Proposta e Experimentação de Modelos de Rotulação para Agrupamentos Hierárquicos de Documentos

Maria Fernanda Moura, Solange Oliveira Rezende

[1]Laboratório de Inteligência Computacional
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Av. Trabalhador São-carlense, 400 – Centro
Caixa Postal: 668 – CEP: 13560-970 – São Carlos – SP – Brazil

{mnanda,solange}@icmc.usp.br

Abstract. *One of the problems in automatic models to generate topic taxonomies is the creation process of the most significant words list that discriminates each document group. In this technical report, there are two proposed methods to find out those lists from hierarchical document groups, as well as the use of two other classical methods in this task. The use of the methods was experimented by some domain specialists, trying to find out a solution that could satisfy their expectations. The experimentation results are all discussed here.*
Key-words*: hierarchical document clustering, cluster labelling, topic taxonomy, domain taxonomy.*

Resumo. *Um dos problemas em modelos automáticos para a geração de taxonomias de tópicos é como gerar a relação de palavras mais significativas para cada grupo de documentos encontrado. Neste trabalho é realtada a proposta de dois métodos para encontrar essa relação de palavras, bem como a utilização de dois outros métodos clássicos, e conduzido um experimento junto a especialistas do domínio de conhecimento, buscando-se um caminho para melhor satisfazer as expectativas desses especialistas. Os resultados dos experimentos são discutidos aqui.*
Palavras-chaves*: cluster hierárquico de documentos, rotulação de clusters, taxonomia de tópicos, taxonomia de domínio.*

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 como os da ferramenta Vivisimo [1]. Na maioria dos casos, exceto pelo agrupamento apresentado pela Vivisimo, 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 ([2], [3]).

Uma forma de organizar o conhecimento, para facilitar a recuperação de informações e navegação, é criando uma estrutura de representação do mesmo dividida em

tópicos hierarquicamente relacionados. Um exemplo é a organização hierárquica do conhecimento de um domínio como na Agência de Informação Embrapa [4]. Nos primeiros níveis da hierarquia estão os conhecimentos mais genéricos e, nos níveis mais profundos, estão os conhecimentos específicos [5]. Cada nó corresponde a um tópico sob o domínio de conhecimento e contém um texto sobre um tema/tópico que é resultante da compilação do conhecimento produzido por pesquisadores, técnicos extensionistas e agricultores, e, referências a outras obras que complementam a informação. Mesmo com uma metodologia consolidada, construir essas hierarquias, atualizá-las e, ainda, disponibilizar um portal completando-as com referências a outras fontes de informação, não é um trabalho trivial, e demanda um bom esforço de uma equipe especializada no tema tratado.

Esses problemas inspiraram um projeto de construção e atualização de taxonomias de tópicos a partir da análise de coleções de textos dinâmicas de um mesmo domínio de conhecimento [6]. Nesse projeto vem sendo construído um ambiente de ferramentas para mineração de textos cujo principal objetivo é auxiliar o especialista do domínio a construir suas taxonomias de tópicos, permitindo-lhe intervir no processo sempre que necessário ou que ele deseje. Para isso vem sendo investigadas, e testadas junto aos especialistas, formas variadas para: gerar os atributos de interesse (palavras simples, palavras compostas, termos do domínio, palavras stemizadas, etc); realizar filtragem de atributos (cortes de Luhn e outros); reduzir a dimensionalidade da matriz atributo-valor a partir de decomposições da estrutura de variância; agrupar os documentos utilizando diferentes ferramentas de cluster; rotular os agrupamentos encontrados; reduzir a hierarquia gerada; ampliar a hierarquia com novos documentos; e, de visualização dos resultados. Futuramente pretende-se, a partir de resultados obtidos, poder chegar a uma organização que permita construir uma estrutura semelhante à da Agência, porém cuja hierarquia representada seja exatamente a dos tópicos cobertos pelas publicações existentes.

Particularmente, neste trabalho são investigados alguns métodos bem conhecidos para a rotulação de clusters procurando adaptá-los a um processo de identificação de tópicos hierárquicos, com o objetivo de selecionar um método ou combinação de métodos que melhor satisfaçam às expectativas dos especialistas do domínio e que sejam de fácil implementação sobre uma hierarquia já existente. Assim, apresentam-se algumas idéias relacionadas ao problema específico de rotulação de agrupamentos, os métodos que foram explorados, melhorados e propostos no trabalho; alguns experimentos para avaliar e validar os possíveis métodos a serem usados no ambiente e seus resultados. Os resultados preliminares são bastante promissores, porém demandam algumas melhorias e trabalho futuro, também apresentado a seguir.

2. Trabalhos Relacionados

Há vários trabalhos que exploram especialmente a geração de rótulos para o agrupamento de documentos textuais, que visam à identificação de palavras-chaves que indiquem os possíveis tópicos aos quais os documentos agrupados se refiram. Em geral, esses métodos dependem diretamente da forma como o agrupamento é obtido, isto é, do método utilizado para agrupar os documentos. Os agrupamentos são calculados a partir da representação matricial da coleção de textos, considerando-se cada texto um elemento a ser agrupado representado por um vetor de atributos. Cada atributo corresponde a uma palavra ou composição de palavras (por exemplo, “inteligência artificial”), para as quais é obtida alguma medida de freqüência, que pode ser qualitativa, representado a pertinência

ou não da palavra no documento, ou quantitativa, como a freqüência ou a freqüência inversa de ocorrência da palavra. Os agrupamentos obtidos refletem tópicos ou subtópicos aos quais os documentos se referem, logo, são representados por conjuntos de atributos mais significativos no grupo, que podem ser interpretados como um conceito associado ao grupo.

O agrupamento de documentos, com base em análise de cluster, é realizado após alguma seleção de atributos que possam discriminar os textos da coleção e, conseqüentemente, auxiliar a delimitação dos agrupamentos por eles formados. O conjunto de atributos é formado pelas palavras utilizadas nos textos, porém não por todas elas, dado que o número de atributos seria gigantesco e não necessariamente significante. Assim, existem vários estudos relativos à seleção das palavras que melhor discriminam a coleção de textos em análise; e, dentre os resultados mais utilizados encontram-se os cortes de Luhn [7]. Esses cortes são feitos sobre o traçado da curva de distribuição das freqüências das palavras na coleção de textos, com os valores ascendentemente ordenados, onde se tomam as aproximações dos pontos de inflexão da curva como estimativas da freqüências mínima e máxima para o corte. O método utiliza a suposição de que as palavras cujas freqüências ficam entre a mínima e a máxima tenham um melhor poder de discriminação sobre os textos da coleção.

Obtidos os agrupamentos, a tarefa de identificar as palavras mais discriminativas em cada qual, também conhecida como rotulação de agrupamentos, pode ser visto como um problema de seleção de atributos [8]. Para seleção supervisionada considera-se cada grupo como uma classe, podendo-se utilizar informação mútua ou ganho de informação; e, para abordagens não supervisionadas a *tf-idf (term frequence - inverse document frequence)* ou a análise fatorial. Um bom exemplo é a utilização da *tf-idf* no módulo *k-means* do ambiente TMSK - *Text Mining Software Kit*[8]. A *tf-idf* corresponde à freqüência observada da palavra na coleção, ou no agrupamento como utilizada nesse caso, ponderada pelo logaritmo da sua freqüência inversa na coleção; a ponderação visa eliminar palavras muito freqüentes e privilegiar as menos freqüentes. No ambiente TMSK, para selecionar as palavras mais discriminativas de cada agrupamento encontrado, o método empregado toma aquelas que possuem as maiores *tf-idfs* médias.

Dentre as idéias mais difundidas e utilizadas para rotular os grupos de documentos estão os métodos baseados em centróides, para agrupamentos obtidos com *k-means* ou *k-medóide*, ou método deles derivado. Esses métodos levam a agrupamentos disjuntos e em alguns casos hierárquicos - aplicando-se métodos semelhantes ao *bissecting k-means* - que divide o conjunto de dados em duas partes, depois cada parte em duas partes e, assim, sucessivamente. Para obter o conjunto de palavras-chaves que melhor discriminem os grupos, em cada grupo calculam-se os centróides, para a base *k-means*, ou as medianas, para os similares ao *k-medóide*; e, então se utilizam os valores de cada célula dos vetores como pesos para cada atributo considerado. Os atributos com maior peso em cada grupo são considerados os mais discriminativos do agrupamento [9].

Trabalhando-se com modelos probabilísticos de agrupamentos e, conseqüentemente, de seleção de atributos mais significativos em cada agrupamento hierárquico, há o modelo CAM - *Cluster Abstraction Model*, desenvolvido por Hoffman [10]. Nesse modelo restringe-se a presença do termo no nó da árvore como ascendente ao agrupamento onde o termo é significativo no documento somada à

distribuição das palavras condicionadas aos documentos, $P(w/d) = \sum_c P(c_d = c/w_d, \theta) \sum_a P(a/c)P(w/a)$, com *w* correspondendo a cada palavra, *d* a cada documento e *c* a uma classe/grupo. Nesse caso, se for considerado *hard assigment* para a distribuição dos grupos dadas as palavras, captura-se a idéia de propagação de termos em uma taxonomia, em que se verifica que um documento é composto por uma mistura de termos que vão dos mais genéricos para os mais específicos. O resultado é bastante bom, embora a forma de se chegar ao mesmo seja muito complexa e envolva o acerto dos critérios de convergência dos métodos empregados. Outro solução, não probabilística e mais simples, é a utilizada no ambiente TaxaMiner [11], que busca identificar taxonomias de tópicos a partir de clusterização hierárquica; onde os documentos são agrupados mediante o cálculo de um fator de coesão dos grupos e a taxa de coesão de cada grupo indica também os atributos mais significativos em cada qual. Nesse método é realizado um procedimento de *pruning* da árvore de cluster com base na idéia de propagação de termos em uma taxonomia. A base do procedimento de *pruning* é que níveis mais específicos da árvore que nada acrescentam aos seus antecedentes podem ser podados; para tomar essa decisão o método utiliza a medida de coesão, nele definida, para maximizar a relação intracluster e também estabelece empiricamente um *threshold*.

Outro trabalho bastante difundido é o proposto por Glover [12], que se baseia exclusivamente nas freqüências observadas para cada palavra em cada agrupamento, ou seja, na 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 $p(w)$ é 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*; o que significa muito comum ou muito raro é experimentalmente determinado. Uma modificação foi proposta a esse método, onde se procura estabelecer um compromisso entre um rótulo simples e uma lista de tópicos, estabelecendo um *descriptive score* ponderado pela *tf-idf* [13]. Embora os resultados sejam mais promissores, o problema de determinar os critérios de convergência experimentalmente foram ampliados também para os novos *cuttoffs* necessários ao *descriptive score*.

O problema comum a todas essas soluções, é que não há um critério específico para se decidir quantos atributos pegar em cada grupo, dado que existem apenas os *rankings* dos mesmos pelos pesos utilizados - coordenada do centróide, *tf-idf* ou probabilidade no grupo. Além disso, há uma grande taxa de repetição das palavras selecionada como mais discriminativas de um grupo para outro, o que numa hierarquia fica ainda mais evidente e confuso - exceto pela solução adotada no ambiente TaxaMiner, tanto de *pruning* da árvore quanto à questão de propagação dos termos na taxonomia.

Uma idéia independente de como a hierarquia é obtida e que tenta evitar duplicações desnecessárias das palavras é a de Popescul e Ungar [14]. Nesse trabalho, foi utilizado um critério de atribuição das palavras mais discriminativas a cada grupo segundo o seu grau de dependência ou associação a cada grupo. Parte-se da raiz para as folhas da árvore, considerando-se os nós pais e filhos; se a hipótese de independência for aceita, então a palavra pertence ao nó pai e é removida dos nós filhos, caso contrário ela é associada aos nós filhos e removida do nó pai. Ainda, as palavras que ficam associadas

aos níveis mais altos da hierarquia, especialmente na raiz, são candidatas a *stopwords de domínio*, isto é, são palavras que poderiam ser cortadas no pré-processamento dos dados, pois não ajudam a discriminar efetivamente os documentos. Um problema desse método é garantir a convergência da distribuição das diferenças dos valores esperados e observados para uma distribuição chi-quadrado, utilizada nos testes, quando os valores esperados são muito pequenos; os valores esperados correspondem à probabilidade esperada de ocorrência da palavra em cada grupo considerado, ou seja, condicionada ao grupo. Dessa forma, conforme os critérios estabelecidos para os valores mínimos esperados e para os intervalos de confiança de aceite das hipóteses, o resultado obtido pode não ser tão confiável. Esse método desenvolve soluções bastante interessantes e poderia ser mais explorado a fim de tentar evitar os problemas inerentes ao uso da estimativa de chi-quadrado.

Neste trabalho, são propostos e avaliados dois modelos; o primeiro pode ser considerado um adendo ao Popescul e Ungar, que também não evita completamente a repetição das palavras nos agrupamentos hierarquizados; e, o segundo é um novo modelo, com base no anterior, que tanto resolve mais diretamente a questão repetição quanto utiliza um estimador mais robusto para testar as hipóteses de independência ou dissociação das palavras aos grupos.

3. Metodologia Desenvolvida

A metodologia aqui apresentada resume-se a encontrar as palavras mais discriminativas para cada agrupamento, considerando-se um agrupamento hierárquico, independentemente de como tenha sido obtido. Porém, para simplificar o desenvolvimento do método, vamos supor que os agrupamentos foram obtidos de algum método de cluster hierárquico, resultando em uma árvore binária. Com essa primeira suposição, tem-se sempre de um a dois nós filhos por pai.

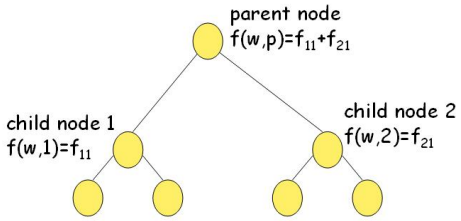

Figure 1. Palavras e freqüências nos nós pai, filho1 e filho2

A idéia geral, utilizada por Popescul e Ungar, e neste trabalho, é que a cada nó corresponde uma totalização da freqüência absoluta de cada palavra pertencente àquele grupo e a partir desta serão testadas hipóteses de independência ou dissociação das palavras a cada grupo, sendo considerados o grupo do pai e os dos filhos. Por exemplo, na Figura 1 pode-se observar uma palavra presente nos nós pai, $filho_1$ e $filho_2$, com as respectivas freqüências de ocorrência em cada nó. Para decidir se a palavra discrimina somente o nó pai, isto é, tanto faz se está em um ou outro filho, ou se a palavra discrimina apenas um dos filhos, faz-se um teste de homogeneidade de distribuição. Para isso, primeiro define-se a Tabela 1, cuja notação será utilizada em todo o trabalho:

- f_{i1}: freqüência absoluta acumulada da palavra no i-ésimo filho;
- f_{i2}: freqüência absoluta acumulada das outras palavras no i-ésimo filho;
- $f_{i.}$: freqüência absoluta total de palavras no i-ésimo filho;
- $f_{.j}$: freqüência absoluta total da j-ésima palavra ou de seu complemento; e,
- $f_{..}$: freqüência total de palavras no grupo pai, isto é, o total de seus filhos.

	$palavra$	$!palavra$	$total$
$filho_1$	f_{11}	f_{12}	$f_{1.}$
$filho_2$	f_{21}	f_{22}	$f_{2.}$
	$f_{.1}$	$f_{.2}$	$f_{..}$

Table 1. Tabela de freqüências positivas e negativas das palavras em cada filho

Para que a distribuição seja homogênea, espera-se que cada casela f_{ij} dependa exclusivamente das freqüências marginais; isto é, que $e_{ij} = f_{i.} * f_{.j} * f_{..}$, que é o valor esperado para cada casela f_{ij} sob a hipótese de independência. Para testar a hipótese pode-se, por exemplo, calcular o estimador chi-quadrado, onde:

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{f_{ij}}$$

A seguir compara-se esse valor ao de uma distribuição chi-quadrado com um grau de liberdade (porque a tabela é $2x2$) e um valor crítico para o aceite do teste (probabilidade de rejeição da hipótese, *p-value*). Se o valor obtido estiver abaixo do tabelado, então se considera a hipótese verdadeira, isto é, a ocorrência da palavra não depende do filho, discrimina apenas o nó pai.

No trabalho de Popescul e Ungar a metodologia desenvolvida utiliza exatamente essas considerações e o teste de chi-quadrado, além disso, é proposto um algoritmo que percorre a hierarquia da raiz para as folhas, decidindo a independência ou dependência das palavras e removendo-as dos ramos onde são menos significativas. Porém, nesse método, as restrições empregadas para a aplicação do cálculo da estimativa de chi-quadrado foram um pouco rígidas demais; o teste foi empregado apenas aos casos onde tanto e_{ij} ou f_{ij} eram maiores que 5 e, também o *p-value* era variado ao longo dos ramos, sendo mais severo para as folhas e menos severo para os ramos interiores. Assim, primeiramente implementou-se esse método, mas, após alguns experimentos, devido à severidade das restrições em muitos ramos das hierarquias não se conseguia tomar decisões, logo, era preciso encontrar alternativas que possibilitassem o uso do método.

As primeiras alternativas foram buscadas em métodos de análise estatística não-paramétrica, de modo que não fosse necessário assumir uma distribuição a *priori*. Porém, em análise não paramétrica considera-se basicamente amostras pequenas, isto é, precisaríamos ter o $f_{..}$ pequeno [15], o que, em problemas de mineração de textos, costuma ser falso. Outra alternativa, recomendada por Everitt [16] para substituir o chi-quadrado é utilizar o teste exato de Fisher, onde se assume uma distribuição hipergeométrica das freqüências da tabela. Porém, também esse teste é melhor aplicado a amostras pequenas, dado que a hipergeométrica envolve o cálculo de fatoriais das freqüências observadas. Além disso, como a distribuição hipergeométrica é discreta, para calcular a distribuição

acumulada até um valor, calcula-se cada valor até chegar ao desejado; logo, mesmo sendo possível calcular o fatorial sem *overflow*, mais essa complexidade de cálculos acumulados seria imposta ao algoritmo.

Voltando ao uso do chi-quadrado e de medidas de associação, verificou-se que haviam recomendações diferentes de outros autores, por exemplo, segundo Kachigan [17] a regra de considerar o valor cinco como limite inferior para as freqüências esperadas e observadas é muito conservadora; Fienberg [18] e Bishop [19] consideram que se o tamanho total da amostra ($f_{..}$) é grande, pode-se utilizar a chi-quadrado sem fatores de correção de continuidade, considerando-se a regra geral $e_{ij} >= 1$; ou, ainda, considerar medidas de associação invariantes para as diferentes combinações de linhas e colunas.

Nos próximos itens são descritos os dois métodos aqui propostos. No primeiro método considera-se as suposições de Fienberg [18] e Bishop [19]; no segundo utiliza-se uma medida de associação mais robusta, invariante para as diferentes combinações de linhas e colunas da tabela de freqüências. No terceiro item são brevemente descritas as implementações realizadas para o protótipo cujo propósito foi avaliar o uso dos métodos. Finalmente, no quarto item, o modelo de avaliação dos experimentos é descrito.

3.1. Método Chi-quadrado Adaptado

Este método é embasado no de Popescul e Ungar, para o qual modificam-se as restrições de aplicação. Assim, o algoritmo original foi redesenhado com as suposições a seguir:

- a hierarquia resulta de um cluster hierárquico: pois, neste primeiro momento, o algoritmo foi implementado para trabalhar com a hierarquia representada em uma árvore binária;
- $e_{ij} > 0.5$: é suficiente para que a hipótese seja testada com o uso da estimativa de chi-quadrado. Note que a restrição de freqüência mínima considerada incide apenas sobre o valor esperado e como ele é calculado, considera-se que a aproximação > 0.5 tende ao valor 1;
- $e_{ij} < 0.5$: assume-se que e_{ij} tende a zero e, conseqüentemente, resolvem-se esses casos, através de regras adotadas para cada qual. Utilizando-se a idéia de associação ou dissociação completa [18], foram estabelecidas as seguintes regras:
  - se ($e_{11} < 0.5$ e $e_{21} < 0.5$) $\Rightarrow$ aceita-se a hipótese de independência, isto é, a palavra discrimina igualmente o $filho_1$ e o $filho_2$, devendo ficar apenas no conjunto de rótulos do pai;
  - se $e_{21} < 0.5 \Rightarrow$ rejeita-se a hipótese de independência, pois a palavra discrimina apenas o $filho_1$;
  - se $e_{12} < 0.5 \Rightarrow$ rejeita-se a hipótese de independência, pois a palavra discrimina apenas o $filho_2$;
  - se ($e_{12} < 0.5$ e/ou $e_{22} < 0.5$) $\Rightarrow$ rejeita-se a hipótese de independência, isto é, a palavra deve discriminar algum ramo sob os $filho_1$ e $filho_2$, devendo ser retirada do conjunto de rótulos do pai.
- se $filho_1$ ou $filho_2$ é um cluster com um único documento: então é verificado se a palavra veio apenas do documento, se sim, ela é removida do cluster pai e o teste de independência pára aqui; se, por ventura, não tenha caído em um dos casos acima de associação ou dissociação completa.

Dessa forma, para vários ramos da hierarquia onde as freqüências observadas são baixas, e o método original de Popescul e Ungar deixa de tomar algum caminho, esta modificação do método toma uma decisão e, ainda, realiza o tratamento de zeros nas medidas esperadas para as freqüências.

3.2. Método Medida Q

Procurando-se por uma estimativa mais robusta, chegou-se à medida Q que tem base na *odds ratio*, que é invariante para as diferentes combinações de linhas e colunas da tabela de freqüências. A estimativa de máxima verossimilhança para a *odds ratio*, ou razão do produto cruzado, em uma tabela $2x2$, é dada por [19]:

$$\alpha = \frac{f_{11} * f_{22}}{f_{12} * f_{21}}$$

- se $\alpha = 1 \Rightarrow$ hipótese de independência é aceita
- c.c $\Rightarrow$ hipótese de independência é rejeitada

Deve-se notar que a troca de ordem entre linhas e colunas não interfere no resultado, exceto em seu sinal; e, também que a medida é invariante para combinações lineares, logo, não exige normalização dos dados; e, ainda, possui uma interpretação clara e é facilmente expansível para tabelas de tamanhos variados [19].

Um problema é que $\alpha \in [0, +\infty]$, logo, precisa-se de um intervalo de confiança para considerar que o α estimado tenda estatisticamente a um, para que a hipótese de independência seja aceita. Com o uso de uma função monotônica crescente de α, onde $f(1) = 1$, tem-se uma medida normalizada de associação baseada em α, cujo valor máximo absoluto é 1, representada pela medida de associação de Yule, $Q = (\alpha - 1)/(\alpha + 1)$, cujos estimadores de máxima verossimilhança são:

$$\widehat{Q} = \frac{\widehat{\alpha} - 1}{\widehat{\alpha} + 1}$$

$$\widehat{\sigma}_Q = \frac{1}{2} * (1 - \widehat{Q}^2) * \sqrt{\frac{1}{f_{11}} + \frac{1}{f_{12}} + \frac{1}{f_{21}} + \frac{1}{f_{22}}}$$

A distribuição da estimativa de Q é normal e o intervalo de confiança, de noventa e cinco por cento, é definido por:

$$\widehat{Q} \approx N(\widehat{Q}, \widehat{\sigma}_Q) \Rightarrow \widehat{Q} \pm 2 * \widehat{\sigma}_Q$$

Deve-se notar que, o valor máximo da função é atingido quando $\widehat{\alpha} = 1$ e $\widehat{Q} = 0$ e, que $\widehat{Q} = 1$ ou $\widehat{Q} = -1$ quando algum $f_{ij} = 0$; logo, se o valor $0 \in \left[\widehat{Q} - 2 * \widehat{\sigma}, \widehat{Q} + 2 * \widehat{\sigma}\right]$ então a hipótese de independência, ou dissociação, é aceita, e caso contrário é rejeitada. Para os valores de f_{ij} tendendo a 0 foram consideradas as mesmas regras utilizadas no método chi-quadrado adaptado. Assim, neste método, além do tratamento de baixas freqüências e de zeros, há a crença de que a estatística seja mais robusta e conseqüentemente mais confiável.

3.3. Protótipos Implementados

O objetivo das implementações realizadas foi permitir a comparação entre os diferentes métodos de obtenção da lista de palavras mais discriminativas para os agrupamentos de documentos. Essa lista de palavras deve se aproximar de uma descrição do tópico ao qual o documento se refere e também permitir indexar os documentos sob o tópico. Foram implementadas ou adaptadas algumas ferramentas para cobrirem as seguintes etapas do processo de análise dos textos, a partir da escolha da coleção de textos até a geração e avaliação dos rótulos:

1. **Pré-processamento**: foi utilizada uma ferramenta, denominda PreText [20] que remove uma dada lista de *stopwords* dos textos, gera os *stems* das palavras e calcula suas freqüências de ocorrência; estejam os textos em português, inglês ou espanhol. A seguir são obtidas algumas estatísticas, calculados alguns cortes e obtidos gráficos das freqüências de ocorrência dos *stems*, com *scripts* em MatLab. Os cortes de atributos (stems, no caso) são escolhidos a partir desses cálculos e gráficos, após o que gera-se uma representação matricial com documentos nas linhas, *stems* nas colunas e freqüências nas caselas. Também nessa etapa gera-se a descrição dos *stems*, que são os atributos selecionados.
2. **Cluster hierárquico**: uma relação hierárquica entre os documentos é gerada. Para não trabalhar com normalizações das freqüências dos *stems* nos documentos, foi utilizada a medida de similaridade com base em distância euclidiana e o algoritmo *complete linkage* para calcular a hierarquia. Ao próximo passo interessa apenas uma representação da hierarquia, que no caso, é matriz de *linkage* gerada (no momento, com MatLab) para todos os nós mais internos do agrupamento;
3. **Gerador de rótulos**: foi implementado um protótipo (em C) para gerar as listas de rótulos de cada agrupamento, segundo cada método implementado, representando-os em tabelas (em HTML para facilitar a posterior visualização). Além dos métodos descritos neste trabalho, foi implementado um quarto método muito utilizado e bastante difundido; logo, os métodos implementados são:
    - Popescul e Ungar: seleção da relação de palavras por chi-quadrado, com a restrição de e_{ij} e f_{ij} maiores que 5;
    - Chi-quadrado adaptado: seleção da relação de palavras por chi-quadrado, com a restrição de $e_{ij} > 0.5$ e tratando por regras os caso de f_{ij} zerados;
    - Método Q: seleção da relação de palavras de acordo com as estimativas do intervalo de confiança de $\widehat{Q}$ e tratando por regras os caso de f_{ij} zerados; e,
    - Mais freqüentes: seleção da relação de palavras de acordo com as freqüências de ocorrência das mesmas em cada agrupamento, isto é, considerando os estimadores de máxima verossimilhança de suas probabilidades de ocorrência condicionadas ao grupo.

    Para os métodos que utilizam chi-quadrado foi estabelecido um p-value de 5%, para compatibilizar com o valor de determinação do intervalo de confiança da estimativa de Q (que é de 95%).
4. **Visualização e avaliação**: também foi implementado um protótipo para visualizar a hierarquia obtida, onde cada para grupo é mostrada a lista de rótulos obtida para um dos métodos. Ainda, abrindo-se as páginas HTML vinculadas aos nós há a relação completa de rótulos, descrição dos *stems* e campo para atribuição de notas à lista de rótulos na taxonomia obtida; dado que a implementação desse protótipo tem o propósito de servir à avaliação dos métodos.

3.4. Método de Avaliação

Para avaliar a aplicabilidade de cada método de construção da lista de rótulos dos agrupamentos é aplicada uma análise subjetiva, realizada por especialistas na área de conhecimento a que a coleção de documentos se refere. Geram-se as listas de rótulos para cada método implementado e permite-se ao avaliador que atribua uma nota à lista de cada grupo, para cada método, considerando a hierarquia como um todo. Foi escolhido um número par de notas para evitar que, em dúvida, o avaliador decida pela nota média. Dessa forma o avaliador, em cada grupo, deve atribuir uma das seguintes notas à lista de rótulos:

- 1: o conjunto de palavras indicado atrapalha a identificação do tópico;
- 2: o conjunto de palavras ajuda pouco na identificação do tópico;
- 3: o conjunto de palavras realmente ajuda a identificação do tópico;
- 4: o conjunto de palavras aproxima-se do ideal na identificação do tópico.

Ao final do processo de avaliação tem-se uma relação de avaliadores, métodos, grupos e notas, em um delineamento hierárquico, pois as notas são dadas dentro de cada grupo e os métodos considerados para cada avaliador. Os dados resultantes são tabelados como na Tabela 2, onde:

	$avaliador$	$método$	$grupo$	$nota$
1	a_1	m_1	c_1	g_{111}
2	a_1	m_1	c_2	g_{112}
...	...	...	...	...

Table 2. Tabela de dados para avaliação dos métodos de rotulação

- a_i: i-ésimo avaliador, i=1,2,...,A, A o número de avaliadores;
- m_j: j-ésimo método, com j=1,2,3,4;
- c_k: k-ésimo grupo da hierarquia, sendo que k depende do maior número de grupos internos encontrados para cada coleção de textos, isto é, o número total de textos menos 1 (tomando-se todos os grupos mais internos de uma clusterização hierárquica *bottom-up*);
- g_{ijk}: nota do i-ésimo avaliador, para o j-ésimo método no k-ésimo grupo.

O modelo de análise de variância e utilizado na comparação múltipla de médias é do tipo hierárquico [21], especificado por dois modelos:

$$\hat{g} = \hat{\mu} + \hat{a} + \hat{m} + \hat{c(m)} + \hat{e}$$

$$\hat{g} = \hat{\mu} + \hat{a} + \hat{m(a)} + \hat{c(m)} + \hat{e}$$

onde,

- $\hat{g}$: estimativa da nota;
- $\hat{\mu}$: estimativa do efeito da média geral do modelo sobre a nota;
- $\hat{a}$: estimativa do efeito do avaliador sobre a nota;
- $\hat{m}$: estimativa do efeito do método sobre a nota;
- $\hat{m(a)}$: estimativa do efeito do método aninhado ao avaliador sobre a nota;
- $\hat{c(m)}$: estimativa do efeito do grupo (cluster) aninhado ao método sobre a nota;
- $\hat{e}$: erro aleatório da estimativa da nota.

O interesse é verificar o quanto os efeitos do avaliador e método, bem como o efeito método aninhado em avaliador, influenciam a nota e realizar as comparações múltiplas de médias sobre esses efeitos do modelo. Com as comparações múltiplas das médias dos efeitos será possível inferir se os métodos atendem às expectativas dos avaliadores e como eles se agrupam diante dessas expectativas.

4. Experimentos e Resultados

Em um primeiro momento foram realizados experimentos com bases de dados relativamente grandes, contendo mais de mil textos de um único domínio de conhecimento; isso foi feito para Gado de Corte em geral, Informática Agropecuária, Manufatura e Melhoramento de Gado, especificamente. Como não há exatamente uma taxonomia com a qual se possa comparar automaticamente a taxonomia de tópicos obtida, fez-se necessário obter uma avaliação subjetiva dos resultados. Seria impossível apresentar extensas taxonomias a especialistas do domínio, para que realizassem uma análise subjetiva minuciosa com os resultados do protótipo obtido; a tarefa não seria trivial e comprometeria a qualidade da avaliação. Então, para viabilizar uma avaliação de qualidade, que permitisse comparar os quatro métodos implementados, optou-se em primeiro lugar por uma base de dados menor, que pudesse ser julgada significativa para os domínios em questão. Primeiramente acertou-se quais seriam os especialistas que julgariam os métodos e em função desses fixou-se o domínio de conhecimento e a amostra de textos a ser utilizada; preprocessaram-se os dados, foram escolhidos cortes e representações; extraíram-se os agrupamentos; geram-se os rótulos e páginas HTML; e, finalmente ocorreu a avaliação e tabulação de resultados, seguida da análise dos mesmos.

4.1. Delimitação das coleções de textos e pré-processamento

Foram escolhidas coleções de textos para o domínio gado de corte, especificamente gado da raça Canchim, e para Informática Agropecuária. Para o domínio Canchim houve a disponibilidade de dois avaliadores especialistas do domínio, um avaliador especialista em informação para o domínio, e 48 textos completos - digitalizados em sua íntegra e em português. Para o domínio de Informática Agropecuária, muito amplo, foram selecionados 31 resumos de assuntos variados e atuais, em português, e, havia um especialista em informação para o domínio e três especialistas em subdomínios e com bom conhecimento do domínio como um todo.

	$CANCHIM$	$InformáticaAgropecuária$
$documentos$	48	31
$oneGrams$	7025	604
$oneGrams - Luhn$	2095	125
$oneGrams - Salton$	5868	514

Table 3. Coleção de textos, número de *onegrams* antes e depois dos filtros utilizados

Para o pré-processamento das coleções de textos foram utilizadas as mesmas listas de *stopwords* e o mesmo procedimento de *stemização*, gerando-se representações *onegram* dos radicais e contabilizando-se freqüência absoluta de ocorrência dos radicais nos textos. Para realizar os filtros foram utilizados cortes de Luhn, como mostrado na Tabela

3. Porém, como a coleção de textos de Informática Agropecuária escolhida é muito pequena e composta por resumos, o número de atributos ficou muito reduzido e obteve-se, na fase seguinte, um agrupamento quase linear dos textos. Assim, fez-se uma nova escolha dos filtros, optando-se por usar os radicais que ocorreram entre um a dez por cento da coleção de textos, que é uma recomendação um trabalho de Salton [22]. Essa escolha aumentou o número de atributos e permitiu obter uma hierarquia mais representativa dos 31 resumos.

4.2. Clusterização e Rotulação

Para a clusterização utilizou-se a medida de dissimilaridade baseada em distância euclidiana e o algoritmo *complete linkage* do MatLab (versão 7). Na Figura 2 é ilustrada a hierarquia para a amostra de textos de informática agropecuária, já com os rótulos calculados para os métodos 4 (mais freqüentes) e 3 (medida Q); os dois outros métodos não apresentam diferenças significativas, pois o método 1 (Popescul e Ungar) aproxima-se muito do 4 e o método 2 (chi-quadrado adaptado) aproximou-se bastante do 3. Deve-se notar que no método 3, cujo resultado é ilustrado na Figura 2, bem como no

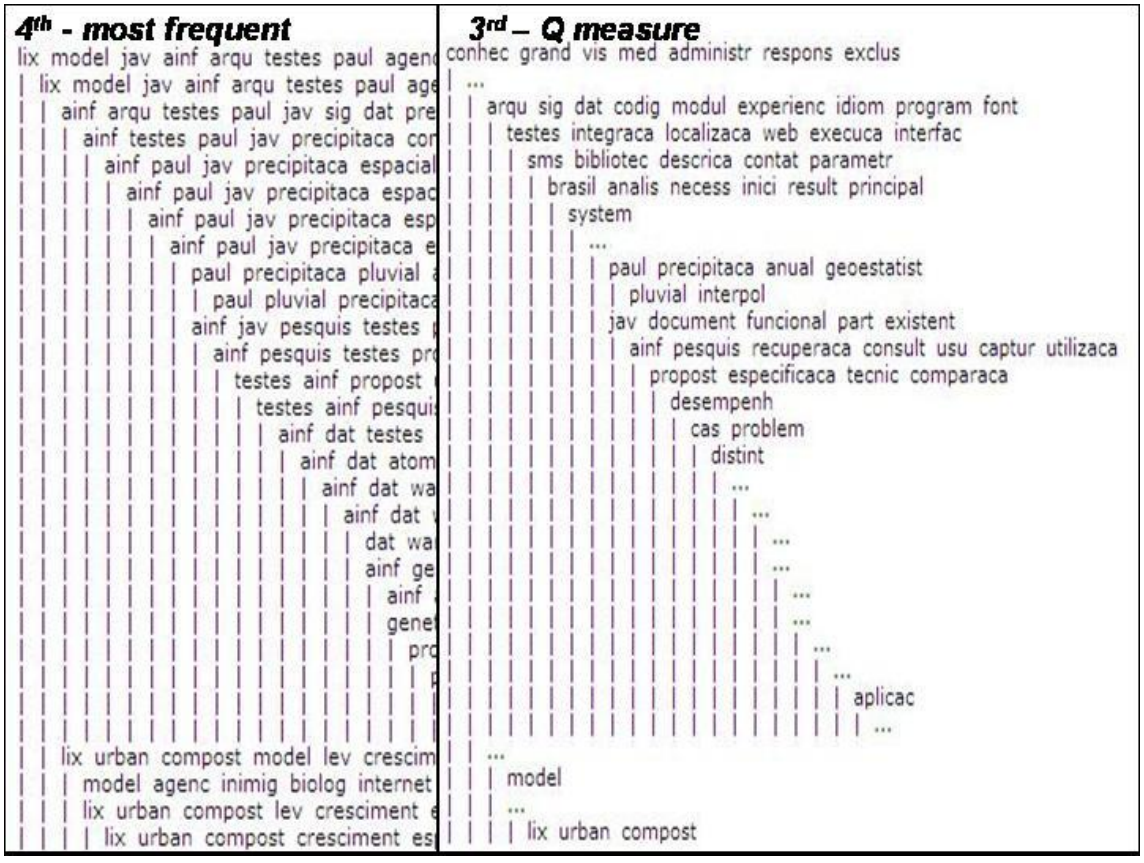

Figure 2. Hierarquia informática agropecuária: comparação dos resultados para os quatro métodos implementados

método 2 (não ilustrado), quando nenhuma das palavras de um grupo é significativamente associada ao grupo, a lista de palavras discriminativas fica vazia; e, na visualização criada, essa lista vazia é representada por "...". Nesses casos, para completar a lista é necessária a intervenção do especialista ou, se todo um ramo não contiver palavras significativas ele,

de fato, poderia ser eliminado, promovendo um *pruning* da árvore (idéia em teste e não implementada nesse primeiro protótipo).

4.3. Avaliação e Resultados

A partir das páginas HTML geradas para cada domínio e método foram realizadas as avaliações, tabulados e analisados os resultados. Além das notas atribuídas a cada nó e método pelos avaliadores, eles também enviaram críticas e sugestões. Dentre esses comentários os mais relevantes referiram-se à forma de gerar as representações das palavras, especialmente entre os especialistas em informação para o domínio, habituados ao uso de *thesaurus*, há a expectativa de encontrar representações ngrams, ou melhor, termos de fato compostos, como por exemplo "composto de lixo" em lugar de uma lista com "lix urban compost cresciment ...", como na última folha das hierarquias mostradas na Figura 2. Outro ponto de confusão foram os nós em que a lista palavras discriminativas eram vazias e representadas pelos ". . ."; para alguns avaliadores o significado disso era extremamente claro e para outros completamente obscuro.

Devido à confusão relativa aos nós rotulados como ". . .", na avaliação dos efeitos do método e do método aninhado a avaliador nos diferentes modelos, foram sempre consideradas dois conjuntos de dados, para cada domínio separadamente: **completo**, isto é, todas as avaliações para todos os nós, em todos os métodos e domínios; e, **sem os ". . ."**, isto é, apenas as observações relativas àqueles nós que não apresentaram os ". . ." em nenhum dos métodos de rotulação. Para cada tabela de resultados, são apresentadas as comparações múltiplas de médias das notas para cada grupo de efeitos estudado. Nessas comparações foi utilizado o teste SNK (Student-Newman-Keuls), com a raiz do erro quadrático do modelo ajustado, os graus de liberdade desse erro e um nível de significância de 5%. Esse teste foi escolhido devido a sua força, pois diferenças reais são mais freqüentemente apontadas por ele. Nas colunas de comparação das tabelas, letras iguais significam que as notas comparadas pertencem aos mesmos grupos; sendo que as letras estão em ordem ascendente de agrupamento de notas. E, em todas as tabelas os números e métodos correspondentes são: 1 para o de Popescul e Ungar, 2 para chi-quadrado adaptado, 3 para a medida Q e o 4 para o dos mais freqüentes.

Na Tabela 4 estão tabuladas as comparações múltiplas da variável nota ajustada pelos efeitos de média geral, avaliador, método e nó aninhado a método. Nota-se que no domínio Canchim houve uma clara preferência pelo método 4, seguido do 1, dado que eles sempre provêm uma lista de palavras claramente interpretável para cada agrupamento, mesmo que com palavras em excesso. No entanto, retirando-se os nós sem rótulos e reajustando-se o modelo, os métodos 1 e 3 têm suas notas mais próximas. Em contrapartida, para o domínio de Informática Agropecuária, os métodos estão claramente divididos em ordem de preferência: 1, 4, 3 e 2; e, quando se retiram os nós sem rótulos, então agrupam-se os que provêm maiores listas de palavras (1 e 4) e os que mais eliminam palavras da hierarquia (3 e 2). Orientando-se somente por essa primeira análise, parece que retirar as palavras menos discriminativas dos rótulos dos nós não ajuda o especialista a identificá-los, sem o auxílio por completo das listas de rótulos dos nós ascendentes.

Na análise seguinte, procurou-se observar o efeito do método aninhado a cada avaliador, isto é, ajustou-se o modelo de nota dependente de média geral, avaliador, método aninhado a avaliador e nó aninhado a método; tabelados nas Tabelas 5 para o

$CANCHIM$						$Agric.Informatics$					
$AllGroups$ $\sqrt{e} = 0.29, df = 290$			$Without - ...$ $\sqrt{e} = 0.27, df = 122$			$AllGroups$ $\sqrt{e} = 0.34, df = 316$			$Without - ...$ $\sqrt{e} = 0.39, df = 133$		
m	g	$comp.$	m	g	$comp.$	m	g	$comp.$	m	g	$comp.$
4	2.58	a	4	2.51	a	1	2.83	a	1	2.65	a
1	2.20	b	1	2.19	b	4	2.67	b	4	2.44	a
3	1.61	c	3	2.13	b	3	1.62	c	3	1.79	b
2	1.48	c	2	1.60	c	2	1.35	d	2	1.75	b

Table 4. Comparação múltipla de médias dos feitos de método sobre a nota

domínio Canchim e 6 para o domínio de Informática Agropecuária. Os efeitos de cada avaliador aninhado ao m-ésimo são representados por $i_k(m)$, o k-ésimo especialista em informação para o domínio, e $d_j(m)$, o j-ésimo especialista no domínio. A principal intenção dessa análise foi verificar se as avaliações dos especialistas em informação para o domínio diferiam muito dos especialistas no domínio. No entanto, as maiores diferenças apresentadas foram entre diferentes especialistas do domínio, ou seja, ficou difícil avaliar o que seria melhor para esse grupo de especialistas. Assim, focou-se a atenção nas avaliações dos especialistas em informação para o domínio.

modelo: $\hat{g} = \hat{\mu} + \hat{a} + \hat{m(a)} + \hat{c(m)} + \hat{e}$

$AllGroups$ $\sqrt{e} = 0.29, df = 284$			$Without - ...$ $\sqrt{e} = 0.26, df = 116$		
$m(a)$	g	$comp.$	$m(a)$	g	$comp.$
$d_1(4)$	3.02	a.....	$d_1(4)$	3.06	a....
$i_1(4)$	2.66	..b....	$d_1(1)$	2.81	a....
$d_1(1)$	2.51	c.b....	$d_1(3)$	2.69	a....
$i_1(1)$	2.30	c.....	$i_1(4)$	2.69	a....
$d_2(4)$	1.92	..d....	$i_1(3)$	2.25	..b..
$d_1(3)$	1.72	e.d....	$i_1(1)$	2.19	..b..
$d_2(1)$	1.67	e.d.f..	$d_1(2)$	1.94	c.b..
$d_1(2)$	1.66	e.d.f..	$d_2(4)$	1.73	c...d
$i_1(3)$	1.65	e.d.f..	$i_1(2)$	1.63	c.e.d
$i_1(2)$	1.40	e...f.g	$d_2(1)$	1.53	c.e.d
$d_2(3)$	1.36	....f.g	$d_2(3)$	1.40	..e.d
$d_2(2)$	1.21	.....g	$d_2(2)$	1.20	..e..

Table 5. Domínio Canchim - resultados da comparação múltipla de notas para os efeitos dos métodos de rotulação aninhados a avaliador

Para o domínio de Canchim, Tabela 5, considerando-se as avaliações do especialista em informação, nota-se que no modelo ajustado para todos os nós há uma nítida preferência pelo método 4 seguido do 1, ficando os métodos 2 e 3 agrupados em uma terceira colocação. Descontando-se os nós que não possuem rótulos e reajustando-se o modelo, o método 4 continua na liderança, porém o método 3 aproxima-se do 1, ficando no mesmo grupo, e o método 2 é empurrado sozinho para a terceira colocação. Isso pode sugerir que o método 3 atende bem às expectativas desse especialista, quando se refere exclusivamente a apontar palavras realmente discriminativas do grupo. Na Tabela **??** encontram-se os resultados para o domínio de informática agropecuária. O especialista em informação, nesse domínio claramente ficou com os métodos 1 e 4 na primeira colocação e os 2 e 3 em outro grupo, quase único; considerando os dois grupos de observações dos dados. Porém, devido à grande discrepância entre os especialistas do domínio, quando se retiram as observações para as quais alguns nós apresentam lista de rótulos vazia, não há diferença

modelo: $\hat{g} = \hat{\mu} + \hat{a} + \hat{m(a)} + \hat{c(m)} + \hat{e}$

AllGroups $\sqrt{e} = 0.30, df = 307$			*Without* − ... $\sqrt{e} = 0.31, df = 124$		
$m(a)$	g	*cmp*	$m(a)$	g	*comp.*
$d_1(1)$	3.10	$a..$	$d_1(1)$	3.00	$a.....$
$d_2(1)$	2.97	$a..$	$d_2(1)$	2.83	$a.....$
$i_1(1)$	2.88	$a..$	$i_1(4)$	2.75	$a.b.....$
$d_1(4)$	2.87	$a..$	$d_1(3)$	2.67	$a.b.c.....$
$i_1(4)$	2.83	$a..$	$d_1(4)$	2.67	$a.b.c.....$
$d_2(4)$	2.80	$a..$	$i_1(1)$	2.67	$a.b.c.d.....$
$d_3(1)$	2.40	$..b$	$d_2(4)$	2.50	$a.b.c.d.e....$
$d_1(3)$	2.27	$..b$	$d_1(2)$	2.33	$a.b.c.d.e.f..$
$d_3(4)$	2.20	$..b$	$d_3(1)$	2.08	$..b.c.d.e.f.g$
$i_1(2)$	1.67	$c..$	$i_1(2)$	2.00	$....c...e.f.g$
$d_1(2)$	1.67	$c..$	$d_3(4)$	1.83	$.....f.g$
$i_1(3)$	1.52	$c..$	$d_3(2)$	1.75	$.....f.g$
$d_3(3)$	1.37	$c.d$	$i_1(3)$	1.58	$h.....g$
$d_2(3)$	1.30	$c.d$	$d_3(3)$	1.50	$h.....g$
$d_3(2)$	1.30	$c.d$	$d_2(3)$	1.42	$h.....g$
$d_2(2)$	1.03	$..d$	$d_2(2)$	1.08	$h.....$

Table 6. Domínio de Informática Agropecuária - resultados da comparação múltipla de notas para os efeitos dos métodos de rotulação aninhados a avaliador

significativa entre as notas dos métodos para os especialistas 1 e 2, enquanto que, para os de número 2 e 3 o método 3 seja o pior de todos.

5. Considerações Finais

Neste trabalho o objetivo foi investigar alguns métodos para a rotulação de clusters hierárquicos, procurando adaptá-los a um processo de identificação de tópicos, com o objetivo de selecionar um método ou combinação de métodos que melhor satisfaçam às expectativas dos especialistas do domínio e que sejam de fácil implementação sobre uma hierarquia já existente. Assim, também foram propostos dois métodos que procuram eliminar o problema de repetição de palavras na lista de rótulos, ou seja, lista de palavras mais significativas, para cada grupo do agrupamento hierárquico.

Esperava-se com esses dois novos métodos um reflexo bastante positivo na análise dos especialistas no domínio, porém isso não ocorreu de forma direta. Suspeita-se que parte do problema seja relativa à forma de visualização dos resultados obtidos, pois cada lista de rótulos, de fato, é uma composição das listas de seus nós ascendentes, mas isso parece não ter sido considerado pelos avaliadores. Como os avaliadores, a partir de cada nó, entravam em uma página HTML que continha apenas a lista de palavras do nó e nela atribuíam as notas, provavelmente perdiam o referencial da hierarquia toda. Assim, eles preferiram ver a lista o mais completa possível em cada nó, como se o nó não pertencesse a uma hierarquia, o que melhor refletem os resultados produzidos pelos métodos 4 e 1. Apesar disso, observando-se individualmente os métodos para os avaliadores especialistas em informação para o domínio, o método 3 encontra bem os termos mais altamente associados ao nó, como se esperava pela robustez da estatística utilizada. E, ainda, houve as observações sobre a questão das listas de rótulos vazias, que precisam ser automaticamente resolvidas, segundo os próprios avaliadores.

Os resultados dos experimentos ajudaram a guiar o trabalho futuro imposto a este, isto é, a visualização de algumas alterações à proposta atual. Primeiramente, implemen-

tar, a partir do método 3 (medida Q), uma ferramenta que construa uma taxonomia mais automaticamente, provendo os *prunings* da árvore conforme os ramos vão ficando vazios. Utilizar esse método como método de *pruning* parece uma boa escolha, devido aos resultados apresentados. A seguir, tentar compor as listas de rótulos obtidos, para a árvore já podada, com o método que seleciona as palavras mais freqüentes (método 4), aplicando os critérios de propagação de termos em taxonomias, a fim de tentar melhorar as interpretações das listas. Finalmente, fazer com que a ferramenta permita a interferência do especialista na construção dos ramos e elaboração da lista de palavras, guiando-o com as estimativas obtidas para cada medida em cada passo do processo.

6. Referências

[1] Koshman, S. Spink, A. Jansen, B. J. Web searching on the Vivisimo search engine, JASIST, Vol. 57 (14), 2006, 1875-1887.

[2] Maedche, A.; Staab, S. Ontology Learning for the Semantic Web., IEEE Intelligent Systems, Vol. 16(2), 2001, 72-79.

[3] Bloehdorn, S. Cimiano, P. Hotho, A. Staab, S. An Ontology-based Framework for Text Mining, LDV Forum, Vol. 20(1), 2005, 87-112.

[4] Empresa Brasileira de Pesquisa Agropecuária - Embrapa, Agência de Informação Embrapa. Copyright @2005-2007 Embrapa. In: http://www.agencia.cnptia.embrapa.br/. Access 07/07/2007.

[5] Evangelista, S. R. M.; Souza, K. X. S.; Souza, M. I. F.; Braga, S. A. C.; Leite, M. A. A.; Santos, A. D.; Moura, M. F. Gerenciador de conteúdos da Agência Embrapa de Informação, International Symposium on Knowledge Managment - ISKM, 2003, 1-12. (Part of CD-ROM)

[6] Moura, M.F. Uma abordagem para a construção e atualização de taxonomias de tópicos a partir de coleções de textos dinâmicas, São Carlos: Inst. Ciências Matemáticas e de Computação - ICMC/USP, 2006. (Phd qualifying monograph).

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

[8] Weiss, M. S. Indurkhya, N. Zhang, T. Damerau, F. J. Text Mining - Predictive Methods for Analyzing Unstructured Information, 1st edn. Springer Science+Business Media, Inc (2005)

[9] Larsen, B. Aone, C. Fast and Effective Text Mining USing Linear-time Document Clustering, In: Proceedings of the Knowledge Discovery and Data Mining, KDD, 1999, San Diego, CA, USA, 1999, 16-22.

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

[11] 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.

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

[13] Treeratpituk, P. Callan, J., Automatically labeling hierarchical clusters, In: Proceedings of the 7th Annual International Conference on Digital Government Research, DG.O 2006, San Diego, California, USA, May 21-24, 2006, 167-176.

[14] Popescul, A. Ungar, L.H. Automatic labeling of document clusters, Unpublished manuscript (2000), available at: http://citeseer.nj.nec.com/popescul00automatic.html.

[15] Lehman, E.L. Nonparametrics: Statistical Methods based on Ranks, Prentice Hall, 1998.

[16] EVERITT, B. S. The analysis of contingency tables. London: Chapman and Hall, 1977. 128 p.

[17] Kachigan, S.K. Statistical Analysis. New York: Radius Press; 1986.

[18] Fienberg, S.E. The analysis of cross-classified categorical data, 2nd.ed, Cambridge: MIT, 1985.

[19] Bishop, Y.M.M. Fienberg, S.E. Holland, P.W. Discrete multivariate analysis : theory and practice, Cambridge, Mass.; London : M.I.T. Press, 1975.

[20] Matsubara, E.T. Martins, C.A. Monard, M.C. Pre-TexT: uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos, 2003, Relatório Técnico n. 209.

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

[22] Salton, G. Yang, C.S. Yu, C.T. A theory of term importance in automatic text analysis, Journal of the American Association Science,Vol. 1(26), 33-34, 1975.