Biblioteca Digital do IFM: uma aplicação para a organização da informação através de agrupamentos hierárquicos

Ricardo Marcondes Marcacini
Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo - USP
Caixa Postal 668 – 13560-970
São Carlos-SP - Brasil
marcacini@grad.icmc.usp.br

Maria Fernanda Moura
Embrapa Informática Agropecuária
Caixa Postal: 6041 – 13083-970
Campinas– SP - Brasil
ICMC-USP
Caixa Postal 668 – 13560-970
São Carlos-SP - Brasil
fernanda@cnptia.embrapa.br

Solange Oliveira Rezende
Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo - USP
Caixa Postal 668 – 13560-970
São Carlos-SP - Brasil
solange@icmc.usp.br

ABSTRACT

In this work is presented a process to organize documents of the same domain base, with text mining techniques, applied on the digital collection of the Instituto Fábrica do Milênio – IFM. The information was structured through hierarchical grouping. The goals cover the generation of keywords descriptors for each topic in the documents collection and the evaluation of this process by domain specialists Additionally, some applications to allow the visualization, recovery and browsing of the documents collection were developed. This process begins the IFM Digital Library organization, an web environment to search and visualize the segmented information, improving the information retrieval and allowing the knowledge management.

Keywords

digital library, text mining, topic taxonomy, hierarchical clustering, information retrieval

RESUMO

Neste trabalho é apresentado o processo de organização dos documentos de um mesmo domínio do conhecimento, com base em mineração de textos, que pertencem ao acervo digital do Instituto Fábrica do Milênio. A informação foi estruturada por meio de agrupamentos hierárquicos. Os objetivos envolvem a geração de palavras-chave para os tópicos de cada agrupamento coberto pela coleção, bem como a avaliação, por especialistas do domínio, das listas de palavras-chave geradas. Desenvolveram-se aplicações para permitir a visualização da hierarquia e recuperação dos documentos. Este processo inicia a organização da Biblioteca Digital do IFM, um ambiente web para consultar e visualizar informação de forma segmentada, facilitando a sua recuperação e a gestão do conhecimento.

Palavras-chaves

biblioteca digital, mineração de textos, taxonomia de tópicos, agrupamento hierárquico, recuperação de informação.

1. INTRODUÇÃO

O Instituto Fábrica do Milênio (IFM) é uma organização em âmbito nacional, com o perfil de atuação focado na pesquisa em manufatura e soluções para as necessidades da indústria [5].

As atividades de pesquisa e desenvolvimento da primeira fase de atuação do IFM resultaram em um grande volume de informação técnico-científica, publicados em diferentes formatos de mídia digital.

A necessidade de consultar e difundir essa informação através de um sistema web, motivou a criação de métodos para a organização dos documentos e o desenvolvimento de aplicativos que permitam visualizar e recuperar o conteúdo dos mesmos. Adicionalmente, foram desenvolvidas ferramentas que possibilitaram a avaliação da organização final por especialistas do domínio, assegurando a qualidade desta organização.

A forma adotada para a organização da informação foi uma estrutura dividida em tópicos hierarquicamente relacionados. Este modelo foi escolhido porque uma organização hierárquica costuma ser de fácil interpretação e satisfaz à premissa que "se um usuário está interessado em um documento específico pertencente a um grupo deve também estar interessado em outros documentos desse grupo" [2]. Além disso, bibliotecas digitais com conteúdo de um mesmo domínio, desenvolvidas de modo semi-automatizado, como a da Agência de Informação Embrapa [13], têm seguido essa linha e são bem aceitas. Nesta estrutura, os primeiros nós da hierarquia contêm informação mais genérica e, conforme a profundidade do nó, a informação se torna mais específica. Cada nó representa um tópico sob o domínio do conhecimento e contém uma lista de palavras-chave, obtida durante o processo, para descrever o nó. O nó ainda pode conter nenhum, um ou mais documentos relacionados. O processo para alcançar esta organização é baseado em uma aplicação de métodos de seleção de palavras-chaves mais significativas para cada grupo de documentos sob um agrupamento hierárquico, independentemente do algoritmo do agrupamento utilizado.

Os métodos e ferramentas descritas neste trabalho estão em estágio de desenvolvimento, porém já é possível aplicar o conhecimento adquirido até o momento e avaliar os resultados da construção da Biblioteca Digital do IFM como experimento prático para aplicação dos conceitos mencionados. Este trabalho está organizado da seguinte maneira: a Seção 2 envolve as etapas principais para a construção da hierarquia de tópicos e descreve a preparação dos documentos (coleta e conversão da mídia digital para um único formato), pré-processamento (obtenção de palavras da coleção), extração do conhecimento (clusterização e seleção de palavras-chave) e pós-processamento (visualização e avaliação da

hierarquia e palavras-chave selecionadas); a Seção 3 mostra os recursos disponíveis ao usuário para navegar sobre a estrutura hierárquica e recuperar documentos; a Seção 4 reforça os resultados obtidos além de comentar sobre outras formas de aplicação desta metodologia; descreve alguns problemas encontrados durante o processo; e cita as ações futuras para a melhoria deste trabalho.

2. CONSTRUÇÃO DA BIBLIOTECA DIGITAL DO IFM

A Biblioteca Digital do IFM é um sistema que está sendo desenvolvido para a recuperação de conteúdo científico gerado por cerca de 800 pesquisadores, em 39 grupos de pesquisa, alocados em 32 instituições de ensino superior [5]. Para viabilizar um primeiro experimento e a análise da qualidade dos resultados, foram utilizados os documentos selecionados para a "I Assembléia Geral do Instituto Fábrica do Milênio", uma coleção de cento e dezenove documentos em português, predominantemente sobre manufatura, que reúnem os principais documentos resultantes das pesquisas pertencentes ao IFM.

2.1 Preparação dos Documentos

Os projetos desenvolvidos na rede de pesquisa do IFM geram conteúdo digital em diversos formatos. Assim, a primeira preocupação para o início do experimento foi padronizar todos os documentos em um único formato para facilitar o processamento de seu conteúdo. Utilizando conversores bem conhecidos em ambiente Unix (*pdf2text*, *doc2tex*, *html2text*) os documentos foram convertidos em formatos de texto simples, com codificação *ISO-8859-1*.

A partir dos documentos em formato texto simples, optou-se por remover todas as acentuações encontradas nos textos, realizando a substituição pelos caracteres correspondentes sem acentuação. Este procedimento foi adotado para solucionar alguns problemas encontrados com a ferramenta utilizada para o pré-processamento dos documentos. Além disso, constatou-se que a remoção de acentuação não teria impacto relevante durante o processo.

Todo o processo de conversão de documentos e remoção de caracteres acentuados atualmente é realizado de forma automática, pois foi desenvolvido um aplicativo que recebe o diretório com a coleção de documentos (*input*) e retorna os documentos convertidos em texto puro e sem acentuação no diretório de saída (*output*).

Os documentos finais contêm apenas os textos dos documentos originais de forma não qualificada, ou seja, sem um conjunto de descritores associados. No entanto, é mantido um identificador através da igualdade dos nomes dos arquivos em disco, para que seja possível estabelecer uma relação entre o arquivo original e o arquivo preparado para pré-processamento. Esta relação é importante, pois será utilizada para recuperação do documento original na etapa de pós-processamento (ver Seção 2.4).

2.2 Pré-processamento dos Documentos

A partir dos documentos obtidos na etapa de preparação, inicia-se a etapa de pré-processamento. Esta etapa tem o objetivo de transformar a coleção de documentos de dados não estruturados, em sua íntegra, em um formato estruturado, representando cada documento como um vetor de palavras que ocorrem neste mesmo documento. No final do processo é gerada uma tabela atributo-valor que contém todas as palavras da coleção dos documentos. Cada linha da tabela representa um documento e cada coluna representa uma palavra da coleção. As células contêm o valor da freqüência em que a palavra é encontrada no documento. Esta representação foi escolhida, neste momento, porque está sendo utilizada a abordagem *bag-of-words*, em que cada palavra é um atributo estatisticamente independente não importando sua ordem de ocorrência. Porém, essas suposições levam à uma matriz atributo-valor altamente esparsa.

O problema da alta dimensionalidade da tabela que representa a coleção dos documentos de forma estruturada é amenizado utilizando algumas técnicas mais amplamente difundidas:

- Lista de *stopwords:* são palavras que pouco caracterizam os documentos da coleção. A lista de *stopwords* abrange artigos, preposições, conjunções e palavras específicas da coleção que os especialistas de domínio considerarem não significativas. Estas palavras são ignoradas no processo de criação da tabela atributo-valor.
- *Stemming:* o processo de *stemming* consiste em reduzir variantes de uma palavra a um termo primitivo (*stem*). Para executar este processo, removem-se os sufixos e inflexões mais comuns das palavras. Por exemplo, palavras como "avaliar", "avaliador", "avaliação", são reduzidas ao termo "avalia". Consequentemente, a dimensão da tabela atributo-valor é diminuída ao se utilizar os *stem* nas colunas da tabela em lugar das palavras.
- Cortes de Luhn [7]: consiste na idéia de que as palavras que aparecem com grande freqüência na coleção dos documentos e as palavras que aparecem com pouca freqüência não são significativas para o processo de organização da informação, pois são pouco descritivas. Desta forma, geram-se intervalos para abranger palavras com muita e pouca freqüência, e eliminam-se os termos pertencentes ao intervalo do corte, diminuindo a dimensão da tabela atributo-valor. A decisão sobre os pontos de corte foi feita de forma arbitrária, por tentativa e erro, até resultar em uma tabela atributo-valor considerável. No entanto, estão sendo desenvolvidos alguns métodos para automatizar a seleção dos pontos de corte.

Para realizar estas tarefas de pré-processamento dos documentos, foi utilizada a ferramenta computacional PreText [9], que remove automaticamente palavras da lista de stopwords, encontra os *stems*, calcula as freqüências de ocorrência e retorna a tabela atributo-valor. Então, escolhem-se visualmente os valores dos cortes de Luhn, observando-se os gráficos de freqüências das palavras. A seguir, reaplica-se a ferramenta PreText aos dados especificando-se os valores de corte. Com essa informação a ferramenta gera uma nova matriz atributo-valor reduzida.

2.3 Extração do Conhecimento

A tabela atributo-valor obtida na etapa de pré-processamento é utilizada para o cálculo de um cluster hierárquico, gerando assim uma relação hierárquica entre os documentos. Existem vários algoritmos e ferramentas para o cálculo de agrupamentos

hierárquicos e a única restrição, do processo utilizado neste trabalho para obter as palavras-chave dos agrupamentos, é manter as medidas originais das freqüências das palavras, sem aplicar-lhes transformações. Como o número de documentos, neste exemplo, não é muito grande, também a complexidade do algoritmo não é muito importante. Assim, optou-se pelo uso do algoritmo *complete linkage* implementado no software MatLab[14] e, para gerar a matriz de similaridade requerida, a dissimilaridade com distância euclidiana, que se comporta bem com dados não normalizados [8]. Para os próximos passos (geração de palavras-chave), é necessário apenas a representação hierárquica do agrupamento, que neste caso é a matriz de *linkage*, gerada pelo MatLab.

Há vários trabalhos que exploram a geração de palavras-chave para o agrupamento hierárquico de documentos. Existem modelos probabilísticos, como o CAM [4] – *Cluster Abstraction Model* – que possui bons resultados, embora sua implementação seja bastante complexa. Um modelo não probabilístico, porém puramente estatístico, e fortemente acoplado à forma de obtenção dos agrupamentos é o do ambiente TaxaMiner [6]. Nesse ambiente é proposto um algoritmo de agrupamento hierárquico *bottom up* a partir de uma medida de coesão entre os grupos; com uma derivação da medida chega-se às palavras mais discriminativas de cada agrupamento e a seguir é realizado um processo de poda da árvore da hierarquia obtida com base na teoria de propagação de termos em uma taxonomia (mais genéricos agrupam-se e os mais específicos vão ficando nas folhas). Os resultados obtidos com o TaxaMiner têm sido bem aceitos na construção de taxonomias. Trabalhos mais simples, baseados apenas nas freqüências observadas de cada palavra do agrupamento, e que, conseqüentemente, independem de como o agrupamento é obtido, seguem a linha do modelo proposto por Glover [3], no qual nas freqüências observadas para cada palavra em cada agrupamento, ou seja, nas estimativas de máxima verossimilhança das suas probabilidades: *p(w/c)* e *p(w)*, com *w* correspondendo à palavra e *c* ao grupo (considerado como uma classe). A hipótese é que se p(w/c) é muito comum e p(w) ´e rara então a palavra discrimina bem a classe c, ou se tanto *p(w/c)* como *p(w)* são comuns então a palavra discrimina melhor a classe pai de c e, finalmente, se *p(w/c)* é muito comum e *p(w)* é relativamente rara na coleção então a palavra discrimina melhor a classe filha de *c*. Os limites muito comum ou muito raro são experimentalmente determinados e muitas vezes este processo é caro. Além disso, um problema comum a todos estes métodos, é descobrir a quantidade de palavras-chave que devem ser selecionados em cada agrupamento.

Neste experimento, optou-se por escolher métodos que facilitam a decisão da quantidade de palavras utilizadas no rótulo do agrupamento e que não dependessem da técnica pela qual a hierarquia foi gerada. Além disso, o objetivo da aplicação exigia que os métodos selecionados fossem de complexidade computacional aceitável e facilmente implementáveis. Desta forma, para gerar a lista de palavras-chave mais discriminativas de cada agrupamento foram implementados quatro diferentes métodos:

1. Popescul e Ungar [11]: A lista de palavras é selecionada por chi-quadrado, em que cada célula possui a restrição de valores observados e esperados das freqüências maiores que cinco.
2. Chi-quadrado adaptado [10]: A lista de palavras é selecionada por chi-quadrado, com a restrição de valores esperados das freqüências maiores que um e tratando por regras os casos de freqüências observadas zeradas.
3. Método Q [10]: A lista de palavras é selecionada de acordo com as estimativas do intervalo de confiança da medida de associação Q de Yule e tratando por regras os casos de freqüências observadas zeradas; e,
4. Mais freqüentes: A lista de palavras é selecionada de acordo com as freqüências das mesmas em cada agrupamento.

Para o método 4 (Mais freqüentes) toma-se até o máximo de quinze palavras por grupo. Este valor foi decidido arbitrariamente, pois muitas das palavras nas listas são repetições de palavras pertencentes aos agrupamentos superiores na hierarquia.

Para os métodos que utilizam chi-quadrado é definido um p-value de 5% para compatibilizar com o valor de determinação do intervalo de confiança de 95% da estimativa da medida de associação Q.

Após a seleção das listas de palavras-chave, elas são armazenadas em arquivos no formato XML para facilitar posterior visualização. A relação hierárquica obtida também é armazenada em um formato XML, e possui em sua estrutura uma *tag* para relacionar cada nó da hierarquia com o nome do arquivo XML que contém sua lista de palavras.

2.4 Pós-processamento dos Documentos

A relação hierárquica e as listas de palavras-chave, obtidas na etapa de extração de conhecimento, precisam passar por um processo de análise pelos especialistas do domínio. A intenção desta etapa é identificar qual o melhor método, ou combinação de métodos, para a geração da lista de palavras-chave dos agrupamentos.

Para facilitar este processo de análise, foi necessário desenvolver uma estrutura de visualização desta hierarquia. Assim, foi desenvolvida uma ferramenta usando a idéia de árvore de diretórios (através dos arquivos XML gerados na etapa anterior). A ferramenta foi implementada com tecnologia *web* para que pudesse ser acessada por um *browser* de qualquer parte da rede. Desta forma, o avaliador pode navegar sobre a hierarquia e acessar cada nó da árvore.

Por limitação da tela, para cada nó são exibidas apenas as primeiras nove palavras da lista, como mostra a Figura 1.

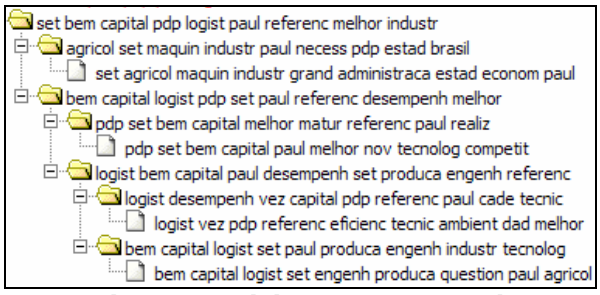

Figura 1. Alguns grupos da hierarquia representada por uma árvore de diretórios.

A lista de palavras-chave completa e os documentos associados em cada nó pode ser obtida clicando no nó desejado. Esta lista é exibida em uma janela separada, permitindo que o usuário possa acompanhar a organização da hierarquia, verificar os documentos do nó selecionado e analisar a lista de palavras-chave gerada. Estas palavras-chave estão na forma de *stems*, mas a ferramenta possibilita que o usuário possa recuperar as palavras originais que foram usadas na geração do *stem* clicando sobre ele, conforme é ilustrado na Figura 2:

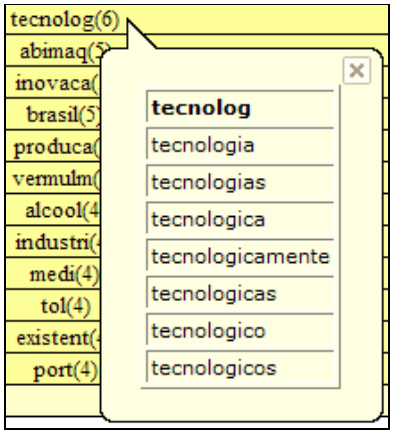

Figura 2. Lista de palavras-chaves de um nó: visualização das palavras derivadas de um *stem*.

Nesta mesma ferramenta, é possível recuperar os documentos originais que pertencem ao nó da hierarquia e realizar o processo de avaliação do método de seleção de palavras-chave, que será descrito a seguir.

2.4.1 Avaliação dos Resultados

Nesta etapa, a hierarquia gerada no pré-processamento dos resultados será rotulada por quatro métodos de seleção de palavras-chave. Quando nenhuma das palavras de um grupo é significante, a lista de palavras-chave fica vazia. Deste modo, é possível intervir manualmente na edição de palavras-chaves de grupo ou então eliminá-lo, promovendo os nós inferiores do grupo. Este processo acaba resultando numa forma de poda da árvore inicialmente gerada, o que pode vir a facilitar sua interpretação.

No caso particular deste experimento, envolvendo uma coleção do acervo digital do IFM, a hierarquia de taxonomia de tópicos obtida possui cento e dezoito nós. Os cento e dezenove documentos estão distribuídos pela hierarquia podendo pertencer a mais de um de seus nós. A aplicação de cada um dos quatro métodos apresentaram os seguintes resultados:

Tabela 1. Relação de cada método pela quantidade de nós que contêm palavras-chave significativas na hierarquia.

Método	Quantidade
1	118
2	41
3	82
4	118

Com base nestes resultados, verifica-se que o método 1 (chi-quadrado) e 4 (mais freqüente) encontram palavras-chave para todos os nós da hierarquia. No entanto, nota-se a repetição de palavras-chave ao longo da hierarquia. O método 2 (chi-quadrado adaptado) e 3 (método Q) conseguem eliminar a propagação de palavras-chave repetidas sobre a hierarquia e algumas vezes alguns nós ficam sem nenhuma palavra-chave significante.

Para a avaliação de cada método de construção de palavras-chave dos agrupamentos hierárquicos foi aplicada uma análise subjetiva, realizada por quatro especialistas na área de manufatura, tema predominante da coleção de documentos.

A ferramenta de avaliação permite que o especialista atribua uma nota para cada nó, para cada método, sempre considerando a hierarquia como um todo. A seguinte escala de notas foi adotada:

Nota 1: O conjunto de palavras indicado atrapalha a identificação do tópico.

Nota 2: O conjunto de palavras ajuda pouco na identificação do tópico.

Nota 3: O conjunto de palavras realmente ajuda na identificação do tópico.

Nota 4: O conjunto de palavras aproxima-se do ideal na identificação do tópico.

Decidiu-se por manter um número par de notas para impedir que o avaliador, em caso de dúvida, decida pela nota média.

A seguir, será comentado o resultado da avaliação realizada sobre a hierarquia gerada com os documentos do IFM, pois estes resultados influenciaram no direcionamento das próximas fases do experimento.

Nas Tabelas 2 e 3 encontram-se as comparações dos efeitos dos métodos sobre as notas atribuídas às listas de rótulos, utilizando o teste SNK (Student-Newman-Keuls) [12] e um nível de significância de 5%. Nas tabelas, a segunda coluna (Nota) corresponde às médias das notas para cada método ou para cada avaliador em cada método, e a terceira coluna indica o grupo ao qual essas médias pertencem, dado o resultado do teste SNK, sendo que letras iguais correspondem aos mesmos grupos. O erro quadrático médio, cuja raiz é apresentada nas tabelas, refere-se ao modelo linear hierárquico ajustado na análise de variância e *gl* aos graus de liberdade do erro.

Nota-se, na Tabela 2, que o método 4 (mais freqüentes) é o preferido, dado que ele sempre provê uma lista mais completa de rótulos. Logo a seguir vem o método 3, embora no mesmo grupo que os métodos 2 e 1; e também aparece como o segundo mais preferido na avaliação de método aninhado a avaliador – como apresentado na Tabela 3. Acredita-se que esse resultado seja conseqüência de ele separar as palavras mais associadas aos nós, de modo mais robusto, devido à estatística utilizada.

Tabela 2. Comparação múltipla de médias dos efeitos de método sobre a nota.

$\sqrt{e} = 0.59, gl = 185$		
Método	**Nota**	**Grupo**
4	2,79	A
3	2,48	B
2	2,13	B
1	2,13	B

Aparentemente os resultados dos métodos 1 (Popescul e Ungar) e do método 2 (chi-quadrado adaptado) não foram muito apreciados pelos avaliadores, como mostrado na Tabela 2. Esses dois métodos têm em comum o uso da mesma estatística para a seleção de palavras, que é a χ^2, porém com diferentes restrições. Logo, o resultado da Tabela 2 pode ser um indicativo que a escolha dessa estatística não tenha sido satisfatória. Embora se destaque a preferência pelo método 4 (mais freqüentes), deve ser salientado que ele promove a repetição das palavras ao longo da hierarquia, enquanto o método 3 produz listas mais enxutas e sem repetição nos níveis ascendentes ou descendentes. Na Tabela 3, onde cada *Aij* significa o *i-ésimo* avaliador do *j-ésimo* método, observa-se que o método 3 encontra palavras candidatas a termos mais altamente associados aos nós, por isso é bem avaliado. No entanto, parece que uma combinação entre os resultados seria mais satisfatória e esta será uma tarefa para trabalhos futuros.

Tabela 3. Comparações múltiplas de médias dos efeitos de método aninhados a avaliador sobre a nota.

$\sqrt{e} = 0.7578, gl = 176$		
aval-método	**Nota**	**grupo**
A1-4	3,33	a
A2-4	3,00	a,b
A2-3	2,92	a,b,c
A2-1	2,92	a,b,c,d
A1-3	2,92	a,b,c,d
A3-4	2,50	a,b,c,d,e
A1-2	2,42	b,c,d,e,f
A4-4	2,33	b,c,d,e,f
A2-2	2,25	b,c,d,e,f
A4-3	2,17	b,c,d,e,f
A4-2	2,17	b,c,d,e,f
A4-1	2,08	b,c,d,e,f
A1-1	2,00	b,c,d,e,f
A3-3	1,92	c e,f
A3-2	1,67	e,f
A3-1	1,50	f

A ferramenta também possui suporte para o envio de críticas ao processo. Dentre as críticas mais relatadas, está o uso de *stems* simples, isto é, dos radicais de palavras únicas (*onegram*), que dificulta a identificação de termos mais usuais como, por exemplo, "control" e "process" em lugar de "controle de processos". A adoção de colocações, como "controle de processos", está sendo estudada para as próximas etapas deste trabalho.

3. NAVEGAÇÃO NA BIBLIOTECA DIGITAL DO IFM

Após a análise da hierarquia e a avaliação dos métodos de geração das listas de palavras-chave, para este experimento foi selecionada a lista de palavras geradas pelo método 3 (Método Q). A partir dela iniciou-se a implantação das ferramentas de navegação, possibilitando a recuperação de grupos de documentos, tanto por navegação na própria hierarquia quanto por busca através de palavras-chave.

A visualização da hierarquia na ferramenta de navegação é feita automaticamente, utilizando os arquivos XML que contém a própria hierarquia e os arquivos XML que contêm a lista de palavras chaves para cada tópico da hierarquia. Desta forma, o usuário final possui dois recursos para navegação e recuperação do acervo da Biblioteca Digital do IFM:

- Busca na hierarquia: através de um sistema web, o usuário pode realizar uma busca simples utilizando uma palavra-chave ou expressões de busca mais complexas. A palavra (ou expressão) informada na busca é comparada com as palavras-chave de cada nó da hierarquia de forma a encontrar os grupos que mais se aproximam da consulta, como ilustrado na Figura 3. Os grupos encontrados são ordenados de forma decrescente por um *score* de semelhança, em que o primeiro da lista possui maior similaridade com a palavra (ou expressão) da busca. Para atender a premissa de que "se um usuário está interessado em um documento específico pertencente a um grupo deve também estar interessado em outros documentos desse grupo" [2], os nós mais internos pertencentes ao nó encontrado também ficam disponíveis no resultado da consulta, permitindo que o usuário possa recuperar os documentos destes grupos.

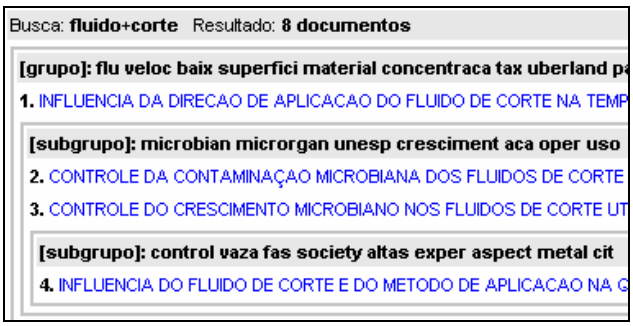

Figura 3. Consulta na hierarquia de tópicos exibindo documentos dos nós mais internos (subgrupo).

- Navegação através da árvore hiperbólica: A árvore hiperbólica é formada por uma rede de nós que se desdobram em suas componentes hierarquicamente dependentes, conforme é ilustrado na Figura 4:

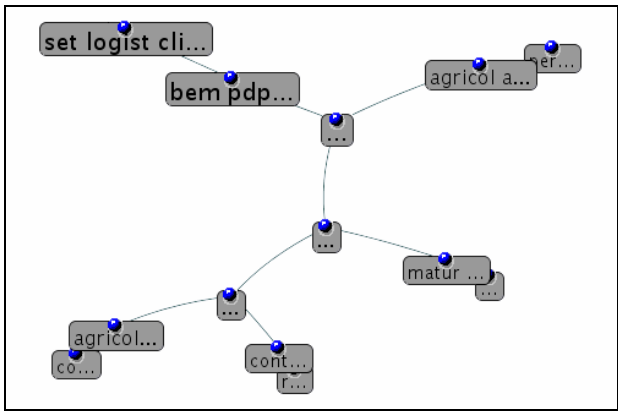

Figura 4. Exemplo de visualização e navegação através de uma árvore hiperbólica.

A hierarquia com a taxonomia de tópicos gerada até o momento é então representada por uma árvore hiperbólica, utilizando os mesmos arquivos XML comentados anteriormente.

Estas estruturas permitem visualizar e navegar sobre os tópicos, e também recuperar os documentos relacionados a um tópico a partir de um clique duplo sobre o nó desejado.

Os dois recursos de navegação e recuperação de documentos são intimamente ligados à organização hierárquica dos grupos de documentos. Assim, o processo de busca de uma certa informação se transforma no trabalho de buscar os agrupamentos que satisfazem à expressão de busca. Este método de busca não tem bons resultados quando a intenção é recuperar um documento específico, visto que neste estágio de desenvolvimento do trabalho a dedicação está voltada ao agrupamento hierárquico da coleção de documentos. No entanto, é possível integrar indexadores para localização de documentos específicos com a hierarquia de tópicos, pois uma vez que o documento é localizado por métodos de agrupamentos disjuntos não hierárquicos, podemos localizar os agrupamentos onde este documento se encontra na hierarquia e assim retornar os outros documentos pertencentes ao mesmo grupo. Existem sistemas em produção adotando práticas similares, ou seja, apresentando resultados de busca organizados em grupos mais significativos como, por exemplo, a ferramenta Vivisimo [15]. No entanto, a organização hierárquica de tópicos para os sistemas de busca mais populares, em geral, é realizada manualmente, por intenso trabalho humano, como do site Yahoo, ou completadas com construção de ontologias auxiliadas por processos semi-automáticos [1].

4. CONSIDERAÇÕES FINAIS

Este trabalho reúne o processo que está sendo empregado para a construção da Biblioteca Digital do IFM. Todas as etapas deste processo estão em estágio de desenvolvimento, porém já foi possível aplicar os conceitos adquiridos até o momento e avaliar a qualidade do resultado final.

A organização da informação e a consolidação da Biblioteca Digital do IFM, aplicados neste momento apenas aos documentos da primeira fase de gestão do IFM, propiciaram o desenvolvimento de um ambiente que estimula a gestão do conhecimento a partir de métodos automáticos e a possibilidade de visualizar e recuperar informação de forma estruturada. A capacidade de integrar os resultados aos serviços de informação já existentes ajuda a aumentar o padrão de qualidade e manter a identidade da instituição. Além de organizar o acervo digital dos documentos de pesquisa de manufatura, existe a possibilidade de empregar este mesmo processo em várias outras áreas (com um domínio específico), em que o objetivo é agrupar características de forma estruturada, como segmentar a base de currículos dos pesquisadores pertencentes à rede do IFM, e assim organizar os pesquisadores por área de habilidade.

Os resultados deste experimento apontaram também a necessidade de alterar algumas etapas do processo. Primeiramente, utilizar uma ferramenta de pré-processamento que permita a geração de colocações sem reduzi-las a seus radicais para facilitar o processo de visualização e compreensão do tópico; isto irá influenciar na dimensão da tabela atributo-valor, o que motivará novas pesquisas para redução de sua dimensionalidade. Quanto à forma de geração das listas de palavras, a partir do método 3 (medida Q) combinar os resultados com os do método 4 (mais freqüentes) e aplicar as propriedades de propagação de termos em taxonomias [6], a fim de melhorar as interpretações da lista de palavras-chave [10]. Finalmente, investir também na melhoria das ferramentas de visualização, avaliação e navegação sobre a hierarquia de tópicos.

5. AGRADECIMENTOS

Os autores deste trabalho agradecem ao CNPq pelo apoio financeiro e ao Instituto Fábrica do Milênio pelas sugestões significativas.

6. REFERÊNCIAS

[1] Bloehdorn, S. Cimiano, P. Hotho, A. Staab, S. An Ontology-based Framework for Text Mining, LDV Forum, v. 20, n. 1, p. 87-112, 2005.

[2] Chakrabarti, S. (2003). Mining the Web: Discovering Knowledge from hypertext data. Morgan Kaufmann Publishers.

[3] Glover, E.J. Pennock, D.M. Lawrence, S. Krovetz, R. Inferring hierarchical descriptions, CIKM, 2002, 507-514.

[4] Hofmann, T. The Cluster Abstraction Model: Unsupervised Learning of Topic Hierarquies from Text Data, International Joint Conferences of Artificial Intelligence (1999), 682-687.

[5] IFM (2007). Instituto Fábrica do Milênio. Disponível em: http://www.ifm.org.br/ [20/08/2007].

[6] Kashyap, V. Ramakrishnan, C. Thomas, C. Sheth, A. TaxaMiner: An Experimentation Framework for Automated Taxonomy Bootstrapping, International Journal of Web and Grid Services, Vol.1(2), 2005, 240-266.

[7] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2):159-165.

[8] Manning, C. e Schütze, H. (2003). Foundations of Statistical Natural Language Processing. MIT Press.

[9] Matsubara, E. T., Margins, C. A., and Monard, M. C. (2003). PreText: uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. Techinical Report 209, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos.

[10] Moura, M. F. and Rezende, S. O. (2007). Proposta e experimentação de modelos de rotulação para agrupamentos hierárquicos de documentos. Technical Report 302, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos.

[11] Popescul, A. and Ungar, L. (2000). Automatic labeling of document cluster. Unpublished manuscript. Disponível em: http://citeseer.nj.nec.com/popescul00automatic.html [17/07/2007]

[12] Snedecor, G.W. Cochran, W.G. Statistical methods, 6th ed, Ames: Iowa State University Press, 1967.

[13] Souza, M.I.F. ; Santos, A. D. ; Moura, M. F. . Agência de Informação Embrapa: uma aplicação para a organização da. In: Workshop de Bibliotecas Digitais, 2006, Florianópolis. Anais do II Workshop de Bibliotecas Digitais.. Florianópolis : Sociedade Brasileira de Computação, 2006. p. 51-56.

[14] The MathWorks - MATLAB and Simulink for Technical Computing. Disponível em: http://www.mathworks.com/ [20/08/2007]

[15] Vivisimo. Web search engine. Disponível em http://www.vivisimo.com [20/08/2007].