Universidade Federal de Pernambuco

Centro de Informática

Mestrado em Ciência da Computação

Dissertação de Mestrado

Seleção Local de Características em Agrupamento Hierárquico de Documentos

Marcelo Nunes Ribeiro (mnr@cin.ufpe.br)

Orientador: Ricardo Bastos Cavalcante Prudêncio

Recife, Fevereiro de 2009.

Marcelo Nunes Ribeiro

Seleção Local de Características em Agrupamento Hierárquico de Documentos

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Curso de Mestrado em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco.

Orientador:

Ricardo Bastos Cavalcante Prudêncio

Recife, Fevereiro de 2009

Resumo

O agrupamento hierárquico de documentos é utilizado para prover interface de navegação em coleções de documentos, ajudando na atividade de recuperação de informação. Como os vetores que representam os documentos possuem uma alta dimensionalidade, a presença de termos irrelevantes confunde o algoritmo de agrupamento. O uso da seleção de características em agrupamento de documentos é capaz de melhorar a precisão e o tempo de execução do agrupamento. Esta dissertação discute vários métodos de seleção de características já aplicados e aborda a forma como a seleção de características interage com o algoritmo de agrupamento, que pode ser classificada de forma global, quando um único subconjunto de características é considerado, ou local, quando cada grupo é descrito por subconjuntos de características distintas. Por conta da diversidade de visões das características proporcionada pela seleção local, o algoritmo de agrupamento é capaz de revelar grupos ocultos nos dados. Nesta dissertação, é aplicado o mesmo princípio de seleção local para o caso de agrupamento hierárquico divisivo de documentos, com a realização de uma nova seleção de características a cada passo de divisão dos grupos. Este método foi batizado de ZOOM-IN. Foram feitos experimentos com as bases de documentos Reuters-21578 e RCV2 e foi comprovado um ganho de precisão no resultado do agrupamento quando a heurística de escolha do número de termos do método ZOOM-IN é capaz de eliminar os termos irrelevantes. Também é desenvolvida uma aplicação dos métodos discutidos para agrupar documentos do resultado de uma consulta ao Google, com etiquetagem e escolha do número de grupos usando amostragem e o conceito de estabilidade do agrupamento. Os resultados mostraram que a execução do algoritmo com diferentes parâmetros é capaz de descobrir diferentes grupos interessantes, o que motiva a pesquisa de uma interface de acesso aos documentos que combine os resultados de diferentes execuções dos algoritmos. Por fim, são apresentadas vantagens e limitações do uso do método ZOOM-IN, além de indicações de trabalhos futuros.

Palavras-chave: Agrupamento de documentos, seleção de características.

Abstract

Hierarchical clustering of documents is used to provide interface for navigating through collections of documents, assisting in the activity of information retrieval. As the vectors representing the documents have a high dimensionality, the presence of irrelevant terms can harm the clustering algorithm. The use of feature selection in text clustering is able to improve the accuracy and execution time of the clustering. This master thesis discusses different methods of feature selection and how feature selection interacts with the clustering algorithm, which can be classified into global, when only one subset of features is considered, or local, when each group is described by different subsets of features. Because of the diversity of views of the features offered by the local selection, the clustering algorithm is able to reveal hidden clusters in the data. This master thesis applied the same principle of local feature selection for partitional hierarchical text clustering, with the application of a new feature selection at each step of partition of the groups. This method is called ZOOM-IN. Experiments were done with the documents of the Reuters-21578 and RCV2 collections and it was shown a gain of precision in the clustering when the heuristic to choose the number of terms of the method ZOOM-IN was able to eliminate the irrelevant terms. It was also developed an application of the methods discussed to clustering the documents returned by a Google query, with labeling and choice of the number of clusters using sampling and the concept of stability of clustering. The results showed that the implementation of the algorithm with different parameters is able to find different interesting clusters, which motivates the research for an interface to access the documents, combining the results of different executions of the algorithms. Finally, we present advantages and limitations of using the ZOOM-IN, and indications of future work.

Keywords: Text clustering, feature selection.

Agradecimentos

Devo agradecer primeiramente a Deus, por ter preparado uma família tão boa para mim. Meu pai Antonio é uma pessoa honesta, trabalhadora, nunca deixou faltar livros em casa e acompanhar o desempenho escolar de seus filhos. Minha mãe Lúcia é uma dona de casa exemplar, que sempre faz de tudo para manter a família feliz e tenho uma grande suspeita que seja a pessoa mais bondosa do mundo. Meu irmão Manoel sempre foi também meu colega de sala de aula até a graduação. Além de sempre estar do meu lado, proporcionou uma competição sadia durante toda minha vida escolar. Eu e estas três pessoas formam o que hoje penso ser uma família próxima da ideal, com respeito, confiança e amor mútuos.

Às tia Marlene e finada tia Fátima, que sempre foram as pessoas externas à casa mais importantes para nossa família. Sempre presentes, sempre nos ajudaram e ainda nos ajudam.

Ao meu finado avô Manoel, que nunca tive a oportunidade de conhecer, mas sei que a partir dele nasceu a sólida base moral de nossa família.

A todos os bons professores de minha vida, em especial aos meus professores de ensino médio do CEFET-AL de Palmeira dos Índios e ao agora meu orientador de mestrado, e futuro doutorado, Ricardo. São pessoas que sei que estão preocupadas comigo e que posso contar para o que for possível.

A outros amigos que sei que posso confiar: Bruno, Diego, Djalma, Fabiano, Geisel, Laudimio, Leonardo, Marcus e Victor. Diego, com sua melhor cabeça para programação, junto do meu irmão, me ajudaram diretamente em diversas implementações realizadas durante o mestrado. Bruno e Victor me ajudaram sobremaneira com relação a minha moradia em Recife. A presença deles tornou o apartamento dividido por nós um ambiente feliz. Bruno, Djalma e Victor também são companheiros de jogos em LAN, e reconhecem que jogo Dota e Call of Duty melhor que todos eles.

Por fim agradeço a todos os pesquisadores que cito durante a dissertação. Mesmo que meus agradecimentos não cheguem a eles, estes pesquisadores sabem o grande favor que fazem a humanidade, tentando entender esta imensa natureza, de conhecimento infinito, deixada por Deus. Agora eu também tento me introduzir neste ramo de garimpagem do conhecimento. Tentarei dar o melhor de mim, para orgulhar minha família e me tornar uma pessoa melhor a cada dia. Que eu possa dar minha contribuição para tornar o mundo mais belo!

Sumário

1 Introdução **1**
1.1 Seleção de características em agrupamento de documentos 2
1.2 Contexto da dissertação 3
1.3 Conteúdo da dissertação 4

2 Agrupamento de documentos **6**
2.1 Representação dos documentos 6
2.1.1 Redução de dimensionalidade 7
2.2 Algoritmos de agrupamento 8
2.3 Critérios de validação 10
2.3.1 Critérios internos 11
2.3.2 Critérios externos 13
2.3.3 Avaliação usando amostragem 15
2.4 Considerações finais 16

3 Seleção de características **18**
3.1 Seleção global de características 20
3.1.1 Filtros 21
3.1.2 *Wrappers* 22
3.2 Seleção local de características 25
3.2.1 Seleção local de características 26
3.2.2 Seleção "glocal" de características 27
3.3 Considerações finais 28

4 Seleção local de características para agrupamento de documentos **29**
4.1 Relevância 29
4.2 Método proposto 30
4.2.1 Análise da complexidade do algoritmo 31
4.3 Escolha do número de grupos usando o conceito de estabilidade 32
4.4 Aspectos de implementação 35
4.4.1 Pré-processamento do texto 36
4.4.2 O algoritmo de agrupamento 36
4.5 Considerações finais 38

5 Experimentos e aplicação **39**
5.1 Experimentos com bases de documentos 39
5.1.1 Descrição 39
5.1.2 Resultados e discussão 40

5.2 Agrupamento de resultados do Google 43
5.2.1 Descrição 43
5.2.2 Resultados e discussão 45
5.3 Considerações finais 50

6 Conclusões **51**
6.1 Resumo das contribuições 51
6.2 Limitações e trabalhos futuros 52
6.3 Considerações finais 52

Lista de Figuras

2.1 Arquitetura básica para agrupamento de documentos 6
2.2 Exemplo de dendograma para 4 documentos 9
2.3 Uso de amostragem e predição para determinar estabilidade do agrupamento 16

3.1 Classificação de características 19
3.2 Características redundantes 19
3.3 Característica irrelevante 20
3.4 Abordagem *wrapper* para aprendizagem não-supervisionada 22
3.5 Características ruidosas 25

4.1 Gráfico do número de grupos pelo valor de critério interno 34
4.2 Pontos que podem ser separados em 2, 3 ou 4 grupos 34

5.1 Micro-média de precisão em relação ao número de termos utilizados para a base Reuters, com seleção de características global e local 41
5.2 Micro-média de precisão em relação ao número de termos utilizados para a base RCV2, com seleção de características global e local 41

Capítulo 1

Introdução

Agrupamento de documentos é a atividade de agrupar documentos similares entre si, de forma a melhor discriminar documentos pertencentes à categorias, assuntos ou contextos distintos. Inicialmente, agrupamento de documentos foi pesquisado como uma maneira implícita de melhorar a precisão ou cobertura de sistemas de recuperação de informação (Steinbach et al., 2000), através da expansão da consulta com termos relevantes de grupos formados por um algoritmo de agrupamento. Inicialmente, os progressos nesse sentido foram pequenos, não se obtendo uma comprovação empírica de melhoria dos resultados (Cutting et al., 1992). A partir do trabalho de Cutting et al. (1992), que propõe uma maneira de navegar em uma coleção de documentos agrupados, a área descobriu um novo caminho, com novos objetivos, entre eles: melhoramento da navegação entre documentos, velocidade e escalabilidade (Zamir et al., 1997).

Com o advento da Web, uma grande quantidade de documentos na forma de texto, imagem ou vídeo está disponível a usuários da Internet. Um usuário em busca de certa informação deseja ter acesso fácil, preciso e abrangente a estes documentos. Isto é obtido tradicionalmente organizando os documentos de forma ordenada em relação a pertinência com a consulta do usuário, como é feito no Google (`http://www.google.com`). Mas esta ordenação dos documentos apresenta pouca informação sobre os vários significados e contextos que estão associados aos documentos retornados pela consulta. Uma forma de discernir entre os diferentes contextos presentes no resultado da consulta é o uso de agrupamento, separando grupos de documentos em diferentes níveis de especificidade.

Zamir et al. (1997) propõe o uso de agrupamento para organizar os resultados de um motor de busca a uma consulta do usuário, princípio implementado pelo site Vivissimo (`http://www.vivissimo.com`). Neste caso, a estrutura de grupos organizada hierarquicamente e apropriadamente etiquetada proporciona uma visão de quais tipos de questões podem ser respondidas pela consulta feita pelo usuário.

Um dos passos geralmente empregados em agrupamento de documentos é a seleção de características relevantes, ou que também pode ser visto como a eliminação de carac-

terísticas irrelevantes, com objetivo de melhora na precisão e do custo computacional. Esta dissertação busca estudar o emprego de seleção de características em agrupamento de documentos de maneira local, ou seja, com a seleção de diferentes subconjuntos de características para a geração de grupos, comparando em ganho de precisão com a abordagem global, onde somente um subconjunto de características é considerado na geração de todos os grupos.

Nas próximas seções, segue uma breve introdução do uso de seleção de características em agrupamento de documentos (seção 1.1) e uma breve discussão do que já foi pesquisado na área de seleção local de características e o que será proposto nesta dissertação (seção 1.2). Na seção 1.3, é apresentada a estrutura da dissertação.

1.1 Seleção de características em agrupamento de documentos

Para realizar agrupamento, os documentos comumente são representados como padrões (vetores) de características. Geralmente as características, no contexto de agrupamento de documentos, são os pesos associados a cada termo presente nos documentos. Como são milhares de termos, a consideração de todos os termos acarreta alguns problemas ao algoritmo de agrupamento:

- O cálculo das proximidades entre os documentos e centróides é a rotina que concentra quase todo o processamento em algoritmo de agrupamento como o K-means. Quanto maior é a quantidade de características nos documentos, assim como quanto maior é a quantidade de documentos, maior é a demanda computacional.

- Ainda quanto ao cálculo das proximidades, quando o tamanho do espaço de características é muito alto, é demonstrado que a distância entre pontos mais similares não é tão diferente de pontos mais distantes entre si. É a conhecida "praga da dimensionalidade" (Duda et al., 2001).

Pensando nas questões de velocidade e precisão, a arquitetura de agrupamento de documentos geralmente contém uma fase de redução de dimensionalidade dos padrões, onde são criadas novas características a partir das existentes, formando padrões em uma dimensionalidade menor. Estas novas características podem estar contidas nas características originais, ou seja, está sendo realizada uma seleção de características. Ou as novas características podem ser criadas a partir de cálculos usando as características originais, ou seja, está sendo realizada uma extração de características. Para o caso de seleção de características, a seleção leva em conta que nem todos os termos são importantes. Alguns são redundantes, irrelevantes ou somente atrapalham o processo de agrupamento, quando

diferentes subconjuntos de características revelam diferentes estruturas (Dy & Brodley, 2004).

O objetivo da seleção de características em problemas de agrupamento é achar o menor subconjunto de características que melhor revela grupos "naturais" dos dados (Dy & Brodley, 2004). Existem diferentes classificações e formas de resolução deste problema. Quanto a interação com o algoritmo de agrupamento, a seleção de características pode ser global ou local, e usando métodos *wrappers* ou filtros. A seleção global de características almeja selecionar um único subconjunto que é relevante para todos os grupos (Li et al., 2008) e a seleção local de características usa diferentes subconjuntos de características associados a cada grupo formado. Quanto aos métodos, uma solução *wrapper* requer o uso do algoritmo de agrupamento para avaliar os subconjuntos de características, guiando a busca das características relevantes. Uma solução filtro escolhe os termos sem o uso do algoritmo de agrupamento, geralmente se baseando em alguma propriedade estatística dos dados. O método filtro vem sendo mais utilizado para seleção de características em agrupamento de documentos, por conta do menor custo computacional em relação a uma solução *wrapper* (Liu et al., 2003; Tang et al., 2005).

1.2 Contexto da dissertação

Apesar da seleção global ser capaz de gerar bons resultados em agrupamento, é possível haver vários subconjuntos de características que revelem bons grupos "naturais" dos dados. Para tirar proveito desta variedade, o uso de subconjuntos de características relevantes a cada partição dos dados pode ser uma forma de melhor revelar e discriminar os grupos. A seleção local de características é proposta neste sentido, pois usa a suposição que cada grupo possui termos mais importantes, que ajudam a discriminar cada grupo dos demais grupos. O trabalho de Li et al. (2008) utiliza esta propriedade para buscar vários subconjuntos de características que revelam diferentes grupos, escolhendo os grupos mais coesos baseado em certo critério.

A abordagem proposta nesta dissertação, como no trabalho de Li et al. (2008), também utiliza diferentes subconjuntos de características para revelar diferentes grupos, mas aqui é proposto o uso de um algoritmo de agrupamento divisivo hierárquico com seleção de características localizada à medida que novos grupos são descobertos. Esta forma é chamada de "glocal" no trabalho de Koller & Sahami (1997) sobre categorização de documentos. Pode ser considerada *local* pois seleciona um subconjunto de características diferente em cada divisão de grupos na hierarquia. Mas funciona de forma *global* recursivamente em cada nó, pois um mesmo subconjunto de características é utilizado para divisão do nó em grupos-irmãos. Esta abordagem difere do método proposto em Li et al. (2008), que busca um subconjunto de características para cada grupo a ser formado.

Neste trabalho, supondo que cada grupo de documentos possui termos mais relevantes, vislumbra-se que a seleção localizada de características em agrupamento hierárquico divisivo revele grupos naturais de documentos, ajudando na precisão do resultado do agrupamento. Serão realizados experimentos que objetivam comprovar a vantagem do uso da seleção local em lugar da global, visando contribuir para o estado da arte sobre o tema de seleção local de características para agrupamento de documentos. Este trabalho proporciona:

- Uma revisão do estado da arte sobre seleção de características para aprendizagem não-supervisionada.
- Implementação e proposta de técnicas de seleção local de características em agrupamento hierárquico de documentos.
- Análise experimental, comparando os resultados do agrupamento com o uso de seleção global e com o uso de seleção local de características, utilizando textos das bases Reuters-21578 (Lewis, 1999) e RCV2 (Lewis et al., 2004).
- Uma aplicação do uso dos métodos discutidos para organizar documentos retornados por uma consulta ao Google (Google, 2008), utilizando escolha do número de grupos com o conceito de estabilidade apresentado em Ben-Hur et al. (2002) e um método simples de etiquetagem.

1.3 Conteúdo da dissertação

Esta dissertação é organizada nos seguintes capítulos:

Capítulo 2 - Agrupamento de documentos: É apresentado a arquitetura básica e funcionamento de agrupamento de documentos, com discussão sobre critérios internos e externos de avaliação.

Capítulo 3 - Seleção de características: Revisão do estado da arte em seleção de características para aprendizagem não-supervisionada, fazendo a distinção entre seleção global, local e "glocal".

Capítulo 4 - Seleção local de características para agrupamento de documentos: É discutido o método proposto de seleção local em agrupamento hierárquico, batizado de método ZOOM-IN. Também é apresentado o método para escolha do número de grupos, que é capaz de fazer esta escolha mesmo com grupos definidos em dimensões diferentes. Por fim, são discutidos aspectos de implementação de um sistema para agrupamento de documentos utilizando os métodos abordados nesta dissertação.

Capítulo 5 - Experimentos e aplicação: São apresentados os experimentos com as bases, usando seleção global e local de características e resultados da aplicação do método de escolha do número de grupos com etiquetagem para organizar os resultados de uma consulta ao Google.

Capítulo 6 - Conclusões: Resumo das contribuições do trabalho realizado e o que pode ser feito em trabalhos futuros.

Capítulo 2

Agrupamento de documentos

Agrupamento de documentos é a atividade de agregar documentos similares, de maneira a melhor discriminar documentos pertencentes a grupos diferentes. O modelo clássico de agrupamento de documentos é apresentado na Figura 2.1. As próximas seções irão explicar cada fase da arquitetura apresentada.

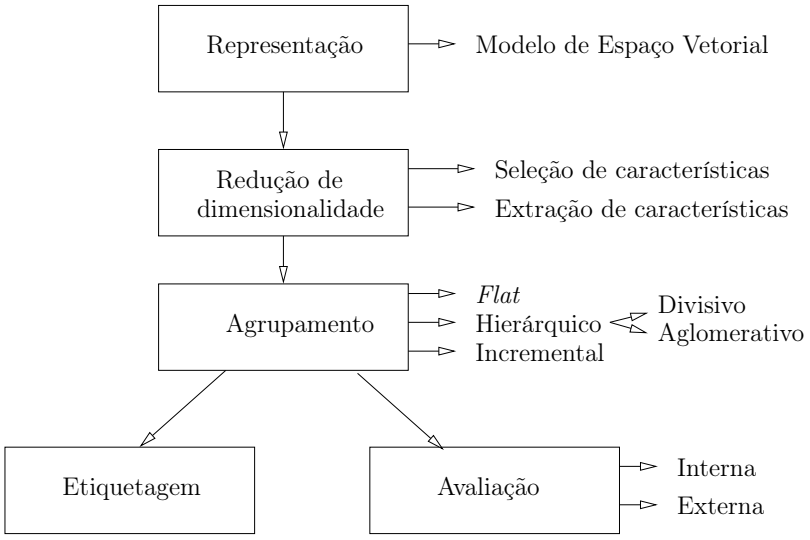


Figura 2.1: Arquitetura básica para agrupamento de documentos

2.1 Representação dos documentos

Pelo Modelo de Espaço Vetorial (Salton et al., 1975), um documento pode ser descrito por um conjunto de palavras-chave, os chamados termos de indexação, que são todo o vocabulário presente na coleção de textos. Neste modelo, um peso é associado a cada termo de indexação, o que define um vetor de termos que representa o documento. Das formas utilizadas para cálculo dos pesos pode-se citar o uso de valores booleanos indicando a presença ou não do termo, ou a freqüência do termo no documento, ou pelo cálculo do

valor TF-IDF, que é a forma comumente utilizada. Para o TF-IDF, cada valor $tfidf_j$ do vetor do documento d_j representa um termo t. Este valor é dado por:

$$tfidf_j = tf_j \log \frac{n}{DF_t}$$

onde

$$tf_j = \begin{cases} 1 + \log t_j & \text{se } t_j > 0 \\ 0 & \text{senão} \end{cases}$$

n é o número de documentos na coleção, DF_t é o número de documentos em que o termo t ocorre e t_j é a freqüência do termo t no documento d_j.

Para definir a proximidade entre dois vetores representados neste modelo, geralmente é utilizada a medida de similaridade *cosseno*, pois seu cálculo ignora a correspondência de valores $0 - 0$ para a mesma característica dos vetores envolvidos como uma forma de similaridade entre vetories, algo bastante comum em se tratando de documentos, que geralmente não possuem muito dos mesmos termos e, sendo levado em conta as várias correspondências $0 - 0$, caso de um medida como a euclidiana, vários documentos serão muito similares entre si. Dados dois vetores $\vec{d_1}$ e $\vec{d_2}$, a medida de proximidade usando cosseno é dada por:

$$\cos\left(\vec{d_1}, \vec{d_2}\right) = \frac{\vec{d_1} \cdot \vec{d_2}}{\left\|\vec{d_1}\right\| \times \left\|\vec{d_2}\right\|}$$

onde $\|\bullet\|$ é a norma de cada vetor.

2.1.1 Redução de dimensionalidade

O cálculo das proximidades entre os documentos e centróides é a rotina que concentra quase todo o processamento em algoritmo de agrupamento como o K-means. Quanto maior é a dimensionalidade, maior é a demanda computacional. E, quando o tamanho do espaço de características é muito alto, é demonstrado que a distância entre pontos mais similares não é tão diferente de pontos mais distantes entre si. É a conhecida "praga da dimensionalidade" (Duda et al., 2001). É necessário o uso de técnicas de *redução de dimensionalidade* para aumentar o desempenho do processo de agrupamento, que pode ser feito de duas maneiras:

1. Seleção de características: um subconjunto de termos é escolhido dentre o vocabulário de termos existentes. Este ponto será abordado no Capítulo 3.

2. Extração de características: novas características são geradas através da transformação ou combinação dos termos do vocabulário. Apesar de técnicas deste tipo, tal

como *Latent Semantic Indexing* (LSI) (Deerwester et al., 1990), *Principal Component Analysis* (PCA) (Hotelling, 1933) e *Independent Component Analysis* (ICA) (Comon, 1994), possuírem também a propriedade de tornar mais similares documentos que são semanticamente associados (Raghavan, 1997), a informação associada aos termos desaparece e um ponto crucial do agrupamento de documentos, que é a compreensão da estrutura dos grupos, fica comprometida.

2.2 Algoritmos de agrupamento

O algoritmo de agrupamento pode ser *flat* ou hierárquico. Um algoritmo de agrupamento *flat* é aquele que particiona todos os dados em grupos, sem organizar estes grupos em uma hierarquia, como é feito no hierárquico. Para realizar agrupamentos *flat* vários algoritmos já foram propostos e aplicados em documentos, tais como os algoritmos baseados em grafos, que usa uma matriz de proximidade dos pontos e conceitos de grafos e hiper-grafos para agregar ou dividir pontos (Guha et al., 2000), algoritmos baseados na teoria da informação (Slonim et al., 2002), algoritmos baseados na densidade dos pontos (Sander et al., 1998) ou algoritmos que usam o conceito de centróide, como o K-means, que será explicado em breve.

Geralmente a tarefa de agrupamento se basea em escolher uma partição que maximize um critério, ou seja, é um problema de otimização. Enumerar todos os possíveis agrupamentos e escolher aquele com melhor valor da função objetivo é um problema intratável (NP completo). O número de configurações possíveis é dado pelo número de Stirling:

$$s = \frac{1}{k!}\sum_{i=1}^{k}(-1)^{k-i}\left(\begin{array}{c} k \\ i \end{array}\right) i^n$$

com k grupos e n padrões (Jung et al., 2003).

É necessário uma heurística para tornar esta busca eficiente. É o que faz o mais conhecido algoritmo *flat*, o *K-means* (Jain et al., 1999). Para entender o *K-means*, um conceito importante é o de centróide de um grupo, que corresponde ao seu centro geométrico, ou seja, a média de todos os vetores do grupo. Formalmente, dado um grupo C_k, então seu centróide é definido como:

$$\vec{c_k} = \frac{1}{|C_k|}\sum_{d_i \in C_k}\vec{d_i}$$

onde $|C_k|$ é o número de padrões pertencentes ao grupo C_k. O *K-means* pode ser resumido pelos seguintes passos:

1. Defina K centróides iniciais, escolhendo K documentos aleatórios da coleção.

2. Afete cada documento para o grupo correspondente ao centróide mais próximo.

3. Recalcule os centróides dos grupos.

4. Repita os passos 2 e 3 até atingir um critério de parada (geralmente até não ocorrer alterações nos centróides).

Já o agrupamento hierárquico associa os dados a uma hierarquia de grupos, formando um dendograma. Para agrupamento de documentos, a solução hierárquica possui mais vantagens em relação a abordagem *flat*, por proporcionar uma melhor visão de quais tipos de questões podem ser respondidas pela coleção de documentos. Ela é capaz de dividir a coleção de documentos em diferentes níveis de granularidade e especificidade, expandindo as opções do usuário, ajudando-o a decidir quais grupos lhe são interessantes na coleção de documentos (Sahoo et al., 2006; Zhao & Karypis, 2002).

Algoritmos de agrupamento hierárquicos podem ser aglomerativos ou divisivos. Aglomerativo é uma abordagem ascendente (*bottom-up*) e começa afetando cada documento a um grupo distinto e prossegue combinando os documentos e grupos mais similares, até que todos os documentos sejam alocados a um único grupo, ou outro critério de parada seja alcançado. Divisivo é uma abordagem descendente (*top-down*) e começa considerando todos os documentos em um único grupo, escolhendo um grupo, particionando-o em outros grupos e prossegue escolhendo e dividindo até que cada grupo terminal da árvore possua somente um documento, ou outro critério de parada seja alcançado. Ambos os métodos hierárquicos formam uma estrutura de documentos conhecida como dendograma, como é visto na Figura 2.2.

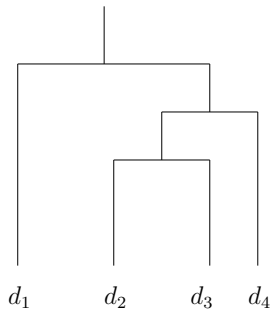


Figura 2.2: Exemplo de dendograma para 4 documentos

A abordagem aglomerativa passou muito tempo conhecida por produzir melhores resultados, mas de acordo com Zhao & Karypis (2002), no agrupamento aglomerativo, qualquer decisão errada de combinação dos grupos no começo da execução do algoritmo tende a multiplicar os erros a medida que o agrupamento é executado. Algoritmos divisivos, por sua vez, tem uma visão global mais privilegiada de possíveis grupos coesos e serão o foco deste trabalho.

Uma desvantagem do agrupamento hierárquico é o fato de que ainda funciona de maneira *flat*, pois os algoritmos de agrupamento não intencionam formar grupos com relação pai-filho, mas em agregar ou dividir grupos (Treeratpituk & Callan, 2006).

O *bisecting K-means* (Steinbach et al., 2000) é um algoritmo divisivo já bastante pesquisado que foi capaz de proporcionar melhores resultados comparado a algoritmos aglomerativos. Ele usa o algoritmo K-means para agir de uma forma divisiva, dividindo os dados em dois sub-grupos em cada iteração. O *bisecting K-means* funciona da seguinte maneira:

1. Escolha um grupo para dividir (começando com um único grupo).
2. Divida-o 2 sub-grupos usando o algoritmo *K-means*.
3. Repita o passo 2 por $ITER$ vezes e fique com a partição com melhor valor do critério interno.
4. Repita os passos 1, 2 e 3 até que o número de grupos requerido é alcançado.

Um critério interno bastante utilizado é o de similaridade geral, definido como:

$$f = \frac{\sum_{i=1}^{N_c} \left\{ \frac{\sum_{j=1}^{p_i} s(c_i, d_{ij})}{p_i} \right\}}{N_c}$$

onde d_{ij} é o j-ésimo vetor, que pertence ao grupo i; c_i é o vetor centróide do i-ésimo grupo, $s\left(c_i, d_{ij}\right)$ é a similaridade entre o documento d_{ij} e o centróide do grupo c_i, p_i é o número de documentos que pertencem ao grupo C_i e N_c é o número de grupos.

Algumas observações importantes ao *bisecting K-means* são:

- A escolha do grupo a ser particionado, no passo 1, pode se dar de várias maneiras. Uma forma usada é escolher o grupo com maior número de documentos. (Zhao & Karypis, 2002) sugere, como critério de escolha, processar o passo 2 para todos os grupos e escolher aquele com melhor resultado do critério interno.
- No passo 3, executar a bisecção $ITER$ vezes é necessário por conta da escolha aleatória dos centróides iniciais muitas vezes levar a resultados pobres, ou seja, convergindo para um ponto de máximo local de baixo valor.
- Cada divisão gera somente dois novos grupos, formando sempre uma árvore binária, o que torna desejável um procedimento de escolha do número de grupos após a execução do agrupamento, através de cortes no dendograma.

2.3 Critérios de validação

É necessário conhecer a importância da validação, ou avaliação, tanto para objetivos de precisão do agrupamento (avaliação externa), quanto para validar os resultados do agrupamento em tempo de execução (avaliação interna). A avaliação pode ser basicamente de duas formas:

- Usando um critério externo, que avalia os resultados de um algoritmo de agrupamento comparando com uma estrutura de classes pré-especificada, que reflete o conhecimento a priori da real estrutura dos dados.

- Usando um critério interno, que geralmente envolve medidas que utilizam os próprios vetores dos dados para mensurar o quão bom é o agrupamento, sem consultar uma estrutura de classes externa. O critério de similaridade geral apresentado na seção anterior é um exemplo de critério interno.

Critérios de validação são aplicados ainda em outros aspectos do agrupamento, tais como: métodos *wrappers* de seleção de características, escolha do número de grupos nos dados e combinação (*ensemble*) de agrupamentos, que é a tarefa de combinar vários resultados de agrupamento para o mesmo conjunto de dados, de vários algoritmos diferentes, com parâmetros diferentes, escolhendo o resultado que mais se assemelha a distribuição real de classes (Patrikainen & Meila, 2006).

A avaliação externa compara o resultado do agrupamento com um caso real de divisão de classes. A interna mede a qualidade do agrupamento sem conhecimento da real distribuição de classes. Uma outra forma de calcular a avaliação interna é o uso de métodos baseados em estabilidade, que mede a estabilidade do agrupamento realizando amostragem dos dados.

2.3.1 Critérios internos

Critérios internos são usados para avaliação de um agrupamento somente com informação dos valores dos vetores dos dados. Um dos critérios mais usados em agrupamento de documentos é a similaridade geral, já apresentado anteriormente, que mede a soma das similaridades do centróide do grupo aos padrões. Quanto maior o seu valor, indica que os pontos similares estão formando subconjuntos similares entre si, ou seja, são mais homogêneos. O caso inverso desta medida é o erro quadrado intra-grupos, que usa medida de dissimilaridades para calcular a separabilidade dos pontos em relação aos seus centróides. A duas medidas medem a proximidade intra-grupos. Supondo que um conjunto $C = \{x_1, \cdots, x_n\}$ de n padrões é agrupado em k subconjuntos disjuntos $C_1, \cdots, C_k$, com $n_1, \cdots, n_k$ sendo o número de padrões em cada grupo. Usando dissimilaridade, o critério de erro quadrado intra-grupos é dado por (Duda et al., 2001; Jung et al., 2003):

$$\Lambda = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2$$

onde c_i é o centróide do i-ésimo grupo. O erro quadrado inter-grupos, que mede o quanto os grupos estão separados, ou seja, mais facilmente discriminados, é dado por:

$$\Gamma = \sum_{i=1}^{k} \|c_i - c\|^2$$

onde c é o centróide global dos dados, calculado por:

$$c = \frac{1}{n} \sum_{C} x = \frac{1}{n} \sum_{i=1}^{k} n_i c_i$$

O agrupamento ótimo é aquele que minimiza a taxa de erro intra-grupos e maximiza a taxa de erro inter-grupos.

Outra forma de obter critérios internos é fazendo uso de matrizes de dispersão, que serão definidas a seguir:

- Matriz de dispersão para o i-ésimo grupo

$$S_i = \sum_{x \in C_i} (x - c_i) (x - c_i)^t$$

- Matriz de dispersão intra-grupo

$$S_W = \sum_{i=1}^{k} S_i$$

- Matriz de dispersão inter-grupo

$$S_B = \sum_{x=1}^{k} n_i (c_i - c) (c_i - c)^t$$

- Matriz de dispersão total

$$S_T = \sum_{x \in C} (x - c) (x - c)^t = S_W + S_B$$

S_W mede o quanto dispersos os padrões estão dos seus centróides. S_B mede o quão dispersos os centróides estão do centróide global. A operação de traço em S_W (soma dos elementos da diagonal principal) é igual ao critério de erro quadrado intra-grupos Λ:

$$tr\left[S_W\right] = \sum_{i=1}^{c} tr\left[S_i\right] = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2$$

$$tr\left[S_B\right] = \sum_{i=1}^{k} n_i \|c_i - c\|^2$$

Como $tr\left[S_T\right] = tr\left[S_W\right] + tr\left[S_B\right]$ e $tr\left[S_T\right]$ independe de como os dados estão particionados, maximizar o erro inter-grupos equivale a minimizar o erro intra-grupos, ou seja,

não vai trazer nenhum benefício adicional ao resultado do agrupamento o uso de um ou de outro critério em qualquer situação (Duda et al., 2001).

É provado que os autovalores $\lambda_1, \cdots, \lambda_d$ da matriz gerada pela operação $S_W^{-1} S_B$ são invariantes a qualquer transformação linear não-singular dos dados (Duda et al., 2001). Pode-se compor vários critérios invariantes através de funções apropriadas destes autovalores, entre elas:

$$tr\left[S_W^{-1} S_B\right] = \sum_{1=1}^{d} \lambda_i$$

$$tr\left[S_T^{-1} S_W\right] = \frac{|S_W|}{|S_T|} = \sum_{1=1}^{d} \frac{1}{1+\lambda_i}$$

2.3.2 Critérios externos

Para critérios externos, sendo $C = \{C_1, ..., C_m\}$ um agrupamento dos dados resultante de um algoritmo de agrupamento e $P = \{P_1, ..., P_s\}$ uma outra partição dos dados. Pode-se classificar um par de pontos (x_v, x_u) em (Halkidi et al., 2002):

- SS: se ambos os pontos pertencem ao mesmo grupo em C e ao mesmo grupo em P.
- SD: se os pontos pertencem ao mesmo grupo em C e a diferentes grupos em P.
- DS: se os pontos pertencem a diferentes grupos em C e ao mesmo grupo em P.
- DD: se ambos os pontos pertencem a diferentes grupos em C e a diferentes grupos em P.

Assumindo que a, b, c, d são respectivamente o número de pares SS, SD, DS e DD, $a+b+c+d = M$, onde M é o número máximo de todos os pares no conjuntos de dados, ou seja, $M = N\,(N-1)\ /\ 2$, onde N é o número de pontos do conjunto de dados.

Para comparar a similaridade de C e P, alguns índices externos utilizados são (Fridlyand & Dudoit, 2001; Halkidi et al., 2002; Law et al., 2002; Meilă, 2007):

- Estatística de Rand:

$$R = \frac{(a+d)}{M}$$

- Coeficiente de Jaccard:

$$J = \frac{a}{(a+b+c)}$$

- Índice de Folkes e Mallows:

$$FM = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

- Estatística de Hubert:

$$\Gamma = \frac{Ma - m_1 m_2}{\sqrt{m_1 m_2 \left(M - m_1\right)\left(M - m_2\right)}}$$

onde $m_1 = a + b$ e $m_2 = a + c$.

O intervalo dos índices de Rand e Jaccard é $[0, 1]$. Já o intervalo do índice de Hubert é $[-1, 1]$. Para todos os índices, quanto maior o valor, mais similares os agrupamentos são, tento como exceção o índice de Hubert, onde os agrupamentos são mais similares de acordo o valor absoluto do índice se aproxima de 1.

Outra métrica é proposta em Meilă (2007), que mede a quantidade de informação que se ganha e perde quando se vai do agrupamento C ao P. Considerando n_i o número de elementos no grupo i de um agrupamento e n_{ij} o número de pontos na intersecção dos grupos C_i e P_j, isto é, $n_{ij} = |C_i \cap P_j|$, a distância de variação de informação é defina por:

$$VI = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{s} n_{ij} \log \frac{n_i n_j^P}{n_{ij}^2}$$

Por fim, outra métrica utilizada para avaliação externa é o *F measure*, que combina as idéias de precisão e cobertura da área de Recuperação de Informação. Para cada grupo j relacionado a uma classe i é calculado:

$$\text{Cobertura}(i, j) = n_{ij}/n_i$$

$$\text{Precisão}(i, j) = n_{ij}/n_j$$

onde n_{ij} é o número de membros da classe i no grupo j, n_j é o tamanho do grupo j e n_i é o número de membros da classe i.

O *F measure* do grupo j e da classe i é dado por:

$$F(i, j) = \frac{2 * \text{Cobertura}(i, j) * \text{Precisão}(i, j)}{\text{Cobertura}(i, j) + \text{Precisão}(i, j)}$$

Para uma hierarquia de grupos o *F measure* de uma classe é o valor máximo considerando qualquer nó da árvore e um valor total de *F measure* pode ser computado por uma média ponderada de todos os valores de *F measure*, como é dado por:

$$F = \sum_i \frac{n_i}{n} \max \left\{F\left(i, j\right)\right\}$$

onde n é o número de documentos.

2.3.3 Avaliação usando amostragem

Avaliação por amostragem faz uso de índices externos para obter uma avaliação interna do agrupamento. O procedimento geral segue os seguintes passos (Fridlyand & Dudoit, 2001):

1. Aleatoriamente dividir todo o conjunto em dois conjuntos disjuntos. Um conjunto de aprendizado L e um conjunto de teste T.
2. Aplicar o algoritmo de agrupamento ao conjunto L para obter a estrutura C^L.
3. Construir um classificador usando a estrutura C^L (o conjunto L e suas etiquetas).
4. Aplicar o classificador resultante ao conjunto de teste T.
5. Aplicar o mesmo algoritmo de agrupamento ao conjunto de teste T para obter a estrutura C^T.
6. Calcular um índice externo comparando os dois conjuntos de etiquetas para T obtido pelo agrupamento e predição (classificador em L).

A Figura 2.3, adaptada de Tibshirani et al. (2001), exemplifica a idéia da amostragem em conjunto com a predição para escolha do número de grupos. Aqueles com melhor pontuação do índice externo são ditos agrupamentos mais estáveis (Levine & Domany, 2001). É desta maneira que se pode usar este critério para escolhas em tempo de execução do agrupamento.

Uma maneira mais eficiente para o cálculo da estabilidade é o uso do mesmo processo enumerado anteriormente, sem o uso do classificador (Ben-Hur et al., 2002), ou seja:

1. Realizar a amostragem, admitindo conjuntos com elementos em comum.
2. Aplicar o agrupamento em ambos os conjuntos.
3. Calcular a intersecção dos dois conjuntos.
4. Usar índices externos para calcular a similaridade das estruturas inferidas para cada conjunto, considerando somente os pontos da intersecção.

É desta maneira que o conceito de estabilidade é usado em Niu et al. (2007) para agrupamento de documentos escolhendo simultaneamente o número de grupos e o subconjunto de características. Também é usado em Ben-Hur et al. (2002) para escolha do número de grupos em agrupamento hierárquico. Neste último caso, o problema se resume a escolher um valor k que corte o dendograma quando k grupos são produzidos. O valor de k é escolhido com a maximização da estabilidade.

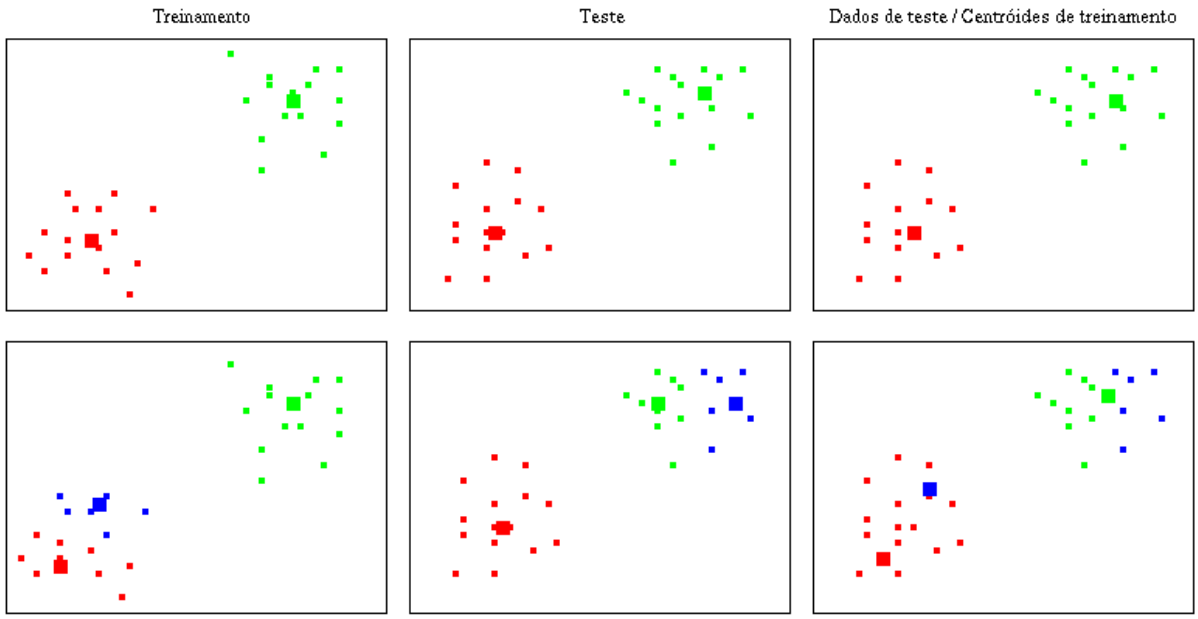


Figura 2.3: Uso de amostragem e predição para determinar estabilidade do agrupamento. Os dados estão em dois grupos bem separados. Na primeira linha, o algoritmo de agrupamento é executado para os dados de treinamento e teste com número de grupos igual a dois. Classificando os dados de teste usando os centróides dos dados de treinamento, os dois grupos são iguais aos obtidos nos dados de teste. O mesmo não acontece na segunda linha, com três grupos, pois quando os três centróides são usados para classificação, os resultados são bem diferentes.

2.4 Considerações finais

Neste capítulo foi introduzido agrupamento de documentos, explicando como representar documentos como vetores numéricos, foi apresentado o algoritmo *bisecting K-means* e como avaliar os resultados de agrupamento de dados. Nesta seção serão citados outros pontos que merecem destaque, que são os algoritmos incrementais e a atividade de etiquetagem dos grupos.

A capacidade de alocar de forma incremental novos documentos após a primeira execução do agrupamento é uma outra propriedade desejada para o algoritmo de agrupamento. Como na Internet existe uma grande quantidade de documentos sendo gerada a todo momento, algoritmos de agrupamento incrementais que realizem esta tarefa eficientemente são necessários. Uma maneira de tornar um algoritmo incremental é simplesmente adicionar os novos documentos ao grupo que proporciona melhor benefício a um critério interno, entretanto como o algoritmo de agrupamento não é executado novamente, a estrutura de grupos irá permanecer a mesma, o que pode não ser aceitável em contextos mais dinâmicos. Soluções para este problema formam um frutífero campo de pesquisa. O trabalho de Sahoo et al. (2006), por exemplo, utiliza o algoritmo ClassIt (adaptação do Cobweb para padrões com valores reais), uma solução já consolidada para realizar agrupamentos

hierárquicos incrementais, que é capaz de mudar a estrutura de grupos quando necessário.

Outro passo essencial ao agrupamento de documentos é uma maneira automática de descrição eficiente dos grupos gerados, ou seja, realizar a *etiquetagem* dos grupos. Ainda existem poucas soluções boas para este problema. O trabalho de Glover et al. (2002) mostra como uma abordagem simples como apresentar uma lista dos termos mais relevantes de cada grupo, ordenando os termos com o uso de algum cálculo estatístico, pode ser uma boa descrição do grupo, diferenciando o grupo de seus irmãos e pais na hierarquia de grupos. E esta é a mesma linha seguida pelo trabalho recente de Treeratpituk & Callan (2006).

É conhecido que documentos representados como vetores de termos possuem uma alta dimensionalidade, o que pode causar alguns problemas ao agrupamento. No próximo capítulo será apresentado a atividade de seleção de características, que aborda este problema.

Capítulo 3

Seleção de características

Seleção de características é a tarefa de desconsiderar termos irrelevantes e redundantes nos vetores de termos que representam os documentos, objetivando achar o menor subconjunto de termos que revelem grupos "naturais" dos documentos. Buscar o menor subconjunto possível irá tornar o tempo de computação da tarefa de agrupamento menor, ao mesmo tempo que evita a praga da dimensionalidade, melhorando a precisão.

Seleção de características é um problema intratável, pois o número de possibilidades de subconjuntos de características aumenta exponencialmente com o número de características e a própria concordância do que são grupos naturais é incerta. Dy & Brodley (2004) conclue que:

1. O número de classes não é necessariamente igual ao número de grupos gaussianos.

2. Diferentes subconjuntos de características revelam diferentes grupos.

Em agrupamento de documentos, geralmente o primeiro ponto é delegado ao usuário, através de uma estrutura hierárquica de documentos. O segundo ponto revela que da mesma maneira que o agrupamento é um problema intratável, seleção de características também o é, pois o número de possibilidades aumenta exponencialmente. Com d características, tem-se 2^d possibilidades de subconjuntos de características. Como as possibilidades de agrupamentos já é enorme, pelo número de Stirling, multiplicando-o por 2^d tem-se um problema ainda maior.

Apropriando os conceitos de Kohavi & John (1997) para aprendizagem não-supervisionada, pode-se classificar as características em dois níveis de relevância: fortes e fracas. Uma característica é fortemente relevante se sua remoção vai degradar o desempenho do agrupamento. Uma característica X é fracamente relevante se não é fortemente relevante e existe um subconjunto de características S, tal que a performance do agrupamento usando S é pior que a performance com $S \cup \{X\}$. Por sua vez, Yu & Liu (2004) propõe que as características fracamente relevantes são divididas em redundantes e não-redundantes. Uma característica é redundante com relação a outra se seus valores estão completamente

correlacionados, mas ainda é necessário definir quando uma característica é correlacionada com um subconjunto de características. A Figura 3.1 mostra como as características podem ser classificadas. As redundantes (ver Figura 3.2) e as irrelevantes (ver Figura 3.3) são as que devem ser eliminadas.

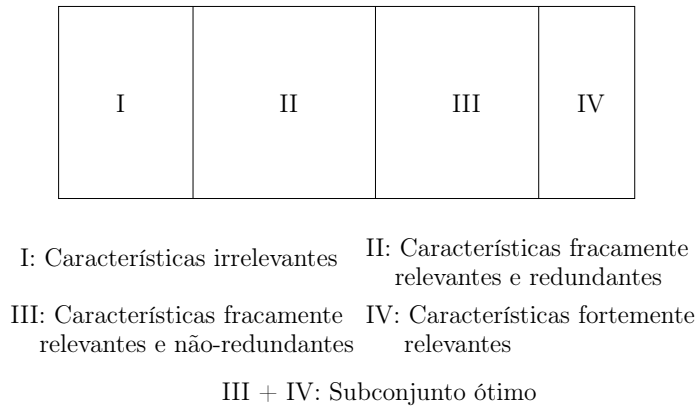


Figura 3.1: Classificação de características

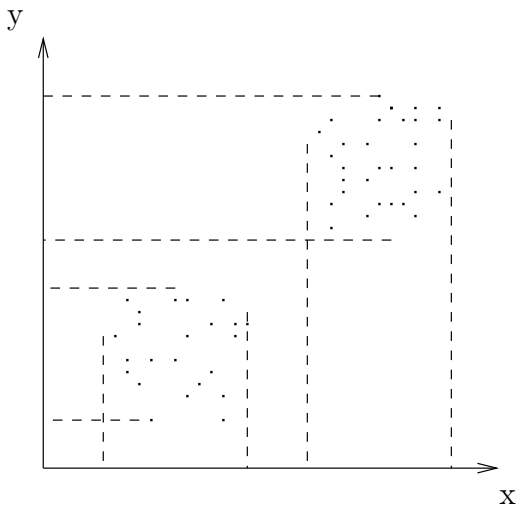


Figura 3.2: Características redundantes. As características x e y são redundantes, pois x provê a mesma informação que y para discriminar os dois grupos de pontos

De acordo com a interação do procedimento de seleção de características com o algoritmo de agrupamento, a seleção de características pode ocorrer de forma global ou local (Dash & Liu, 2000). A *seleção global de características* é o processo que seleciona uma vez as características e considera sempre as mesmas características no processo de descoberta dos grupos, sendo a forma mais pesquisada até então (Dash et al., 2002; Dy & Brodley, 2004; Law et al., 2004; Tang et al., 2005). As características, além de redundantes e irrelevantes, também podem somente atrapalhar o processo de agrupamento, quando diferentes subconjuntos de características revelam diferentes partições dos dados (Dy & Brodley, 2004). Isto motiva o uso da *seleção local de características*, onde um subconjunto

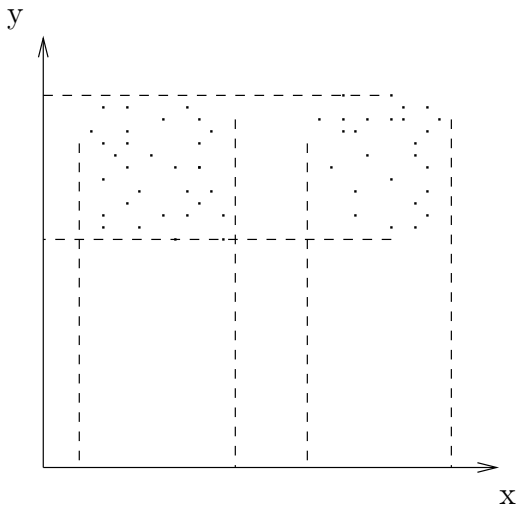


Figura 3.3: Característica irrelevante. A característica y é irrelevante, porque se omitido x, tem-se um único grupo, que não é um grupo "natural" dos dados

de características é escolhido para cada grupo, usando a suposição que cada grupo possui termos mais importantes, que ajudam a discriminar cada grupo dos demais grupos. A próxima seção aborda a seleção global de características, onde serão resumidos alguns métodos utilizados para seleção de características, que também são aproveitados para a arquitetura de seleção local de características.

3.1 Seleção global de características

Seleção global de características é o processo que seleciona uma vez as características e considera sempre as mesmas características no processo de descoberta de grupos. Isto ajuda a conhecer as características importantes antes de realizar o processo de agrupamento (Dash & Liu, 2000), tornando o processo mais eficiente e focado somente nas características relevantes.

Os métodos para seleção de características podem ser divididos em filtros ou *wrappers* (Dy & Brodley, 2004):

- Filtros: usam alguma propriedade intrínseca dos dados para selecionar características com o uso de um limiar, sem usar o algoritmo de agrupamento como forma de avaliação.

- *Wrappers*: aplicam o algoritmo de agrupamento para cada subconjunto de características selecionado no espaço de busca de características e então avaliam este subconjunto por funções de avaliação sobre o resultado do agrupamento. Na Seção 2.3, foram apresentadas diversas formas de efetuar esta avaliação.

3.1.1 Filtros

A maneira mais básica para realizar métodos filtros é usar alguma propriedade estatística dos dados para ordenar as características e escolher quantas são desejadas, com o uso de um limiar ou um número fixo de características desejadas. Métodos filtros são os mais utilizados para agrupamento de documentos pelo menor custo computacional em relação a um método *wrapper*. Alguns dos critérios já utilizados em agrupamento de documentos são citados a seguir (Liu et al., 2003; Tang et al., 2005):

- **Freqüência em documentos (DF)**

  O valor DF do termo t é definido como o número de documentos em que o termo t ocorre ao menos uma vez na coleção de documentos. Este simples método é conhecido como a base para seleção de características, pois geralmente é utilizado para pré-processamento dos dados.

- **Média do TF_IDF (TI)** (Tang et al., 2005)

  TI é um método proposto em (Tang et al., 2005), definido pela média do valor $tfidf_j$ para todos os documentos ($j = 1, \ldots, n$), considerando um termo t. Mostrou ter uma perfomance superior ao DF e similar ao TfV.

- **Variância de freqüência do termo (TfV)** (Dhillon et al., 2003)

  Sendo tf_j a freqüência do termo t no documento d_j, a qualidade do termo t é calculada por:

$$TfV_t = \sum_j^n tf_j^2 - \frac{1}{n}\left[\sum_j^n tf_j\right]^2$$

  O uso da variância como critério de relevância segue a mesma intuição do uso da média. Nos experimentos feitos em (Dhillon et al., 2003; Tang et al., 2005), o método TfV mostrou ser bem competitivo comparado com outros critérios de seleção de características, mantendo a precisão do processo de agrupamento com até 15% do número total de características presentes na coleção de documentos.

- **Intensidade do termo (TS)**

  É igual a probabilidade condicional que o termo ocorra na segunda metade dos pares de documentos onde a similaridade é maior que uma constante β, dado que ele ocorre na primeira metade. É representado por:

$$TS_t = p\left(t \in d_j | t \in d_i\right), \quad sim\left(d_i, d_j\right) > \beta$$

- **Similaridade entre características**

  O método proposto em Mitra et al. (2002) usa medida de similaridade entre duas variáveis aleatórias (as características) para eliminar características redundantes. A medida utilizada é o índice de compressão máximo de informação, calculado por:

$$\lambda_2(x, y) = \text{menor auto-valor de } \Sigma$$

  onde Σ é a matriz de covariância das variáveis aleatórias x e y. Então λ_2 é da forma:

$$2\lambda_2 = (var(x) + var(y)) - \sqrt{(var(x) + var(y))^2 - 4\, var(x)\, var(y)\left(1 - \rho(x, y)^2\right)}$$

  onde ρ é o coeficiente de correlação calculado por:

$$\rho(x, y) = \frac{cov(x, y)}{\sqrt{var(x)\, var(y)}}$$

  O valor de ρ é zero quando as características são linearmente dependentes e aumenta quando a depêndencia decresce. $cov(\cdot)$ é a covariância entre duas variáveis e $var(\cdot)$ é a variância de uma variável. λ_2 possui uma série de propriedades interessantes, tais como: simetria, sensibilidade a escala e invariância à rotação. Usando esta medida, Mitra et al. (2002) desenvolve um algoritmo de seleção de características que particiona as características em um certo número de grupos homogêneos e seleciona uma característica representativa de cada grupo.

3.1.2 *Wrappers*

Na Figura 3.4, é apresentada a idéia do método *wrapper* para seleção de características em aprendizagem não-supervisionada. É usada uma busca com um critério sobre o resultado do algoritmo de agrupamento, que atua simultaneamente como função de avaliação e heurística para guiar o processo de busca.

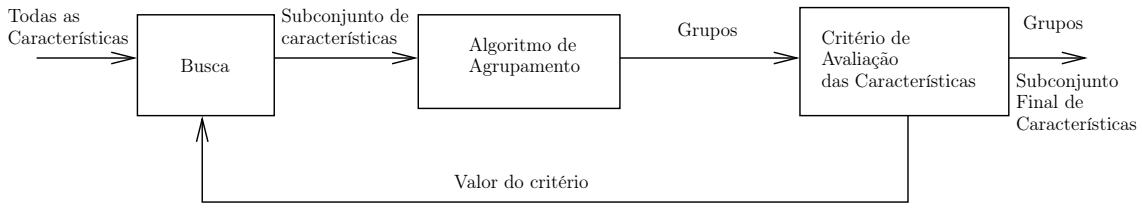


Figura 3.4: Abordagem *wrapper* para aprendizagem não-supervisionada

Através deste método, achar um subconjunto ótimo reflete o real benefício ao processo

de agrupamento, evitando um critério de corte arbitrário com o uso de um limiar (Law et al., 2002). No entanto algumas desvantagens e ponderações emergem desta abordagem:

- O espaço de busca é muito grande. É necessário um método de busca eficiente, como será mostrado posteriormente.
- É necessário a execução do algoritmo de agrupamento para avaliar cada subconjunto de características, tornando o processo muito custoso computacionalmente e não indicado para agrupamento *on-line* de documentos.
- O critério usado para validação do agrupamento não é um consenso, e várias medidas já foram propostas (ver seção 2.3).

Busca em espaço de características

O critério de validação do agrupamento nem sempre é monotônico com relação ao número de características utilizadas. Ou seja, adicionando uma característica, nem sempre o valor da função de critério é igual ou cresce. Então não se pode usar isto como base heurística para o método de busca. Esta não-monoticidade decorre da propriedade das características irrelevantes degradarem o resultado do agrupamento. Pudil et al. (1994) propõe seleção flutuante para tratar estes casos, no lugar de uma simples busca seqüencial que iria ser fadada a uma máximo local indesejável. Este método é um melhoramento do algoritmo *plus l - take away r* (Pudil et al., 1994), proposto para uso em critérios não-monotônicos, que consiste em avançar l passos no espaço de busca e retirar r elementos. No entanto, com $l > r$, esta busca se torna progressiva e com $l < r$, regressiva. No método flutuante seqüencial progressivo, após cada passo para frente é efetuado passos para trás até que o subconjunto de características seja melhor que os avaliados previamente. O algoritmo é apresentado no Algoritmo 1.

A forma regressiva deste algoritmo (quando se começa com todas as características) é obtida de maneira direta substituindo as inclusões por exclusões. As diferenças entre seleção progressiva ou eliminação regressiva fica por conta que o primeiro é mais rápido e o segundo é mais robusto por detectar mais facilmente características dependentes entre si (as fracamente relevantes) (Kohavi & John, 1997).

Como uma maneira de ordenação das características, necessário ao método de seleção flutuante, pode-se utilizar algum método filtro (Seção 3.1.1) como critério de ordenação. Portanto, neste sentido, o método *wrapper* pode ser visto como uma adição ao método filtro, evitando uma escolha arbitrária do número de termos selecionados com o uso de um limiar.

Outros métodos para realizar a busca no espaço de características são propostas, tais como algoritmos genéticos (Kim et al., 2000) e otimização por enxame de partículas (Wang

Algoritmo 1 Algoritmo de busca flutuante seqüencial progressiva

1: $Y = \{y_j \mid j = 1, ..., D\}$ {Todas as características
2: $X_k = \{x_j \mid j = 1, ..., k, \quad x_j \in Y\}, \quad k = 0, 1, ..., D$
3: $J(\bullet)$ {Critério de relevância da característica
4: $X_0 \leftarrow \emptyset$
5: $k \leftarrow 0$
6: **while** $k \neq$ número de características requeridas **do**
7: $x^+ \leftarrow \arg\max_{x \in Y - X_k} J(X_k + x)$ { A característica mais relevante ainda não pertencente a X_k
8: $X_{k+1} \leftarrow X_k + x^+$
9: $k \leftarrow k + 1$
10: $x^- \leftarrow \arg\max_{x \in X_k} J(X_k - x)$ { A característica menos relevante em X_k
11: **if** $J(X_k - \{x^-\}) > J(X_{k-1})$ **then**
12: $X_{k-1} \leftarrow X_k - x^-$
13: $k = k - 1$
14: Vá para linha 9
15: **else**
16: Vá para linha 6
17: **end if**
18: **end while**
return X_k

et al., 2007), também com bons resultados, mas sem maiores ganhos em relação à seleção flutuante.

Normalização do critério de avaliação em relação a dimensão dos dados

O uso direto dos critérios apresentados na Seção 2.3 para o *framework wrapper* de seleção de características não é adequado, pois um valor de critério como o traço da matriz $\left[S_W^{-1} S_B\right]$ varia em relação a dimensão dos dados, sendo necessário uma normalização (Dy & Brodley, 2004). Dado $CRIT(S_i, C_j)$ (função critério com a estrutura de grupos C_j, considerando as características S_i) e dois subconjuntos de características S_1 e S_2 (agrupando os dados usando S_1, obtém-se o resultado C_1 e agrupando usando S_2, obtém-se C_2), o valor normalizado do critério é dado por (Dy & Brodley, 2004):

$$ValorNormalizado(S_1, C_1) = NV(S_1, C_1) = CRIT(S_1, C_1) + CRIT(S_2, C_1)$$

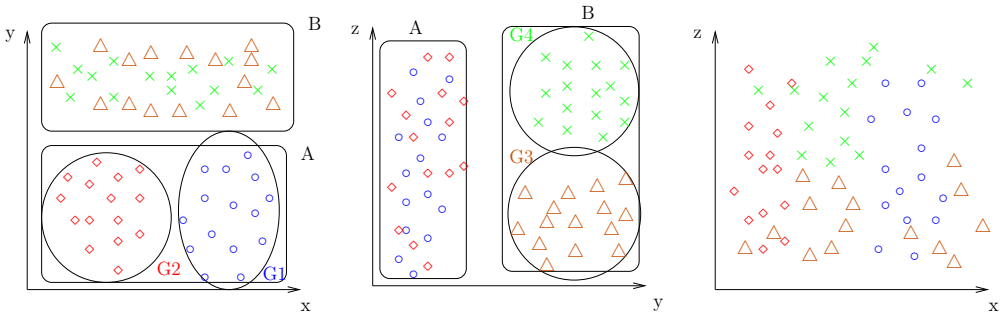
$$ValorNormalizado(S_2, C_2) = NV(S_2, C_2) = CRIT(S_1, C_2) + CRIT(S_2, C_2)$$

Se $NV(S_i, C_i) > NV(S_j, C_j)$, é escolhido o subconjunto de características S_i. Se $NV(S_i, C_i) = NV(S_j, C_j)$, é escolhido o menor subconjunto de características. Com isto a dimensionalidade não mais afetará o critério de seleção de características usado.

3.2 Seleção local de características

Sabe-se que diferentes subconjuntos de características podem revelar estruturas diferentes. Os trabalhos atacam este problema de uma maneira global (um subconjunto de características é escolhido para todos os grupos de documentos) ou de maneira localizada (um subconjunto de características é escolhido para cada grupo de documentos) (Dash et al., 2002).

A Figura 3.5 ilustra um conjunto de objetos pertencentes a quatro diferentes grupos, que são descritos pelas características x, y e z, em que se quer distinguir os grupos presentes nos dados com o uso de somente 2 características. Os grupos G1 e G2 são somente revelados quando os atributos x e y são considerados, ou seja, o atributo z é irrelevante para distingüir entre G1 e G2 (caso da Figura 3.5(a)). A Figura 3.5(b), por sua vez, mostra que as características y e z são relevantes para identificar os grupos G3 e G4, ou seja, a característica x é irrelevante neste contexto. Finalmente, os atributos x e z são um sub-conjunto irrelevante de características (Figure 3.5(c)). Nesta situação, qualquer sub-conjunto com duas características eventualmente retornado por um método global não será capaz de identificar os quatro grupos existentes. Também pode ser concluído que é necessário examinar uma característica no contexto de diferentes sub-conjuntos antes de afirmar que a característica é realmente irrelevante (Yu & Liu, 2004).



(a) em x e y (b) em y e z (c) em x e z

Figura 3.5: Dados dos grupos G1, G2, G3 e G4 para diferentes características.

Para solucionar questões neste sentido, pouca pesquisa já foi feita para a abordagem local de seleção de características em agrupamento tradicional de dados. O que já foi feito pode ser dividido em três linhas de pesquisa:

- O trabalho de Li et al. (2008) busca por vários subconjuntos de características e os resultados do algoritmo de agrupamento usando estas características são comparados. São escolhidos os melhores grupos formados baseado em uma função objetivo, que é um critério de avaliação interna do agrupamento.

- O agrupamento considerando diferentes subespaços de atributos para agrupamentos baseados em densidade. É o caso em que se particiona o espaço de características

em regiões, e se busca as células (hiper-retângulos) de alta densidade, que formarão os grupos (Agrawal et al., 2005; Parsons et al., 2004; Patrikainen & Meila, 2006).

- O agrupamento com a chamada seleção "glocal" de características, caso pesquisado até então somente para a classificação hierárquica de documentos (Esuli et al., 2008; Koller & Sahami, 1997). Neste caso, cada execução da tarefa de seleção de características se comporta de maneira global, pois a seleção de características é feita para cada nó interno da árvore, que geram nós filhos que foram divididos usando as características selecionadas para o nó pai. O fator local desta abordagem é o fato de que cada conjunto de nós irmãos da árvore foi gerado por subconjuntos de características diferentes.

3.2.1 Seleção local de características

O trabalho de Li et al. (2008), propõe um algoritmo com seleção local de características, ou seja, onde se busca grupos formados por diferentes subconjuntos de características a cada divisão de grupos. De forma geral, este algoritmo funciona da seguinte maneira:

1. Primeiramente, agrupa-se os dados em k grupos, baseando-se em todas as características.

2. Para cada grupo, acha-se o subconjunto de características relevantes considerando somente os pontos de cada grupo, podendo usar qualquer método proposto para seleção de características.

3. Usando os k subconjuntos de características encontrados no passo 2, agrupa-se todos os dados novamente, uma vez para cada subconjunto de característica, formando k estruturas de agrupamento.

4. Constrói-se k listas de grupos mais similares de uma estrutura para a outra (os que possuem mais pontos em comum), e é selecionado um grupo de cada lista de grupos baseado em um critério de avaliação normalizado.

Os grupos formados poderão possuir pontos repetidos (em mais de um grupo) e pontos não afetados para nenhum grupo, já que no passo 3, todos os dados são agrupados novamente. Li et al. (2008) propõe como critério usado no passo 4 para escolha dos melhores grupos, uma forma ajustada do valor normalizado apresentado na Seção 3.1.2. Conside que uma medida para pontos repetidos entre grupos é dada por:

$$O = \sum_{i \neq j}^{k} \frac{|C_i \cap C_j|}{\text{média}\left(|C_i|, |C_j|\right)}$$

onde C_i e C_j são dois grupos. Conside que outra medida para pontos não afetados a algum grupo é dada por:

$$U = \frac{n_u}{n}$$

onde n é o número total de pontos e n_u o número de pontos sem grupo. Estas duas medidas são usadas para criar uma penalidade ao valor normalizado do critério, ajustando-o, resultando nos valores:

$$ValorNormalizadoAjustado\left(C_i, S_i\right)|_C = ANV\left(C_i, S_i\right)|_C = NV\left(C_i, S_i\right)|_C \cdot \mathrm{e}^{-\alpha\Delta O - \beta\Delta U}$$

$$ValorNormalizadoAjustado\left(C'_i, S'_i\right)|_{C'} = ANV\left(C'_i, S'_i\right)|_{C'} = NV\left(C'_i, S'_i\right)|_{C'} \cdot \mathrm{e}^{\alpha\Delta O + \beta\Delta U}$$

Onde (C'_i, S'_i) é o grupo da estrututa de grupos C' com mais pontos em comum com o grupo (C_i, S_i) pertencente a estrutura C, ΔO e ΔU são as diferenças das medidas O e U de cada grupo, no caso em que o grupo (C_i, S_i) é trocado pelo grupo (C'_i, S'_i). α e β são duas constantes, onde valores altos respectivamente desencorajam a ocorrência de pontos repetidos e não afetados a algum grupo. Esses valores são atribuídos empiricamente, de acordo com a aplicação.

Quando dois grupos são comparados, se $ANV\left(C_i, S_i\right)|_C > ANV\left(C'_i, S'_i\right)|_{C'}$, é escolhido o grupo (C_i, S_i). Se $ANV\left(C_i, S_i\right)|_C = ANV\left(C'_i, S'_i\right)|_{C'}$, é escolhido o grupo com menor dimensionalidade.

Para afetar os pontos não-afetados no final do agrupamento, pode-se usar um critério de similaridade com os grupos, através de uma distância normalizada, dada por:

$$d = \frac{\left\| X_i|_{S_j} - \mu_j \right\|}{\sigma_j^2}$$

onde X_i é o ponto não-afetado, μ_j é o centróide do grupo C_j, S_j é o subconjunto de características de C_j e $X_i|_{S_j}$ é a projeção de X_i em S_j.

Com este método, os autores obtiveram uma melhora na precisão do algoritmo K-means agrupando dados do repositório do UCI (UC Irvine, 2007). Isto mostra que a pesquisa da seleção local de características é relevante.

3.2.2 Seleção "glocal" de características

A abordagem "glocal", como é chamada no trabalho de Esuli et al. (2008), foi primeiramente aplicada no trabalho de Koller & Sahami (1997) para classificação hierárquica de documentos. Nesta abordagem a atividade de classificação é dividida em um conjunto de sub-tarefas menores, onde cada uma corresponde a uma parte da hierarquia. A motivação é que cada sub-tarefa vai lidar com um conjunto menor de categorias, podendo tomar decisões de classificação com um subconjunto menor de características.

Dado qualquer documento, muitas das características são irrelevantes para a atividade de classificá-lo e somente servirão para confundir o classificador. Gerar um classificador para cada nó da hierarquia, considerando somente características relevantes à categoria relacionada ao nó, irá melhorar a capacidade de generalização de todo o processo de classificação, pois cada classificador estará focado para a parte da hierarquia em questão, onde o documento é testado e, se dada uma resposta negativa, não mais será testado para os filhos do classificador. Com isto, por exemplo, um documento sobre computadores provavelmente não irá encontrar um classificador usando a palavra "carne".

A seleção "glocal" de características é uma definição, até então, própria a atividade de classificação hierárquica, sendo uma forma intermediária entre a seleção global e local. Esta idéia será utilizada nesta dissertação para o caso de agrupamento divisivo, onde a cada nova divisão é realizada uma nova seleção de características. A seleção de características permanece global a cada divisão de grupo, mas a seleção passa a ser executada recursivamenente em cada nó interno formado. A intuição leva a entender, se cada grupo possui termos relevantes que ajudam a discriminá-lo, haverá uma melhoria na precisão. Isto será comprovado nos experimentos realizados nesta dissertação.

3.3 Considerações finais

Nesta seção foi motivada e abordada a atividade de seleção de características, com alguns métodos utilizados e suas classsificações. Foi visto que o algoritmo de seleção de características pode interagir com o algoritmo de agrupamento de maneira global, local ou de uma forma intermediária, chamada de "glocal".

No andamento do trabalho, foi vislumbrado que, até onde foi pesquisado no levantamento bibliográfico, a abordagem "glocal" para seleção de características não foi aplicada em nenhum trabalho de agrupamento hierárquico de documentos. Foi decidido então investigar seu uso para agrupamento de documentos. No próximo capítulo, será discutido o método proposto para tal e suas propriedades.

Capítulo 4

Seleção local de características para agrupamento de documentos

Como visto anteriormente, a seleção de características é um aspecto importante do processo de agrupamento. No contexto, geral de agrupamento, vimos que a seleção de característica pode ser dividida entre global e local. A seleção global no contexto de agrupamento foi investigada por diferentes autores. A seleção local, por outro lado, ainda foi pouco explorada tanto no contexto geral de agrupamento como no contexto específico de agrupamento de texto. Assim, nesse capítulo propomos um método de agrupamento hierárquico divisivo que irá explorar conceitos de seleção local de características.

Este capítulo é dividido em seção 4.1, com motivação e objetivos do uso da seleção local de características, em seção 4.2, com a descrição do método proposto, em seção 4.3 com apresentação de um método de escolha do número de grupos usando o conceito de estabilidade e em seção 4.4, com os aspectos de implementação, onde é feita uma análise de parâmetros e propriedades de um sistema de agrupamento de documentos.

4.1 Relevância

É conhecido que, para agrupamento de documentos, uma boa solução é aquela que organiza os documentos em uma hierarquia de grupos etiquetados, para que o usuário possa escolher quais grupos são relevantes e em qual nível da hierarquia. Esta organização tenta passar a noção de que documentos em nós-pais tratam de um assunto mais geral e os nós-filhos de casos específicos do mesmo assunto. Isto é desejado, mas não uma realidade, pois a variedade de grupos que podem ser gerados por conta da característica aleatória do algoritmo de agrupamento, o que pode ser uma vantagem para navegação entre documentos (Treeratpituk & Callan, 2006), nem sempre reflete uma relação pai-filho entre grupos. Isto acaba dificultando a atividade de etiquetagem, pois esta é tão boa quanto seja a coesão dos grupos.

Basicamente, uma forma de avaliar o uso de um algoritmo de agrupamento de documentos é a sua precisão, que é o resultado da comparação de um agrupamento com uma partição realizada por um especialista humano (que representa uma intuição da real distribuição dos documentos) e questões de escalabilidade, como custo de tempo de execução e uso de memória. A seleção global de características foi proposta neste sentido, pois é capaz de melhorar a precisão (atenuando a praga da dimensionalidade) e diminui o tempo de processamento, pois com menos características, menos processamento é necessário para realizar os cálculos de similaridade entre os documentos. No entanto, com a escolha de um único sub-conjunto de características em todo o processo, grupos coesos que possam ser identificados por outros subconjuntos deixam de ser revelados.

Para classificação hierárquica de documentos, no trabalho de (Koller & Sahami, 1997), é proposto uma abordagem que divide a atividade de classificação em problemas mais simples, um para cada nó da árvore de classificação. Então cada um desses problemas menores pode usar um subconjunto de características menor. A divisão da tarefa de categorização em sub-tarefas para cada nó da hierarquia torna capaz a realização da classificação com poucos termos relevantes de cada nó, e ainda com um melhor desempenho, pois os termos não-relevantes de cada nó somente "confundem" o classificador. Nesta dissertação, foi vislumbrado que este mesmo conceito pode ser apropriado para o agrupamento e foi elaborado um algoritmo de agrupamento hierárquico divisivo com seleção local de atributos. É esperado que a privilegiada visão global em algoritmos divisivos possa tomar proveito de uma visão local proporcionada pela seleção local de características. A variedade de subconjuntos de características selecionados a cada divisão dos grupos poderá revelar grupos ocultos por características ruidosas, com o objetivo de melhorar ainda mais a precisão em relação a seleção global. Também pode haver um ganho no custo computacional, pois a cada divisão um subconjunto diferente é escolhido e como os grupos vão diminuindo de tamanho, menos termos são necessários.

Na próxima seção o método proposto é apresentado, junto de uma variante, aqui batizada de método ZOOM-IN.

4.2 Método proposto

O algoritmo proposto para seleção de características local para agrupamento hierárquico divisivo consiste em adicionar um passo ao *bisecting K-means*:

1. Escolha um grupo para dividir, considerando um grupo inicial contendo todos os dados.

2. Faça a seleção de características do grupo escolhido, usando como critério a escolha de n termos ou de um limiar τ **(passo de seleção local)**.

3. Construa 2 sub-grupos usando o algoritmo *K-means* (**passo de bisecção**).

4. Repita o passo 3 por $ITER$ vezes e fique com a partição com melhor valor do critério interno.

5. Repita os passos 1, 2, 3 e 4 até que o número de grupos requerido seja alcançado.

Este algoritmo pode ser utilizado para resolver o problema da Figura 3.5, simplesmente começando por um subconjunto de características que melhor revela grupos nos dados (pode ser o caso da Figura 3.5(a)) e particionando todos os dados em grupo A (constituem os dados de G1 e G2) e grupo B (os dados de G3 e G4). Realizado uma nova seleção de características para os dados de cada grupo, o grupo A permanece com as características x e y e grupo B com y e z. Os grupos A e B agora podem ser facilmente discriminados em G1 e G2 (filhos do grupo A), G3 e G4 (filhos do grupo B), revelando assim todos os grupos naturais presentes para estes dados.

A medida que o algoritmo é executado, os grupos criados são cada vez menores e, com isto, o número de termos distintos presentes nos documentos também diminui. Desta forma, um critério constante de escolha do número de termos a cada escolha do grupo a ser particionado tende a perder seu potencial seletivo, pois o número de termos selecionados irá se aproximar do número de termos distintos. Uma alternativa é escolher um número pequeno de termos, mas assim se perde informação necessária quando os grupos ainda são grandes. A saída é usar um limiar normalizado ou uma escolha do número de termos variável de acordo com o tamanho dos grupos ou o número de termos distintos. De maneira simples, neste trabalho é proposto a escolha do número de termos n_i para o grupo i igual a:

$$n_i = \left\lfloor \frac{N_T}{N_C} \cdot m_i \right\rfloor$$

onde N_T é o número de termos distintos em toda coleção de documentos, N_C é o tamanho da coleção de documentos e m_i é o tamanho do grupo i. N_T/N_C é a proporção de termos revelados distintos em cada documento da coleção. Com isto o número de termos selecionados diminue a cada divisão de grupo, lembrando o ajuste de um binóculo. Este método será referido neste trabalho como método ZOOM-IN.

4.2.1 Análise da complexidade do algoritmo

Considerando que todas as operações básicas consomem a mesma unidade de tempo t, a complexidade de tempo do *bisecting K-means* com seleção local de características é dada por:

$$T_{BK} = (f(\bar{n_j}, 2, \bar{d}) + 2\bar{d})\bar{L}I(K-1) + S(K-1)t$$

onde $\bar{n_j}$ é a média de tamanho dos grupos escolhidos em cada passo da bisecção, $\bar{d}$ é a média do tamanho da dimensão dos dados (escolhido em cada execução do passo de bisecção), $\bar{L}$ é a média do número de *loops* do K-means tradicional para cada execução do passo de bisecção, I é o número de iterações para cada passo de bisecção, S é definido segundo o método de seleção de características utilizado e K é o número de grupos desejado. A função f define o número de operações necessárias para o K-means tradicional afetar pontos ao centróide mais próximo e $2\bar{d}$ é o número de operações necessárias para recalcular os centróides. Em f, os três parâmetros equivalem respectivamente ao tamanho dos dados, número de grupos desejado e tamanho da dimensão dos dados. Com o uso da função de similaridade cosseno, esta função é expressa por (Li & Chung, 2007):

$$f(n, k, d) = 2nkd + nk + nd$$

Para o *bisecting K-means* com seleção local de características, K é sempre igual a 2, d passa a ser a média de dimensões escolhidas antes de cada passo de bisecção (é o $\bar{d}$ da primeira fórmula) e n passa a ser a média do tamanho dos grupos escolhidos para serem submetidos à bisecção (é o $\bar{n_j}$ da primeira fórmula). O que muda em relação a complexidade do *bisecting K-means*, sem seleção local de características, é a adição de $S(K-1)$ e o parâmetro d é calculado pela média das dimensões, passando a ser representado por $\bar{d}$. Como para grupos menores a quantidade de termos selecionados tende a ser também menor, como é o caso da seleção utilizando o ZOOM-IN, é esperado que $\bar{d}$ sempre seja significantemente menor que a quantidade total de termos d. Com isto existe um ganho de tempo em relação ao *bisecting K-means* sem seleção local, que é contrabalanceado pela parcela $S(K-1)$. Se o método usado para seleção de características não for demasiadamente custoso, o tempo gasto pode ser igual ou até melhor em relação ao *bisecting K-means* tradicional, e ainda é esperado um ganho na precisão do agrupamento.

4.3 Escolha do número de grupos usando o conceito de estabilidade

O algoritmo *bisecting K-means* não prevê a escolha do número de grupos e se limita a dividir todos os documentos sempre em dois grupos, cada vez menores, resultanto em uma hierarquia de nós internos sempre com dois nós-filhos. Esta não é uma forma proveitosa de navegação, em um primeiro momento, e é necessário uma maneira para se inferir o número de grupos.

No caso de agrupamento hierárquico, a escolha do número de grupos geralmente é feita, dada uma árvore que representa a hierarquia de grupos, com um corte na árvore, onde K grupos são gerados. Com o uso do *bisecting K-means*, um corte para cada nível

da árvore nem sempre vai resultar em um número de grupos que cresça de 1 em 1, por exemplo, para os cortes dos níveis (1,2,3,4) em uma árvore binária completa, resultam (1,2,4,8) grupos. Para conseguir uma correspondência 1 em 1, uma forma é guardar na execução do *bisecting K-means* uma lista com a seqüência de todas as escolhas de grupos divididos. Desta forma, caso seja necessário selecionar K grupos, basta consultar esta seqüência de grupos e ficar com os grupos que aparecem até a posição K da lista.

Com a seleção local de características, os pares de grupos-irmãos da hierarquia são formados por características diferentes selecionadas pela análise do grupo-pai, portanto, uma vez definida toda a hierarquia, os vetores de características não podem ser diretamente utilizados para definir um critério interno para otimização da escolha do número de grupos, que é a forma utilizada por boa parte dos trabalhos na área (Jung et al., 2003). Uma maneira de obter um critério interno sem o uso dos vetores de características, é o uso do conceito de estabilidade, apresentado na seção 2.3.3. Isto é proposto no trabalho de Ben-Hur et al. (2002), com o algoritmo:

1. Faça duas amostragens com uma fração α dos dados X.
2. Faça o agrupamento em K grupos para cada amostra.
3. Calcule a similaridade, através de um critério externo, dos pontos comuns a ambos os agrupamentos resultantes das amostras.
4. Repita I vezes os passos 1, 2 e 3 e calcule a média das similaridades obtidas no passo 3.
5. Repita os passos 1, 2, 3 e 4 para uma faixa de número de grupos $K = 2$ até $K = K_{max}$.

A medida de similaridade usada no passo 3 pode ser qualquer uma das apresentadas na seção 2.3.2. Para simplificar o processo, pode-se usar uma única amostragem com um parâmetro real α, que varia de 0 a 1, que indica a razão de padrões dos dados originais que serão utilizados na amostragem. Desta forma, os pontos comuns do algoritmo anterior podem ser vistos como a própria amostragem. A média dos valores de similaridade calculados no passo 4 equivale a um critério interno, utilizado para inferir o número de grupos.

Existe um consenso na literatura estatística, que os melhores valores de K para o número de grupos são aqueles onde ocorrem "cotovelos" no gráfico da relação número de grupos por um valor de critério interno, como pode ser visto na Figura 4.1. O motivo para a identificação de vários cotovelos é que de fato um conjunto de dados pode ser dividido em várias possíveis configurações de número de grupos, pois a definição do que é um grupo é imprecisa (ver Figura 4.2). Para identificar estes cotovelos no gráfico, como o valor deste

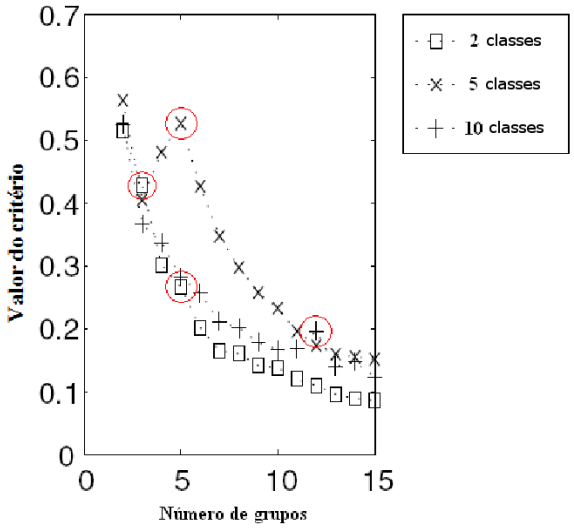


Figura 4.1: Gráfico do número de grupos pelo valor de critério interno. Figura extraída do trabalho de Niu et al. (2007). Os círculos vermelhos indicam possíveis "cotovelos" no gráfico, que são pontos do gráfico que formam uma angulosidade.

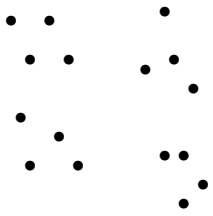


Figura 4.2: Pontos que podem ser separados em 2, 3 ou 4 grupos

critério tende a decrescer a medida que o número de grupos aumenta, não basta escolher o valor máximo do critério, que geralmente vai ocorrer para valores mínimos de K. É necessário alguma forma de normalização do critério ou de análise dos "cotovelos". Para normalização, o trabalho de Niu et al. (2007) propõe o uso de um preditor aleatório de K grupos, que afete cada documento do conjunto original e use este resultado do preditor para classificar cada documento da amostragem, calculando outro valor de similaridade com os resultados deste preditor. A subtração deste valor do valor obtido pela amostragem usando agrupamento vai gerar o valor normalizado que vai ser usado para avaliar o mérito da escolha de K. Outra forma que pode ser utilizada é simplesmente analisar o gráfico das possíveis escolhas do número de grupos pelo valor do critério sem normalização e inferir os possíveis "cotovelos", o que evita o custo de uso do preditor aleatório (além do seu uso ser inapropriado para agrupamentos com dados definidos em dimensões diferentes). Um algoritmo simples para realização desta análise, elaborado nesta dissertação, é:

1. Para todos os conjuntos de 2 em 2 valores seguidos do critério (aqui referidos como "pontos"), calcular a maior de todas as distâncias de cada ponto para outro ponto adjacente.

2. Dos pontos com maior distância entre si, escolher como cotovelo o de valor máximo.

É esperado que pontos distantes entre si sejam os pontos onde ocorrem as maiores angulosidades, e, destes pontos, o ponto com maior valor seja um possível "cotovelo". Com um número K inferido, tendo um dendograma já construído com uma estrutura de grupos básica, que é o caso do resultado do *bisecting K-means*, basta cortar a árvore onde é formado K grupos ou cortar usando a lista de seqüência de partição dos grupos, formando uma nova hierarquia de grupos. Mas para os grupos escolhidos ainda é possível a escolha do número de grupos para as sub-árvores formadas, até que um critério de parada seja alcançado. Um algoritmo simples é da forma:

1. Empilhar todos os grupos formados pela escolha de K grupos, que não sejam grupos com um tamanho menor que um valor t definido.

2. Enquanto a pilha não for vazia, faça:

   (a) Desempilhe um grupo G, itere por todas as escolhas possíveis de número de grupos que sejam filhos de G.

   (b) Escolha a melhor alternativa de K filhos de G e empilhe os novos grupos gerados, que tenham um tamanho maior que t, também os colocando na árvore que forma a nova hierarquia de grupos.

Além deste método de escolha do número de grupos por amostragem proporcionar bons resultados (Ben-Hur et al., 2002), este é um método que não demanda grande custo computacional, pois se resume a um passo adicional ao agrupamento, usando os mesmos grupos do resultado do agrupamento, mudando somente a organização de sua estrutura baseado em uma medida que não faz uso do cálculo de similaridades entre padrões, tornando este método apropriado para uso junto da seleção local de características.

Na próxima seção, serão apresentados alguns aspectos de implementação de um sistema de agrupamento de documentos, prevendo o uso de seleção local de características e escolha do número de grupos.

4.4 Aspectos de implementação

Nesta seção, serão resumidas algumas observações necessárias na implementação de um sistema de agrupamento de documentos com seleção local de características, conhecimento este obtido na implementação do sistema para realização dos experimentos desta dissertação. A seção será dividida em subseções sobre observações das fases de pré-processamento do texto e do algoritmo de agrupamento.

4.4.1 Pré-processamento do texto

Em geral, documentos são representados por conjuntos de termos de indexação, com valores númericos associados. O procedimento de cálculo destes valores mais utilizado é o TF-IDF, como já apresentado na seção 2.1. Para o cálculo destes valores, é necessário considerar a eliminação de *stopwords* e *stemming*.

Stopwords são termos que usualmente ocorrem em grande parte dos documentos e são geralmente constituídos por artigos, preposições e conjunções. A eliminação de *stopwords* é necessária pois estes termos não são úteis para a atividade de recuperação de informação já que não servem para discriminar documentos.

O processo chamado de *stemming* é o de extração do radical das variantes de uma mesma palavra, assim as variantes passam a ser representadas por um mesmo termo. O algoritmo mais utilizado para realização de *stemming* é o de Porter (Porter, 1997), que usa uma lista de sufixos e aplica uma série de regras para extrair os sufixos das palavras. Em alguns sistemas de agrupamento de documentos que buscam uma melhor formulação das etiquetas dadas aos grupos, o processo de *stemming* não é realizado (Treeratpituk & Callan, 2006). Assim as variantes do mesmo termo são preservadas.

Para formar etiquetas de texto também é preferencialmente usado, na fase de representação, a identificação de *grupos nominais*. Por exemplo, em uma coleção de documentos sobre Inteligência Artificial, é esperado que uma etiqueta de um grupo de documentos venha a ser "Redes Neurais", algo impossível em modelos de representação que usem somente arranjos com uma única palavra.

Vale ressaltar que cada língua possui suas próprias *stopwords* e regras de *stemming*. Estes são os únicos fatores de um sistema de agrupamento de documentos que mudam de língua para língua, pois o resto do sistema somente processa vetores numéricos.

Outro passo muito usado no pré-processamento do texto é a desconsideração de termos que não ocorram em pelo menos uma pequena porcentagem dos documentos da coleção, que da mesma maneira que as *stopwords*, não são úteis para o propósito de recuperação de informação.

4.4.2 O algoritmo de agrupamento

A primeira preocupação básica na implementação de um algoritmo de agrupamento é qual será a estrutura que armazena os dados dos documentos. Mais de uma cópia de toda a estrutura na memória deve ser evitada, e com esta preocupação é necessário preparar a estrutura para que ela preveja as possibilidades de cópia presentes no algoritmo de agrupamento. Por exemplo, quando se for efetuar a seleção local de características, basta marcar cada termo dos documentos do grupo a ser dividido com "*flags*" booleanas e especializar a função de similaridade para que ela atente a esta *flag* e condicione o cálculo de similaridades para valores em que a *flag* marca verdadeiro. Isto evita a cópia de valores

da estrutura de documentos para a formação de uma nova estrutura.

Outra especialização necessária à função de similaridade é prever a ocorrência de vetores de documentos que somente possuam valores nulos, pois com a seleção local, especialmente em casos onde são selecionados poucos termos, alguns documentos poderão ter selecionados somente termos com valores nulos. Isto levará o denominador da medida de similaridade cosseno a ter valor nulo, o que indica que não se tem informação sobre ou um ou os dois documentos envolvidos na comparação. Uma solução seria ignorar tal documento no processo de agrupamento ou mantê-lo no grupo-pai onde ocorreu a seleção de somente valores nulos.

Para se obter um menor tempo de processamento o algoritmo *bisecting K-means* pode ser paralelizado em vários processadores. Um algoritmo simples de paralelização, proposto em Li & Chung (2007) (aqui é adicionado um passo para seleção local de características) é:

1. Distribua n pontos de dados para p processadores.
2. Selecione um grupo C_j para dividir, faça a seleção de características para os pontos do grupo C_j e notifique estas informações para todos os processadores.
3. Crie 2 sub-grupos de C_j usando o algoritmo K-means (passo de bisecção):
   (a) Selecione 2 pontos de C_j como centróides iniciais e notifique isto para os p_j processadores que possuam pontos membros de C_j;
   (b) Cada processador deve calcular a função de critério do agrupamento para seus pontos de C_j com os 2 centróides conhecidos, afetando cada ponto para a melhor escolha.
   (c) Colete toda a informação necessária para atualizar 2 centróides e notifique isto para os p_j processadores que participaram deste passo de bisecção.
   (d) Repita os passos (b) e (c) até a convergência.
4. Repita os passos 2 e 3 I vezes e selecione a partição que produz os grupos que melhor satisfaçam a função global de avaliação.
5. Repita os passos 2, 3 e 4 até que K grupos sejam obtidos.

Este algoritmo, com a adição de um passo de predição, apresentado na trabalho de Li & Chung (2007), que visa melhorar a carga de trabalho para cada processador, mostra, em experimentos realizados no mesmo trabalho, possuir um *speed-up* (razão do tempo de execução do algoritmo seqüencial com o paralelo) linear para o número de processadores e o número de pontos.

4.5 Considerações finais

Neste capítulo foi motivado e proposto o uso da seleção local de características para o problema de agrupamento de documentos, sendo feita uma análise do que esta escolha acarreta ao resultado do algoritmo, tais como a necessidade de um critério apropriado para otimização do número de grupos. Na seção 4.4, foram apresentados alguns aspectos que podem ser levados em conta na implementação de um sistema que use os conceitos aqui apresentados.

No próximo capítulo, serão apresentados e discutidos os resultados dos experimentos realizados com o método e será apresentada uma aplicação do método para agrupar resultados de uma consulta do usuário à ferramenta de busca Google.

Capítulo 5

Experimentos e aplicação

Neste capítulo, na seção 5.1, serão apresentados os experimentos com as bases de documentos Reuters-21578 (Lewis, 1999) e RCV2 (Lewis et al., 2004), para avaliar o método de seleção local apresentado na seção 4.2 com os critérios de ordenação de características TfV, DF e TI. Também é apresentado, na seção 5.2, uma aplicação do algoritmo de agrupamento para o caso de organização do resultado de uma consulta ao Google (Google, 2008), com o objetivo de testar o desempenho da aplicação dos métodos de agrupamento discutidos na dissertação com seleção global e local de características e escolha do número de grupos usando o conceito de estabilidade.

5.1 Experimentos com bases de documentos

5.1.1 Descrição

Para realização dos experimentos foram preparados dois subconjuntos das bases. São bases já bastante usadas pela comunidade científica por possuir uma grande quantidade de documentos já classificados. Para a base Reuters, foi considerado como classes os assuntos marcados como TOPICS. Destas classes, foram selecionadas as que possuiam no mínimo 10 documentos e aleatoriamente selecionados no máximo 30 documentos para cada classe, o que resultou em 1228 documentos selecionados em 42 classes. Para a base RCV2 também foram selecionadas as classes com no mínimo 10 documentos e aletoriamente selecionados no máximo 100 documentos para cada classe, o que resultou em 2023 documentos particionados em 23 classes. A base RCV2 fica portanto com uma maior quantidade de documentos por classe. Quanto aos assuntos abordados pelos documentos nas bases, a base Reuters geralmente trata de assuntos relacionados a economia, tais como dinheiro, exportações, petróleo, etc. Já a base RCV2 trata de assuntos gerais e variados, como esportes, religião e política.

A razão, neste trabalho, do uso de um subconjunto de cada base é pelo fato das bases possuirem muitos documentos (Reuters com 20 mil e RCV2 com 800 mil), o que torna

complicado o uso de todos os documentos, pois um computador caseiro não possui recursos de memória e processamento para executar os algoritmos em um tempo praticável. As bases estão disponíveis da maneira como foram processadas neste trabalho, em formato SQL, na página `http://www.cin.ufpe.br/~mnr`.

Os resultados foram comparados usando a micro-média de precisão, também utilizada no trabalho de (Slonim et al., 2002). Um grupo tem uma boa precisão em relação a uma classe, se no grupo a proporção de documentos da classe é maior que a de outras classes. A micro-média de precisão assume que cada grupo formado pelo algoritmo de agrupamento possui uma classe c representante que é majoritária. Considerando T o conjunto de grupos e C o conjunto de classes, a micro-média de precisão é dada por (Slonim et al., 2002):

$$P(T) = \frac{\sum_{c \in C} \alpha(c, T)}{\sum_{c \in C} \alpha(c, T) + \beta(c, T)}$$

onde $\alpha(c, T)$ é o número de documentos corretamente afetados para c e $\beta(c, T)$ é o número de documentos incorretamente afetados para c.

Os métodos comparados serão o *bisecting K-means* com seleção global para uma faixa de vários valores do número de termos selecionados, o *bisecting K-means* com seleção local para a mesma faixa de termos selecionados e o *bisecting K-means* com escolha do número de termos usando ZOOM-IN.

Para avaliação do agrupamento hierárquico, os grupos considerados são aqueles presentes nas folhas do dendograma, que com o uso do *bisecting K-means* é garantido que todos os documentos estão em algum grupo-folha do dendograma. Os números de grupos especificados para execução dos algoritmos foi igual ao número de classes da base. O número ITER do algoritmo *bisecting K-means* foi atribuído com valor 5 e todos os valores de precisão apresentados neste trabalho são a média de 30 execuções dos algoritmos. Foi utilizado a remoção de *stopwords* (preposições e palavras comuns), uso de *stemming* com o algoritmo de Porter (Oleander Solutions, n.d.) e remoção de termos que ocorrem menos que 5 vezes para cada base.

Na próxima seção, são apresentados e discutidos os resultados obtidos.

5.1.2 Resultados e discussão

Na Figura 5.1, são apresentados os resultados obtidos para a base Reuters utilizando *bisecting K-means* com seleção de características global, que é a maneira tradicional, e local, utilizando o método proposto, com o número de características constante, para os critérios de ordenação DF, TfV e TI.

Para esta base, todos os métodos globais conseguem manter a precisão quando até uma certa porcentagem dos termos é desconsiderada. Para poucos termos, a precisão cai e o critério de ordenação que melhor mantém a precisão é o TfV, com desempenho

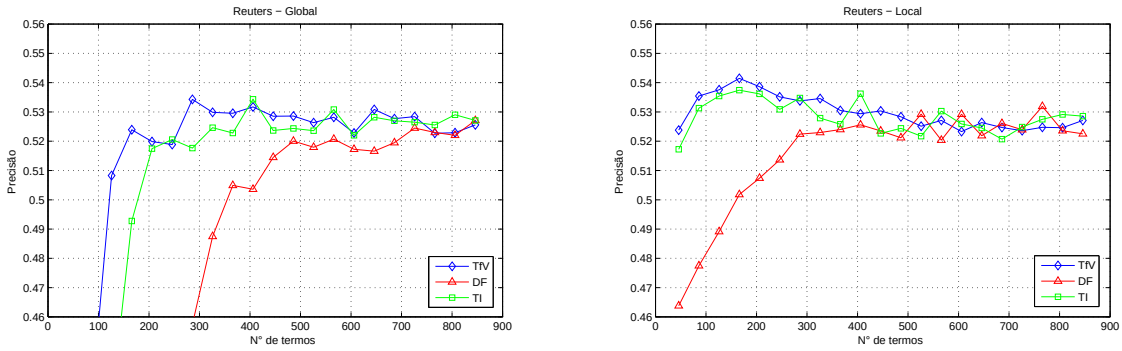


Figura 5.1: Micro-média de precisão em relação ao número de termos utilizados para a base Reuters, com seleção de características global e local

similar ao TI e melhor que o DF, tal como já observado no trabalho de Tang et al. (2005). Pode-se concluir que para poucos termos, existe pouca informação sobre os documentos, o que deteriora a precisão do agrupamento.

Por outro lado, a abordagem local consegue manter a precisão até com uma quantidade de termos muito pequena, com exceção do método DF. E acontece um fenômeno interessante para um número relativamente pequeno de termos, onde a precisão aumenta. Isto se deve ao fato que para grandes valores do número de termos selecionados, o potencial seletivo (capacidade de selecionar somente termos relevantes) diminui localmente, pois de acordo com que os grupos vão ficando menores, existem menos termos distintos. Com um valor pequeno de termos selecionados, o potencial seletivo consegue se manter durante as divisões dos grupos e, de maneira interessante, a precisão melhora, o que indica que a seleção local ajuda para que as divisões dos grupos não venha a separar grupos coesos.

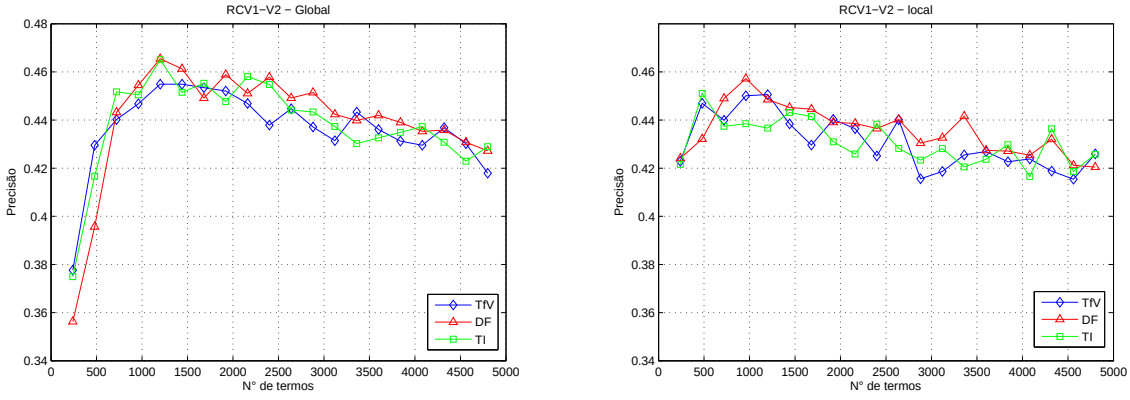


Figura 5.2: Micro-média de precisão em relação ao número de termos utilizados para a base RCV2, com seleção de características global e local

Os resultados com a base RCV2 mostram outro cenário (ver Figura 5.2). Nesta base, o método DF (talvez a forma mais simples para realizar seleção de características), foi mais competitivo com a seleção global de aproximadamente 1200 termos de 4800. Para uma quantidade ainda menor de termos ele tem um desempenho pior em relação ao TfV

e TI. Ou seja, o TfV mostra novamente que é capaz de manter a precisão para uma quantidade muito pequena de termos, mas desta vez foi o DF e o TI que conseguiram os melhores valores de precisão com seleção global. Também se percebe que a seleção global proporciona uma ganho significativo na precisão, algo que não ocorre na base Reuters. Isto indica a existência, nesta base, de uma grande quantidade de termos irrelevantes, que confundem o algoritmo de agrupamento. Como a base Reuters possui documentos de um único tópico geral, dividido em classes específicas, e por conta de outros fatores próprios a base (como a forma de redação dos documentos), uma proporção menor de termos são irrelevantes. Em todo caso, a principal conclusão destes resultados com seleção global é que para classificar documentos, bastam relativamente poucos termos relevantes.

Para a seleção local na base RCV2 acontece o mesmo fenômeno visto na base Reuters, de melhora da precisão quando poucos termos são selecionados. Mas desta vez, a seleção local não foi capaz de proporcionar melhores resultados em relação à seleção global. Uma possível explicação é que com uma proporção tão grande de termos irrelevantes (cerca 70% dos termos são irrelevantes) e com um valor absoluto de termos relevantes para todos os grupos também grande (cerca 1200 termos possuem alguma relevância para diferenciar todos os grupos), um critério de escolha constante do número de termos não consegue balancear o nível de relevância dos termos escolhidos para a divisão de cada grupo. Ora são muitos termos irrelevantes selecionados, ora se tem pouca informação. Para seleção local, o melhor valor constante selecionado ficaria numa faixa um pouco menor que o melhor valor de termos relevantes para todos os grupos, e é o que ocorre na base RCV2, onde os melhores valores de precisão se encontram numa faixa de valores um pouco menor que 1200 termos.

É necessário selecionar os termos que refletem um real benefício ao agrupamento, função bem desempenhada por métodos *wrappers*. Como o custo computacional do uso de *wrappers* pode não ser admissível para agrupamento *on-line* de textos, o uso de uma quantidade variável de termos selecionados localmente, que é o proposto método ZOOM-IN, objetiva manter o mesmo potencial seletivo durante o agrupamento.

A Tabela 5.1 apresenta na coluna 2 o resultado sem uso de seleção de características, na coluna 3 o melhor valor de precisão usando seleção local, na coluna 4 o melhor valor de precisão usando seleção global, e nas colunas 5, 6 e 7 os resultados com o procedimento ZOOM-IN para cada método de ordenação de características. Nos resultados com o ZOOM-IN, é comprovado um bom desempenho desta abordagem para a base Reuters, com valores de precisão compatíveis com a seleção local com número de termos constante. O que mais uma vez não ocorre com a base RCV2, o que mostra que uma heurística para escolha do número de termos baseado no tamanho dos grupos não é capaz de proporcionar bom desempenho quando a proporção de termos irrelevantes na base de documentos é muito grande.

A escolha do número de termos para seleção local de características levando em conta

Tabela 5.1: Micro-médias de precisão sem uso de seleção, com uso de seleção global e local com valor constante, e com o método ZOOM-IN para os critérios TfV, DF e TI

Base	**Sem método**	**Local**	**Global**	**TfV**	**DF**	**TI**
Reuters	0,527117	0,541477	0,534311	0,540988	0,527362	0,541395
RCV2	0,42722	0,457242	0,465447	0,436794	0,450338	0,440715

o tamanho dos grupos não é uma boa heurística para todos os contextos. Com proporções variadas de termos distintos revelantes ou irrelavantes, este método pode não funcionar bem. Como hipótese levantada, é esperado que o método ZOOM-IN funcione bem quando, na seleção prévia a atividade de divisão dos grupos, tenha sido eliminado a maior parte dos termos irrelevantes para discriminar todos os documentos. Outro experimento foi realizado com a base RCV2, agora fazendo uma seleção global prévia a atividade de agrupamento, com a consideração somente de termos que ocorrerem mais que 30 vezes em toda base, resultando em torno de 1500 termos selecionados, valor próximo ao ponto de máximo obtido com a seleção global. Com o uso do ZOOM-IN, para os métodos TfV, DF e TI, os respectivos valores de precisão obtidos foram 0,451326, 0,448707 e 0,451936. São valores, com exceção do DF, maiores que os obtidos com o método ZOOM-IN na base RCV2, mas ainda menores que valores obtidos com o uso de seleção local constante e seleção global. O método ZOOM-IN mostrou que é capaz de melhorar a precisão em relação a seleção global, e para isto é preciso assegurar que os termos selecionados, a cada passo de divisão, sejam de fato os mais relevantes, o que estimula a pesquisa de métodos *wrappers* de seleção de características.

Mesmo com os problemas apresentados, o método ZOOM-IN serve para o propósito, de demonstrar nos experimentos realizados, que poucos termos são necessários para discriminar grupos pequenos de documentos.

Na próxima seção, é apresentado um sistema de agrupamento de documentos retornados por uma pesquisa ao Google, desenvolvido como um exemplo de aplicação dos métodos discutidos nesta dissertação.

5.2 Agrupamento de resultados do Google

5.2.1 Descrição

A título de demonstração da aplicação dos algoritmos discutidos durante esta dissertação, foi desenvolvida uma biblioteca para agrupamento de documentos retornados por uma consulta ao Google. Escrita em `C++`, inspirada na biblioteca CGoogle (Konen, 2005), esta dependente da plataforma MFC da Microsoft, a biblioteca aqui desenvolvida não é dependente da plataforma MFC e faz uso da biblioteca `cURL` (cURL, 2008) para acesso

ao protocolo HTTP. A biblioteca é basicamente um *parser* dos arquivos em formato HTML retornados pela pesquisa ao Google, inserindo o título, *url* e um breve trecho do documento retornado pelo Google (conhecido como *snippet*) a uma estrutura de vetor. Este vetor é utilizado para representação dos documentos, usando novamente a `cURL` para carregar as páginas de todas as *urls* no vetor. O resultado final é o conhecido formato de representação de documentos em vetores numéricos que serão submetidos aos algoritmos de agrupamento e escolha do número de grupos discutidos durante este trabalho.

A novidade agora é a introdução de um simples algoritmo de etiquetagem, que constrói cada etiqueta usando uma listagem dos termos mais relevantes de cada grupo, buscando diferenciar cada etiqueta de um grupo, dos seus pais e irmãos na hierarquia de grupos (Treeratpituk & Callan, 2006). Para gerar etiquetas inteligíveis foi necessário, após o processo de *stemming*, manter a variante majoritária como representante de todas as variantes de uma mesma palavra. O algoritmo de etiquetagem é da forma:

1. Para cada grupo ordenar os termos segundo algum critério de relevância. Nesta aplicação foi utilizado o critério TfV.

2. Escolher os n termos melhores colocados, como as pré-etiquetas de cada grupo. Assim são selecionados termos importantes para cada grupo. Ainda é necessário escolher dentre estes termos, os que são significantes para diferenciar o grupo de seus irmãos e pais na hierarquia de grupos.

3. Fazendo uso de uma estrutura de árvore do dendograma de grupos, percorra a árvore de nível em nível, começando da raiz, sempre seguindo o mesmo sentido em cada nível. Criar uma lista de exclusão de termos com os termos das pré-etiquetas de cada nó, excluindo sempre estes termos no próximo nó visitado. Desta forma as etiquetas de cada grupo não terão termos dos pais e irmãos na hierarquia.

4. Alguns grupos podem ficar sem etiquetas, devido a exclusão de todos os termos. Possíveis soluções seriam alterar o valor de n no passo 2 de forma apropriada, ou escolher um método mais elaborado de construção da lista de exclusão (com muitos termos selecionados para exclusão nos primeiros níveis e diminuindo conforme a árvore é percorrida), ou recursivamente juntar (*merge*) os grupos sem etiquetas aos seus irmãos e pais na árvore. Esta última foi a alternativa adotada neste trabalho.

Para avaliar o sistema, são comparados a seleção global de características e o método ZOOM-IN, usando o conceito de estabilidade para escolha de número de grupos, com etiquetagem da forma apresentada há pouco, para três consultas ao Google. Serão apresentados os tempos de execução de cada abordagem, para se ter uma idéia do tempo absoluto de execução dos algoritmos, mas uma análise relativa do tempo de processamento já foi realizada na Seção 4.2.1. Também será apresentada a hierarquia de grupos gerada, seguida de uma discussão sobre a coesão dos grupos e a qualidade das etiquetas geradas.

5.2.2 Resultados e discussão

A seguir são apresentadas três consultas ("iraq war", "Jesus Christ" e "Artificial intelligence") que foram organizadas em duas hierarquias de grupos etiquetados. Uma hierarquia foi calculada usando seleção global com 1000 termos selecionados e outra usando o método ZOOM-IN, ambas com o uso do critério de ordenação TfV, uso do conceito de estabilidade para escolha do número de grupos, conforme é descrito na seção 4.3, e etiquetagem. Os algoritmos foram executados em um PC Intel Core 2 Duo com 2,2 GHz e 2 GB de RAM. As consultas são apresentadas a seguir, com o tempo de execução do agrupamento, seguido da hierarquia de grupos etiquetada (o tamanho de cada grupo é apresentado entre colchetes):

```
Método global, 1000 termos:

"iraq war" (215 segundos)
-----
0 [200] Iraqi,Iraq,States,News,Bush,report,war
  1 [186] Saddam,Post,United
    3 [105] comments,Obama,time
      22 [2] Please,petition,copy,signed,sample,send,Peace,message,list
      23 [37] troops,day
      24 [17] Veterans,military,cost,billion
      28 [2] topic,Library,subject,Guide,Research,Visual,trillion,middot
      33 [2] IMDb,Movies,CNN,Uncovered,truth,title,rate,forces,Coalition
      34 [45] Blog,Video,Lord,book,UK,Keegan,discussion,Review
    4 [81] Iran,oil,weapons
      11 [2] country,Muslim,terrorism,view,correct,European,Turkey,move
      13 [5] American,soldiers,Biden,Pictures,withdrawal,agreement
      15 [3] companies,Business,profit,July,contracts
      17 [3] Merkley,Jeff,action,Oregon,article,plan,escalation
      19 [50] Iranian,March,Baghdad
      20 [16] dollar,administration,interests,world
  2 [14] photos
-----

"Jesus Christ" (84 segundos)
-----
0 [200] God,Christian,Jesus,life,Church,Christ,John,believe
  1 [73] Son,Father,sins
    9 [10] circumcision,law,children,love
```

```
    10 [63] Spirit,Gospel
      11 [24] receive,resurrection
      12 [39] only,divine
        21 [2] Chapter,Section,Notovitch,Issa,Christine,spiritual
        27 [31] Age,Luke
        28 [5] story,disciples,Medical,death,Roman,Easter
  2 [127] Bible,day
    3 [87] comment,Read
      7 [16] religion,many,years,living,King
      8 [60] Mormon,News
        13 [7] teachings,study,Episode,inspired,download,page
        17 [37] Saints,Answer
        18 [15] Click, Posted
      6 [11] DVD,movie,Reviews,Blu,ray,customer,star
    4 [40] Superstar,Lyrics,font
      19 [6] soundtrack,musical,image,Judas,album,man,Broadway
      29 [30] Winking,Web,healing,work,unity
      30 [4] website,Creator,Passion,public,Lord
-----

"Artificial intelligence" (45 segundos)
-----
0 [200] AI,Computer,Intelligence,systems,Science,research,learning
  1 [24] AIML,chat,frameset,Artificial
  3 [59] Conference,University,paper
  5 [17] program,reasoning,human,knowledge,information
  7 [31] Volume,Issue,book,Price,Journal,Edited
    29 [26] Applications
    30 [5] Customer,Review,product,Amazon,items,star
  9 [5] wiki,Reference
  10 [64] Robotics,problems
    25 [2] simulation,project,view,game,neurons,developed,brain,team
    26 [35] networks,algorithms
      27 [14] symbolic
      28 [21] planning,agent
  14 [25] post,years,comments
-----

Método ZOOM-IN:
```

```
"iraq war" (290 segundos)
-----
0 [200] Iraqi,States,Post,Bush,Iraq,News,American,United
  1 [184] Saddam,security
    4 [21] Lord,UK,legal,law,story
    8 [21] comments,PM,November,troops,time
    9 [3] deaths,killed,wounded,total,Oct,Mar,Jun
    10 [137] Iran
      12 [3] record,interview,wrote,Search,political,MoJo,lies,senior
      13 [11] Veterans,IVAW,people,military,soldiers,home
      15 [2] Lateline,Maliki,terrorist,meeting,many,Baghdad
      17 [115] oil
        20 [2] liberal,Daily,hours,country,cost,world,September
        21 [2] false,statement,officials,intelligence,public,National
        24 [2] Online,Babylon,pound,Arts,Travel,Tate
        25 [105] weapons,report,downloads,collection,resistance
        26 [4] add,votes,friend,free,send,Explore,subscription
      18 [6] CBS,photos,civilians
    6 [2] Salon,Nov,McCain,letter,Sep,Obama
  2 [16] war,youtube,Watch,Video
-----

"Jesus Christ" (129 segundos)
-----
0 [200] God,Jesus,Christian,life,Christ,Church,John,believe
  1 [136] Lord
    3 [96] sins,Son
      5 [16] receive,love,only
      6 [80] Father
        9 [58] Bible
          19 [4] trust,grave,laid,suffering,poor,fire,heart
          35 [5] Read,Click,China,Elvis,India,King
          36 [32] comment
          30 [9] UPI,inspired,Download,Assemblies
          26 [2] JCSM,Phone,Gastrich,free,Jason,week,Web,skeptical
          14 [6] body,Cf,Eucharist,bread,blood,wine,present
        10 [15] Mormon,view
        8 [7] Tickets,Salon,time,links,movie,man,Judas
```

```
    4 [40] table,Lyrics,shy,font,google,Statue,height,map,show,object
  2 [64] people,Gospel,day
    15 [55] world
      27 [9] Reply,hours,customer,religion
      28 [39] book
        33 [36] Age
        34 [3] death,crucifixion,concerning,Medical,Roman,source
      22 [2] Jews,word,Why,stone,identity,good,religious
      18 [4] resurrection,toast,tomb,Posted,Mar,dead,PM,friend,Mon
    16 [2] English,edit,Usage,note,derived,proper,page,li,Latin
-----

"Artificial intelligence" (80 segundos)
-----
0 [200] AI,computer,Intelligence,systems,learning,Science,machine
  1 [21] AIML,chat,frameset,bot
  3 [6] Generation,post,game,Idiap,Email,articles,Robotics
  4 [173] human
    5 [67] knowledge,problems,program
      15 [10] Vision,Proceedings,recognition
      19 [17] information,agent,language,planning
      28 [9] comp,cs,Google,links,Philosophy
      12 [5] superintelligence,theme,comments,Wordpress,Reply,goal
    6 [106] Conference,Artificial,Volume
      7 [38] University,paper,book
        31 [3] page,code,Please,Home,Correct,site
        32 [27] Group,Society
          35 [8] JSAI,provide,SGAI,Japanese,Events,English,Workshop
          36 [19] AAAI
      8 [68] Journal,Issue
        9 [1] Zone,zombies,Zoeken
        13 [12] Number,October
        23 [4] part,PDF,course,aima,students,Online
        29 [30] reviews,input,MSc,IDSIA,Master,project,accredited
        30 [3] submission,Hellenic,AIAI,product,areas
        18 [18] Price,Applications,Edited
-----
```

Como o TfV é um critério de cálculo mais custoso em relação ao DF e TI, as várias execuções deste método e de um algoritmo de ordenação a cada passo de bisecção acarreta um maior tempo de processamento ao método ZOOM-IN, mesmo com a progressiva redução da dimensionalidade.

Quanto à coesão dos grupos, cada método parece ter sua própria visão de quais são os grupos presentes dos dados. As etiquetas da mesma consulta para diferentes métodos possuem uma grande diferença e na maioria das vezes não se pode falar que algum grupo foi criado equivocadamente. Por exemplo, na consulta "iraq war", com o método global, os documentos foram divididos em dois grupos majoritários que podem ser classificados em documentos sobre pontos de vista de jornalistas e personalidades sobre a guerra (grupo 3) e em documentos constituídos por notícias sobre como a guerra vem afetando a economia (grupo 4). Já a pesquisa com o método ZOOM-IN consegue identificar grupos sobre feridos e mortos na guerra (grupo 9) e artigos que contam a história da guerra (grupo 4), não presentes com o uso da seleção global. No entanto, o resultado com o método ZOOM-IN não consegue discriminar os documentos de pontos de vista sobre a guerra, e só há um grupo majoritário sobre como a guerra vem afetando a economia e as relações internacionais (grupo 10). Isto decorre do fato de um único documento poder ser classificado em várias categorias e estas, por suas vez, podem ser categorias inteiramente distintas. O algoritmo de agrupamento somente produz uma hierarquia de grupos por vez. Com os resultados aqui apresentados, uma conclusão preliminar é que não se pode avaliar o agrupamento de documentos somente por sua precisão, mas deve ser levado em conta também a sua capacidade de gerar etiquetas que possam sugerir diversas classificações e a interface de navegação deve prover a capacidade de sugerir outros grupos, talvez até com uma nova execução do agrupamento.

Uma boa etiqueta pode ser avaliada pelo o quanto ela é capaz de descrever um grupo e o quão fácil, através da visualização de etiquetas de grupos diferentes, um usuário poderá conhecer os diversos contextos presentes nos documentos retornados pela pesquisa. Nas etiquetas geradas pelo sistema, estes fatores são dificilmente tratados pela listagem de termos fornecida, exigindo que o usuário possua já um certo conhecimento sobre o tema da pesquisa para que ele possa dicernir sobre os diferentes assuntos que cada grupo trata.

As etiquetas têm como grande fator limitador de qualidade a maneira como os documentos são representados. A fase de representação que use conhecimento lingüístico para unir significados iguais em um mesmo termo significativo no contexto dos documentos analisados é um primeiro passo para se obter etiquetas de melhor qualidade (Caraballo, 1999). Isto pode ser preliminarmente proporcionado por um banco de dados tal como o WordNet (Miller et al., 2008), que é uma famosa ferramenta para união de significados, com um amplo dicionário de sinônimos. Mas a obtenção de expressões que venham a ser boas etiquetas exige ainda um maior rigor do processamento lingüístico, e todos estes fatores não foram trabalhados nos experimentos deste trabalho.

5.3 Considerações finais

Este capítulo apresentou os resultados obtidos com os métodos de seleção de características propostos, concluindo que é possível melhorar a precisão do agrupamento através da abordagem local, principalmente quando o método de seleção consegue selecionar somente termos relevantes a cada passo de divisão. A aplicação do método para organizar resultados do Google foi capaz de gerar grupos coesos, mas as etiquetas, por razão de uma fase de representação dos documentos ainda muito básica, não apresentaram uma boa qualidade.

O próximo capítulo segue com a conclusão do trabalho, onde serão expostas algumas intuições sobre o assunto e propostos possíveis trabalhos futuros que possam estender o trabalho feito até aqui.

Capítulo 6

Conclusões

Nesta dissertação, foi abordado o uso de seleção local de características em agrupamento hierárquico de documentos e estudado os métodos de seleção de características. Foi proposto o método ZOOM-IN, que seleciona termos a cada passo de divisão de um algoritmo hierárquico divisivo. Visando selecionar uma grande proporção de termos relevantes a cada divisão, a quantidade de termos a ser selecionada é calculada a partir do tamanho de cada grupo. Os experimentos realizados mostram um ganho da precisão quando a abordagem local é capaz de eliminar as características irrelevantes, mas a heurística de escolha do número de características baseada no tamanho de cada grupo não mostrou bons resultados em uma das bases utilizadas. Para demonstrar uma aplicação dos métodos abordados nesta dissertação para escolha do número de grupos e etiquetagem, também foi desenvolvido um sistema para agrupar documentos retornados por uma pesquisa ao Google. No decorrer deste estudo, pode-se enumerar algumas contribuições realizadas, que serão apresentadas na próxima seção.

6.1 Resumo das contribuições

Pode-se destacar como contribuições deste trabalho:

- Discussão de uma metodologia para realização de um sistema de agrupamento de documentos com seleção de características, escolha do número de grupos e etiquetagem.
- Demonstração que a abordagem de seleção local ajuda na precisão do agrupamento hierárquico, com a publicação do artigo Ribeiro et al. (2008).
- Desenvolvimento de uma biblioteca para extração de resultados do Google via *parser* dos arquivos HTML.

Com os resultados dos experimentos realizados, foram percebidas limitações do trabalho, que geraram intuições de possíveis trabalhos futuros. Estas limitações serão enumeradas na próxima seção.

6.2 Limitações e trabalhos futuros

A primeira limitação a ser comentada é referente ao método ZOOM-IN. O uso do tamanho dos grupos para calcular a proporção de termos a serem considerados em cada divisão de grupo demonstrou não funcionar bem para contextos variados. Existe a necessidade de selecionar somente os termos relevantes em cada divisão, para que o método local possa alcançar um desempenho melhor que a seleção global. Isto pode ser obtido através do uso de *wrappers* para seleção de características, obtendo assim o subconjunto de características que realmente trazem benifícios ao critério interno de avaliação do agrupamento. Mas para uso em agrupamento de documentos, onde geralmente é necessário sua aplicação *on-line*, o custoso método *wrapper* pode não ser aceitável e sua aplicação para agrupamento de documentos ainda merece um maior estudo.

Quanto ao sistema discutido na seção 5.2, as etiquetas geradas através de uma listagem dos termos mais relevantes de cada grupo não foi capaz de gerar etiquetas de fácil compreensão. Ainda é necessário um processamento lingüístico mais rigoroso para representar cada documento e um melhor método de escolha destes termos, soluções já propostas em trabalhos sobre etiquetagem, tais como Treeratpituk & Callan (2006) e Glover et al. (2002), mas ainda é uma área que deve evoluir bastante.

O dendograma de grupos etiquetados gerado por uma única execução do sistema é capaz de separar grupos em diversos assuntos. Cada execução do algoritmo de agrupamento, por sua característica aleatória e pelo uso de parâmetros diferentes, consegue gerar uma estrutura capaz de revelar ainda outros grupos coesos. Como escolher e organizar esta variedade de grupos em uma interface de uso também é um ponto interessante de futura pesquisa.

Por fim, o algoritmo de agrupamento com seleção local gera uma hierarquia de conjuntos de grupos-irmãos onde cada conjunto foi formado por um subconjunto de características diferente. Isto compromete o uso de critérios internos tradicionais para decisões que envolvam vários grupos, como um exemplo, o uso de algoritmos incrementais. Este é um problema a ser contornado que precisa ser trabalhado futuramente.

6.3 Considerações finais

Agrupamento de documentos é um tema bastante diverso, com poucos pontos já consolidados e os trabalhos sobre a área sempre trazem novos algoritmos e métodos para

tratar o problema. Esta dissertação teve como objetivo mostrar um cenário propício de agrupamento de documentos e como melhorar a precisão com o uso de seleção local de características. É esperado que este texto possa servir como uma boa fonte de consulta e estudo para pessoas interessadas nas áreas de agrupamento hierárquico de documentos e seleção de características.

Referências Bibliográficas

Agrawal, R., Gehrke, J., Gunopulos, D. & Raghavan, P. (2005), 'Automatic subspace clustering for high dimensional data', *Data Mining and Knowledge Discovery* **11**, 5–33.

Ben-Hur, A., Elisseeff, A. & Guyon, I. (2002), A stability based method for discovering structure in clustered data, *in* 'Pacific Symposium on Biocomputing', pp. 6–17.

Caraballo, S. A. (1999), Automatic construction of a hypernym-labeled noun hierarchy from text, *in* 'Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics', Association for Computational Linguistics, Morristown, NJ, USA, pp. 120–126.

Comon, P. (1994), 'Independent component analysis, a new concept?', *Signal Processing* **36**(3), 287–314.

cURL (2008), 'curl and libcurl', `http://curl.haxx.se/`.

Cutting, D. R., Karger, D. R., Pedersen, J. O. & Tukey, J. W. (1992), Scatter/gather: a cluster-based approach to browsing large document collections, *in* 'Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, New York, NY, USA, pp. 318–329.

Dash, M. & Liu, H. (2000), Feature selection for clustering, *in* 'Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining', Springer-Verlag, London, UK, pp. 110–121.

Dash, M., Choi, K., Scheuermann, P. & Liu, H. (2002), Feature selection for clustering - a filter solution, *in* 'Proceedings of the 2002 IEEE International Conference on Data Mining', IEEE Computer Society, Washington, DC, USA, pp. 115–122.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990), 'Indexing by latent semantic analysis', *Journal of the American Society for Information Science* **41**, 391–407.

Dhillon, I., Kogan, J. & Nicholas, C. (2003), Feature selection and document clustering, *in* M. W. Berry, ed., 'Survey of Text Mining', Springer, pp. 73–100.

Duda, R. O., Hart, P. E. & Stork, D. G. (2001), *Pattern Classification*, 2 ed., John Wiley & Sons, Inc, New York.

Dy, J. G. & Brodley, C. E. (2004), 'Feature selection for unsupervised learning', *Journal of Machine Learning Research* **5**, 845–889.

Esuli, A., Fagni, T. & Sebastiani, F. (2008), 'Boosting multi-label hierarchical text categorization', *Information Retrieval* **11**(4), 287–313.

Fridlyand, J. & Dudoit, S. (2001), Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method, Technical report, Division of Biostatistics, University of California.

Glover, E., Pennock, D. M., Lawrence, S. & Krovetz, R. (2002), Inferring hierarchical descriptions, *in* 'CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management', ACM, New York, NY, USA, pp. 507–514.

Google (2008), 'Google', `http://www.google.com`.

Guha, S., Rastogi, R. & Shim, K. (2000), 'Rock: a robust clustering algorithm for categorical attributes', *Information Systems* **25**(5), 345–366.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2002), 'Cluster validity methods: part I', *SIGMOD Record* **31**(2), 40–45.

Hotelling, H. (1933), 'Analysis of a complex of statistical variables into principal components', *J. Educational Psych* **24**, 417–441.

Jain, A. K., Murty, M. N. & Flynn, P. J. (1999), 'Data clustering: a review', *ACM Computing Surveys* **31**(3), 264–323.

Jung, Y., Park, H., Du, D.-Z. & Drake, B. L. (2003), 'A decision criterion for the optimal number of clusters in hierarchical clustering', *Journal of Global Optimization* **25**(1), 91–111.

Kim, Y., Street, W. N. & Menczer, F. (2000), Feature selection in unsupervised learning via evolutionary search, *in* 'Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining'.

Kohavi, R. & John, G. H. (1997), 'Wrappers for feature subset selection', *Artificial Intelligence* **97**(1-2), 273–324.

Koller, D. & Sahami, M. (1997), Hierarchically classifying documents using very few words, *in* 'ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning', Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 170–178.

Konen, S. (2005), 'Cgoogle: A google search class', http://www.codeproject.com/KB/IP/cgoogle.aspx.

Law, M. H. C., Figueiredo, M. A. T. & Jain, A. K. (2002), Feature saliency in unsupervised learning, Technical report, Department of Computer Science and Engineering, Michigan State University.

Law, M. H. C., Figueiredo, M. A. T. & Jain, A. K. (2004), 'Simultaneous feature selection and clustering using mixture models', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(9), 1154–1166.

Levine, E. & Domany, E. (2001), 'Resampling method for unsupervised estimation of cluster validity', *Neural Computation* **13**(11), 2573–2593.

Lewis, D. D. (1999), 'Reuters-21578 text categorization test collection distribution 1.0', http://www.daviddlewis.com.

Lewis, D. D., Yang, Y., Rose, T. G. & Li, F. (2004), 'A new benchmark collection for text categorization research', *Journal of Machine Learning Research* **5**, 361–397.

Li, Y. & Chung, S. M. (2007), 'Parallel bisecting k-means with prediction clustering algorithm', *Journal of Supercomputing* **39**(1), 19–37.

Li, Y., Dong, M. & Hua, J. (2008), 'Localized feature selection for clustering', *Pattern Recognition Letters* **29**(1), 10–18.

Liu, T., Liu, S., Chen, Z. & Ma, W.-Y. (2003), An evaluation on feature selection for text clustering, *in* 'Proceedings of the Twentieth International Conference on Machine Learning', pp. 488–495.

Meilă, M. (2007), 'Comparing clusterings—an information based distance', *Journal of Multivariate Analysis* **98**(5), 873–895.

Miller, G. A., Fellbaum, C. & Tengi, R. (2008), 'Wordnet: a lexical database for the english language', http://wordnet.princeton.edu/.

Mitra, P., Murthy, C. A. & Pal, S. K. (2002), 'Unsupervised feature selection using feature similarity', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(3), 301–312.

Niu, Z.-Y., Ji, D.-H. & Tan, C. L. (2007), 'Using cluster validation criterion to identify optimal feature subset and cluster number for document clustering', *Information Processing Management* **43**(3), 730–739.

Oleander Solutions (n.d.), 'Oleander Stemming Library', `http://www.oleandersolutions.com/stemming/stemming.html`.

Parsons, L., Haque, E. & Liu, H. (2004), 'Subspace clustering for high dimensional data: a review', *SIGKDD Explorations Newsletter* **6**(1), 90–105.

Patrikainen, A. & Meila, M. (2006), 'Comparing subspace clusterings', *IEEE Transactions on Knowledge and Data Engineering* **18**(7), 902–916.

Porter, M. F. (1997), 'An algorithm for suffix stripping', *Readings in information retrieval* pp. 313–316.

Pudil, P., Ferri, F. J., Novovicova, J. & Kittler, J. (1994), Floating search methods for feature selection with nonmonotonic criterion functions, *in* 'Proceedings of the 12th IAPR International. Conference on Computer Vision & Image Processing', Vol. 2, pp. 279–283 vol.2.

Raghavan, P. (1997), Information retrieval algorithms: a survey, *in* 'Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms', Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, pp. 11–18.

Ribeiro, M. N., Neto, M. J. R. & Prudêncio, R. B. C. (2008), Local feature selection in text clustering, To appear in the Proceedings of the 15th International Conference on Neural Information Processing.

Sahoo, N., Callan, J., Krishnan, R., Duncan, G. & Padman, R. (2006), Incremental hierarchical clustering of text documents, *in* 'Proceedings of the 15th ACM international conference on Information and knowledge management', ACM, New York, NY, USA, pp. 357–366.

Salton, G., Wong, A. & Yang, C. S. (1975), 'A vector space model for automatic indexing', *Communications of the ACM* **18**(11), 613–620.

Sander, J., Ester, M., Kriegel, H.-P. & Xu, X. (1998), 'Density-based clustering in spatial databases: The algorithm gdbscan and its applications', *Data Min. Knowl. Discov.* **2**(2), 169–194.

Slonim, N., Friedman, N. & Tishby, N. (2002), Unsupervised document classification using sequential information maximization, *in* 'SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, New York, NY, USA, pp. 129–136.

Steinbach, M., Karypis, G. & Kumar, V. (2000), A comparison of document clustering techniques, Technical report, Department of Computer Science and Engineering, University of Minnesota.

Tang, B., Shepherd, M., Milios, E. & Heywood, M. I. (2005), Comparing and combining dimension reduction techniques for efficient text clustering, *in* 'International Workshop on Feature Selection for Data Mining'.

Tibshirani, R., Walther, G., Botstein, D. & Brown, P. (2001), Cluster validation by prediction strength, Technical report, Department of Biostatistics, Stanford University.

Treeratpituk, P. & Callan, J. (2006), Automatically labeling hierarchical clusters, *in* 'Proceedings of the 2006 international conference on Digital government research', ACM, New York, NY, USA, pp. 167–176.

UC Irvine (2007), 'UCI Machine Learning Repository', `http://archive.ics.uci.edu/ml/`.

Wang, X., Yang, J., Teng, X., Xia, W. & Jensen, R. (2007), 'Feature selection based on rough sets and particle swarm optimization', *Pattern Recognition Letters* **28**(4), 459–471.

Yu, L. & Liu, H. (2004), 'Efficient feature selection via analysis of relevance and redundancy', *Journal of Machine Learning Research* **5**, 1205–1224.

Zamir, O., Etzioni, O., Madani, O. & Karp, R. M. (1997), Fast and intuitive clustering of web documents, *in* 'Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining', pp. 287–290.

Zhao, Y. & Karypis, G. (2002), Evaluation of hierarchical clustering algorithms for document datasets, *in* 'CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management', ACM, New York, NY, USA, pp. 515–524.