# Uma Abordagem para Seleção de Grupos Significativos em Agrupamento Hierárquico de Documentos

**Ricardo M. Marcacini**[1], **Maria F. Moura**[2], **Solange O. Rezende**[1]

[1]Instituto de Ciências Matemáticas e de Computação – ICMC
Universidade de São Paulo – USP
Caixa Postal: 668 – CEP: 13560-970 – São Carlos – SP

[2]Embrapa Informática Agropecuária
Caixa Postal: 6041 – CEP: 13083-970 – Campinas – SP

marcacini@grad.icmc.usp.br, mnanda@icmc.usp.br, solange@icmc.usp.br

***Abstract.*** *Hierarchical document clustering usually generates many cluster and subclusters, making the analysis and interpretation of the results difficult. In this paper an approach to obtain a reduced hierarchy of documents from the original hierarchies is presented, selecting only significant clusters. The selection is supported by quality measures of cluster, adapted to the high dimensionality of textual data and by considering the hierarchical relation among the clusters. An experimental evaluation was carried out through 10 textual collections and three different hierarchical clustering algorithms; which presented good results.*

***Resumo.*** *O agrupamento hierárquico de documentos geralmente fornece muitos grupos e subgrupos, dificultando a análise e interpretação dos resultados. Neste trabalho é apresentado uma abordagem para obtenção de hierarquias de documentos reduzidas, a partir das hierarquias originais, selecionando-se apenas os grupos mais significantes. A seleção é apoiada por medidas de qualidade de grupos, adaptadas para a alta dimensionalidade de dados textuais e para considerar o relacionamento hierárquico entre os grupos. Uma avaliação experimental foi realizada em 10 coleções de documentos e três diferentes algoritmos de agrupamento hierárquico; apresentando bons resultados.*

## 1. Introdução

Uma forma eficiente para gerenciar o conhecimento implícito em coleções textuais é por meio de agrupamento hierárquico de documentos. Uma estrutura hierárquica permite a visualização e exploração intuitiva dos dados em diferentes níveis de abstração, permitindo que o usuário analise, de forma interativa, grandes coleções de documentos. Além disso, uma organização hierárquica satisfaz à premissa de que se um usuário está interessado em um documento específico pertencente a um grupo deve também estar interessado em outros documentos desse grupo ou de seus sub-grupos.

Os algoritmos de agrupamento hierárquico são classificados em aglomerativos (*botton-up*) [Sneath and Sokal 1973, Guha et al. 1998] ou divisivos (*top-down*) [Jain and Dubes 1988, Boley 1998]. Tanto os algoritmos divisivos como aglomerativos, constroem uma estrutura que descreve uma hierarquia de agrupamentos sobre os dados. O *dendrograma* é a estrutura mais freqüentemente utilizada para representar essa hierarquia,

e consiste de um tipo especial de árvore, na qual os nós pais agrupam os exemplos representados pelos nós filhos. O *dendrograma* não é apenas um conjunto de agrupamentos, mas uma estrutura com toda a hierarquia dos agrupamentos gerados sobre um conjunto de dados. Na Figura 1, é ilustrado um exemplo de *dendrograma*, em que uma hierarquia significante está indicada pelas linhas tracejadas.

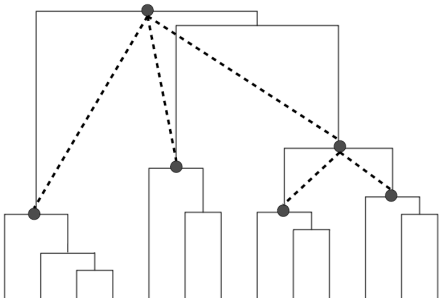

**Figura 1. Exemplo de *Dendrograma* e a Seleção de Grupos Significativos.**

A análise, interpretação e seleção de grupos significativos de um *dendrograma*, em geral, é um grande desafio para os usuários. Quando o conjunto de dados é pequeno, uma simples inspeção visual dos resultados é suficiente para a análise. No entanto, à medida que o conjunto de dados aumenta, fica inviável realizar este processo de forma subjetiva, sendo necessário o uso de técnicas que automatizam a seleção dos grupos significantes.

O objetivo da abordagem apresentada neste trabalho é obter, a partir de um *dendrograma*, uma hierarquia reduzida e com grupos coesos, facilitando sua exploração e interpretação, além de permitir que os subgrupos também sejam visualizados, sem perda de informação. Para isto, é realizada a seleção de grupos significativos de um *dendrograma* baseada em medidas de qualidade de agrupamento. Foram definidas e comparadas três medidas de qualidade, adaptadas para a alta dimensionalidade encontrada em dados textuais, e de forma a considerar também a qualidade das relações hierárquicas entre os grupos.

Assim, na Seção 2 são apresentados os trabalhos relacionados. A descrição da abordagem utilizada para seleção de grupos significantes em *dendrogramas* é feita na Seção 3. Na Seção 4 é apresentada uma avaliação experimental e a discussão dos resultados. Finalmente, as considerações finais são apresentadas na Seção 5.

## 2. Trabalhos Relacionados

As técnicas para apoiar a seleção de grupos significantes, ou "verdadeiros", a partir dos resultados de algoritmos de agrupamento, são chamadas de técnicas de validação de agrupamentos [Jain and Dubes 1988, Halkidi et al. 2002a, Halkidi et al. 2002b]. Esta validação, em geral, é feita com base em medidas estatísticas, que expressa alguma informação a respeito da qualidade do agrupamento. Estimar a qualidade dos grupos é importante, pois um algoritmo de agrupamento sempre encontrará grupos em um conjunto de dados, independentemente dos conceitos por eles representados.

Espera-se que bons grupos sejam compactos de modo que seus elementos apresentem alta similaridade (medida *intra-grupo*), enquanto que a similaridade com os elementos de outros clusters seja a menor possível (medida *inter-grupo*). A maioria das

abordagens para obter a qualidade de um agrupamento, utiliza técnicas baseadas nas medidas *intra-grupo* e *inter-grupo*, adaptando-as para alguma necessidade específica. No geral, estas ténicas procuram obter partições, que dividem a coleção de dados original, ignorando-se a relação hierárquica entre os grupos. Além disso, as medidas utilizadas geralmente são influenciadas na presença de alta dimensionalidade. Por exemplo, [Milligan and Cooper 1985] avaliaram trinta métodos de seleção de grupos significantes. Essa avaliação foi baseada em conjuntos de dados pequenos e bem separados, e as medidas melhores avaliadas são apropriadas apenas para agrupamentos compactos, segundo [Sarle and Kuo 1993]. Além disso, o objetivo das medidas avaliadas era selecionar um nível do *dendrograma* que contém grupos significantes, extraí-los, e considerá-los como partições, perdendo-se a relação hierárquica entre os grupos [Everitt et al. 2001].

No *TaxaMiner Framework* [Kashyap et al. 2005], os grupos significantes são selecionados a partir de uma medida *intra-grupo* durante a própria execução do algoritmo de agrupamento, obtendo uma hierarquia de grupos, porém não se considera a qualidade da ligação hierárquica entre os grupos. Além disso, a técnica é dependente do algoritmo de agrupamento utilizado. Um método semelhante é descrito em [Fung et al. 2003], utilizando uma medida *inter-grupo* como critério de seleção. Já no *TaxGen* [Müller et al. 1999], da *IBM*, os grupos são selecionados a partir de valores predefinidos de profundidade máxima da hierarquia e quantidade de documentos por grupo, visando uma árvore de fácil navegação para o usuário. Outros utilizam aprendizagem semi-supervisionada, ou seja, é necessário que o usuário informe a estrutura desejada, e a partir disto o restante da hierarquia é construído [Huang et al. 2006].

O diferencial da abordagem apresentada neste trabalho é a seleção de grupos significantes de um *dendrograma*, obtido por qualquer algoritmo de agrupamento hierárquico, por meio de medidas de qualidade definidas para obterem melhor desempenho em um ambiente de alta dimensionalidade, considerando também as relações hierárquicas entre os grupos. Além disso, uma vez calculadas as medidas para um *dendrograma*, o usuário pode explorar os resultados em diferentes níveis de detalhamento, apenas configurando os valores das medidas de qualidade, sempre que necessário, e sem a necessidade de executar novamente o agrupamento da coleção, que é uma tarefa de alto custo computacional.

## 3. Seleção de Grupos Significativos

Neste trabalho, é utilizado o modelo espaço-vetorial para representação da coleção textual, em que cada documento *d* é expresso como um vetor de atributos (termos ou palavras), e cada posição do vetor possui um valor relacionado ao atributo, como, a freqüência de ocorrência no documento. Então, calcula-se a similaridade entre os documentos.

A similaridade de cosseno é indicada, pois sofre pouca influência na presença de alta dimensionalidade, comum quando se trata de coleções textuais [Feldman and Sanger 2006]. O valor da similaridade cosseno fica no intervalo [0,1]; quanto mais próximo de 1 se encontra o valor, mais similares são os dois documentos. Assim, a similaridade entre dois documentos, $d_i$ e $d_j$, pode ser calculada como:

$$cos(d_i, d_j) = \frac{{d_i}^t d_j}{\|d_i\| \, \|d_j\|} \tag{1}$$

Um grupo de documentos também possui representação no espaço-vetorial. Seja um grupo *G*, com um total de *k* documentos *d*, o *centróide* é um vetor representante do

grupo *G* e é definido como:

$$\textit{centróide(G)} = \frac{\sum_{i=1}^{k} d_i}{k} \tag{2}$$

Para obter a medida *intra-grupo* em um determinado grupo *G*, com *k* documentos *d*, neste trabalho, calcula-se a variância entre as similaridades cosseno de cada documento com o centróide de *G*, conforme a seguinte expressão:

$$\textit{intragrupo(G)} = 1 - var(cos(d_1, \textit{centróide(G)}), ..., cos(d_k, \textit{centróide(G)})) \tag{3}$$

Considera-se um grupo significante quando a medida *intra-grupo* se aproxima de 1, pois indica que os documentos de um mesmo grupo são similares, com baixa variabilidade. Ao aproximar-se de 0, o grupo é considerado não significante.

A medida *inter-grupo* de um grupo *G*, é definida, neste trabalho, como o valor da similaridade cosseno entre o centróide de *G* e o centróide do pai de *G* na hierarquia, de acordo com a expressão:

$$\textit{intergrupo(G)} = 1 - cos(\textit{centróide(G)}, \textit{centróide(pai(G))}) \tag{4}$$

Neste caso, um valor próximo de 0 para *inter-grupo* indica alta similaridade entre o grupo pai e filho e, portanto, não é significante, uma vez que o filho não especializa a informação contida no grupo pai. Por outro lado, um valor próximo de 1 indica que a informação entre o grupo pai e filho estão bem divididas e o grupo é significativo.

Em [Lin and Chen 2005], é proposta uma medida de similaridade baseada no conceito de "*joinability*", vista como a "intenção" de um exemplo pertencer a um determinado grupo. A seguir, é realizada uma adaptação deste conceito no cenário de agrupamentos hierárquicos, e uma nova medida, nomeada *fusão*, é definida. Dado um grupo *G*, com *k* documentos *d*, calcula-se o *raio* de *G* em relação ao centróide do pai de *G*:

$$\textit{raio(G)} = \sqrt{\frac{\sum_{i=1}^{k} cos(d_i, \textit{centróide(pai(G))})^2}{k}} \tag{5}$$

Assim, a medida de fusão pode ser obtida com a expressão:

$$\textit{fusão(G)} = \exp\left(-\frac{\textit{intergrupo(G)}}{raio(G)}\right) \tag{6}$$

A medida de fusão combina os conceito *intra-grupo*, obtido pelo *raio*, e o *inter-grupo*. Sua principal característica é possuir grande tolerância a ruídos e alta dimensionalidade dos dados. Quando o valor da *fusão* de um grupo *G* se aproxima de 0, indica que não é significante na hierarquia; caso contrário o valor da *fusão* se aproxima de 1.

Definidas as medidas *intra-grupo*, *inter-grupo* e *fusão*, estas são calculadas para todos os grupos de um *dendrograma*. Como existem diferenças relativamente grandes no número de documentos de cada grupo, os valores obtidos em cada medida são normalizadas por meio da *z-score*. Na seqüência, é realizada uma varredura no *dendrograma*, verificando se o grupo visitado satisfaz a um valor definido pelo o usuário. Se o grupo não for aceito, ele é removido e seus grupos filhos são promovidos ao parente mais próximo acima da hierarquia. No final deste processo, uma hierarquia reduzida é obtida, contendo apenas os grupos selecionados. O cálculo das medidas é realizado apenas uma vez por

*dendrograma*. Assim, o processo de seleção de grupos significativos pode ser repetido sempre que necessário, variando-se as medidas de qualidade de acordo com as necessidades do usuário, já que o processo tem custo computacional linear em relação ao número de grupos do *dendrograma*.

É importante observar que, geralmente, o *dendrograma* construído pelo algoritmo de agrupamento já informa algumas destas medidas. A vantagem de recalculá-las é que a validação realiza-se por critérios diferentes dos utilizados pelo algoritmo, ou seja, o critério do algoritmo não fica superestimado.

## 4. Experimentos e Resultados

A abordagem apresentada neste trabalho foi avaliada experimentalmente sobre 10 coleções textuais de diferentes características. A menor coleção possui 390 documentos e a maior coleção 4069 documentos. Cada coleção foi pré-processada conforme processo detalhado em [Nogueira et al. 2008], com remoção de *stopwords* e realização de *stemming*. Foram considerados apenas os atributos que ocorrem em dois ou mais documentos. Os detalhes das coleções textuais estão apresentados na Tabela 1.

| Coleção | Origem | N° Documentos | N° Atributos | N° Classes |
|---|---|---|---|---|
| 20ng | UCI | 2000 | 12376 | 20 |
| comp | ACM | 408 | 14895 | 4 |
| k1b | WebACE | 2340 | 13457 | 6 |
| la2 | TREC | 3075 | 14540 | 6 |
| nsf1 | UCI | 559 | 3295 | 21 |
| physics | ACM | 390 | 16265 | 4 |
| re0 | UCI | 1504 | 2708 | 13 |
| reviews | TREC | 4069 | 22511 | 5 |
| tr41 | TREC | 878 | 7027 | 10 |
| wap | WebACE | 1560 | 8060 | 20 |

**Tabela 1. Detalhes das coleções textuais utilizadas na avaliação experimental.**

As coleções *20ng*, *nsf1* e *re0* foram obtidas do repositório da *UCI KDD - http://kdd.ics.uci.edu*. A primeira consiste de 2000 mensagens eletrônicas distribuídas em 20 grupos de discussão. A segunda é uma coleção de resumos sobre resultados de pesquisa básica em 21 países, mantida pela *National Science Foundation - http://www.nsf.gov*. A terceira pertence à base *Reuters-21578*, utilizada para avaliar categorização de documentos. Artigos completos das áreas de computação e física foram obtidos do repositório da *ACM - http://portal.acm.org*, originando-se as coleções *comp* e *physics*, respectivamente. As coleções *k1b* e *wap* foram disponibilizadas pelo projeto Webace [Han et al. 1998], e consistem de documentos do *Yahoo! Directory - http://dir.yahoo.com/*. As coleções *la2*, *reviews* e *tr41* pertencem ao repositório da *Text Retrieval Conference* (TREC) - *http://trec.nist.gov*, e tratam de artigos jornalísticos de diversas áreas.

Todas as coleções estão divididas em classes predefinidas. Desta forma, é possível quantificar a qualidade do agrupamento hierárquico, verificando o quanto as classes podem ser representadas pelos grupos existentes no *dendrograma*.

### 4.1. Critérios para Avaliação do Experimento

As coleções textuais foram submetidas a três tradicionais algoritmos de agrupamento hierárquico: *Average Linkage*, *Complete Linkage* e *Ward* [Zhao and Karypis 2002, El-Hamdouchi and Willett 1986]. Os resultados do agrupamento foram avaliados pela medida *FScore*, um critério de qualidade de agrupamento proposto em [Larsen and Aone 1999], que utiliza o conhecimento prévio sobre as classes das coleções. Nesta medida, dada uma classe $L_r$ de tamanho $n_r$ e um grupo $S_i$ de tamanho $n_i$, a *FScore* é definida como:

$$F(L_r,S_i) = \frac{2 * R(L_r, S_i) * P(L_r, S_i)}{R(L_r, S_i) + P(L_r, S_i)} \tag{7}$$

na qual as expressões $R(L_r,S_i)$ e $P(L_r,S_i)$ representam, respectivamente, valores de *recall* e *precision* para a classe $L_r$ no grupo $S_i$. O *FScore* da classe $L_r$ é o maior valor obtido por algum grupo da hierarquia $H$:

$$F(L_r) = \max_{S_i \in H} F(L_r,S_i) \tag{8}$$

Assim, o valor *FScore* global de uma hierarquia, com um número $c$ de classes, é calculada como a soma da *FScore* de cada classe ponderada pelo tamanho da classe:

$$FScore = \sum_{r=1}^{c} \frac{n_r}{n} F(L_r) \tag{9}$$

Conforme o agrupamento hierárquico reconhece as classes predeterminadas de uma coleção, o valor de *FScore* se aproxima de 1. Caso contrário, a *FScore* tem valor 0. O quadro geral da avaliação *FScore* sobre os agrupamentos nas coleções textuais são apresentadas na Tabela 2. Os valores *FScore* obtidos foram calculados sobre os *dendro-*

| Coleção | Average | Complete | Ward |
|---|---|---|---|
| 20ng | 0.45 | 0.37 | 0.43 |
| comp | 0.66 | 0.59 | 0.67 |
| k1b | 0.89 | 0.85 | 0.89 |
| la2 | 0.56 | 0.51 | 0.57 |
| nsf1 | 0.64 | 0.61 | 0.60 |
| physics | 0.76 | 0.70 | 0.75 |
| re0 | 0.61 | 0.53 | 0.68 |
| reviews | 0.78 | 0.54 | 0.74 |
| tr41 | 0.78 | 0.69 | 0.81 |
| wap | 0.64 | 0.57 | 0.62 |

**Tabela 2. Avaliação *FScore* dos algoritmos em cada coleção textual.**

*gramas* gerados em cada algoritmo, e expressam a capacidade do algoritmo em encontrar as classes pré-existentes. Ao realizar a seleção de grupos significativos, o objetivo é obter uma hierarquia compacta e mais natural ao usuário, ou seja, que os relacionamento entre os grupos não seja estritamente binário; e que o valor *FScore* após a seleção não se distancie muito do original.

### 4.2. Avaliação dos Resultados

Para analisar o comportamento do valor de cada medida na seleção de grupos significantes, foram realizadas sucessivas execuções, com os valores variando entre 0 e 1. Um total

de 30 *dendrogramas* foi analisado. Nas Figuras 2, 3 e 4 é apresentado o comportamento das medidas em três destes *dendrogramas*. Em cada figura, o gráfico da esquerda representa a variação do valor *FScore* em relação à variação das medidas. No gráfico da direita é apresentada a variação do tamanho da hierarquia em relação à variação das medidas. É importante observar que a redução da hierarquia apresentada nos gráficos é obtida apenas com a aplicação de cada medida. Na prática, as medidas são utilizadas junto com outros critérios, como a remoção de grupos considerados muito pequenos pelos usuários.

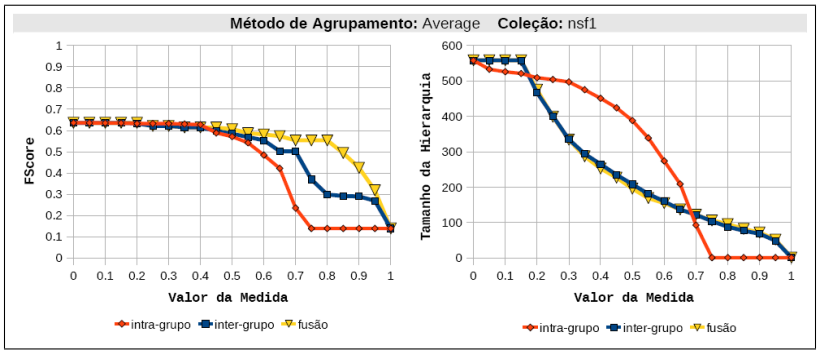


**Figura 2. Comparação das Medidas na Coleção *nsf1*.**

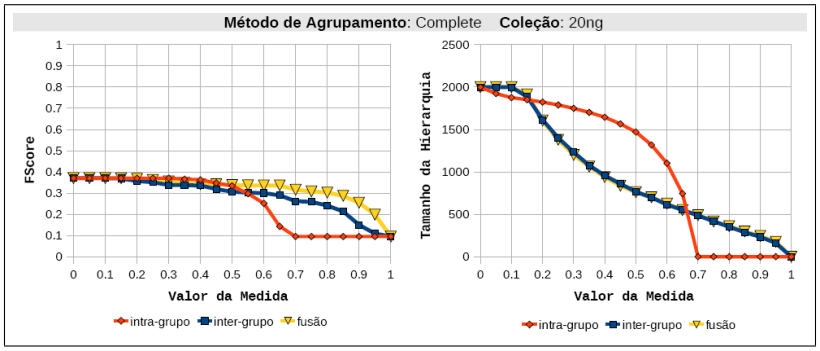


**Figura 3. Comparação das Medidas na Coleção *20ng*.**

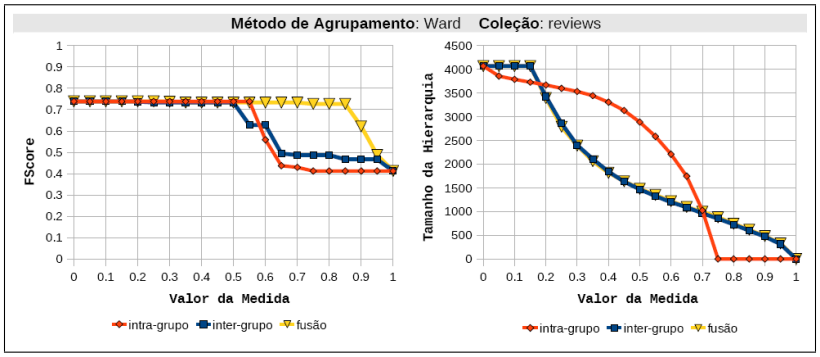


**Figura 4. Comparação das Medidas na Coleção *reviews*.**

A medida *fusão* obteve melhor desempenho que a medida *inter-grupo* e *intra-grupo* em todos os *dendrogramas* analisados, obtendo hierarquias mais compactas e significativas. Seu maior destaque ocorreu, principalmente, na detecção e remoção de grupos em efeito cadeia (*outliers*), que são uma divisão, ou junção, desbalanceada ocorrida

durante o agrupamento. A medida *inter-grupo* tem o segundo melhor desempenho, com comportamento semelhante à *fusão*, porém, mais vulnerável na presença de ruídos e *outliers*. Já a medida *intra-grupo* obteve pouca eficiência na obtenção de hierarquias compactas. Isto ocorre porque ela atua principalmente em grupos próximos ao topo da hierarquia, que apresentaram alta variabilidade. Como a quantidade destes grupos é relativamente menor, a medida *intra-grupo* obteve uma menor taxa de redução da hierarquia.

A análise dos gráficos permitiu observar que quando as medidas *intra-grupo* e *inter-grupo* ficam muito rigorosas, os grupos selecionados já não representam as classes naturais das coleções, diminuindo-se o valor *FScore*. A medida *fusão* é mais robusta, e consegue manter bons valores *FScore* com um número menor de grupos, principalmente quando o cojunto de dados é bem separado (alto valor *FScore*). O comportamento das medidas apresentado nas Figuras 2, 3 e 4 se repete nos 30 *dendrogramas* analisados.

Na próxima etapa do experimento, avaliou-se a qualidade das hierarquias obtidas utilizando a medida *fusão* para a seleção de grupos significativos. Foi escolhido o valor 0.3 para a medida, pois apresentou bons resultados de acordo com observações empíricas. Conforme é realizado em outros trabalhos relacionados, também fixou-se a seleção de grupos que possuíam um número mínimo de documentos, uma vez que esta técnica ajuda na performance do processo, eliminando grupos muito pequenos. Foi utilizado o valor mínimo de 5 documentos por grupo, que é relativamente baixo, e não interfere na avaliação das outras medidas, uma vez que todas as classes das coleções textuais deste experimento possuem mais que 5 exemplos.

| | Average | | | Complete | | | Ward | | |
|---|---|---|---|---|---|---|---|---|---|
| | F | F' | N | F | F' | N | F | F' | N |
| 20ng | 0.45 | 0.37 | 240 | 0.37 | 0.33 | 326 | 0.43 | 0.40 | 318 |
| comp | 0.66 | 0.63 | 54 | 0.59 | 0.54 | 71 | 0.67 | 0.51 | 50 |
| k1b | 0.89 | 0.76 | 256 | 0.85 | 0.76 | 306 | 0.89 | 0.74 | 288 |
| la2 | 0.56 | 0.54 | 420 | 0.51 | 0.49 | 493 | 0.57 | 0.48 | 511 |
| nsf1 | 0.64 | 0.57 | 66 | 0.61 | 0.47 | 72 | 0.60 | 0.47 | 76 |
| physics | 0.76 | 0.72 | 57 | 0.70 | 0.62 | 55 | 0.75 | 0.70 | 51 |
| re0 | 0.61 | 0.44 | 211 | 0.53 | 0.44 | 212 | 0.68 | 0.54 | 197 |
| reviews | 0.78 | 0.68 | 514 | 0.54 | 0.46 | 583 | 0.74 | 0.73 | 665 |
| tr41 | 0.78 | 0.67 | 139 | 0.69 | 0.60 | 165 | 0.8 | 0.79 | 163 |
| wap | 0.64 | 0.54 | 167 | 0.57 | 0.51 | 202 | 0.62 | 0.53 | 190 |

**Tabela 3. Avaliação *FScore* após seleção de grupo significativos.**

A medida *FScore* foi recalculada para todas as hierarquias obtidas após processo de seleção de grupos significantes. Na Tabela 3 é apresentado o resultado final para os três algoritmos de agrupamento utilizados, em que *F* representa o *FScore* do *dendrograma*, conforme a Tabela 2, e *F'* e *N* são, respectivamente, o *FScore* e o número de grupos da hierarquia obtida após a seleção de grupos significantes. O número de grupos da hierarquia original é *K-1*, em que *K* é o número de documentos da coleção textual (ver Tabela 1), já que o *dendrograma* tem relação binária entre os documentos da coleção. É importante observar que o tamanho final da hierarquia, apresentado na Tabela 3, é influenciado pela remoção de grupos que não atingem o número mínimo de documentos, pois são os grupos mais profundos na hierarquia e em maior quantidade e, geralmente, triviais. No entanto, a medida *fusão* é responsável por considerável redução do número de

grupos, além de permitir que as hierarquias obtidas não sejam essencialmente binárias, e sim balanceadas de acordo a qualidade dos grupos e subgrupos selecionados, facilitando a exploração pelos usuários.

Para verificar se, em um dado algoritmo de agrupamento, os valores de $F$ e $F'$ apresentam diferenças significativas, foi aplicado o teste estatístico não paramétrico de *Mann-Whitney*. O teste não encontrou diferença estatística significativa entre a qualidade dos *dendrogramas* e a qualidade da hierarquia obtida após a seleção de grupos, em nenhum dos três algoritmos de agrupamento, com intervalo de confiança de 95%, indicando que a hierarquia obtida continua representativa em relação ao agrupamento original.

## 5. Considerações Finais e Trabalhos Futuros

Neste trabalho, três medidas de qualidade de grupos, *intra-grupo*, *inter-grupo* e *fusão* foram definidas e utilizadas na tarefa de seleção de grupos significativos de *dendrogramas*. A medida *fusão* foi proposta a partir da adaptação do conceito de "*joinability*", apresentado em [Lin and Chen 2005]. Uma avaliação experimental foi realizada, envolvendo 10 coleções textuais, de diversas características, para comparar a eficiência de cada medida.

A medida *fusão* obteve melhor desempenho na compactação da hierarquia sem afetar muito a capacidade de representação da hierarquia original. Além disso, constatou-se que a seleção de grupos significativos foi realizada com sucesso nos três algoritmos de agrupamento hierárquico abordados, indicando que a aplicação possa ser independente do algoritmo, necessitando-se apenas de um *dendrograma* inicial como entrada. O resultado final mantém as relações hierárquias principais entre os grupos, ao contrário da maioria dos critérios de validação da literatura, que objetivam obter partições *flats* dos *dendrogramas*. O uso da abordagem apresentada é interessante, pois permite que, após o cálculo das medidas, o usuário possa explorar os resultados sob vários cenários, em tempo real, apenas ajustando os critérios de qualidade.

Como trabalho futuro, espera-se investigar a obtenção de valores ideais para a medida *fusão*, de forma automática e específicos para cada hierarquia, baseando-se nas propriedades do próprio agrupamento. No entanto, com os resultados já obtidos, é possível visualizar a hierarquia de um modo reduzido, mais coeso e em tópicos de níveis mais altos, facilitando sua análise e interpretação.

## Referências


Boley, D. (1998). "Principal direction divisive partitioning". *Data Mining and Knowledge Discovery, v.2, n.4*, pages 325–344.

El-Hamdouchi, A. and Willett, P. (1986). "Hierarchic document classification using Ward's clustering method". In *Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 149–156.

Everitt, B. S., Landau, S., and Leese, M. (2001). *Cluster Analysis*. Arnold Publishers.

Feldman, R. and Sanger, J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Fung, B., Wang, K., and Ester, M. (2003). "Hierarchical document clustering using frequent itemsets". In *Proceedings of the SIAM International Conference on Data*, pages 59–70.

Guha, S., Rastogi, R., and Shim, K. (1998). "CURE: an efficient clustering algorithm for large databases". *ACM SIGMOD Record, v.27, n.2*, pages 73–84.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002a). "Cluster validity methods: Part I". *ACM SIGMOD Record, v.31, n.2*, pages 40–45.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002b). "Clustering validity checking methods: Part II". *ACM SIGMOD Record, v.31, n.3*, pages 19–27.

Han, E., Boley, D., Gini, M., Gross, R., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., and Moore, J. (1998). "WebACE: A web agent for document categorization and exploration". In *Proceedings of the second international conference on Autonomous agents*, pages 408–415.

Huang, R., Zhang, Z., and Lam, W. (2006). "Refining Hierarchical Taxonomy Structure Via Semi-supervised Learning". In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 653–654.

Jain, A. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall.

Kashyap, V., Ramakrishnan, C., Thomas, C., and Sheth, A. (2005). "Taxaminer: An experimentation framework for automated taxonomy bootstrapping". *International Journal of Web and Grid Services, v.1, n.2*, pages 240–266.

Larsen, B. and Aone, C. (1999). "Fast and effective text mining using linear-time document clustering". In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 16–22.

Lin, C. and Chen, M. (2005). "Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging". *IEEE Transactions on Knowledge and Data Engineering, v.17, n.2*, pages 145–159.

Milligan, G. and Cooper, M. (1985). "An examination of procedures for determining the number of clusters in a data set". *Pshychometrika, v.50, n.2*, pages 159–179.

Müller, A., Dorre, J., Gerstl, P., and Seiffert, R. (1999). "The TaxGen Framework: Automating the Generation of a Taxonomy for a Large Document Collection". In *Proceedings of the 32nd Annual Hawaii International Conference on System Sciences*, page 2034.

Nogueira, B. M., Moura, M. F., Conrado, M. S., Rossi, R. G., Marcacini, R. M., and Rezende, S. O. (2008). "Winning Some of the Document Preprocessing Challenges in a Text Mining Process". In *IV Workshop em Algoritmos e Aplicações de Mineração de Dados*, pages 10–18.

Sarle, W. S. and Kuo, A. H. (1993). "The MODECLUS procedure". Technical Report 256, NC: SAS Institute Inc.

Sneath, P. H. and Sokal, R. R. (1973). *Numerical Taxonomy*. Freeman, London, UK.

Zhao, Y. and Karypis, G. (2002). "Evaluation of hierarchical clustering algorithms for document datasets". In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 515–524.