Documentos

Dezembro, 2004 47

ISSN 1677-9274

Proposta de Utilização de Mineração de Textos para Seleção, Classificação e Qualificação de Documentos

Embrapa

ISSN 1677-9274
Dezembro, 2004

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura, Pecuária e Abastecimento

Documentos 47

Proposta de Utilização de Mineração de Textos para Seleção, Classificação e Qualificação de Documentos

Maria Fernanda Moura

Campinas, SP
2004

Embrapa Informática Agropecuária
Área de Comunicação e Negócios (ACN)
Av. André Tosello, 209
Cidade Universitária "Zeferino Vaz" Barão Geraldo
Caixa Postal 6041
13083-970 - Campinas, SP
Telefone (19) 3789-5743 - Fax (19) 3289-9594
URL: http://www.cnptia.embrapa.br
e-mail: sac@cnptia.embrapa.br

Supervisor editorial: *Ivanilde Dispato*
Normalização bibliográfica: *Marcia Izabel Fugisawa Souza*
Editoração eletrônica: *Área de Comunicação e Negócios (ACN)*

1ª. edição on-line - 2004

Moura, Maria Fernanda.
Proposta de utilização de mineração de textos para seleção, classificação e qualificação de documentos / Maria Fernanda Moura. – Campinas : Embrapa Informática Agropecuária, 2004.
29 p. : il. (Documentos / Embrapa Informática Agropecuária ; 47).

ISSN 1677-9274

1. Mineração de textos. I. Título. II. Série.

CDD – 006.3 (21st ed.)

Autora

Maria Fernanda Moura
M.Sc. em Engenharia Elétrica, Pesquisadora da Embrapa Informática Agropecuária, Caixa Postal 6041,
Barão Geraldo - 13083-970 - Campinas, SP
e-mail: fernanda@cnptia.embrapa.br

Apresentação

As mudanças ocorridas na sociedade nestes últimos anos devido às novas tecnologias de informação e de comunicação exigem da Embrapa novos procedimentos para organização da informação que auxiliem na efetiva transferência de tecnologia. A Agência de Informação Embrapa é uma importante iniciativa de organização e disponibilização de informação para o público em geral, na forma de um portal *web*.

A Agência constitui-se no elemento de congregação das diversas agências de informação especializadas por temas ou produtos que tem como elemento orientador a cadeia produtiva. Sua informação é estruturada em uma árvore do conhecimento, para qual cada conteúdo de nó deve ser elaborado, citando referências de interesse, especialmente que possam ser localizadas via internet – sítios (*sites*) ou documentos eletrônicos.

Várias são as dificuldades envolvidas na organização de uma Agência, desde como melhor representar a hierarquia do conhecimento para cada domínio do conhecimento, seleção de material adequado e atual, digitalização do material, catalogação do material; contribuindo para isso, também o fato de as fontes de recurso de informação serem muito variadas. Logo, faz-se necessário investir na identificação e/ou desenvolvimento de técnicas, métodos e ferramentas que auxiliem partes e o todo desse processo.

Neste trabalho é realizado um levantamento bibliográfico de técnicas de mineração de texto e a exposição de uma proposta de utilização das mesmas na Agência de Informação.

José Gilberto Jardine
Chefe-Geral

Sumário

Proposta de Utilização de Mineração de Textos para Seleção, Classificação e Qualificação de Documentos

Maria Fernanda Moura

Introdução

Um estudo recente do grupo Delphi indica que 80% da informação das corporações no mundo é representada por documentos textuais, dado que essa é a forma mais natural de armazenar informações. Muitos desses documentos são armazenados em meio eletrônico e uma boa parte é lançada diariamente na *web*, formando grandes coleções de dados, tais como: relatórios diversos, especificações de produtos, relatórios de erros e as mensagens de advertências de software, resumos, notas, correspondência eletrônica, documentos diversos (boletins, jornais, revistas, etc.) e toda sorte de publicações eletrônicas textuais (bibliotecas virtuais, acervos documentais variados, etc.) (Han & Kamber, 2001). Até pouco tempo esse fato não era visto como uma vantagem competitiva, ou como suporte à tomada de decisão, com indicativos de sucessos e fracassos.

A mineração de textos vem tornando possível transformar esse grande volume de dados, geralmente não-estruturados, em conhecimento útil, muitas vezes inovador, para as empresas. O seu uso permite extrair conhecimento a partir de dados textuais brutos (não-estruturados), fornecendo elementos de suporte à Gestão do Conhecimento, que é o modo de reorganizar como o conhecimento é criado, usado, compartilhado, armazenado e avaliado. Conhecimento inclui dados históricos de todos os tipos, modelos, metodologias, equipes e suas habilidades. O sucesso na gestão deste é colocar uma alta prioridade em seu compartilhamento, porém são mudanças culturais internas à organização que efetivamente podem garantir esse sucesso. Tecnologicamente, o apoio de mineração de textos à gestão do conhecimento se dá na transformação do conteúdo de repositórios de informação em conhecimento a ser analisado e compartilhado pela organização (Velickov, 2004).

A Embrapa possui um grande acervo documental e várias iniciativas para organizá-lo, estruturá-lo e armazená-lo de forma adequada, permitindo a recuperação de metadados e muitas vezes, também de seus conteúdos. Os sistemas de apoio ao gerenciamento de projetos de pesquisa, como o SINSEP, o SIGER e o SEG, armazenam informações relevantes da memória de pesquisa dos últimos doze anos, muitas vezes não-exploradas por não fazer parte do conjunto de requisitos desses sistemas. Dentre os projetos estruturantes e integrativos da Embrapa, previstos pelo seu quarto plano diretor (Embrapa, 2004), o projeto de Desenvolvimento Organizacional prevê o resgate da memória técnica institucional, com a disponibilização de maneira organizada das informações que se encontram dispersas em arquivos físicos e eletrônicos, em diferentes formas e locais da empresa. Essa ação deverá alimentar uma nova base, provavelmente rica em textos, contendo informações, que talvez, isoladamente, pouco contribuam para o efetivo resgate e compartilhamento do conhecimento nela embutido. O sistema de controle do acervo documental das bibliotecas da empresa bem como da produção científica interna, denominado AINFO, guarda metadados de documentos, entre os quais resumos e palavras-chaves – elementos de fundamental importância para cruzar referências entre publicações, que poderiam vir a ser um bom indicativo de tendências internas e/ou externas de pesquisa em vários domínios do conhecimento. Além dos acervos internos a Embrapa tem acesso a várias bibliotecas digitais, como o Portal da CAPES e o de Plantas Medicinais da ESALQ/USP, que poderiam servir de complementos ao cruzamento de informações. Uma outra iniciativa interessante, é a Agência de Informação Embrapa, ou simplesmente Agência, um portal *web* composto por várias Agências de Produtos, que organizam informação técnica relevante para o agronegócio, especializada por produto e estruturada basicamente sob a ótica da cadeia produtiva do agronegócio, atendendo a perfis diversificados de consumidores de informação, tais como: produtores rurais, extensionistas, pesquisadores, técnicos, professores, estudantes, etc. (Evangelista et al., 2003).

A metodologia da Agência traz um diferencial importante na organização da informação, pois implementa uma solução desenvolvida especialmente para ela, que estrutura o conhecimento de uma cadeia produtiva de uma forma hierárquica, chamada de *árvore do conhecimento.* Nos primeiros níveis desta árvore estão os conhecimentos mais genéricos e, nos níveis mais profundos, estão os conhecimentos específicos. Cada nó desta árvore contém um texto sobre um tema que é resultante da compilação do conhecimento produzido por pesquisadores, técnicos extensionistas e agricultores, e, referências a outras obras que complementam a informação. Essa solução metodológica aplica-se muito bem às Agências de Produtos, tais como Feijão, Bovino de Corte, Bovino de Leite, Coco e outros, porque as árvores são extraídas de um subconjunto da cadeia produtiva – embora, ainda existam alguns conflitos quanto a melhor forma de realizar essa extração. No entanto, a Agência de Informação deve também contemplar Agências Temáticas, tais como: meio ambiente, monitoramento ambiental, agroclimatologia, cerrados, etc.; e, para estas não existe, ainda, uma solução única

ou tendências de consenso quanto à metodologia a ser utilizada, apenas algumas iniciativas nos temas citados. Por exemplo, no tema agroclimatologia (Embrapa Informática Agropecuária, 2004b) foi montada uma árvore de zoneamento agrícola para cada cultura específica em cada região contendo nos nós as características agrícolas e climáticas da região. Mesmo com uma metodologia mais consolidada, construir essas árvores e ainda disponibilizar um portal completando-as com referências a outras fontes de informação, não é um trabalho trivial, demanda uma boa equipe que precisa trabalhar com um grande volume de informação disponível em bases de dados textuais, mas cujo processo de identificação, seleção, classificação e síntese é bastante oneroso. Essa dificuldade de atingir consenso metodológico e a grande quantidade de informação disponível, nem sempre de qualidade e de real interesse, cria a demanda por ferramentas, quer automatizadas ou semi-automatizadas, que analisem os textos originais, de modo a filtrar o que é de fato útil para o portal.

Ainda, durante todo o processo de organização e atualização de uma Agência é realizada a seleção de recursos de informação. Esses recursos correspondem à informação bibliográfica existente a ser referenciada na Agência, ou seja as referências a outras obras que complementam a informação, e também àquela utilizada para auxiliar a construção/atualização de seus conteúdos. Essa seleção de recursos pode ser feita tanto pelos especialistas em informação quanto pelos editores da Agência, que são os seus especialistas do domínio. Os recursos eletrônicos associados ao assunto, além de textos, podem ser vídeos, figuras, fotos, eventos, bases de dados, mapas, imagens de satélite, *sites*, etc. Após, e/ou durante, a elaboração da árvore, esses recursos são referenciados nos conteúdos de nós da árvore, e também precisam ser catalogados. O processo de catalogação provê uma relação de metadados desses recursos, entre os quais categoria e palavra-chave, o que lhes confere uma boa qualificação. Dessa forma, a classificação dos recursos na Agência de Informação Embrapa ocorre em três grandes eixos:

- árvore do conhecimento: que corresponde a um modelo da cadeia produtiva;
- palavra-chave: fornece um tratamento detalhado da informação, microcategoria bastante personalizada, com uma lista de palavras-chaves para acesso. Atualmente, as versões de software disponíveis para a Agência recomendam a escolha das palavras-chaves nos tesauros NAL Agricultural Thesaurus e Thesagro, e se outras fontes forem utilizadas devem ser especificadas; e
- categoria: divisão em grandes categorias do conhecimento. Hoje a Agência recomenda o uso do tesauro NAL Agricultural Thesaurus para escolha da categoria, e, se, outra for utilizada, deve ser especificada.

Tanto a seleção quanto a classificação desses recursos não é um processo trivial; salvo as principais publicações de interesse nos assuntos, existe um grande número de textos e outros recursos que poderiam servir de apoio à construção de conteúdos e/ou serem referenciados, mas que podem não estar em formatos facilmente recuperados por ferramentas de busca, espalhados por toda sorte de publicações eletrônicas, quer sejam bibliotecas virtuais, boletins eletrônicos diversos, acervos documentais variados, páginas de empresas e instituições, páginas pessoais de pesquisadores e outros profissionais da área do domínio, arquivos digitalizados e não-publicados, etc. Para melhorar a produtividade e aumentar a qualidade desse trabalho, seria interessante construir uma solução integrada de ferramentas de software que minerem esses recursos – especialmente os textos, podendo auxiliar a transformação desse grande volume de dados em conhecimento e permitindo identificar agrupamentos e extrair-lhes informação, o que pode dar subsídios ao profissional de informação e ao editor para decidir como melhor categorizar/classificar e utilizar os recursos de informação na construção das Agências.

Com base nesses problemas de organização da informação, e também em soluções hoje mundialmente aplicadas, o objetivo deste trabalho é a partir de um levantamento bibliográfico na área de mineração de textos, identificar possíveis soluções, técnicas e ferramentas, que auxiliem o processo de levantamento de tendências em meio à fonte de material textual de um domínio do conhecimento, auxiliando a hierarquização, a identificação e a seleção da informação potencialmente mais relevante para a Agência. Pretende-se com isso, futuramente propor uma metodologia de uso integrado das ferramentas e técnicas no auxílio à organização de informação da Agência.

Assim, este trabalho divide-se em um resumo do levantamento bibliográfico realizado, uma proposta de aplicação das técnicas à Agência, que pode ser executada em um curto prazo, e a relação de alguns trabalhos futuros visualizados a partir deste.

Revisão de Literatura

A Mineração de Textos é uma área de pesquisa tecnológica cujo objetivo é a busca por padrões, tendências e regularidades em textos escritos em linguagem natural. Normalmente refere-se ao processo de extrair informações interessantes e não-triviais de textos não-estruturados; tendo sido inspirada na Mineração de Dados, que é "a extração não-trivial de informação implícita, previamente desconhecida e potencialmente útil de dados", ou busca por relações e padrões globais existentes em bases de dados (Frawley et al., 1992). Trata-se de uma área interdisciplinar, relativamente nova, que engloba: Processamento de Linguagem Natural, mais particularmente a Lingüística Computacional; Aprendizado de Máquina; Recuperação de Informação; Mineração de Dados; Estatística; e Visualização de Informação.

A mineração de textos e a recuperação de informação possuem uma alta dependência de processamento de linguagem natural, especialmente de lingüística computacional. O processamento de linguagem natural corresponde ao uso de computador para interpretar e manipular palavras como parte da linguagem. A Lingüística Computacional é o ramo que lida com a gramática e a lingüística, onde é desenvolvido o ferramental necessário para investigar textos e extrair informação sintática e gramaticalmente classificada dos mesmos (Williams, 2000). Esse ferramental também se utiliza de análise estatística de grandes coleções de textos para descobrir padrões úteis, e em especial para resolver problemas de desambigüidade no uso de termos da língua. Por esse motivo, também costuma ser considerado ferramenta de mineração de textos, dado que o objetivo é descobrir informação ainda não-conhecida, que não poderia ter sido diretamente documentada (Hearst, 2003; 1999). Por outro lado, na recuperação de informação, ou busca como é comumente chamada, o usuário tipicamente está procurando por algo que já é conhecido e que foi documentado (escrito, catalogado, armazenado, etc.). Nesse caso o problema é a razão entre a precisão e a revocação, isto é, a razão entre a percentagem de documentos recuperados, que são relevantes para a busca especificada, e a percentagem de documentos que realmente são relevantes para a busca especificada e foram de fato recuperados (Han & Kamber, 2001).

Um processo de mineração de textos pode ser esquematicamente representado pela Fig. 1 (inspirada no *framework* proposto por Tan (2004)). Deve-se notar que a fonte de dados, ou amostras para análise, são obtidas de grandes coleções de textos, em vários meios digitais incluindo-se a *web*, que podem ser não-estruturados, semi-estruturados (como o caso daqueles representados em XML) ou completamente estruturados – geralmente implementados sob um modelo relacional (Han & Kamber, 2001).

A primeira etapa do processo de mineração de textos é o refinamento (Fig. 1), cujo objetivo é obter uma amostra de textos sobre a qual a mineração irá trabalhar. Para isso, utiliza-se ferramentas de recuperação ou de extração de informação, especificando-se um critério de interesse; por exemplo um conjunto de palavras-chaves ou uma expressão de busca. A partir da amostra obtém-se uma forma de representação intermediária – sobre a qual serão realizadas as análises propriamente ditas. Na forma intermediária baseada em documentos cada entidade obtida corresponde a um documento analisado pelo filtro e sua análise (via a etapa de destilar conhecimentos) infere padrões e relações entre documentos. No agrupamento (ou *clustering*) de documentos, o relacionamento entre documentos é mais explícito, enquanto a categorização identifica tópicos-chaves dos documentos. Na forma intermediária baseada em conceitos, ocorre um processo de extração de conceitos dos documentos originais, ou da forma intermediária baseada em documentos, o que significa que as análises obtidas a partir dela são dependentes do domínio de aplicação (Tan, 2004). Logo, determinar o que será feito em termos de análise de dados, quer puramente estatística quer a partir de um

algoritmo de aprendizado, é um processo que se estende desde a preparação dos dados, passa pela análise semântica dos textos e freqüências dos termos em análise (Ebecken et al., 2003).

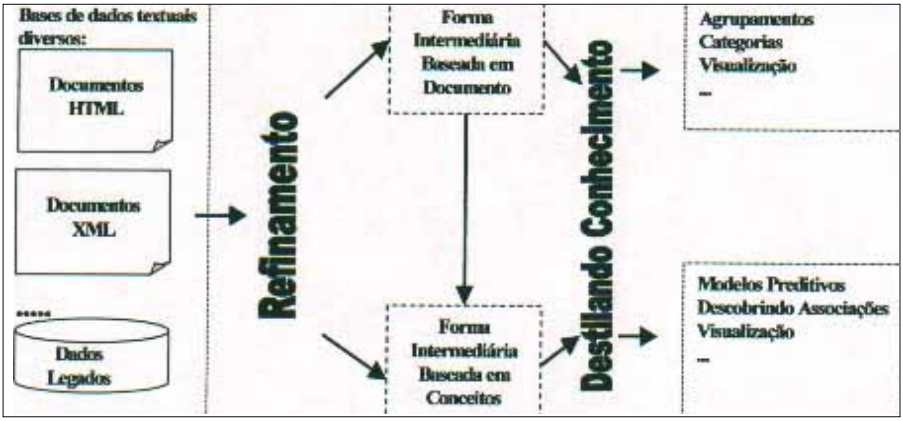

FIG. 1. Processo de mineração de textos.
Fonte: Tan (2004), adaptada pelo autor.

Para análises independentes do domínio de aplicação, como a categorização baseada em palavras-chaves, utiliza-se geralmente uma matriz de freqüências da ocorrência dos termos (palavras-chaves) nos textos. Nessa matriz cada coluna corresponde a um documento e cada linha a um termo, sendo cada elemento$_{ij}$ a freqüência do termo$_i$ no documento$_j$. Essa representação é uma das generalizações do modelo de espaço vetorial, que é dos mais utilizados por ferramentas de busca e de mineração de textos (Ebecken et al., 2003), dado a sua simplicidade e a complexidade computacional de se trabalhar com vetores ser geralmente baixa. Se o objetivo for procurar por termos conjuntos, por exemplo "estatística e inferência e bayesiana", continuará-se com a independência da língua, pois pode-se utilizar o classificador Naïve Bayes. Para esse classificador – assume-se que o texto é um conjunto de palavras independentes, logo a probabilidade de ocorrência de um conjunto de palavras simultaneamente é a produtória das probabilidades individuais (freqüências estimadas) (Domingos & Pazzani, 1996). Um processo que costuma ser utilizado nesses casos, é o *stemming*. Nele as extrações de palavras do texto, para contar sua freqüência, é feito isolando-se palavras com prefixos comuns, porém desconsiderando-se os contextos, em quase todas as variações do método. Essa desconsideração do contexto faz com que ele não se diferencie muito dos métodos que consideram as palavras com probabilidades de ocorrência independentes – também conhecidos como abordagem *bag of words*. Uma aplicação interessante de categorização de documentos é a sua utilização para obter metadados estruturados, como no trabalho de Pierre (2002). Nesse trabalho, técnicas de mineração de textos são utilizadas para auxiliar a extração de

metadados, para categorizar os textos e então técnicas de mineração de dados tradicionais são utilizadas sobre a base estruturada. Existem outros trabalhos nessa linha, como a ferramenta de software DiscoTEX que também usa um módulo de extração de informação, sobre bases de dados textuais não-estruturadas, para obter indicativos de metadados e suas associações e a partir dessas um formato estruturado para a base (Nahm & Mooney, 2000). Após essa estruturação, a DiscoTEX aplica técnicas de mineração de dados tradicionais à base; porém, para identificar os metadados e relações são utilizadas máquinas de aprendizado.

Deve-se notar que a freqüência de relações entre termos pode ser considerada uma medida de similaridade, proximidade entre termos, com base em "possui ou não uma característica", mesmo quando suas probabilidades de ocorrência são consideradas independentes. Uma medida de similaridade, ou proximidade, é a base para agrupar elementos quaisquer, e pode ser: coeficiente de correlação, basicamente ângulo entre dois vetores representativos dos elementos; distância euclidiana, baseada nos valores das variáveis em estudo; mapeamento de similaridades, geralmente binário, se os elementos possuem ou não alguma característica (Kashigan, 1986). Uma outra medida bastante utilizada, especialmente para estimar similaridade entre textos, é a *edit distance*, que, a grosso modo, define um número mínimo de inserções, deleções ou substituições necessárias para transformar uma cadeia de caracteres em outra, podendo-se utilizar um algoritmo de programação dinâmica para o seu cálculo ou outras abordagens, como o modelo estocástico proposto por Nahm et al. (2002). Em geral a similaridade entre palavras tem sido baseada em suas distribuições em grandes coleções de documentos denominada *corpora* (que servem de base às experimentações de lingüística computacional); e, aceitando-se a afirmação que palavras que ocorrem em um mesmo texto tem tendência em apresentar similaridades (Pantel & Lin, 2003). No trabalho de Fang et al. (2001) foi utilizada a freqüência de coocorrência de termos, com probabilidades de ocorrência independentes, para a obtenção de agrupamentos (*clusters*) de textos. A técnica de *cluster* utilizada é baseada na bisseção com k-médias, o que dá um resultado muito próximo ao *cluster* hierárquico, segundo Steinbach et al. (2000).

Outras abordagens usam *thesaurus* para ajudar a contabilizar freqüências entre termos relacionados. Um *thesaurus* pode ser definido como um vocabulário controlado que representa sinônimos, hierarquias e relacionamentos associativos entre termos (Ebecken et al., 2003). Os *thesaurus* são úteis para obter uma indexação temática dos textos, que é montada através das suas hierarquias. Com esse tipo de indexação as ferramentas de mineração de textos encontram rapidamente generalizações e especializações de termos específicos. Esse tipo de indexação pode tanto ser utilizada para auxiliar as análises que independem do domínio quanto as que dele dependem. A ferramenta de busca Vivisimo (2004) utiliza esse tipo de abordagem para construir *clusters on the fly*; isto é, para cada busca realizada pelo usuário ela cria *clusters* que representam os agrupamentos dos documentos obtidos na busca. A Fig. 2 ilustra o resultado da busca pela expressão "gado + de + corte".

As análises dependentes do domínio de aplicação, por exemplo as que buscam identificar taxonomias ou ontologias a partir de mineração de textos, recaem em problemas de análise de termos e suas relações, ou seja análise semântica, e dependem da língua. Uma taxonomia é um princípio de divisão e classificação sistemática de grupos em categorias, enquanto que uma ontologia é uma estruturação hierárquica de conhecimentos sobre coisas que podem ser subcategorizadas de acordo com sua essência, ou ainda, é uma compreensão compartilhada de algum domínio de interesse (Chandrasekaran et al., 1999). Nesses casos a análise dos textos não se restringem a reconhecer símbolos e suas freqüências, o significado é que deve ser reconhecido e palavras similares em significado normalmente são morfologicamente diferentes; essas similaridades semânticas precisam ser aprendidas dos contextos de uso das palavras e de seus padrões de ocorrências nos textos (Willians, 2000).

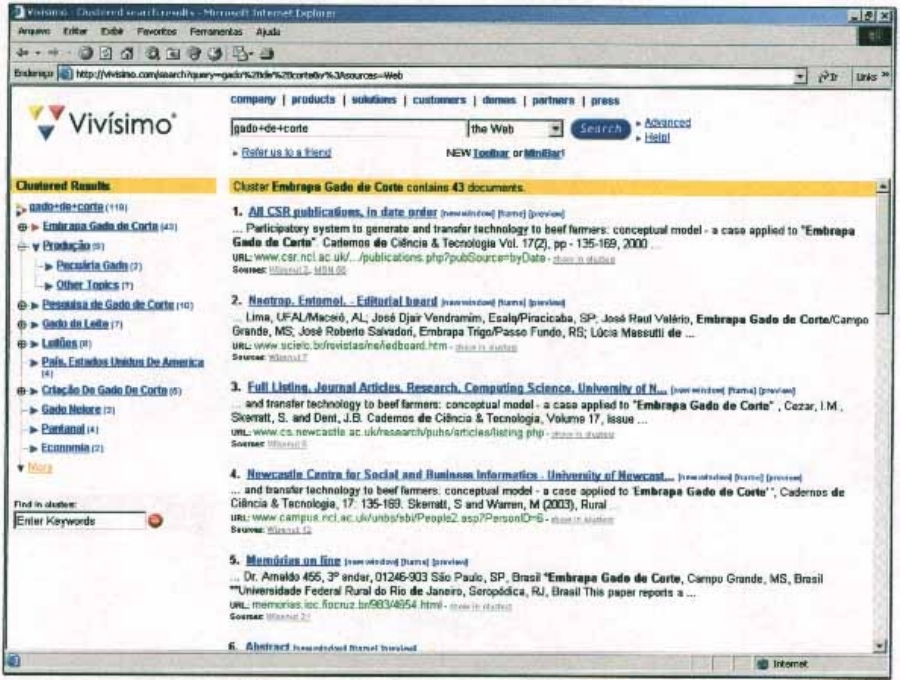

FIG. 2. *Clusters* resultantes da busca por "gado + de + corte" com a ferramenta Vivisimo.

Um tipo de indexação de textos que representa bem essa idéia é a indexação semântica latente (LSI - Latent Semantic Indexing), que aplica decomposição de valor singular à matriz de freqüências de palavras nos documentos de modo a obter a projeção de ambos em outro espaço – chamado espaço latente. Porém, sua

complexidade computacional é muito alta o que dificulta seu uso com grandes volumes de documentos – cada documento corresponde a uma coluna da matriz. O primeiro nível de complexidade refere-se à identificar palavras-chaves nos textos, para o quê utilizam-se gramáticas, com seus analisadores sintático/semântico (*parsers*) e suas expressões. Identificadas as palavras-chaves, elas serão marcadas nos textos para serem utilizadas como índices – que serão chamados de *tags*. Finalmente, essas *tags*, que são partes do texto, passarão por um processo de indexação que será a base da categorização dos textos.

Após um processo de indexação dos textos, atribuem-se métricas de importância aos termos selecionados, das quais as mais usadas quando a dependência de domínio está em jogo, são as de ganho de informação, informação mútua e independência de distribuição. O ganho de informação reflete o número de partes da informação obtida para predição da categoria pela presença ou ausência de um termo em um documento, e costuma alimentar métodos de aprendizado de máquina; ela envolve as estimativas das probabilidades condicionais de uma categoria dado um termo e o cálculo da entropia, ou seja, a quantidade de informação necessária para que a definição do termo seja convergente (Haykin, 1999), e geralmente alimentam redes neurais bayesianas (Han & Kamber, 2001). A informação mútua é muito usada em modelagem estatística da linguagem em associações de palavras, deriva-se de uma relação entre as probabilidades de ocorrência mútua de dois termos com as suas probabilidades de ocorrências isoladas (Haykin, 1999). Um problema dessa medida é que ela é fortemente influenciada pelas probabilidades marginais dos termos, se vista em uma tabela de contingência, logo não se pode comparar termos com medidas muito diferentes. E, ainda, colocando-se as medidas de freqüência em tabelas de contingência, uma das medidas utilizadas é a estimativa de quiquadrado, para medir o quão independente as palavras estão umas das outras (Mood et al., 1974).

No trabalho de Missikoff et al. (2003), uma ontologia para o domínio de turismo, definida de modo formal, é utilizada como uma base de conhecimento, para alimentar um método de aprendizado junto a estatísticas que a suportam e realimentam. Grosso modo, a ontologia é utilizada como o *thesaurus* na indexação, colocando as relações entre os termos em três categorias semânticas: termos abrangentes, similares e relacionados. Na captura da ontologia, identificação de conceitos chave no domínio, os termos são marcados com suas freqüências de ocorrência e sua categoria semântica. Após esse processo é calculada a informação mútua e o fator Dice (duas vezes a freqüência de ocorrência mútua sobre a produtória das freqüências isoladas). Os autores também utilizaram uma medida de relevância no domínio e entropia; termos com essa medidas consideradas irrelevantes, não plausíveis, são considerados ruídos e desprezados. Os autores fizeram um bom número de experimentos, utilizando várias medidas de importância entre termos, e consideram os resultados encorajadores, apesar de não estarem completamente validados.

Outro trabalho onde o objetivo era enriquecer uma ontologia é o de Agirre et al. (2000) onde eles constroem *clusters* hierárquicos de palavras, a fim de obter uma análise léxica de termos, e utilizaram os resultados como uma medida de desambigüidade entre termos. Os resultados foram muito bons atingindo 90% de precisão. Um dos pontos levantados nesse trabalho é justamente a falta de métodos que trabalhem sobre coocorrência de conceitos e não apenas coocorrência de palavras.

O trabalho de Maedche & Staab (2001), onde o objetivo é enriquecer ontologias a fim de viabilizar a *web* semântica, utilizou um processo de extração de informação dos textos, com base em análise léxica e semântica, e utilizam *cluster* hierárquico para ajudar a definir a taxonomia; ou seja, os resultados do *cluster* realimentam as medidas de similaridade e a taxonomia por si mesma, até que seja satisfeito um critério de prumo colocado por um especialista. Na Fig. 3 é ilustrada a hierarquia de documentos obtida, após *clusterização* de documentos, para o domínio de turismo com o objetivo de auxiliar a especificação da ontologia desse domínio, no trabalho de Hotho et al. (2001). A conclusão, relativa ao uso do *cluster*, é que de fato, extraindo-se a hierarquia de textos em linguagem natural, os termos adjacentes ou as relações sintáticas entre termos carregam um considerável poder descritivo para inferir a semântica de uma hierarquia de conceitos relacionados a esses termos.

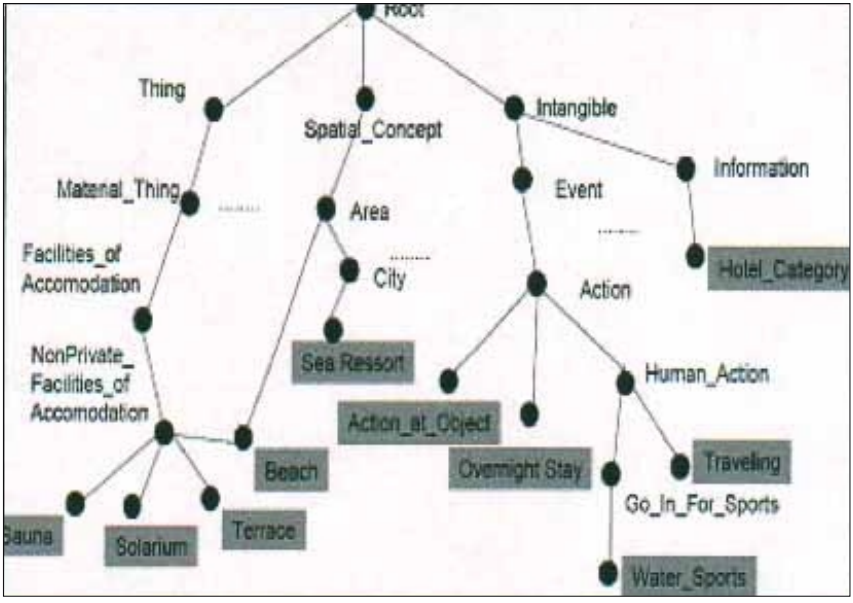

FIG. 3. *Cluster* de documentos obtido para o domínio de turismo.
Fonte: Missikoff et al. (2003).

No trabalho de Aizawa (2004) foram utilizados micro-*clusters*, a partir das submatrizes mais significativas da matriz de coocorrências de palavras nos textos, e medida de entropia para a avaliação do *cluster*; considerando-se a máxima entropia como medida de qualidade da obtenção dos agrupamentos.

Dentre as formas de visualização dos resultados de *cluster* hierárquico a mais utilizada é o dendograma (Fig.4), onde cada agrupamento é visto junto às suas intersecções com os demais; as intersecções ocorrem devido ao fato de os elementos poderem fazer parte de mais de um grupo – isso também deve ser uma característica que se busque minimizar ao identificar grupos. Ainda, em um dendograma deve-se escolher um ponto de corte, isto é, onde, na hierarquia, deseja-se quebrar os grupos, ou manter toda a hierarquia até a raiz da mesma.

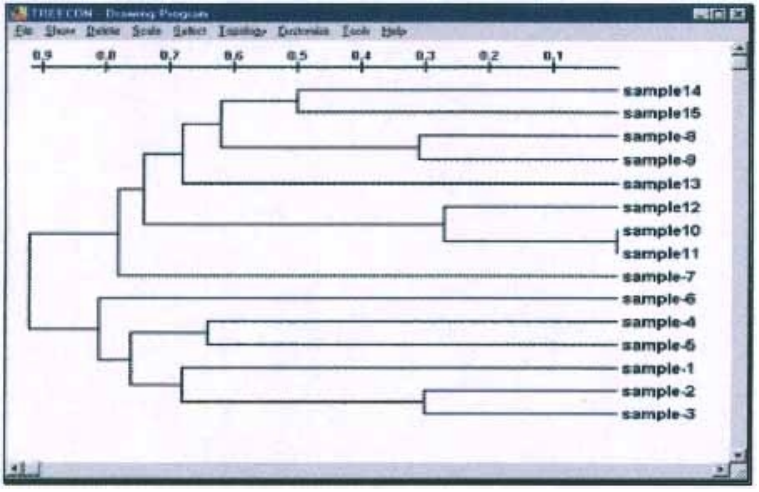

FIG. 4. Exemplo de Dendograma.
Fonte: Scanalytics Inc. (2004).

Outras representações da hierarquia são árvores simples (como na Fig. 3 e na parte esquerda da Fig.2) e também árvores hiperbólicas. Com relação à árvore hiperbólica, no trabalho de Kampanya (2004), foi realizado um experimento de formas de visualização dos resultados de *clusters* hierárquicos e proposta uma evolução para a forma de visualização dada pela ferramenta CitiViz (com base na árvore hiperbólica da empresa Inxight Software Incorporated (2004)). Na evolução proposta o autor parte de um *scatter plot* (gráfico com os *ranks* de ocorrência em cada *cluster*) para construir torres de documentos e mapeia as torres para uma árvore hiperbólica. Na árvore hiperbólica, em cada nó, é mostrado quantos documentos o compõem e as referências a cada documento – veja ilustração na Fig. 5.

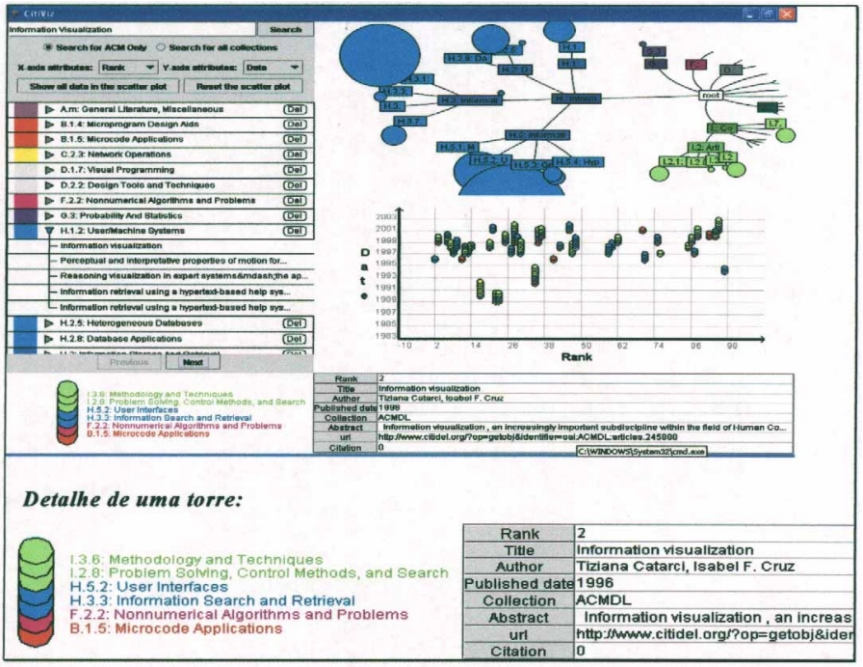

FIG. 5. Representação com árvore hiperbólica.
Fonte: Kampanya (2004).

Para avaliar os resultados dos *clusters* obtidos são utilizados vários métodos, inclusive a avaliação de um especialista do domínio de conhecimento, carregando toda a subjetividade esperada. Para avaliar os resultados obtidos pelo *cluster* sistematicamente, além da entropia, são utilizadas também a relação entre precisão e revocação (utilizadas em recuperação de informação) e a medida F proposta por Larsen & Aone (1999). Nessa proposta, estabelece-se para cada tópico T do *cluster* o número N1 de documentos julgados relevantes por um especialista do domínio, o número N2 de documentos no *cluster*, e o número N3 de todos os documentos julgados pertinentes no tópico T em toda a hierarquia daquele *cluster*. Então, calcula-se a precisão pela razão de N1 e N2 e a revocação pela razão de N1 e N3. A medida F proposta é a razão de duas vezes o produto da precisão e da revocação e as suas somas, ou seja, é a média harmônica entre a precisão e revocação. Utiliza-se então, a medida F como o *score* de cada *cluster* e calcula-se uma F geral pela média ponderada das F´s de cada *cluster* pelo número de tópicos (cada T). O desvio padrão médio fica tipicamente próximo de 0.01 e ocasionalmente atinge 0.02; procura-se deixar os clusters o mais próximo possível dos valores médios nos

múltiplos caminhos, desconsiderando-se nesse balanceamento o tempo de processamento dos algoritmos utilizados.

O que se observa é que a acuracidade dos resultados dos *clusters* depende da medida de similaridade escolhida, e não existe ainda hoje uma medida representativa para similaridade entre conceitos. Deve-se entender como conceito o seguinte conjunto: um termo, que rotula o conceito; uma descrição do significado do conceito; e, um conjunto de relações com outros conceitos (Missikoff et al., 2003). O conjunto de conceitos e suas inter-relações formam uma rede semântica. Faz-se necessário estabelecer uma medida adequada ao problema a ser atacado, seu domínio e talvez, alguma característica dependente da língua. Nessa linha de pesquisa há alguns grupos como o da Universidade do Texas, liderado por Mooney (University of Texas, 2004; Bilenko, 2004) que vem tentando desenvolver novos métodos de aprendizado de funções de similaridade e suas aplicações a diferentes domínios, em métodos de *cluster* e na identificação de duplicidade de registros em bases de dados. Outras frentes de pesquisa concentram-se nas transformações de espaço vetorial, sobre a matriz com os vetores de características utilizados no cluster, pois quanto mais termos ou conceitos e documentos são considerados na análise, maiores as matrizes (cada documento corresponde a uma coluna e cada termo/conceito a uma linha). Por exemplo, pode-se utilizar análise de componentes principais, para diminuir a dimensão do espaço vetorial tratado, como proposto em outras abordagens de mineração de dados (Han & Kamber, 2001).

Proposta de Aplicação à Agência

Embora a área de mineração de textos seja relativamente nova, existem métodos bastante aceitáveis, alguns inclusive automatizados por ferramentas de software comerciais e algumas tímidas iniciativas em software livre para o agrupamento hierárquico de textos. A maioria desses métodos não é dependente do idioma, pois tomam como base a independência da probabilidade de ocorrência das palavras, bastando alimentá-los com uma lista de palavras a serem desprezadas na análise (como artigos, interjeições, advérbios, etc.).

A proposta é utilizar ferramentas para obter aglomerados, categorias e suas formas de visualização, utilizando-se também algumas outras ferramentas que permitam identificar associações, e combinando-se esses resultados com o processo manual de catalogação em uso, de modo a procurar estabelecer um novo processo semi-automatizado que lhe configure maior produtividade, e maior confiabilidade de resultados; e que também auxilie a construção das árvores da Agência. O processo automatizado deve fornecer parâmetros para a tomada de decisão, e a partir disso deve-se investir no levantamento de técnicas e ferramentas de inteligência artificial que melhorem as estimativas desses parâmetros.

O método de automatização do processo deve envolver um conjunto de ferramentas de software, preferencialmente livres e que adotem padrões abertos, que devem receber como entrada um conjunto de arquivos textuais já digitalizados (formatos HTML, XML, txt, PDF convertido para HTML, etc.), previamente filtrados por uma ferramenta de recuperação de informação. Além das medidas de similaridade utilizadas nos agrupamentos, deverão ser integradas às ferramentas métricas pertinentes aos *clusters*, como a medida F e entropia entre eles. Também deve ser gerada uma representação hierárquica desses agrupamentos que possa ser utilizada por uma ferramenta de visualização, por exemplo, pela ferramenta HiperVisual disponível na Rede AgroLivre (Embrapa Informática Agropecuária, 2004a).

Para atingir esse propósito propõe-se:

- estabelecer uma base de documentos, definir um domínio do conhecimento e um corpo de textos para analisar, simulando um espaço amostral de documentos desse domínio, mas que também contenha documentos que não pertencem ao domínio ou que sabidamente não o representem de forma adequada; deste corpo serão retiradas as amostras da fase de desenvolvimento/integração do ferramental. O resultado deverá ser um espaço amostral de trabalho controlado, isto é, delimitado pelo domínio de conhecimento e mantido durante as fases de refinamento do processo, a fim de que se possa estar repetindo os experimentos sobre o mesmo conjunto de documentos;

- estabelecer uma parceria com um especialista do domínio escolhido, a fim de contar com a sua participação nas validações dos resultados obtidos em cada fase do desenvolvimento da proposta, bem como o de um especialista da área de informação – com experiência em seleção, classificação e qualificação de documentos;

- estabelecer uma base de ferramentas de *cluster* de domínio público para ser utilizada. Se necessário, estabelecer parcerias com outras instituições que possam fornecer ferramentas. Essas ferramentas, com base em métodos convencionais, serão utilizadas para avaliar a massa de documentos selecionada, obtendo-se as estimativas iniciais dos agrupamentos e avaliando-se os algoritmos de *cluster* utilizados. Se as ferramentas não proverem uma medida de entropia e a medida F, suas formas de cálculo deverão ser integradas a elas; e

- utilizar as ferramentas para obter os aglomerados (*clusters*) para auxiliarem o processo de catalogação, e, junto a um especialista da área de informação e pelo menos um do domínio, chegar a uma metodologia específica para utilizar esses resultados.

Hipóteses

Supõe-se a priori, que este trabalho efetivamente auxilie a identificação de importantes restrições em ferramentas de *cluster*, relativas a particularidades da língua portuguesa, que mereçam ser melhor estudadas e que seus resultados intermediários auxiliem o processo de catalogação. Logo, supõe-se que:

- o corpo de documentos estabelecido é satisfatório, para que se possa utilizá-lo em comparações de métodos e melhorias que venham a ser empregadas aos métodos. Para isso, o especialista do domínio e o de informação que estarão trabalhando junto ao projeto deverão opinar neste quesito; que leva a um julgamento bastante subjetivo;
- as ferramentas de *clusters* escolhidas fornecem agrupamentos confiáveis. Para validar esta hipótese utilizar-se-ão as estatísticas F e entropia obtidas. Somar-se-á, também, um julgamento subjetivo do especialista do domínio categorizando os agrupamentos obtidos em: ótimos, bons, ruins ou péssimos;
- a metodologia de auxílio à catalogação é válida de acordo com um julgamento subjetivo do especialista do domínio e do especialista de informação. Além da utilização da metodologia com o corpo de documentos formado, novas amostras de documentos deverão ser colhidas de outras fontes, tais como: internet, bibliotecas virtuais, e outros documentos digitalizados que venham a ser fornecidos pelos parceiros. Os especialistas deverão classificar o auxílio ao processo em: facilitou muito, facilitou ou não ajudou.

Resultados Esperados

O objetivo da proposta é melhorar o processo de seleção, classificação e qualificação de recursos de informação, provendo ferramentas automáticas de identificação de agrupamentos e extração de conhecimentos dos dados brutos para fornecer parâmetros de auxílio à tomada de decisão nesse processo.

O primeiro resultado esperado é um método eficiente de agrupamento de textos para um domínio de conhecimento. Esse método deverá ser automatizado e transformado em uma ferramenta de software livre, que deverá ser disponibilizada via a Rede AgroLivre de acordo com as licenças de uso e evolução por ela previstas e de acordo com a licença dos softwares livres utilizados originalmente. Ainda, a ferramenta deve gerar uma forma de representação hierárquica dos *clusters* obtidos que possa ser integrada a uma ferramenta de visualização gráfica de resultados; a princípio considera-se o uso da ferramenta HiperVisual, que gera uma representação de árvore hiperbólica, e também está disponível na Rede AgroLivre.

Além da ferramenta automatizada, deve ser obtida uma metodologia de uso dos resultados dos agrupamentos para auxílio à seleção e catalogação dos recursos de informação para a Agência. Essa metodologia deverá ser validada, não só com o corpo de documentos formado para teste, mas também em paralelo à construção de uma nova Agência.

Trabalhos Futuros

O agrupamento de dados textuais permite facilitar a construção de taxonomias do domínio de conhecimento, e de ontologias também, de modo a auxiliar a resolução de conflitos de consenso entre equipes que estejam trabalhando em atividades desse tipo. Por exemplo, no projeto Agência de Informação, esse método poderia auxiliar também a tomada de decisão sobre a metodologia de construção das chamadas árvores de conhecimento da Agência, especialmente para agências que não são de produto; pois, as de produto têm tido como base uma simplificação da cadeia produtiva, o que tem sido satisfatório. Ainda, em projetos como a Agência de Informação, que organizam seus dados por domínio de conhecimento, pode-se partir de coleções de textos já disponíveis na própria empresa, e montar toda a organização desses documentos com base apenas nos agrupamentos automaticamente obtidos, para então disponibilizá-los em um portal.

Uma outra aplicação importante de agrupamentos de dados é para auxiliar tarefas de *dataware housing*. Deve-se entender por *dataware housing* o processo de limpeza e integração de dados, provendo arquitetura e ferramentas para sistematicamente organizar, entender e usar os dados em processos estratégicos de tomada de decisões. A definição dos agrupamentos simplifica a identificação de relações e tendências variadas entre documentos. Pode-se utilizá-la para estabelecer, por exemplo, uma modelagem relacional de dados, para organizá-los em um banco de dados convencional, assim como para identificar a repetição de documentos – que podem sofrer um processo de eliminação; ou, ainda, serve como um identificador de informação útil e não-útil para a empresa. As tendências indicadas nos agrupamentos podem alimentar processos de tomada de decisão. No caso da Embrapa, o método poderia ser usado junto aos documentos pelos quais pretende-se recuperar a memória técnica de pesquisa da empresa, tentando-se analisar grupos, tendências, métodos utilizados em experimentação, experimentos utilizados, dados tabulados nos documentos, etc.; auxiliando a sua organização e uma efetiva recuperação de informações a partir dessa análise.

O método aplicado à produção científica da Embrapa poderia indicar tendências de áreas de concentração, pontos fortes e fracos, a partir de uma análise completamente imparcial das publicações constantes do acervo, sem vícios de julgamento, cruzando também referências. Os resultados obtidos poderiam ser comparados a outras fontes de publicações de outros órgãos de pesquisa,

repetindo-se os mesmos critérios de análise, com maior rapidez e confiabilidade. O mesmo poderia ser feito com os relatórios de projetos de pesquisa constantes das bases do SINSEP e SEP, e as novas bases do SEG. Essa ferramenta seria muito útil ao gerenciamento da pesquisa, pois permitiria obter indicativos de tendências e citações, permitindo uma análise imparcial.

Obtendo-se maior rapidez e qualidade na organização e análise de bases de dados textuais, consegue-se melhorar o processo de transferência de tecnologia de informação dependente dessas bases, e conseqüentemente levar essas vantagens aos clientes da Embrapa, que vão desde pequenos produtores rurais a renomadas instituições de produção e de pesquisa, englobando praticamente toda a sociedade brasileira.

Agradecimentos

À Sílvia Maria Fonseca Silveira Massruhá e à Maria Angelica de Andrade Leite, pelos incentivos ao trabalho, indicações e empréstimos de bibliografia. E, também a Sílvio Roberto Medeiros Evangelista e Álvaro Seixas Neto pelas críticas que foram responsáveis por várias melhorias neste trabalho.

Referências Bibliográficas

AGIRRE, E.; ANSA, O.; HOVY, E.; MARTÍNEZ, D. Enriching very large ontologies using the WWW. In: STAAB, S.; MAEDCHE, A.; NEDELLEC, C.; WIEMER-HASTINGS, P. (Ed.). **ECAI´2000 Workshop on Ontology Learning**: proceedings of the first Workshop on Ontology Learning OL´2000, Berlin. Karlsruhe: University of Karlsruhe: Université Paris Sud: University of Edinburgh, 2000. (CEUR Workshop Proceedings, v. 31). Disponível em: <http://ol2000.aifb.uni-karlsruhe.de/final/EAgirre_14.pdf>. Acesso em 19 jul. 2004.

AIZAWA, A. A. Method of cluster-based indexing of textual data. In: BIRD, S. (Ed.). **ACL anthology**: a digital archive of research papers in computational linguistics: COLING 2002: the 17th International Conference on Computational Linguistics. [Melbourne]: Melbourne University: Linguistic Data Consortium, 2004. ACL Disponível em: <http://acl.ldc.upenn.edu/C/C02/C02-1045.pdf>. Acesso em: 19 jul. 2004.

BILENKO, M. Learnable similarity functions and their applications to clustering and record linkage. In: AAAI/SIGART DOCTORAL CONSORTIUM, 9th, 2004, San Jose, CA. Disponível em: <http://www.cs.utexas.edu/users/ml/papers/marlin-aaaidc-04.pdf>. Acesso em: 19 jul. 2004. (To appear in Proceedings of the Ninth AAAI/SIGART Doctoral Consortium, San Jose, CA, July, 2004).

CHANDRASEKARAN, B.; JOSEPHSON, J. R.; BENJAMINS, V. R. What are ontologies, and why do we need them? **IEEE Intelligent Systems**, v. 14, n. 1, p. 20-26, Jan./Feb. 1999.

DOMINGOS, P.; PAZZANI, M. Beyond independence: conditions for optimality of the simple bayesian classifier. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 13., 1996, Bari, Italia. **ICML´96 papers** ... [Bari: s.n.: 1996]. p .105-112.

EBECKEN, N. F. F.; LOPES, M. C. S.; COSTA, M. C. A. Mineração de textos. In: REZENDE, S. O. (Org.). **Sistemas inteligentes**: fundamentos e aplicação. Barueri: Manole, 2003. cap. 13, p. 337-370.

EMBRAPA. **IV Plano Diretor da Embrapa – 2004-2007**. [Brasília, DF], 2004.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **AgroLivre - Rede de Software Livre para Agropecuária**. [Campinas], 2004. Disponível em: <http://www.agrolivre.gov.br/>. Acesso em: 15 jul. 2004a.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. Zoneamento agrícola: SP. In: EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Agritempo – sistema de monitoramento agrometeorológico**. [Campinas]: Embrapa Informática Agropecuária: Cepagri/Unicamp, 2004. Disponível em: <http://www.agritempo.gov.br/publish/zoneamento/SP.html>. Acesso em: 19 jul. 2004b.

EVANGELISTA, S. R. M.; SOUZA, K. X. S. de; SOUZA, M. I. F.; BRAGA, S. A. C.; LEITE, M. A. de A.; SANTOS, A. D. dos; MOURA, M. F. Gerenciador de conteúdos da Agência Embrapa de Informação. In: INTERNATIONAL SYMPOSIUM ON KNOWLEDGE MANAGEMENT - ISKM = SIMPÓSIO INTERNACIONAL DE GESTÃO DO CONHECIMENTO, 2003, Curitiba. **Anais**... Curitiba: Pontifícia Universidade Católica do Paraná, 2003. p.1-12. CD-ROM.

FANG, Y. C.; PARTHASARATHY, S.; SCHWARTZ, F. Using clustering to boost text classification. In: IEEE INTERNATIONAL CONFERENCE ON DATA MINING, 2001, San Jose, CA. **Proceedings of IEEE ICDM-2001**. [San Jose, CA, 2001].

FRAWLEY, W.; GIATETSKY-SCHAPIRO, G.; MATHEUS, C. Knowledge discovery in databases: an overview. **AI Magazine**, Fall, p. 213-228, 1992.

HAN, J.; KAMBER, M. **Data mining concepts and techniques**. San Diego, CA: Academic, 2001. 550 p.

HAYKIN, S. **Neural networks**: a comprehensive foundation. Upper Saddle River, NJ: Prentice-Hall, 1999. 842 p.

HEARST, M. **What is text mining**? 2003. Disponível em: <http://www.sims.berkeley.edu/~hearst/text-mining.html>. Acesso em: 19 jul. 2004.

HEARST, M. A. **Untangling text data mining**. 1999. Trabalho apresentado no Proceedings of ACL´99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999. Disponível em: <http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html>. Acesso em: 19 jul. 2004.

HOTHO, A.; STAAB, S.; MAEDCHE, A. Ontology-based text clustering. In: IJCAI-2001, 17.; WORKSHOP ON TEXT LEARNING: BEYOND SUPERVISION, 2001, Seattle. **Proceedings...** Disponível em: <http://www.aifb.uni-karlsruhe.de/WBS/sst/Research/Publications/hothoetal-ijcaiws2001.pdf>. Acesso em: 15 jul. 2004.

INXIGHT SOFTWARE INCORPORATED. **Inxight products**. Disponível em: <http://www.inxight.com/products>. Acesso em: 15 jul. 2004.

KAMPANYA, N. **User study of the digital libray clustering result visualization**. Blacksburg: Virginia Polytechnic Institute and State University, 2004. 20 p. (Independent Study Final Report. CS5974). Disponível em: <http://eprints.cs.vt.edu:8000/archive/00000747/01/nkampany-cs5974.pdf>. Acesso em: 15 jul. 2004.

KASHIGAN, S. K. **Statistical analysis**: an interdisciplinary introduction to univariate & multivariate methods. New York: Radius Press, 1986. 589 p.

LARSEN, B.; AONE, C. Fast and effective text mining using linear-time document clustering. In: CONFERENCE ON KNOWLEDGE DISCOVERY IN DATA, 1999, San Diego, CA. **Proceedings of the 5th ACM SIGKDD international conference on knowledge discovery and data mining**. New York: ACM Press, 1999. p. 16-22. Disponível em: <http://delivery.acm.org/10.1145/320000/312186/p16-larsen.pdf?key1=312186&key2=4347430901&coll=GUIDE&dl=ACM&CFID=24651906&CFTOKEN=4559939>. Acesso em: 20 jul. 2004.

MAEDCHE, A.; STAAB, S. Ontology learning for the semantic Web. **IEEE-Intelligent Systems,** v. 16, n. 2, p. 72-79, Mar./Apr. 2001.

MISSIKOFF, M.; VELARDI, P.; FABRIANI, P. Text mining techniques to automatically enrich a domain ontology. **Applied Intelligence**, v. 18, n. 3, p. 323-340, May, 2003.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. Tests of independence in contigency tables. In: MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics.** 3rd ed. Auckland: McGraw-Hill Book, 1974. p. 452-461. (McGraw-Hill Series in Probability and Statistics).

NAHM, U. Y.; BILENKO, M.; MOONEY, R. J. Two approaches to handling noisy variation in text mining. In: ICML-2002 WORKSHOP ON TEXT LEARNING, 2002, Sydney. **Proceedings...** [Sidney: s. n.], 2002. p. 18-27.

NAHM, U. Y.; MOONEY, R. J. Using information extraction to aid the discovery of prediction rules from text. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD-2000)., 6th, 2000, Boston. **Proceedings of KDD-2000 Workshop on Text Mining**. [Boston: s. n.], 2000. p. 51-58.

PANTEL, P.; LIN, D. Automatically discovering word senses. In: HUMAN LANGUAGE TECHNOLOGY CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (HLT-HAACL), 2003, Edmonton. **Proceedings...** [Edmonton, 2003]. p. 21-22. Disponível em: <http://acl.ldc.upenn.edu/N/N03/N03-4011.pdf>. Acesso em: 20 jul. 2004.

PIERRE, J. M. Mining knowledge from text collections using automatically generated metadata. In: KARAGIANNIS, D.; REIMER, U. (Ed.). **Practical aspects of knowledge management**: 4th International Conference - PAKM 2002: proceedings. Vienna: Springer, 2002. (Lectures Notes in Computer Science, v.2569). Disponível em: <http://www.sukidog.com/jpierre/pakm2002.pdf>. Acesso em: 19 jul. 2004.

SCANALYTICS INC. **Scanalytics**: gene profiler: dendogram analysis. Fairfax, 2004. Disponível em: <http://www.scanalytics.com/product/genePro/dendogram.shtml>. Acesso em: 15 jul. 2004.

STEINBACH, M.; KARYPIS, G.; KUMAR, V. A comparison of document clustering techniques. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD-2000)., 6th, 2000, Boston. **Proceedings of KDD-2000 Workshop on Text Mining**. [Boston: s. n.], 2000. p. 70-71.

TAN, A.-H. **Text mining**: the state of the art and the challenges. Disponível em: <http://www.ewastrategist.com/papers/text_mining_kdad99.pdf>. Acesso em: 4 jul. 2004.

UNIVERSITY OF TEXAS. Machine Learning Research Group. **UT ML Group**: text data mining. Disponível em: <http://www.cs.utexas.edu/users/ml/publication/text-mining.html>. Acesso em: 2 jul.2004.

VELICKOV, S. **TextMiner theoretical background**. Disponível em: <http://www.delft-cluster.nl/textminer/theory/>. Acesso em: 19 jul. 2004.

VIVISIMO INC. **Vivisimo clustering - automatic categorization and meta-search software**. Pittsburgh, 2004. Disponível em: <http://www.vivisimo.com>. Acesso em: 16 jul. 2004.

WILLIAMS, S. **A survey of natural language processing techniques for text data mining**. Adelaide: CSIRO Mathematical and Information Sciences, 2000. Disponível em: <http://www.bi.cmis.csiro.au/reports/2000_127.pdf>. Acesso em: 18 jul. 2004.

Ministério da Agricultura,
Pecuária e Abastecimento

BRASIL
UM PAÍS DE TODOS
GOVERNO FEDERAL