M. W. Berry (Ed.), *Survey of Text Mining*, pp. 73–100. Springer.

Dupret, G. e B. Piwowarski (2005). Deducing a term taxonomy from term similarities. In *Proceedings of Second International Workshop on Knowledge Discovery and Ontologies - KDO2005*, KDO, pp. 11–22.

Duval, E., W. Hodgins, S. Sutton, e S. L. Weibel (2002). Metadata principles and practicabilities. *D-Lib Magazine 8*(4), x–y. Disponível em: http://www.dlib.org/dlib/april02/weibrl/04-weibel.html. Consultado em 08/2006.

Ebecken, N. F. F., M. C. S. Lopes, e M. C. d. Aragão (2003). Mineração de textos. In S. O. Rezende (Ed.), *Sistemas Inteligentes: Fundamentos e Aplicações* (1 ed.)., Chapter 13, pp. 337–364. Manole.

Evangelista, S. R. M., K. X. S. Souza, M. I. F. Souza, S. A. B. Cruz, M. A. A. Leite, A. D. Santos, e M. F. Moura (2003). Gerenciador de conteúdos da agência embrapa de informação. In . Curitiba: Pontifícia Universidade Católica do Paraná (Ed.), *International Symposium on Knowledge Managment-ISKM*, Volume CD-ROM of *ISKM*, pp. 1–12.

Frigui, H. e O. Nasraoui (2004). Simultaneous clustering and dynamic keyword weighting for text documents. In M. W. Berry (Ed.), *Survey of Text Mining Clustering, Classification and Retrieval* (1 ed.)., Chapter 1, pp. 45–72. Springer-Verlag.

Gonzalez, M. A. I., V. L. S. de Lima, e J. V. de Lima (2006, May). Tools for nominalization: An alternative for lexical normalization. *Proceedings of the Seventh Workshop on Computational Processing of Written and Spoken Portuguese (PROPOR) - Springer Berlin / Heidelberg 3960*, 100–109.

Harman, D. (1991). How effective is suffixing? *Journal of of the American Society for Information 42*, 7–15.

Hotho, A., A. Maedche, e S. Staab (2002). Ontology-based text document clustering. *KI 16*(4), 48–54.

Jiang, X. e A.-H. Tan (2005). Mining ontological knowledge from domain-specific text documents. In *ICDM*, pp. 665–668.

Kashyap, V., C. Ramakrishnan, C. Thomas, e A. Sheth (2004a). Taxaminer: An experimentation framework for automated taxonomy bootstrapping. Technical report, computer Science Department - University of Georgia. Disponível em: http://lsdis.cs.uga.edu/ cthomas/resources/taxaminer.pdf. [17/07/2005].

Kashyap, V., C. Ramakrishnan, C. Thomas, e A. Sheth (2004b). Taxaminer: Improving taxonomy label quality using latent semantic indexing. Technical report, computer Science Department - University of Georgia. Disponível em: http://lsdis.cs.uga.edu/ cthomas/resources/taxaminer.pdf. [17/07/2005].

Lawrie, D. e W. B. Croft (2000). Discovering and comparing topic hierarchies. In *Proceedings of the 6th RIAO´2000*, pp. 314–330.

Librelotto, G. R., J. C. Ramalho, e P. R. Henriques (2004, Dezembro). Tm-builder : um construtor de ontologias baseado em topic maps. *CLEI Electronic Journal 2*.

Liu, L., J. Kang, J. Yu, e Z. Wang (30 Oct.-1 Nov. 2005). A comparative study on unsupervised feature selection methods for text clustering. *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on*, 597–601.

Liu, T., S. Liu, Z. Chen, e W.-Y. Ma (2003). An evaluation on feature selection for text clustering. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, pp. 488–495. AAAI Press.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics 11*, 22–31.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal os Research and Development 2*(2), 159–165.

Maedche, A., V. Pekar, e S. Staab (2002). Ontology learning part one - on discovering taxonomic relations from the web. In N. Zhong (Ed.), *Web Intelligence*. Springer.

Manning, C. D., P. Raghavan, e H. Schütze (2008, February). Language models for information retrieval. In *An Introduction to Information Retrieval*, Chapter 12. Cambridge University Press.

Marcacini, R. M. (2008). Um ambiente interativo para análise visual de agrupamentos hierárquicos. Monografia conclusão de curso de graduação, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos.

Marcacini, R. M., M. F. Moura, e S. O. Rezende (2007). Biblioteca digital do ifm: uma aplicação para a organização da informação através de agrupamentos hierárquicos. In . Porto Alegre, RS : Universidade Federal do Rio Grande do Sul (Ed.), *Workshop de Bibliotecas Digitais, 2007, Gramado - RS. XIII Brazilian Symposium on Multimedia and the Web - WDL 2007 - III Workshop on Digital Libraries*, Volume CD-ROM of *WDL*, pp. 1–16.

Mardia, K. V., J. T. Kent, e J. M. Bibby (1979). *Multivariate Analysis*. Academic Press.

Martins, C. A. (2003). *Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado*. Ph. D. thesis, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos.

Martins, C. A., E. T. Matsubara, e M. C. Monard (2003). Um estudo de caso utilizando uma ferramenta computacional que auxilia na redução da dimensão da representação de documentos em tarefas preditivas de mineração de textos. *IV Workshop on Advances and Trends in AI, Chillan - In Proceedings of Fourth Workshop on Advances & Trends in AI for Problem Solving*, 21–27.

Matsubara, E. T., C. A. Martins, e M. C. Monard (2003). Pre-text: uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. Technical Report 209, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos.

Mazzari, C. A. (2007). Gestão de pessoas e identificação de competências estratégicas em unidades descentralizadas da embrapa - o caso embrapa pecuária sudeste. Projeto de Pesquisa e Desenvolvimento, Situação: Em andamento; Natureza: Pesquisa. Embrapa Pecuária Sudeste.

Moura, M. F. e S. O. Rezende (2007). Choosing a hierarchical cluster labelling method for a specific domain document collection. In J. Neves, M. F. Santos, e J. M. Machado (Eds.), *New Trends in Artificial Intelligence.* (1 ed.)., Chapter 11, pp. 812–823. Lisboa, Portugual: APPIA - Associação Portuguesa para Inteligência Artificial. EPIA-Encontro Portugues de Inteligência Artificial, 2007, Guimarães, Portugual.

Neto, J. L., A. D. Santos, C. A. A. Kaestner, e A. A. Freitas (2000). Document clustering and text summarization. In L. T. P. A. Company (Ed.), *Proceedings of the 4th International Conference on Practical Applications of Knowledge Discovery and Data Mining - PADD2000*, pp. 41–55.

Nogueira, B. M., M. F. Moura, M. S. Conrado, e S. O. Rezende (2008). Avaliação de métodos não-supervisionados de seleção de atributos para mineração de textos. In *Proceedings of the First Workshop on Web and Text Intelligence (WTI) - Nineteenth Brazilian Symposium on Artificial Intelligence (SBIA) - A ser publicado*, Volume CD-ROM.

Nunes, M. (1996). The design of a lexicon for brazilian portuguese: Lessons learned and perspectives. *Proceedings of the Second Workshop on Computational Processing of Written and Spoken Portuguese*, 61–70.

Orengo, V. M. e C. Huyck (2001). A stemming algorithm for the portuguese language. *Eighth International Symposium on String Processing and Information Retrieval (SPIRE)*, 183–193.

Porter, M. (1980, July). An algorithm for suffixing stripping. *Program 14*(3), 130–137.

Punera, K., S. Rajan, e J. Ghosh (2005). Automatically learning document taxonomies for hierarchical classification. In ACM (Ed.), *Proceedings of the WWW 2005, Chiba, Japan*, WWW, pp. 1010–1011.

Ratnaparkhi, A. (1996, March). A maximum entropy model for part-of-speech tagging. *Proceedings of the Empirical Methods in Natural Language Processing Conference, University of Pennsylvania*, 491–497.

Rezende, S. O., R. Oliveira, L. C. M. Félixa, e C. Rocha (1998). Visualization for knowledge discovery in database. In N. F. F. Ebecken (Ed.), *Data Mining* (1 ed.)., pp. 81–95. WIT Press Computational Mechanics Publications.

Salton, G., C. S. Yang, e C. T. Yu (1975). A theory of term importance in automatic text analysis. *Journal of the American Association Science 1*(26), 33–44.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*, 44–49.

Silberschatz, A. e A. Tuzhilin (1995). On subjective measures of interestingness in knowledge discovery. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pp. 275–281. AAAI Press.

Silla Jr, C. N. e C. A. A. Kaestner (2002). Estudo de métodos automáticos para sumarização de textos. *Anais do Simpósio de Tecnologias de Documentos - STD, São Paulo-SP. 1*, 45–48.

Souza, M. I. F., M. F. Moura, e A. D. Santos (2005). Similaridade e complementaridade dos metadados nso sistemas agência de informação embrapa e acervo documental do ainfo. In C. A. B. do Paraná: FEBAB 2005 (Ed.), *Congresso Brasileirod e Biblioteconomia, Documentação e Ciência da Informação, 21st, 2005 - Curitiba*, pp. 12.

Sphinx Brasil (2008, Março). Sphinx. http://www.sphinxbrasil.com.br/po/.

Tang, L., J. Zhang, e H. Liu (2006). Acclimatizing taxonomic semantics for hierarchical content classification from semantics to data-driven taxonomy. In *KDD*, pp. 384–393.

Zhao, Y. e G. Karypis (2005). Topic-driven clustering for document datasets. In *Proceedings of the SIAM International Conference on Data Mining 2005*, pp. 358–369.