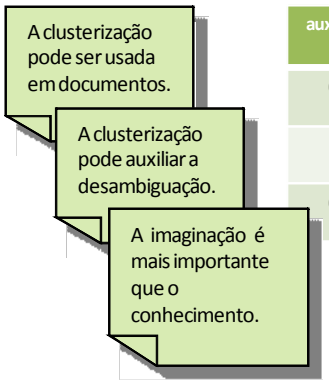
4
Agrupamento de documentos

"É a teoria que decide o que podemos observar."
Albert Einstein

Um dos métodos mais utilizados de mineração de dados descritiva é conhecido como **análise de grupos**[23]. Com ele, dada uma matriz composta de n linhas (correspondendo às amostras) e p colunas (correspondendo às variáveis que compõem as respectivas amostras), objetiva-se agrupar as amostrar em grupos internamente homogêneos (coesão interna) e com heterogenia entre os grupos (dispersão externa). O mesmo raciocínio pode ser aplicado a documentos, pois conforme apresentado no capítulo 3 é possível representar documentos através de um modelo vetorial e conseqüentemente pode-se utilizar a forma matricial sendo cada documento uma linha, ou seja, uma amostra. As colunas devem respeitar a ordem e a existência das variáveis, que no caso dos documentos são os termos obtidos na fase de pré-processamento pela *tokenização* com o uso ou não de outras técnicas conforme apresentado no capítulo 2. Cada *token* existente no léxico deve estar associado a uma única dimensão e por isso é bastante comum que a matriz se torne esparsa, o que acaba se tornando um problema de implementação, que será abordado no capítulo 6.



auxiliar	cluste-rização	conhe-cimento	desam-biguação	docu-mentos	imagi-nação	impor-tante	pode	usada
0	1	0	1	1	0	0	1	1
1	1	0	0	0	0	0	1	0
0	0	1	1	0	1	1	0	0

Cada termo corresponde a uma dimensão.

Figura 16 – Matriz representativa do corpus constituído pelos três documentos apresentados

[23] Do inglês, *cluster analysis*.

A figura 16 apresenta um exemplo de matriz dada construída com a saída de um *tokenizador* simples, que usa artigos, advérbios e preposições e palavras de tamanho menor ou igual a três letras como *stopwords*, a partir de três documentos.

O agrupamento de documentos é um processamento de textos que resulta em grupos de documentos com conceitos similares. É normalmente considerada uma abordagem de aprendizado não-supervisionado porque não existe um guia para conduzir o processo de treinamento, já que os tópicos são, na maioria das vezes, considerados indisponíveis (veja seção 1.3.2 – Agrupamento de Documentos). Em contraste com a classificação de documentos, que é normalmente considerada uma abordagem de aprendizado supervisionado devido ao uso de informações pré-classificadas que são utilizadas para guiar o processo de treinamento (veja seção 1.3.1 – Classificação de Documentos). Existem abordagens que usam técnicas de agrupamento e classificação de forma conjunta para tirar vantagem adicional do relacionamento entre palavras com intuito de diferenciar melhor os tópicos (HUNG, 2003).

Existem diversas formas de realizar agrupamento, porém, de maneira geral, todas são constituídas pelas etapas a seguir (GIUDICI, 2003): escolha das variáveis a serem utilizadas, método de formação dos grupos, tipo de medidas de similaridade ou proximidade e escolha dos critérios de avaliação.

Agrupamento de documentos será empregado no presente trabalho com o intuito de formar, de certa maneira, a taxonomia semântica capaz de representar as categorias associadas aos diversos textos e assim auxiliar o processo de desambiguação. A taxonomia, na verdade, pode ser encarada como uma analogia já que as categorias não são rotuladas e sequer existe interação humana durante a montagem. Por outro lado é exatamente essa falta de interação humana que contribui para a maior automatização e robustez dos procedimentos propostos no trabalho.

4.1
Escolha das variáveis

Na escolha das variáveis a serem usadas como parâmetros de agrupamento é preciso considerar todos os aspectos relevantes ao resultado que se deseja obter.

Em geral, durante uma análise de grupos o uso de variáveis de pouca importância tende a piorar os resultados. No caso de documentos as variáveis são os termos *tokenizados* e o uso de uma *stoplist* e técnicas adicionais de pré-processamento (veja seção 2.2 - pré-processamento), podem ser aplicadas para reduzir a dimensionalidade garantindo assim que os termos selecionados possuam maior importância para o processo de agrupamento. Ter maior importância para o processo significa os termos serem mais relevantes sob o aspecto de diferenciação entre documentos e forte caracterização do documento ao qual pertence. De forma simples, pode-se dizer que um termo importante para o processo é o que aparece com grande freqüência em um documento (ou grupo de documentos similares) e com freqüência muito baixa nos outros documentos (ou outros grupos de documentos).

4.2
Método de formação dos grupos

Os métodos de agrupamentos são normalmente diferenciados pela estrutura formada no agrupamento e dessa forma costumam ser divididos entre hierárquicos e não-hierárquicos.

Os métodos hierárquicos geram uma árvore hierarquizada de grupos de documentos que costuma ser chamada de dendrograma (Figura 7). Esses métodos são bastante flexíveis quanto ao número de grupos desejados. A flexibilidade consiste no uso de uma estrutura que armazene o co-relacionamento entre os grupos, permitindo que usuário possa identificar grupos mais específicos ou mais abrangentes, conforme suas necessidades. Os métodos hierárquicos podem ser aglomerativos ou divisivos. Nos aglomerativos, inicialmente, cada documento forma um grupo distinto. Então, por meio de iterações, cada par de documentos mais similares são unidos em um único grupo até que algum critério de parada estabelecido seja satisfeito. Nos divisivos todos os documentos compõem, inicialmente, um mesmo e único grupo. Então, de forma recursiva, um grupo é selecionado e dividido em grupos menores até que se alcance uma condição de término pré-concebida. Ambos os métodos se utilizam de alguma técnica para que os co-relacionamentos sejam recalculados a cada iteração, de forma que cada nova

formação de grupos possua valores que melhores representem as novas similaridades. Os métodos mais conhecidos de atualização das similaridades são ligação completa (*complete link*), ligação simples (*single link*) e ligação por média dos grupos (*group average link*) (CHANG, 1998). Alguns trabalhos tentam comparar o desempenho de diferentes técnicas de ligação (STEINBACH, 2000).

Os métodos não-hierárquicos, também conhecidos como métodos de particionamento alocam os documentos em um número fixo de grupos não vazios. Todos os grupos se localizam em um mesmo nível, não existindo portando nenhuma forma de hierarquia de similaridades. O número de grupos é um pré-requisito do processo de agrupamento. Os métodos mais utilizados dessa categoria são o k-means e seus variantes *k-medoids* (HAN, 2001). O *k-means* básico se inicia pela distribuição, de forma aleatória, dos documentos no número pré-estabelecido de grupos. A cada iteração a média de cada grupo é calculada e cada documento é designado ao grupo de média mais próxima. As iterações cessam quando não houver mais nenhuma mudança de documentos entre os grupos. O uso do *k-means* aplicado a documentos pode ser encontrado em (BELLOT, 1999) e (ILIOPOULOS, 2001). O método *k-means* também pode ser aplicado recursivamente para geração de grupos hierárquicos por meio de um método conhecido como *bisection k-means* (STEINBACH, 2000). Nesse método, todo o conjunto de documentos é inicialmente considerado um único grupo. Então o algoritmo seleciona, recursivamente, o maior grupo e usa o *k-means* básico para dividi-lo em dois grupos menores até que o número desejado de grupos seja alcançado ou que se tenham sido gerados grupos de apenas um documento, caso não haja outro critério de parada.

Para o caso específico de documentos textuais como sendo objetos de agrupamento alguns algoritmos específicos têm sido propostos. (ZAMIR, 1999) descreve o uso de uma árvore de sufixos para agrupar documentos e batiza a e esse método de agrupamento como STC (*Suffix Tree Clustering*). O método STC trata um documento como sendo uma *string* e não simplesmente como sendo um conjunto de palavras. Utiliza uma árvore de sufixos para, de maneira bastante eficiente, identificar conjuntos de documentos que compartilham frases inteiras. Com isso, além de criar os grupos é capaz ainda de sumarização de documentos, ou seja, extrair pequenos trechos com alta representatividade semântica do

documento. (BEIL, 2002) propõem dois métodos de agrupamento baseados na freqüência do conjunto de termos, o FTC (*Frequent Term-based Clustering*) e o HFTC (*Hierarchical Frequent Term-based Clustering*). A idéia central dos métodos é considerar apenas um subconjunto dos termos no modelo de representação do documento utilizado para fazer o agrupamento. A seleção dos termos candidatos é baseada na freqüência apresentada por estes nos respectivos documentos que os contém. (FUNG, 2002) propôs uma forma de incrementar o método HFTC, batizado como HIFC (*Frequent Itemset-based Hierarchical Clustering*). O método HIFC utiliza a mesma idéia de trabalhar com um subconjunto de termos, porém dessa vez trabalhar com a freqüência de um conjunto de palavras que aparecem juntas em uma fração mínima de documento. (HAMMOUDA, 2002) propôs um algoritmo de agrupamento incremental pela representação de cada grupo com um histograma de similaridades. (WEISS, 2000) descreve um método de agrupamento de documentos usando os vizinhos mais próximos. E, finalmente, (LIN, 2001) descreve um método chamado WBSD (*Word-based Soft Clustering*).

4.3
Métricas para agrupamento

Existem muitas maneiras de se agrupar *n* objetos em *k* grupos, o que pode ser calculado por um número de Stirling do segundo grau, dado por:

$$\left(\frac{1}{k!}\right)\sum_{i=1}^{k}(-1)^{k-i}\binom{k}{i}i^{n}$$

e, dessa forma, por meio da adição de todos os valores para $k = 1, 2, 3, \cdots, n$ grupos tem-se o total de possibilidades de agrupamento.

Existe, em geral, um elevado grau de subjetividade com relação à escolha de uma medida de similaridade ou distância entre documentos. Essa escolha deve levar em consideração o modelo de representação dos documentos, a natureza das variáveis (discreta, contínua, binária), a escala de medida (nominal, ordinal, intervalo) e etc. A seguir serão apresentadas, de maneira resumida, algumas das métricas mais utilizadas.

4.3.1
Métricas de Distância

Conforme apresentado no capítulo 3 documento um corpus de documentos pode ser representado através de uma matriz. Com base em uma matriz contendo valores numéricos e sendo *i* e *j* duas linhas dessa matriz, ou seja, dois documentos, existe uma função $d(i,j)$ que define a distância entre dois documentos se satisfizer as seguintes propriedades (GIUDICI, 2003):

- Não-negatividade: $d(i,j) \geq 0$ para todo *i* e para todo *j*;
- Identidade: $d(i,j) = 0$ se e somente se $i = j$ para todo *i* e para todo *j*;
- Simetria: $d(i,j) = d(j,i)$ para todo *i* e para todo *j*;
- Desigualdade triangular: $d(i,j) \leq d(i,k) + d(j,k)$ para todo *i*, *j* e *k*.

A medida de distância mais utilizada é Euclidiana, que pode ser definida, para dois documentos indexados *i* e *j*, como sendo a raiz quadrada da diferença entre os vetores correspondentes em um espaço t-dimensional, ou seja:

$$d(i,j) = \sqrt{\sum_{n=1}^{t} (i_n - j_n)^2}$$

A distância de bloco, também conhecida na literatura como Distância de Manhattan é dada por uma fórmula semelhante, ou seja:

$$d(i,j) = \sum_{n=1}^{t} (i_n - j_n)$$

Essas fórmulas de distância podem ser generalizadas pela métrica de Minkowsky, defina por:

$$_m d(i,j) = \left[\sum_{n=1}^{t} (i_n - j_n) \right]^{1/m}$$

e, dessa forma, com *m* igual 1 tem-se a distância de bloco, com *m* igual a 2 tem-se a distância Euclidiana. O aumento do valor de *m* dá maior ênfase ao efeito da distância.

Outra forma de cálculo da distância é baseada na sobreposição das áreas definidas pelas duas representações vetoriais dos documentos em questão (BEUSEKON, 2006). Para isso utiliza-se a analogia dos *pixels* ou dos blocos que

as componentes do vetor conforme apresentado na figura 17. Para cada par de documentos, que formam um *layout* específico assim como exemplo apresentado para duas dimensões (figura 17) a distância baseada em sobreposição de áreas, de forma geral, tem sua fórmula dada por:

$$d(i,j) = 1 - \frac{2 \times area(i \cap j)}{area(i) + area(j)}$$

No exemplo da figura 17 essa distância seria $d(i,j) = 1 - \frac{2 \times 15}{40 + 57} = 0{,}69.$

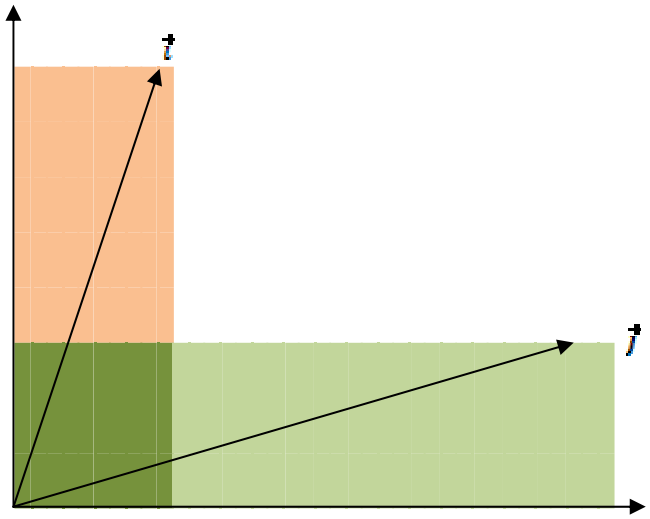


Figura 17 – Distância baseada na sobreposição de áreas.

4.3.2
Métricas de Similaridade

De forma análoga às medidas de distância é possível se estabelecer as métricas de similaridade, dadas por uma função genérica $s(i,j)$ que gozam das seguintes propriedades (GIUDICI, 2003) :

- Não-negatividade: $s(i,j) \geq 0$ para todo *i* e para todo *j*;
- Normalização: $s(i,i) = 1$ para todo *i*;
- Simetria: $s(i,j) = s(j,i)$ para todo *i* e para todo *j*;

Nesta categoria se enquadram diversas métricas tendo sua grande maioria se originado de estudos com intuito de organizar o fluxo produtivo e diminuir os tempos improdutivos em células de manufatura (SHAFERS, 1993). Desde então têm sido empregadas nos mais diversos métodos de agrupamento, tendo destaque

especial os coeficientes de Dice, Jaccard, Sokal e Michener, Russel e Rao, dentre outros (GIUDICI, 2003).

Alguns métodos são baseados na junção de pares de itens similares. Então o cálculo de todas as similaridades dois a dois pode ser exigido e assim montada a matriz similaridade. Uma matriz triangular é suficiente, já que ela é simétrica, ou seja, $s(i,j) = s(j,i)$. Esses métodos, que juntam pares de documentos mais similares a cada passo, conhecidos como hierárquicos aglomerativos se diferenciam pela maneira com que ligam seus objetos a cada iteração para formação dos grupos. As formas mais utilizadas de ligação são:

- *Single link*: junta, a cada passo, os pares de objetos mais similares que ainda não estão no mesmo grupo. É de simples implementação, porém pode gerar grupos não muito compactos.
- *Complete link*: usa o para menos similar entre cada dois clusters para determinar a similaridade intercluster. Gera estruturas pequenas e compactas.
- *Group average link*: usa os valores médios dos pares ligados dentro de um cluster para determinar a similaridade. Todos os objetos contribuem para a similaridade intercluster, resultando em uma estrutura intermediária entre a amarração dispersa do método single link e a amarração compacta do método complete link.
- *Ward's method*: também conhecido como método da variância mínima, pois une a cada estágio o par de clusters cuja fusão minimize o aumento do total das somas dos quadrados dos erros dentro do novo grupo formado. Tende a produzir grupos homogêneos e uma hierarquia simétrica.