Boletim de Pesquisa 26
e Desenvolvimento
ISSN 1677-9266
dezembro, 2010

Metodologia para a comparação de diferentes métodos de descrição de agrupamentos hierárquicos de documentos independentes do algoritmo de agrupamento

Embrapa

ISSN 1677-9266
Dezembro, 2010

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

Boletim de Pesquisa e Desenvolvimento 26

Metodologia para a comparação de diferentes métodos de descrição de agrupamentos hierárquicos de documentos independentes do algoritmo de agrupamento

Maria Fernanda Moura
Fabiano Fernandes dos Santos
Ricardo Marcondes Marcacini
Solange Oliveira Rezende

Embrapa Informática Agropecuária
Campinas, SP
2010

Embrapa Informática Agropecuária
Av. André Tosello, 209 - Barão Geraldo
Caixa Postal 6041 - 13083-886 - Campinas, SP
Fone: (19) 3211-5700 - Fax: (19) 3211-5754
www.cnptia.embrapa.br
sac@cnptia.embrapa.br

1ª edição on-line 2010

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Informática Agropecuária

Metodologia para a comparação de diferentes métodos de descrição de agrupamentos hierárquicos de documentos independentes do algoritmo de agrupamento / Maria Fernanda Moura... [et al.]. - Campinas : Embrapa Informática Agropecuária, 2010.

37 p. : il. - (Boletim de pesquisa e desenvolvimento / Embrapa Informática Agropecuária , ISSN 1677-9266 ; 26).

1. Descritores de agrupamento. 2. Agrupamento hierárquico. 3. Mineração de texto. 4. Modelos lineares generalizado. 5. Analise de variância. 6. Recuperação de informação. I. Moura, Maria Fernanda. I. Titulo. II. Série.

CDD 006. 3 (21. ed.)

Sumário

Metodologia para a comparação de diferentes métodos de descrição de agrupamentos hierárquicos de documentos independentes do algoritmo de agrupamento

Maria Fernanda Moura[1]
Fabiano Fernandes dos Santos[2]
Ricardo Marcondes Marcacini[3]
Solange Oliveira Rezende[4]

Resumo

Para facilitar a compreensão de uma coleção de documentos, pode-se organizá-la em grupos hierárquicos e obter descritores para cada um dos grupos automaticamente. O problema que se apresenta é decidir entre métodos de agrupamentos e de descrição dos grupos, que sejam eficientes e apresentem bons resultados. Particularmente, este trabalho apresenta uma proposta para a comparação entre resultados obtidos a partir de métodos de seleção de descritores em agrupamentos hierárquicos de documentos, especificamente para métodos independentes do algoritmo de agrupamento utilizado. Para esses métodos, dado um agrupamento hierárquico, o objetivo é selecionar descritores (palavras ou sentenças) discriminativos dos grupos, preferencialmente sem repetição de descritores ao longo dos ramos, utilizando-se de um vocabulário variado e de qualidade, que seja

[1] *Doutora em Ciências de computação e matemática computacional, Pesquisadora da Embrapa Informática Agropecuária, fernanda@cnptia.embrapa.br*

[2] *Universidade de São Paulo, fabianof@icmc.usp.br*

[3] *Universidade de São Paulo, rmm@icmc.usp.br*

[4] *Universidade de São Paulo, solange@icmc.usp.br*

efetivamente representativo da coleção de textos agrupada. Dessa forma, torna-se imperativo encontrar uma medida que verifique a efetividade da discriminação para os descritores selecionados, bem como uma medida de qualidade destes. Nessa proposta, a discriminação é medida por meio da avaliação dos resultados de um processo de recuperação de informações, que utiliza os descritores para formar as expressões de busca. A qualidade é medida pela variabilidade do vocabulário obtido e sua representatividade em relação ao conjunto total de atributos utilizado para o agrupamento dos documentos. Essas medidas e processo de validação respeitam a hierarquia produzida pelo agrupamento, bem como padronizações e validações confiáveis do ponto de vista estatístico. Os experimentos e os resultados obtidos mostram que essa metodologia é capaz de avaliar seguramente a diferença de efetividade entre métodos de descrição de agrupamentos hierárquicos, tendo sido aplicada a dezesseis coleções de textos e quatro diferentes métodos de descrição.

Termos para indexação: descritores de agrupamentos hierárquicos de documentos, revocação, precisão, informação mútua média, comparação múltipla de médias, modelos lineares generalizados.

A methodology to compare different labeling methods for hierarchical document clusters which are independent from cluster algorithm

Abstract

To facilitate a document collection understanding it can be automatically organized in hierarchical groups followed by a description. The problem arises in deciding among clustering methods and description methods that are efficient and effective. Particularly, this work presents a proposal for the comparison among results from hierarchical document clusters labeling methods, specifically methods that are independent of the clustering algorithm. For these methods, given a hierarchical clustering, the goal is to select descriptors (words or sentences, also called labels) to discriminate the groups, preferably without repetition of descriptors along the hierarchy branches, using a varied and high quality vocabulary which is text collection representative. Consequently, it becomes imperative to find a measure to check the effectiveness of the selected descriptors for discrimination and a measure of their quality. In this proposal discrimination is measured by assessing the results of an information retrieval process, which uses the keywords to form search expressions. Quality is measured by the variability and the representativeness of the selected descriptors in relation to the original attribute, which is used for grouping the text collection. These measures and their validation process respect the hierarchy produced by the group as well as data standardization in a reliable statistical point of

view. The experiments and results show that this methodology is certainly capable of evaluating the difference in effectiveness between methods for describing hierarchical clustering; it was applied in sixteen collections of texts and four different methods of hierarchical document cluster labeling.

Index terms: hierarchical document clusters descriptors, precision, mutual information mean, multiple mean comparisons, generalized linear models.

1 Introdução

Grande parte dos dados disponíveis no universo digital encontra-se em um formato não-estruturado, principalmente no formato textual, em forma de relatórios, boletins, artigos, notícias de jornais, emails e todo tipo de documentos (GANTZ; REINSEL, 2009). Esse elevado volume de dados e informações faz com que seja necessária sua organização de forma mais rápida e eficiente, sendo um trabalho praticamente impossível de se fazer manualmente. Tal organização traz, em geral, vantagem competitiva para as empresas e pessoal, possibilitando a recuperação de informação e a descoberta de conhecimento para suporte à tomada de decisões.

Um modo natural de organizar esses documentos, para então começar a analisá-los e utilizá-los, é dispô-los em uma organização hierárquica de tópicos relacionados. Sendo que, 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 (VIVISIMO, 2006). Ainda, 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 (MAEDCHE, 2002). Outro exemplo de organização hierárquica do conhecimento de um domínio pode ser encontrado na Agência de Informação Embrapa (EMBRAPA, 2010).

Frente a esse cenário, a gestão do conhecimento para um domínio pode ser apoiada pela mineração de textos com a construção de taxonomias de tópicos de forma semiautomática ou automática. Uma forma simples, que envolve apenas modelos estatísticos de aprendizado de máquina, é organizar a coleção de documentos a partir de um agrupamento hierárquico, por meio de algoritmos aglomerativos ou divisivos, e obter descritores para cada um dos grupos automaticamente.

Os agrupamentos de documentos são calculados a partir da representação matricial de uma coleção de textos, considerando-se cada texto como um

elemento a ser agrupado, que é representado por um vetor de atributos. Cada atributo corresponde a uma palavra ou composição de palavras (por exemplo, “informática agropecuária”, “agricultura”), para as quais é obtida uma medida em relação ao texto. Essa medida pode ser qualitativa, representando a pertinência ou não da palavra ao documento, ou quantitativa, como a frequência ou a frequência inversa de ocorrência da palavra no texto. Os agrupamentos obtidos refletem tópicos e subtópicos aos quais os documentos se referem, logo, para melhor interpretá-los selecionam-se conjuntos de atributos mais significativos no grupo, denominados **descritores** ou **rótulo** que, por sua vez, podem ser interpretados como um conceito associado ao grupo.

Pode-se decidir explorar um grupo de documentos guiado pelo seu rótulo; porém, resumir um agrupamento com base em um pequeno conjunto de termos é uma tarefa bastante difícil. Logo, a questão de abstração de dados em problemas de agrupamento de documentos envolve obter conjuntos de descritores concisos e significativos para os grupos e, a seleção desses descritores chega a ser tão importante quanto o próprio agrupamento (FELDMAN; SANGER, 2007).

Embora os resultados apresentados por métodos de obtenção de descritores dos grupos baseados nos algoritmos utilizados para obter o agrupamento (exemplo o Taxaminer) (KASHYAP et al., 2005) ou o de Lamirel et al. (2008) apresentem bons resultados, a maioria dos algoritmos propostos para agrupamento de documentos não são acompanhados pela geração de descritores de grupos (KONCHADY, 2006). Investe-se muito na melhoria dos algoritmos de agrupamento de documentos, complexos devido ao tratamento de matrizes altamente esparsas, pois vários atributos não ocorrem em vários textos, mas não se investe da mesma forma na solução de descrição dos grupos obtidos.

Quando se dispõe do resultado do agrupamento, a tarefa de identificar os descritores pode ser vista como um problema de seleção supervisionada de atributos (WEISS et al., 2005), considerando-se cada grupo como uma classe. Mesmo assim, alguns autores utilizam seleção não supervisionada, com base na frequência de ocorrência dos atributos, tomando-se a moda de cada qual como descritor; a medida *tf-idf* (*term frequency - inverse document frequency*) considerando-se os seus maiores valores; ou, alguma

análise fatorial. Um bom exemplo de seleção não supervisionada é dado no ambiente WEKA, no qual o módulo kmeans utiliza as modas de cada atributo nominal como descritores dos grupos obtidos. Outro bom exemplo é o uso da medida *tf-idf* no módulo *k-means* do ambiente Text Mining Software Kit (TMSK) (WEISS et al., 2005). A medida *tf-idf* corresponde à frequência observada da palavra na coleção, ou no agrupamento como utilizada nesse caso, ponderada pelo logaritmo da sua frequência inversa na coleção; a ponderação visa eliminar palavras muito frequentes e privilegiar as menos frequentes. No ambiente TMSK, para selecionar as palavras mais discriminativas de cada agrupamento encontrado, o método empregado considera aquelas que possuem as maiores *tf-idfs* médias.

Geralmente, quando o objetivo é encontrar descritores para agrupamentos hierárquicos, de forma independente de como o agrupamento tenha sido obtido, usa-se a distribuição de frequências das palavras ao longo dos grupos. Glover et al. (2002) propuseram um modelo baseado exclusivamente nas frequências observadas para cada palavra em cada grupo, isto é, nas estimativas de máxima verossimilhança de suas probabilidades: $p(w|c)$ e $p(w)$, com w correspondente à palavra e c ao grupo, considerado como uma classe. A hipótese assumida é que, fixando-se uma classe c (um grupo), se $p(w|c)$ é muito comum e $p(w)$ é raro, então a palavra é um bom discriminador para a classe c; caso contrário, se $p(w|c)$ e $p(w)$ são comuns, então a palavra discrimina a classe ascendente de c e, finalmente, se $p(w|c)$ é muito comum e $p(w)$ é relativamente rara na coleção, então a palavra é um melhor discriminador para a classe descendente de c. Determinam-se faixas de valores para *muito comum* e *raro* experimentalmente.

Nessa linha, caminham métodos que se baseiam em teste de dependência entre um atributo e uma divisão de uma classe em subclasses, isto é, de um grupo em subgrupos, considerando todos os grupos como diferentes populações e que o atributo respeita a uma distribuição multinomial. Dessa forma, um atributo é discriminativo ou não em um dado nó da hierarquia se ele depende estatisticamente do nó. A seleção é realizada da raiz a partir de um nó ascendente, testando-se se a distribuição de cada atributo é homogênea (comum) ou heterogênea (rara) ao longo dos filhos desse nó. Se a hipótese de independência é aceita, então o atributo tem distribuição homogênea (é comum) nos filhos e está associado apenas ao nó pai, caso

contrário, o atributo deve estar associado a algum dos nós descendentes, pois é raro em relação a algum dos nós descendentes. Métodos com base nessas hipóteses são implementados, de forma simples e eficiente, por alguns autores (MOURA; REZENDE, 2010; MOURA et al., 2008; POPESCUL; UNGAR, 2000).

Popescul e Ungar valeram-se do uso do estimador qui-quadrado para testar a independência dos atributos em relação às classes, considerando as restrições de aplicação do estimador mais recomendadas pela literatura estatística, que prevê um valor observado ou estimado mínimo maior ou igual a cinco (cinco) para as frequências dos atributos (BISHOP, 1975; EVERITT; LANDAU, 2001); e, não se preocuparam em tratar as possíveis correções dos *p-values* no processo de decisão (recomendadas por Bonferroni ou Dunett (MOURA et al., 2008; RAMSEY, 1993) propuseram um algoritmo que se utiliza de estimadores mais robustos (Q de Yule e o U da diferença entre as proporções explicadas da variância), bem como trataram exceções, correspondentes às frequências e/ou as suas esperanças estarem próximas a zero; obtendo conjuntos de descritores bastante enxutos. A vantagem do uso de estimadores como o Q de Yule e a proporção explicada da variância (BISHOP et al., 1975) é a não dependência da distribuição de probabilidades das frequências observadas, pois em mineração de textos, essas frequências são sempre muito esparsas. Em um segundo trabalho, Moura e Rezende (2010), propuseram um novo método com baixa complexidade computacional, que além de utilizar estimadores robustos e tratar exceções, diminui a propagação de erros de decisão sobre a seleção de descritores ao longo da hierarquia dos textos, de modo que as correções dos *p-values* podem ser desconsideradas; com isso obtiveram descritores únicos para cada grupo na hierarquia.

Recentemente, Zhang et al. (2009) propuseram um método, também independente de como o agrupamento tenha sido obtido, extremamente simples, cuja base é a utilização do vetor central de cada grupo, tomando-lhe as cinco palavras de maior peso, seguido de uma escolha heurística das melhores palavras; o problema é que a heurística é um tanto quanto subjetiva - envolve escolha humana. A validação dessa proposta é realizada sobre um agrupamento, manualmente anotado, e usando diferentes classificadores. Assim, esse método não toma decisões sem intervenções humanas e para ser validado também depende destas.

Logo, um problema é decidir por um método de descrição de grupos, hierarquicamente obtidos, que tome decisões sem a necessidade de intervenções humanas, e, que seja capaz de fornecer uma boa discriminação para cada grupo; pois, quando se realizam comparações subjetivas de resultados de diferentes métodos, sejam computacionais ou não, existem várias influências nas respostas, que podem sofrer um forte efeito do *background* de cada avaliador, como o avaliador se sente no dia, como o questionário foi elaborado, e outras questões difíceis de serem isoladas, de forma que se torna difícil focar a avaliação nos pontos de interesse. Avaliações objetivas, a partir de medidas sobre o objeto avaliado, isolam as características que se deseja avaliar das demais.

Assim, a metodologia de avaliação, aqui proposta, compara os resultados obtidos por métodos de descrição de agrupamentos, que independam de como o agrupamento tenha sido obtido e que tomem decisões sem intervenções subjetivas, por meio de medidas objetivas. A proposta, descrita na próxima seção, consiste em utilizar os resultados como expressões de busca, verificar a variabilidade de vocabulário utilizada pelos métodos e a representatividade destes em relação aos atributos iniciais. Complementando a proposta, inicialmente há um processo de padronização dos fatores que influenciam os resultados dos métodos, e, por fim, a forma de avaliar as medidas obtidas é dada por meio de um modelo linear generalizado para análise de variância. A proposta foi aplicada a vários casos, ilustrando-se na seção “Resultados e Discussão” os resultados obtidos para quatro métodos de descrição de agrupamentos hierárquicos aplicados a dezesseis coleções de textos. Finalmente, conclui-se, na seção “Conclusões e Trabalhos Futuros”, que a metodologia de avaliação cobre, com segurança, as comparações que levam às melhores escolhas de descritores para agrupamentos hierárquicos de documentos.

2 Metodologia proposta

Nesta proposta, buscou-se por algumas medidas que, aplicadas sob determinadas condições, conseguissem classificar os resultados obtidos por diferentes métodos de descrição de agrupamentos hierárquicos de documentos.

Idealmente, para que se encontrem bons discriminadores para cada grupo em uma hierarquia, os atributos escolhidos como descritores não devem ser replicados ao longo dos mesmos ramos na hierarquia; pois, em uma hierarquia espera-se que os conjuntos de descritores mais genéricos estejam em um nó mais próximo à raiz da árvore e os mais específicos em seus descendentes. Dessa forma, um conjunto de descritores completo, ou mais abrangente, para um determinado nó seria a união de seus descritores e todos os descritores de seus ascendentes.

Por exemplo, considere a hierarquia apresentada na Figura 1, na qual cada nó é descrito por um conjunto de palavras *stemmizadas*[5]. Essa hierarquia foi obtida com um algoritmo aglomerativo e depois automaticamente descrita pelo método RLUM (MOURA; REZENDE, 2010). Note que a hierarquia refere-se a produtos agrícolas, como peixes e gado, e sistemas de produção de gado. Note que para o nó "*reproduct*" existe um conjunto de descritores específicos, {*reproduct*} (corresponde à reprodução), e os conjuntos de descritores dos ascendentes são {*cattl, beef*} (corresponde a gado e corte) e {*pantan*} (corresponde a Pantanal). Logo, se o objetivo for localizar documentos que se refiram "à reprodução de gado de corte na região do Pantanal", provavelmente esses documentos devem estar sob esse grupo. Assim, para recuperar esses documentos podem-se unir os descritores desse nó e de seus ascendentes.

Logo, buscaram-se medidas que refletissem essas características, como mostrado nas próximas subseções. Ainda, para comparar quaisquer carac-

Figura 1. Hierarquia rotulada com o método RLUM.

Fonte: Moura e Rezende (2010).

[5] Stemmizar *corresponde a remover todas as inflexões das palavras; não é uma radicalização, trata-se de um processo mais simples.*

terísticas medidas, os dados a serem medidos ou utilizados nas funções que as meçam precisam estar padronizados, para que as influências sobre as medidas possam ser consideradas como dependentes apenas dos diferentes métodos utilizados. Por fim, a comparação das medidas pode ser realizada de acordo com um modelo generalizado de análise de variância, que permite que se observe e compare cada diferença medida para cada método, levando-se em conta efeitos de fatores como em que altura da hierarquia a medida é obtida. As próximas subseções tratam cada uma dessas questões.

2.1 Padronização dos dados

Deve-se lembrar que, o objetivo é comparar métodos de descrição de agrupamentos hierárquicos que trabalhem a partir do resultado de um agrupamento. Assim, obtém-se o agrupamento hierárquico da coleção de textos de interesse, então uma cópia dessa hierarquia e da matriz de incidência é fornecida a cada método. Desse modo, documentos, o conjunto de atributos e a hierarquia são exatamente os mesmos.

Como alguns métodos causam mudanças na configuração da hierarquia, para compará-los de modo justo, não se devem comparar os conjuntos de rótulos obtidos para a hierarquia toda a partir dos resultados dos métodos, mas sim apenas os conjuntos de rótulos dos grupos (nós) mantidos nas hierarquias após a aplicação dos métodos. Desse modo, as comparações são realizadas nó a nó, de acordo com a sua presença dos mesmos na hierarquia. Logo, devem-se comparar apenas os resultados obtidos para os mesmos nós - nós conservados ao longo da aplicação dos métodos. Para simplificar as comparações nó a nó, garantindo que sejam feitas dessa forma, e que se possa comparar apenas os efeitos dos diferentes métodos nos resultados obtidos, a proposta é ajustar um modelo linear generalizado (SEARLE, 1971) aos dados obtidos, como descritos na subseção seguinte.

Por fim, para que as comparações sejam justas nó a nó, o ideal é que os conjuntos de rótulos sejam de tamanhos semelhantes. Pode-se forçar essa situação, colocando-se um corte de número máximo de rótulos em cada conjunto, como costuma ser feito para o método dos Mais Frequentes - se-

leciona os atributos de maior frequência no grupo como descritores deste. Utilizando essa ideia, basta aplicar o método dos mais frequentes a cada resultado apresentado pelos demais métodos; ou seja, toma-se cada rótulo específico (produzido para um único nó), ordenam-se os elementos por frequência e, então, se limita o conjunto aos k (a ser determinado) elementos mais frequentes. Assim, neste trabalho, em todas as comparações, serão utilizados os mesmos grupos, que são os nós mantidos pelos diferentes métodos, e conjuntos de rótulos de tamanhos semelhantes.

2.1.1 Modelo linear generalizado e comparações múltiplas de médias

Assim, os experimentos que forem conduzidos precisam medir alguma característica dos resultados do conjunto de descritores naquele nó, e, então, tabular os dados com: a medida no nó, o número do nó (ou nome dele) e, nome do método empregado. Então, ajusta-se um modelo linear da seguinte forma:

$$\hat{c}_i = \hat{\mu} + \hat{n}_i + \hat{m}_{ij} + \hat{\epsilon}_i$$

Sendo,

$\hat{c}_i$: valor estimado da característica no *i-ésimo* nó,

$\hat{\mu}$: estimativa da média geral da característica,

$\hat{n}_i$: estimativa da influência do efeito do *i-ésimo* nó na característica medida, sendo que o valor de *i* varia do primeiro nó numerado ao último,

$\hat{m}_{ij}$: estimativa da influência do efeito do *j-ésimo* método de descrição de agrupamentos hierárquicos de documentos sobre a característica medida no *i-ésimo* nó, e

$\hat{\epsilon}_i$: estimativa do erro aleatório sobre a característica no i-ésimo nó.

O ajuste desse modelo, que pode ser feito por vários softwares estatísticos, é exemplificado neste trabalho com o uso do MODLIN do SOC (desenvolvido pela Embrapa Informática Agropecuária, repositório da Rede Agrolivre[6]. Após o ajuste, o MODLIN apresenta uma tabela de análise de variância (ANOVA), da seguinte forma:

[6] *Disponível em: <http://repositorio.agrolivre.gov.br/>.*

Tabela 1. ANOVA para o modelo linear da característica medida.

Fonte	Graus de liberdade	Soma de quadrados	Quadrado médio	VALOR F	P_value
Nó	n-1	SQ(nó)	QM(nó)	QM(nó)/QMRes	1-P(f≤F)
Método	m-1	SQ(método)	QM(método)	QM(métod)/QMRes	1-P(f≤F)
Resíduo	(k-1)-(n-1)-(m-1)	SQRes	QMRes		
Total	k-1	SQTotal			

Na Tabela 1, os quadrados médios são as variâncias relativas a cada fonte de variação, sequencialmente colocadas no modelo, *n* corresponde ao número de nós nas hierarquias; *m* o número de métodos comparados; *k* o número total de observações tabuladas como dados. O primeiro teste que se faz é da representatividade da proporção da variância explicada pelo efeito do nó sobre a proporção de variância explicada apenas pela média geral (o valor da estimativa de μ). O valor de F obtido nessa primeira linha, indica se a proporção explicada pelo nó é ou não significava no modelo em relação à explicada pelo erro aleatório (resíduo). Para interpretar o *p_value* fornecido, estabelece-se uma probabilidade de aceitação de que o efeito seja significativo, por exemplo 95%; nesse caso, se o *p-value* for menor ou igual a 5%, o efeito é significativo no modelo. A segunda hipótese refere-se à proporção da variância explicada pelo método de descrição dos agrupamentos hierárquicos de documentos em presença da média geral e da proporção de variância explicada pelo nó; e, interpreta-se o *p_value* da mesma forma.

Supondo que os efeitos no modelo tenham sido significativos[7], considera-se este modelo e pode-se calcular a comparação entre as médias obtidas para os efeitos dos métodos sobre as características medidas. Para isso, utilizam-se de comparações múltiplas de médias (MORETTIN; BUSSAB, 2006), que corrigem as variações pelos cálculos das variâncias apresentadas no modelo linear ajustado. Um bom teste, comum a vários softwares estatísticos, é o Student-Newman-Keuls (SNK), e também devido à sua força, pois diferenças reais são mais frequentemente apontadas por ele (SNEDECOR; COCHRAN, 1967). Para obter as comparações múltiplas das diferentes médias, existem várias diferentes formas, sugeridas na lite-

[7] *Em todos os experimentos realizados o modelo teve essa mesma configuração e seus efeitos foram todos significativos.*

ratura, de acordo com as distribuições dos dados observados, por exemplo, o MODLIN também implementa Least Significant Difference (LSD), Duncan, Scheffé, Modified Least Significant Difference (MODLSD) e Tukey.

Para executar o teste SNK, no MODLIN, com a raiz do erro quadrático do modelo ajustado, os graus de liberdade desse erro e um nível de significância de 5%, simplesmente executa-se, logo após o ajuste do modelo, o comando "compara'' seguido do nome do efeito que se quer testar. Então, é fornecida uma tabela do tipo da Tabela 2, considerando-se testar o efeito dos métodos como no modelo do exemplo.

Tabela 2. Comparação múltipla de médias.

Teste "SNK" para a variável: "nome dela." g.l qme = Raiz(QMRes) alfa = 0.05, médias ligadas com uma mesma letra não são significativamente diferentes			
Efeito comparado	**Num. observações**	**Valor médio da característica**	**Grupo, representado por letras**
$Método_x$	n_x	c_x	l_x
$Método_y$	n_Y	c_y	l_y
...	...	...	...
$Método_z$	n_z	c_z	l_z

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. Ou seja, componentes do efeito agrupados com a mesma letra indica que não há diferença estatisticamente significante entre eles. E, ainda, a ordem das letras significa que os componentes daquele efeito provocam desvios maiores que os dos demais.

2.1.2 Medidas para a discriminação dos descritores

A proposta é tratar todos os rótulos, conjuntos de descritores de um nó, como expressões de busca. Por exemplo, seja o rótulo específico R_{ei}={*husbandri,product*} de certo nó *i* de uma hierarquia; ele pode ser usado como uma expressão de busca da seguinte forma: "*husbandri and product*"'. Desse modo, espera-se que os documentos sob esse nó *i* sejam recuperados dentre os demais documentos da coleção de textos agrupada sob a hierarquia.

Para implementar um processo de recuperação de informações, para testar essas características, pode-se utilizar a matriz atributo-valor utilizada pelo algoritmo de agrupamento. Nessa matriz, as linhas correspondem aos documentos e as colunas aos atributos utilizados, sendo que cada célula corresponde à frequência de ocorrência do atributo no documento. Para decidir se um documento foi ou não recuperado, pela expressão, todos os termos da busca devem ter frequência maior que zero no documento. Após a aplicação do processo de recuperação, para cada grupo (nó da hierarquia) calculam-se os valores da Tabela 3.

Tabela 3. Proporções de documentos recuperados em relação a um grupo.

	recuperados	!recuperados	
(nó n_i)	t_p	f_n	d_c
!(nó n_i)	f_p	t_n	n_c
	r_d	n_r	#d

As medidas de interesse sobre os valores da Tabela 3 que serão utilizadas são:

- **precisão**: a proporção de documentos dentre os recuperados que está na classe correta, $p = t_p/r_d$;
- "***recall***" (revocação): a proporção de documentos dentre os do grupo, que foram corretamente recuperados, $r= t_p/d_c$, também chamada de cobertura de recuperação;
- "***false alarm***" (alarme falso): a proporção dentre os documentos recuperados que, de fato, não pertence ao grupo, $f_a = f_p/n_c$;
- "***F measure***" (medida F): a média harmônica entre precisão e "*recall*", $F = 2*p*r/(p+r)$. O valor ideal de *F* é igual a um, porque o ideal é ter *p=1* e *r=1*.

Particularmente, a medida de maior interesse é a cobertura de recuperação (*recall*), porque ela representa bem a qualidade do conjunto de descritores, quando usado como expressão de busca. Quando os descritores são usados como palavras-chave de busca, ligados pelo operador "*and*", se forem realmente representativos do grupo, o valor de "*recall*" deverá se aproximar de um, isto é, ter-se-ia um número muito baixo de falsos negativos (f_n) e um número de positivos verdadeiros (t_p) bastante alto, próximo ao total de documentos no grupo.

Para verificar experimentalmente uma hipótese sobre o comportamento de uma dessas medidas, em relação a diferentes modelos de descrição de agrupamentos hierárquicos, independentes do algoritmo de agrupamento, calculam-se as medidas para cada um dos nós sobre os resultados de cada método. Tabulam-se essas medidas junto ao número do nó e ao método que as gerou, então se ajusta um modelo linear generalizado para analisar a variância e obter estimativas para cada medida. O modelo deve considerar os efeitos do nó e do método de rotulação, além da média geral:

$$\hat{m}_e = \hat{\mu} + \hat{n}_i + \hat{l}_m + \hat{\varepsilon}$$

$\hat{m}_e$: estimativa da medida de avaliação após ajuste do modelo;
$\hat{\mu}$: média geral do modelo, sem qualquer outro efeito;
$\hat{n}_i$: estimativa do efeito do nó n_i, que é o quanto a medida desvia-se da média geral devido a esse efeito;
$\hat{l}_m$: estimativa do efeito do método de descrição, que é o quanto a medida desvia-se da média geral devido à aplicação de um dado método;
$\hat{\varepsilon}$: efeito aleatório associado a cada estimativa.

Com o uso dessas estimativas para cada medida, para cada método de descrição espera-se obter comparações mais estatisticamente confiáveis. Então, aplicam-se testes de comparações múltiplas de médias a cada modelo, e, também se utilizam essas estimativas para construir gráficos das medidas para cada coleção de textos e método de descrição.

2.1.3 Medida para a variabilidade dos descritores

Independentemente de como os atributos tenham sido obtidos, não há garantias de que todos eles sejam utilizados nos conjuntos finais de descritores. Uma boa maneira de avaliar o quanto eles são utilizados é verificar a diferença entre as cardinalidades do conjunto de atributos e da união de todos os descritores selecionados ao final do processo de descrição dos agrupamentos por algum método.

É esperado que se uma fatia maior do conjunto de atributos inicial tenha sido utilizada na produção dos descritores, então o método utilizado não tenha repetido muito o uso dos mesmos descritores para diferentes nós da hierarquia.

Para verificar o comportamento dessa característica, entre diferentes métodos de descrição de agrupamentos hierárquicos de documentos, seria interessante utilizar diferentes coleções de textos, de diferentes tamanhos, diferentes idiomas ou domínios; e, então, aplicar cada método, calcular as uniões entre os conjuntos de descritores de cada método, calcular as cardinalidades dos conjuntos de atributos iniciais e as cardinalidades das uniões dos descritores. Para realizar uma comparação múltipla de médias, pode-se ajustar o seguinte modelo linear aos dados tabulados para cada coleção de textos:

$$\hat{card}_j = \hat{\mu} + <\hat{card}_A> + \hat{m}_j + \hat{\epsilon}_j$$

Sendo,

$\hat{card}_j$: cardinalidade final inferida para cada união de descritores produzida por cada *j-ésimo* método,

$\hat{\mu}$: a média geral de dispersão estimada ao inferir as cardinalidades;

$<\hat{card}_A>$: estimativa do efeito da cardinalidade do conjunto de atributos inicial, em cada caso, para inferência da cardinalidade final. Essa cardinalidade deve ser colocada no modelo para que o efeito de diferentes números de atributos seja devidamente tratado e considerado nas inferências finais. Desse modo, ao se considerar os efeitos relativos a diferentes métodos de descrição dos agrupamentos, esse efeito do tamanho do conjunto de atributos inicial já tenha sido retirado do modelo de inferência;

$\hat{m}_j$: estimativa do efeito do método de descrição aplicado, isto é, quanto à estimativa final desvia-se da média geral de dispersão em relação a esse efeito;

$\hat{\epsilon}_j$: erro aleatório do ajuste do modelo linear.

2.1.4 Medida para a representatividade dos descritores

Uma maior cardinalidade da união do conjunto de descritores não garante a sua qualidade, pois verificar essa propriedade pode ser apenas reflexo de uma série de más escolhas. Para avaliar as escolhas, que podem ser vistas como quão bem os rótulos selecionados predizem o conteúdo dos documentos em cada grupo, além do uso da medida recall, pode-se utilizar a esperança da informação mútua entre o conjunto de atributos e o vocabulário final. Para a coleção completa de documentos após a rotulação,

a Esperança da Informação Mútua Expected Mutual Information Mean (**EMIM**) (LAWRIE et al., 2001; MANNING; SCHÜTZE, 1999,) é definida como:

$$\text{EMIM}n_i(L_s(n_i),A) = \sum_{u=1}^{\text{card}(\text{Ls}(ni))} \sum_{k=1}^{\text{m}} p(\text{l}_\text{u}, \text{a}_\text{k}) * \log\left(\frac{p(\text{l}_\text{u}, \text{a}_\text{k})}{p(\text{l}_\text{u}) * p(\text{a}_\text{k})}\right)$$

Sendo,

n_i : *i-ésimo* nó da hierarquia, *i*=1,...,N

$Ls(n_i)$: conjunto de descritores do *i-ésimo* nó, L={l1, l2, ..., $l_{card(Ls(ni))}$}, u=1, .., card(Ls_{ni})

A : conjunto inicial de atributos, $A=\{a_1, a_2, ..., a_m\}$, *k*=1, ..., *m*,

Considerando-se cada $p(l_u, a_k)$ como a probabilidade de ocorrerem simultaneamente o descritor l_u no conjunto com a união dos descritores e a_k no conjunto de atributos *A*, e, essa estimativa da probabilidade ser o estimador de máxima verossimilhança dado pelo número de ocorrências de lu e ak em cada documento da coleção. Como definida aqui, a EMIM reflete estatisticamente a qualidade do conjunto total de descritores selecionados por um método de descrição de agrupamentos hierárquicos de documentos. Como esses valores são calculados para os mesmos grupos de documentos, os valores obtidos para a EMIM podem ser comparados utilizando-se comparação múltipla de médias.

Como a EMIM não tem um intervalo de valores definidos ou mesmo um máximo, os maiores valores obtidos refletem as melhores predições. E, embora a EMIM seja sensível ao tamanho dos conjuntos envolvidos, essa condição já foi contornada ao se considerar apenas os nós (grupos) comuns às hierarquias e limitar o tamanho de cada conjunto de rótulos ao máximo de x elementos cada.

Dessa forma, pode-se ajustar um modelo linear para a análise de variância e comparação múltipla de médias, da seguinte forma:

$$\hat{emim}_{ij} = \hat{\mu} + \hat{n}_i + \hat{m}_j + \hat{\epsilon}_{ij}$$

Sendo,

$\hat{emim}_{ij}$: EMIM final inferida para cada nó em cada método;

$\hat{\mu}$: a média geral de dispersão estimada para inferir a EMIM;
$\hat{n}_i$: estimativa do efeito de cada nó ni na inferência da EMIM;
$\hat{m}_j$: estimativa do efeito do método aplicado, isto é, quanto à média geral de dispersão desvia-se da EMIM estimada em relação a esse efeito;
$\hat{\epsilon}_{ij}$: erro aleatório de ajuste do modelo linear.

3 Resultados e discussão

Para aplicar e validar esse tipo de avaliação objetiva de métodos de descrição de agrupamentos hierárquicos de documentos, foram selecionados quatro diferentes métodos, para os quais se dispunha das implementações e havia o interesse na sua avaliação, a saber: Robust Labelling Up Method (RLUM) (MOURA; REZENDE, 2010), Robust Labelling Down Method (RLDM) (MOURA et al., 2008), Popescul e Ungar (2000) e o método dos Mais Freqüentes.

Para se ter uma noção de como esses métodos produzem descritores para um agrupamento hierárquico, pode-se observar o exemplo ilustrado na Figura 2. Nota-se que o método dos Mais Freqüentes repete muito os descritores em mesmos ramos da hierarquia, assim como, nesse exemplo, isso também ocorre com o método de Popescul e Ungar; já os métodos RLDM e RLUM não apresentam esse comportamento.

Para comparar os métodos, o ideal é utilizar várias coleções de textos. Quatro coleções de textos foram selecionadas para os experimentos, cada qual dividida em quatro categorias, de tamanhos diferentes, que foram utilizadas como bases isoladas, logo, dezesseis coleções de textos foram utilizadas nos testes.

As coleções passaram por processos semelhantes de pré-processamento, seguido por processos idênticos de agrupamento hierárquico e depois pelos quatro métodos de descrição. As análises dos resultados foram conduzidas em relação a cada uma das características consideradas: (i) cobertura para a recuperação de informações a partir dos conjuntos de rótulos, (ii) variabilidade dos descritores encontrados para cada método e (iii) repre-

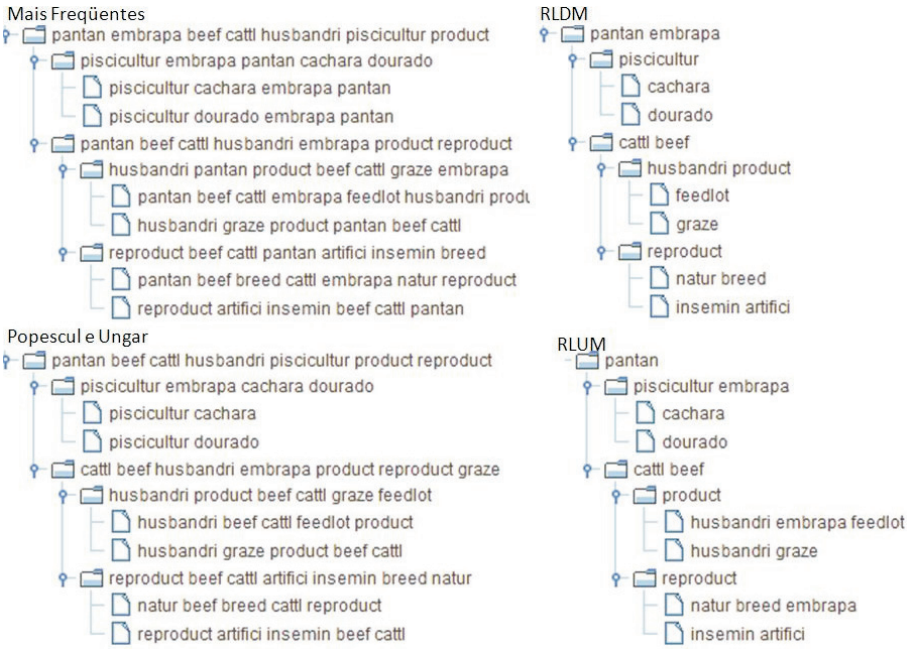

Figura 2. Resultados da descrição dos agrupamentos para quatro diferentes métodos

Fonte: Moura (2009).

sentatividade dos descritores em relação ao conjunto inicial de atributos. As subseções, a seguir, apresentam cada passo dos experimentos.

3.1 Pré-processamento, agrupamento e padronizações

Na Tabela 4, é apresentado, de forma resumida, o resultado do pré-processamento das coleções utilizadas no experimento. A primeira coleção de textos corresponde a artigos completos, em português, da segunda assembléia do Instituto Fábrica do Milênio - IFM[8], que é uma coleção de textos em português. No primeiro momento, essa coleção tinha 614 documentos, dividida em quatro categorias, com 198 documentos na maior classe; as

[8] *O IFM é uma organização brasileira cujas ações são focadas na busca por soluções de necessidades da manufatura nas indústrias; veja: http://www.ifm.org.br/.*

Tabela 4. Resumo do pré-processamento das coleções de textos do experimento.

Coleção	#docs	#stems	Filtro com df	Card(A)
WP01	80	10268	2≤df≤10	4799
WP02	65	11630	2≤df≤9	5000
WP03	198	19904	2≤df≤21	9302
WP04	48	13870	2≤df≤7	5252
Hard	89	11014	2≤df≤10	3453
AI	94	10128	2≤df≤11	4445
HCI	98	8915	2≤df≤12	3971
Sec	127	18780	2≤df≤14	6253
AnaC	100	18408	2≤df≤12	6645
InoC	97	18436	2≤df≤12	8864
OrgC	99	20116	2≤df≤12	5238
PolyC	100	11628	2≤df≤12	5634
BioP	95	13179	2≤df≤12	6187
GeoP	97	12011	2≤df≤12	5100
Mech	117	12660	2≤df≤14	4378
Quan	82	9586	2≤df≤10	3798

classes correspondem a projetos de trabalho do instituto (WP01, WP02, WP03 e WP04). A segunda coleção de textos corresponde a artigos em inglês, 408 documentos, sobre computação, divididos nos temas: inteligência artificial (AI), interface homem máquina (HCI), hardware (Hard) e segurança e criptografia (SEC). A terceira coleção corresponde a 396 artigos sobre química, em inglês, divididos nos temas: orgânica (OrgC), inorgânica (InoC), analítica (AnaC) e ciência dos polímeros (PolyC). A quarta coleção são artigos em inglês sobre física, divididos em biofísica (BioP), geofísica (GeoP), mecânica (Mech) e física quântica (Quan).

Na segunda coluna da Tabela 4, encontram-se os números de documentos utilizados em cada coleção após a conversão dos textos do formato original (PDF) para texto sem formatação (formato txt). Quando esse processo é aplicado, algumas vezes, alguns documentos são descartados, devido à conversão apresentar algum tipo de problema; fica-se apenas com os textos cujo conteúdo possa ser utilizado no processo. A seguir, aplicou-se *stem-*

mização às coleções, obtiveram-se os unigramas (número total em cada coleção na terceira coluna da tabela, *#stems*), com a ferramenta PreText (MATSUBARA et al., 2003). Para simplificar e padronizar o procedimento de filtro, foi utilizado aproximadamente o proposto por Salton et al. (1975) para palavras discriminativas. Esse filtro consiste em usar as palavras que possuem frequência de ocorrência nos documentos (*df: document frequency*) entre um a dez por cento da coleção. Adicionalmente, como as coleções serão submetidas a um processo de agrupamento aglomerativo, usou-se o mínimo de *df=2* e, para compensar a retirada de *df=1*, acresceu-se um ou dois pontos ao corte de 10% da *df*. Após a aplicação desse filtro, na quarta coluna da tabela tem-se a cardinalidade do conjunto A de atributos.

Ainda, para manter uma padronização de resultados, um agrupamento aglomerativo é aplicado a cada coleção, utilizando-se similaridade de cosseno e o algoritmo *average linkage*. Após esse procedimento, são aplicados os métodos de descrição dos agrupamentos hierárquicos de documentos. Como eles geram diferentes conjuntos específicos de descritores, em cada nó, de diferentes cardinalidades, fixou-se o número máximo de descritores por nó em catorze (14) - resultado de uma escolha completamente subjetiva. Para fazer isso, os descritores obtidos em cada nó, para cada método, foram ordenados por frequência de ocorrência, como no método dos mais frequentes. Desse modo, cada método de descrição gera um número qualquer de descritores por nó, mas, no máximo, consideram-se os catorze com maior frequência acumulada de ocorrência no nó. Essa frequência corresponde à soma das frequências do atributo (selecionado como descritor) em todos os documentos da coleção que se encontram sob o nó.

3.2 Medidas de discriminação

Para essa avaliação, consideraram-se os conjuntos de rótulos genéricos e os conjuntos de rótulos específicos separadamente, em diferentes experimentos. A ideia é avaliar a utilização dos conjuntos específicos como expressão de busca, isto é, cada $\boldsymbol{L_s(n_i)}$ como se o grupo fosse resultado de um *flat cluster* e, então, comparar os resultados obtidos por cada método de descrição. A seguir, utilizaram-se os rótulos genéricos $\boldsymbol{L_g(n_i)}$ (união de todos os L_s dos nós ascendentes de n_i) unido ao $L_s(n_i)$. Essa segunda

expressão permite que se avalie a rotulação hierárquica propriamente dita, pois reflete o fato de que os descritores genéricos, descritores dos antecessores, discriminam o nó atual junto ao seu próprio descritor.

Na Tabela 5, é apresentado o resumo dos resultados desse primeiro experimento para a coleção de textos WP01. Nessa Tabela 5, estão presentes os resultados das comparações múltiplas de médias para cada medida de interesse: F, média harmônica de precisão e revocação; *prec*, precisão; r, revocação ou cobertura (*recall*); e, *F_a*, falso alarme. Para obter as tabe-

Tabela 5. Resultados da medida *F*, revocação (*rec*), precisão (*prec*) e alarme falso (Fa) para *Lg* e *Ls* na coleção Wp01.

modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{\hat{e}} = 0, gl = 1$

método	n	F	*grupo*
$MostFreq$	1	0.697248	a
$PopeUngar$	1	0.697248	a
$RLDM$	1	0.367816	b
$\overline{F}_{RLUM}$	30	1.000000	*singular*

$L_s(n_{ij})$, $\sqrt{\hat{e}} = 0.0282, gl = 60$

método	n	F	*grupo*
$RLDM$	30	0.807975	a
$RLUM$	30	0.701492	b
$PopeUngar$	17	0.401681	c
$MostFreq$	17	0.401681	c

modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{\hat{e}} = 0.0179, gl = 64$

método	n	*prec*	*grupo*
$RLUM$	30	0.590959	a
$RLDM$	6	0.166667	b
$PopeUngar$	31	0.032258	c
$MostFreq$	31	0.032258	c

$L_s(n_{ij})$, $\sqrt{\hat{e}} = 0.0371, gl = 89$

método	n	*prec*	*grupo*
$RLDM$	31	0.801459	a
$RLUM$	30	0.579107	b
$PopeUngar$	31	0.197366	c
$MostFreq$	31	0.197366	c

modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{\hat{e}} = 0.0007, df = 89$

método	n	*rec*	*grupo*
$RLUM$	30	1.000000	a
$PopeUngar$	31	0.017265	b
$MostFreq$	31	0.017265	b
$RLDM$	31	0.007269	b

$L_s(n_{ij})$, $\sqrt{\hat{e}} = 0.0541, df = 89$

método	n	*rec*	*grupo*
$RLUM$	30	1.000000	a
$RLDM$	31	0.848667	b
$PopeUngar$	31	0.326405	c
$MostFreq$	31	0.326405	c

modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{\hat{e}} = 0.0042, df = 89$

método	n	F_a	*grupo*
$MostFreq$	31	0.064665	a
$PopeUngar$	31	0.064665	a
$RLUM$	30	0.056270	a
$RLDM$	31	0.003314	b

$L_s(n_{ij})$, $\sqrt{\hat{e}} = 0.0041, df = 89$

método	n	F_a	*grupo*
$MostFreq$	31	0.065600	a
$PopeUngar$	31	0.065600	a
$RLUM$	30	0.059169	a
$RLDM$	31	0.009922	b

las foi ajustado um modelo linear generalizado, como explicado na seção 1.1.2. Os mesmos cálculos e avaliações foram realizados para as dezesseis coleções de textos; e, o comportamento observado para as medidas nas dezesseis coleções foi semelhante.

Deve-se notar na Tabela 5 que, para algumas medidas, como a F e a precisão (*prec*) não há o mesmo número de pontos calculados, para cada nó (n_i, na tabela); o que ocorre devido à alguma divisão por zero. Ainda, na Tabela, 5 para a expressão de busca *Lg(ni) U Ls(ni)*, a medida F do RLUM não pode ser comparada às demais, por insuficiência de pontos nos resultados dos demais métodos. Porém, nota-se que o resultado de F para RLUM foi *1*, nos trinta pontos totais, ou seja, o resultado atingiu o valor ideal. Nota-se, na Tabela 5, que o RLUM só perde para o RLDM na precisão, quando utilizada a expressão de busca $L_s(n_i)$, mas em todos os demais casos seu desempenho é melhor que dos outros três. Quanto ao falso alarme, o comportamento do RLUM é sempre aceitável, comparável ao outros; porém o RLDM erra mais.

Outros resultados interessantes referem-se, por exemplo, aos métodos dos mais frequentes e de Popescul e Ungar, que sempre apresentam um comportamento semelhante. O método RLDM tem um melhor desempenho, com respeito à *precisão* e *F measure*, quando se usam os conjuntos de rótulos específicos como expressão de busca, obtendo-se resultados melhores que os demais métodos, enquanto o método RLUM sempre tem um excelente desempenho no valor de *recall*, próximo a um, um bom índice da *F measure* e um falso alarme dentro do esperado. Seu desempenho é sempre superior ou, no mínimo, equivalente aos demais, quando essas medidas são boas para os demais e a expressão de busca *Lg(ni) U Ls(ni)* é usada; logo, estatisticamente, ele obtém os conjuntos de rótulos mais representativos quando toda a hierarquia é considerada.

Reforçando essas observações, os resultados das medidas de revocação (*recall*), F e precisão (*precision*) para as dezesseis coleções de textos estão resumidos nos gráficos ilustrados nas Figuras 3, 4 e 5. Para esses gráficos, foram utilizadas as médias obtidas pelos modelos lineares ajustados aos dados, pois é a única média que pode ser comparada de forma justa, por terem sido descontados os demais efeitos que podem interferir na medida (veja seção 1.1.1).

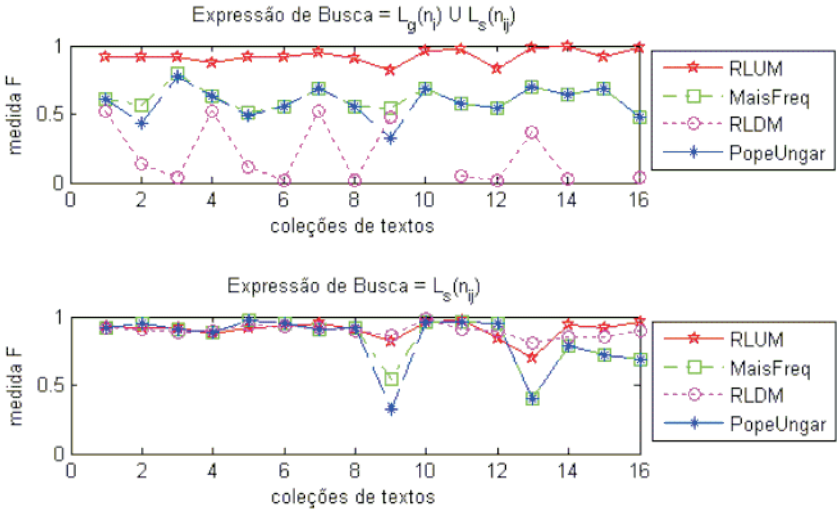

Figura 3. Gráficos das médias da medida F para as dezesseis coleções de textos, utilizando descritores específicos ou genéricos como expressão de busca.

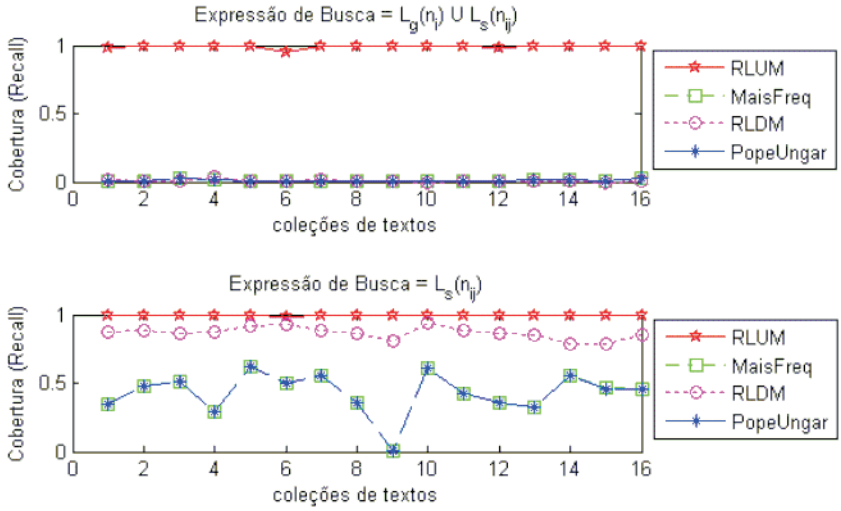

Figura 4. Gráficos das médias da medida *recall* para as dezesseis coleções de textos, utilizando descritores específicos ou genéricos como expressão de busca.

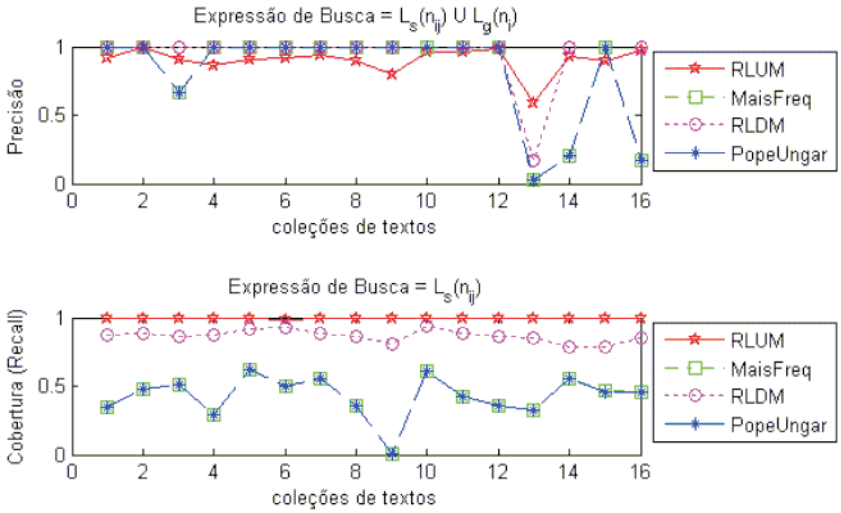

Figura 5. Gráficos das médias da medida precisão para as dezesseis coleções de textos, utilizando descritores específicos ou genéricos como expressão de busca.

Observando-se os gráficos da Figura 3, nota-se que ao utilizar os conjuntos de descritores específicos como expressão de busca, isto é, considerando cada grupo como um *flat cluster*, os métodos RLDM e o RLUM têm um bom desempenho, nesses casos, os dois métodos têm seus valores de F próximos a um. Porém, nos mesmos casos, os métodos de Popescul e Ungar e o Mais Frequentes apresentam uma grande variação da F para algumas coleções de textos. Por outro lado, usando a expressão de busca como a união dos conjuntos de descritores dos anos ascendentes e de um nó específico, observa-se que o RLUM conserva o mesmo comportamento para a F; o RLDM apresenta um grande decréscimo da F. O Popescul e Ungar e o Mais Frequentes novamente apresentam comportamento similar, embora também apresentem um decréscimo da F. Tomando apenas essa medida de interesse (a F), tem-se uma boa evidência de que o RLUM produza os descritores mais discriminativos.

Os métodos RLDM, Popescul e Ungar e Mais Freqüentes não apresentaram resultados tão ruins para a medida F devido ao comportamento da precisão, de acordo com os resultados apresentados na Figura 5. Embora a precisão para esses métodos pareça ser melhor que para o RLUM,

exceto em algumas das coleções de textos, isso não mostra que produzam conjuntos de descritores mais significativos. A precisão pode ser boa mesmo quando a cobertura (*recall*) apresenta-se ruim, pois recuperar um único documento correto significa ter uma precisão de cem por cento, por exemplo, se o total de documentos no grupo tiver sido 100, recuperou-se apenas um por cento do grupo.

3.3 Medida de variabilidade

Na Tabela 6, são apresentadas as cardinalidades calculadas para os conjuntos de atributos de cada coleção e depois para as uniões dos conjuntos de descritores selecionados para método de descrição de agrupamentos hierárquicos de documentos, com, no máximo, 14 descritores por nó. Nota-se, apenas observando a tabela, que o RLUM utiliza um número mais variado de atributos na seleção dos descritores, o que é uma evidência de que repita menos os descritores ao longo da hierarquia.

Tabela 6. Cardinalidades do conjunto de atributos e descritores selecionados.

Coleções	Cardinalidade do conjunto de atributos	Popescul e Ungar	Mais Frequentes	RLDM	RLUM
WP01	4799	645	645	488	927
WP02	5000	566	566	690	867
WP03	9302	1383	1384	1595	2245
Wp04	5252	495	495	558	739
Hard	3453	836	824	478	1000
AI	4445	747	746	665	928
HCI	3971	808	801	597	937
Sec	6253	1116	1116	766	1359
AnaC	6645	915	914	838	1244
InoC	8864	803	798	974	1120
OrgC	5238	1051	1030	512	1309
PolyC	5634	786	784	757	967
BioP	6187	759	754	799	977
GeoP	5100	825	824	793	1032
Mech	4378	937	937	691	1179
Quan	3798	719	716	607	912

Após o ajuste do modelo linear generalizado (veja seção 1.1.3) e realizada a comparação múltipla de médias das cardinalidades, obteve-se o resultado apresentado na Tabela 7. Pode-se notar que houve o mesmo número de observações para cada método isto é, a cardinalidade foi observada e considerada para as dezesseis coleções de textos. A cardinalidade média para cada método foi maior para o RLUM, com 0.95 de certeza. Como esperado, as cardinalidades do Popescul e Ungar e dos Mais Frequentes são estatisticamente iguais. Porém, estatisticamente, já não se observa que o RLDM apresente uma cardinalidade mais baixa que os demais, ou seja, ele pode ser considerado como obtendo um vocabulário tão variado quanto os métodos de Popescul e Ungar e o dos Mais Freqüentes.

Tabela 7. Teste SNK para a cardinalidade final, com rejeição=0.05, 59 gruas de liberdade e variância média de 42482.62.

Método	#observações	média	grupo
RLUM	16	1108.875000	a
Popescul e Ungar	16	836.937500	b
Mais Frequentes	16	833.375000	b
RLDM	16	738.000000	b

3.4 Medida de representatividade

A cardinalidade de um método ser superior à dos demais não garante que a qualidade do vocabulário produzido seja boa, pois ele pode ser resultado de uma série de más seleções. Embora a medida de revocação (*recall*) já indique se as escolhas foram ou não boas para a seleção dos conjuntos de descritores de cada grupo. A medida EMIM é mais um teste para evidenciar se a qualidade dos descritores é ou não boa.

Para verificar a diferença entre as medidas de EMIM, em cada nó, para cada método de descrição dos agrupamentos, calcula-se a EMIM em cada nó e depois se ajusta o modelo linear da seção 1.1.4. Realiza-se a análise de variância e a comparação múltipla das médias com o teste SNK, com 0.95 de certeza. Na Tabela 8, encontram-se os resultados obtidos para essa comparação, a partir das coleções de textos do domínio de conhecimento de computação. Os demais resultados para as outras bases repetem um comportamento muito similar a estes.

Tabela 8. Comparação múltipla de médias para as EMIM´s calculadas em quatro das coleções de textos

Col. de Textos AI $gl = 147, var.media = 160140.0445$				*Col. de Textos HCI* $gl = 108, var.media = 286839.4677$			
método	*n*	*emim*	*grupo*	*método*	*n*	*emim*	*grupo*
RLUM	50	1089.899965	*a*	*RLUM*	37	885.627493	*a*
PopeUngar	50	893.817625	*b*	*PopeUngar*	37	765.692200	*a*
MostFreq	50	893.742870	*b*	*MostFreq*	37	762.961843	*a*
RLDM	50	557.042183	*c*	*RLDM*	37	373.175243	*b*
Col. de Textos Hard $gl = 36, var.media = 185914.7144$				*Col. de Textos Sec* $gl = 159, var.media = 487370.7896$			
método	*n*	*emim*	*grupo*	*método*	*n*	*emim*	*grupo*
RLUM	13	1104.943006	*a*	*RLUM*	54	1613.500940	*a*
PopeUngar	13	909.675011	*a*	*PopeUngar*	54	1337.548719	*a*
MostFreq	13	906.329243	*a*	*MostFreq*	54	1337.383964	*a*
RLDM	13	77.376233	*b*	*RLDM*	54	636.411828	*b*

Deve-se observar, na Tabela 8, que todos os valores são calculados, tem-se sempre o mesmo número de grupos (ou nós), *n*, para cada média calculada. Percebe-se também que o agrupamento esperado de médias se confirma nas quatro coleções de textos, ou seja, a hipótese de qualidade dos descritores, de acordo com a EMIM, é confirmada com 0.95 de certeza. Assim, além do RLUM apresentar a maior variabilidade de termos no vocabulário produzido, ele também garante a melhor qualidade desses termos, em relação aos três métodos comparados. Também se deve observar que os métodos de Popescul e Ungar e dos Mais Frequentes encontram-se nos mesmos grupos, o que mostra que seus resultados são sempre muito próximos. Ainda, o RLDM tende a produzir um conjunto de descritores mais pobre, em termos de cardinalidade e de qualidade.

4 Conclusões e trabalhos futuros

Neste trabalho foram propostas medidas objetivas para a comparação de métodos automáticos de descrição e agrupamentos hierárquicos de documentos, independentes dos algoritmos de agrupamentos utilizados. Visavam-se métodos simples e eficientes, para comparar as descrições quanto a não repetição de descritores ao longo dos mesmos ramos da hie-

rarquia, maior variabilidade de escolhas dos descritores e alguma medida de qualidade complementar a essas.

Assim, a melhor discriminação dos conjuntos de descritores foi entendida como a não repetição excessiva dos mesmos descritores ao longo da hierarquia e a composição dos descritores dos nós antecedentes com o descritor do mesmo nó como uma descrição genérica ou completa desse nó. Para verificar essa propriedade, os conjuntos de descritores em nós da hierarquia foram utilizados como expressões de busca sobre as coleções de textos, a fim de recuperar os textos agrupados sob o nó. Nesse quesito, verificou-se especialmente a cobertura (*recall*) dos resultados.

Para verificar a ordem de grandeza dos conjuntos de descritores produzidos, ou seja, o quão variadas (não repetitivas) foram as escolhas, propôs-se comparar as cardinalidades dos vocabulários produzidos por diferentes métodos de descrição de agrupamentos. E, para testar a qualidade das escolhas de descritores, foi proposto o uso da medida da Esperança da Informação Mútua (EMIM) entre os descritores selecionados e o conjunto de atributos inicial, após a padronização dos conjuntos de dados, pois a EMIM é uma medida influenciável pela grandeza das amostras utilizadas.

Os experimentos mostraram que o uso combinado dessas medidas, com uma comparação justa, na qual os demais efeitos que influenciem os resultados sejam considerados ao se calcular as variâncias (por meio de um modelo linear generalizado), leva a resultados estatisticamente confiáveis. Ou seja, de fato, a metodologia proposta auxilia na escolha de um modelo de descrição adequado a agrupamentos hierárquicos de documentos, desde que os modelos partam de princípios semelhantes, como a não dependência do algoritmo de agrupamento.

Seria interessante, como trabalho futuro, complementar essa metodologia com um processo de comparação entre os descritores produzidos e os descritores fornecidos por uma taxonomia de tópicos padrão (ou *gold*, como referenciada por alguns autores), isto é, uma taxonomia pré-existente e já bem aceita, como realizado por outros autores (KASHYAP et al., 2005; KRISHNAPURAM; KUMMAMURU, 2003). Espera-se com esse tipo de validação conseguir simular uma validação subjetiva exclusivamente do método de descrição dos grupos, complementando a metodologia atual.

5 Referências

BISHOP, Y. M. M.; FIENBERG, S. E.; HOLLAND, P. W. **Discrete multivariate analysis**: theory and practice. Cambridge; London: MIT Press, 1975. 557 p.

EMBRAPA. **Agência de Informação Embrapa**. [Home page]. Disponível em: <http://www.agencia.cnptia.embrapa.br/>. Acesso em: 02 nov. 2010.

EVERITT, B. S.; LANDAU, S.; LEESE, M. **Cluster analysis**. 4th ed. London: Oxford: University Press, 2001. 237 p.

FELDMAN, R.; SANGER, J. **The text mining hand book**: advanced approaches in analysing unstructured data. Cambridge: University Press. 2007. 410 p.

GANTZ, J. F.; REINSEL, D. **As the economy contracts, the digital universe expands**: Framingham, MA: IDC Analyze the Future, 2009. 10 p. External publication of IDC (Analyse the Future) information and data.

GLOVER, E.; LAWRENCE, D. M. P. S.; E. KROVETZ, R. Inferring hierarchical descriptions. In: NTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 11., **Proceedings**... New York: ACM, 2002. p. 507-514. ICIKM, 2002.

KASHYAP, V.; RAMAKRISHNAN, C.; THOMAS, C.; SHETH A. TaxaMiner: an experimentation framework for automated taxonomy bootstrapping. **International Journal of Web and Grid Services**, v. 1, n. 2, p. 240-266, 2005.

KONCHADY, M. Text mining application programming. Boston: Charles River Media, 2006. 412 p. (Charles River Media programming series).

KRISHNAPURAM, R. E.; KUMMAMURU, K. Automatic taxonomy generation: Issues and possibilities. In: INTERNATIONAL FUZZY SETS AND SYSTEMS, 10., 2003, Istanbul, **Proceedings**… New York: Springer-Verlag, 2003. p. 52- 63. (Lecture notes in computer science, v. 2715). IFSA 2003.

LAMIREL J-C.; TA, A. P.; ATTIK, M. Novel labeling strategies for hierarchical representation of multidimensional data Analysis. In: NTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND APPLICATIONS, 26., 2008. Proceedings... ACM, 2008. DOI: http://doi.acm.org/10.1145/223904.223956.

LAWRIE, D.; CROFT, W. B.; ROSENBERG, A. Finding topic words for hierarchical summarization. In: ANNUAL INTERNATIONAL CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 24., 2001, New York. **Proceedings**... New York : ACM, 2001. p. 349-357. SIGIR.

MAEDCHE, A. D. **Ontology learning for the semantic web**. Boston: Kluwer Academic Publishers, 2002. 244 p. (Kluwer international series in engineering and computer science, 665).

MANNING, C. D.; SCHÜTZE, H. **Foundations of statistical natural language processing**. Cambridge: MIT Press, 1999. 680 p.

MATSUBARA, E. T.; MARTINS, C. A.; MONARD, M. C. **Pre-TexT**: uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. São Carlos, SP: USP, São Carlos, SP, Instituto de Ciências Matemáticas e de Computação, 2003. Relatório técnico 209.

MORETTIN P. A.; BUSSAB, W. O. **Estatística básica**. São Paulo: Saraiva, 2006. 526 p.

MOURA, M. F.; MACACINI, R. M.; REZENDE, S. O. Easily labelling hierarchical document clusters. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 23.; SIMPÓSIO BRASILEIRO DE ENGENHARIA DE SOFTWARE, 22.; WORKSHOP EM ALGORITMOS E APLICAÇÕES DE MINERAÇÃO DE DADOS, 4., 2008, Campinas. **Anais**... Campinas: Unicamp, Instituto de Computação, 2008, p. 37-45.

MOURA, M. F. **Contribuições para a construção de taxonomias de tópicos em domínios restritos utilizando aprendizado estatístico**. 2009. 137 f. Tese (Doutorado) em Ciências de Computação e Matemática Computacional, Instituto de Ciências Matemáticas e de Computação, USP, São Carlos, SP, 2009. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-05042010-162834/>. Acesso em: 15 nov. 2010.

MOURA, M. F.; REZENDE, S. O. A simple method for labeling hierarchical document clusters. In: IASTED INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND APPLICATIONS, 10., 2010, Innsbruck. **Proceedings**... Anaheim: ACTA Press, 2010. p. 363-371. AIA 2010.

POPESCUL, A. E.; UNGAR, L. **Automatic labeling of document clusters**. 2000. Disponível em: <http://citeseer.nj.nec.com/popescul100automatic.html>. Acesso em: 15 nov. 2010.

RAMSEY, P. H. Multiple comparisons of independent means. In: EDWARDS, L. K. (Ed.). **Applied analysis of variance in behavioral science**. New York: Marcel Dekker, 1993.

SALTON, G.; YANG, C. S.; E YU, C. T. A theory of term importance in automatic text analysis. **Journal of the American Association Science**, v. 1, n. 26, p.33-44, 1975.

SEARLE, S. R. **Linear models**. New York : J. Wiley, 1971. 532 p.

SNEDECOR, G. W.; COCHRAN, W. G. **Statistical methods**, 6th ed. Iowa State: University Press, 1967. 272 p.

VIVÍSIMO. **Vivísimo clustering - automatic categorization and meta-search software**. 2006. Disponível em: <http://vivisimo.com/>. Acesso em: 02 nov. 2010.

WEISS, S. M.; INDURKHYA, N.; ZHANG, T.; DAMERAU, F. J. **Text mining**: predictive methods for analizing unstructured information. New York : Springer, 2005. 237 p.

ZHANG, C.; WANG, H.; LIU, Y.; XU, H. Document clustering description extraction and its application. In: INTERNATIONAL CONFERENCE ON THE COMPUTER PROCESSING OF ORIENTAL LANGUAGES, 22., 2009, Hong Kong. **Proceedings**... Berlin, Heidelber: Springer-Verlag 2009. p. 370-377. (Lecture notes in computer science, v. 5459). ICCPOL 2009.

Embrapa

Informática Agropecuária

Ministério da
Agricultura, Pecuária
e Abastecimento

BRASIL
UM PAÍS DE TODOS
GOVERNO FEDERAL

CGPE 9154