AGRUPAMENTO AUTOMÁTICO BASEADO EM AUTORIDADE E CONTEÚDO

Ana Carolina do Prado

UFU - Universidade Federal de Uberlândia

Faculdade de Computação

Programa de Pós-Graduação em Ciência da Computação

AGRUPAMENTO AUTOMÁTICO BASEADO EM AUTORIDADE E CONTEÚDO

Ana Carolina do Prado

Dissertação apresentada ao Programa de Pós-Graduação da Faculdade de Computação da Universidade Federal de Uberlândia como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Ilmério Reis da Silva

09 de Agosto de 2005

Universidade Federal de Uberlândia

Faculdade de Computação

Os abaixo-assinados, por meio deste, certificam que leram e recomendam à Faculdade de Computação a aceitação da dissertação intitulada "Agrupamento Automático baseado em Autoridade e Conteúdo", de autoria de Ana Carolina do Prado, como parte dos requisitos exigidos para a obtenção do título de Mestre em Ciência da Computação.

Uberlândia(MG), 09 de Agosto de 2005.

Avaliador:

Prof. Dr. Wagner Meira Júnior (UFMG)

Avaliador:

Prof. Dr. Carlos Roberto Lopes (UFU)

Orientador:

Prof. Dr. Ilmério Reis da Silva(UFU)

Co-Orientador:

Prof. Dr. João Nunes de Souza(UFU)

Aos meus dois
pais queridos.

Agradecimentos

Agradeço a Deus pela graça da vitória após a luta inevitável.

Ao meu orientador Prof. Dr. Ilmério Reis da Silva. Grande instrumento de Deus nessa minha trajetória.

Ao meu co-orientador Prof. Dr. João Nunes de Souza pela ajuda prestada.

A minha mãe Maria Euci, pelo amor e cuidado. Aproveito e peço perdão pelas tantas faltas em grandes momentos.

Ao meu pai, Álvaro Pereira do Prado, pelo apoio e incentivo. Fonte inesgotável de coragem que tantas vezes precisei. Sempre tão sensato e compreensivo. Muito do que sou, ele é o responsável.

Ao meu outro pai "Dida", Belgides Pereira do Prado, que além da obrigação é e sempre será muito mais que um pai.

Aos meus irmãos: Alcione, Alvacir, Alessandro e respectivos cônjuges, que me permitiram crescer amparada em amor e me fizeram tia de lindos cinco sobrinhos.

Aos amigos que se demonstraram solidários às minhas lutas, tropeços e desafios:

Marcos Aurélio Batista pela a sua ajuda insubstituível nos vários momentos dessa pesquisa. Tornou-se braço forte, quando eu não conseguia mais ter

força. Parte disso tudo também é dele.

Ângela pela paciência, cuidado e tolerância nos momentos tão difíceis.

Dona Conceição, Sr. Ramiro, Rosemary e filhos pelo apoio e cuidado.

Giovanna, Amir, Cecília (*in memorian*), Gabi e Laura, pela força, incentivo e amizade.

Em especial aos grandes amigos: Fernanda, com quem compartilhei os maiores e menores momentos desse mestrado, cada um deles foi uma vitória. Rosana, me escutou, e levantou meu ânimo em muitas vezes. Graça, que tanto se preocupou comigo e foi sempre uma irmã. Ao Humberto, pela capacidade de se doar na ajuda. Amo muito vocês!

A querida secretária da Faculdade de Computação Maria Madalena que com sua competência e carisma conquista a todos que chegam em Uberlândia e torna nossa jornada mais tranqüila.

Cada momento foi uma vitória! Obrigada.

"As aplicações práticas de uma ciência geralmente antecedem o desenvolvimento da própria ciência"

N. K. Jerne

Resumo

Esta dissertação apresenta uma técnica de análise de agrupamentos que combina conceitos de similaridade de documentos por conteúdo com informações de ligações. O Modelo Vetorial Clássico é utilizado para efetuar o cálculo de similaridades entre os documentos e um algoritmo de análise de ligações é utilizado para obter o valor de autoridade de cada documento. Calculam-se os valores das autoridades locais dos documentos pertencentes a cada um dos grupos e, nomeando a maior autoridade local como representante do grupo, temos uma redistribuição dos documentos nos grupos. Essa combinação possibilita a obtenção de grupos onde, quem melhor o representa é a maior autoridade daquele assunto. Esse algoritmo, chamado de Agrupamento por Autoridade Local (AAL), foi proposto, implementado e a qualidade do agrupamento resultante foi avaliada através de comparação com o método de agrupamento tradicional k-médias. O AAL possui a estrutura de ligações da *Web* como definidora das características que serão utilizadas para agrupar os documentos, trazendo consigo inúmeras aplicações nesse ambiente, como identificação de grupos em uma grande coleção de páginas com o intuito de minimizar o escopo da busca, ou até mesmo agrupar o resultado de pesquisa realizada, gerando grupos distintos de documentos.

Palavras-chave:

Agrupamento não supervisionado, Agrupamento por autoridade local, Grupo, Recuperação de Informação, *AAL*

Abstract

This dissertation introduces a technique of clustering analysis that combines concepts of document similarities by contents with link information. The Classic Vector Model is used to carry out the calculation of the similarities between the documents and a link analysis algorithm that is used to get the value of the authority of each document. Calculating the values of the local authorities from the documents belonging to each one of the groups and employing the biggest local authority as the reassign the cluster, we have redistribution of the documents to the clusters. This combination provides clusters represented by the best authority in that subject. This algorithm, called Local Authority Clustering, was proposed, introduced and the quality of its results was evaluated through comparison with the traditional K-means. The AAL has the link structures of the Web as definite from the characteristics that will be used to clustering the documents with several applications in this environment, as the identification of the clusters in a large collection of pages to minimize the search or even to gather together the result of the search generating different clusters of documents.

Keywords:

AAL, Local Authority Clustering, Clustering, Cluster, Information Retrieval, Link Analysis.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Neste capítulo é apresentada a motivação para o desenvolvimento da técnica de agrupamento automático aqui proposta, bem como os trabalhos relacionados e a descrição dos capítulos que compõe esta dissertação.

1.1 Motivação

Com o desenvolvimento de tecnologias de comunicação cada vez mais eficientes e, especialmente com o advento da Internet, o volume de informações que uma pessoa tem acesso cresce diariamente de forma exponencial. Alguém que se disponha a buscar documentos com o objetivo de satisfazer determinada necessidade de informação, encontra dificuldades ao acessar um ambiente que é composto por uma grande quantidade de informações, como esse da Web, e consequentemente enfrenta a sobrecarga gerada pelo excesso de informações. Portanto é importante que o usuário tenha a sua disposição ferramentas para auxiliá-lo na tarefa de encontrar a informação que lhe é útil.

Diante do desafio de auxiliar o usuário a encontrar a informação de seu interesse, tem-se a possibilidade de organizar os documentos em uma estru-

tura hierárquica de assuntos, o que daria ao usuário uma possibilidade de navegação por meio de ligações (*links*) entre os documentos, podendo assim selecionar os que são de seu interesse. Essa organização hierárquica é chamada categorização e geralmente a identificação dos assuntos, em classes pré-definidas é feita manualmente, o que causa um grande atraso entre a confecção do documento e a disponibilização do mesmo na estrutura hierárquica. Além disso, o custo deste procedimento é alto, principalmente quando se trata de uma base de dados extensa como a da Internet. Diante disso, tem sido feitos vários esforços para a classificação automática dos documentos [Halkidi, Nguyen & Varzigiannis 2003].

O foco deste trabalho é o agrupamento de documentos, onde as classes não são pré-definidas, o que é um grande problema de organização de documentos. É definida uma técnica de agrupamento (não supervisionado) de documentos que considera não só a similaridade entre os documentos, mas também as informações da estrutura de ligações da *Web*, Assim, essa proposta define grupos representados pelo documento que é considerado como a maior autoridade, o índice de autoridade é calculado considerando as ligações que referenciam o documento. Para viabilizar a formação não supervisionada dos grupos seguindo essas características, tem-se como referência o algoritmo de agrupamento k-médias considerando que, no momento de calcular a similaridade entre os documentos é utilizado o modelo vetorial.

Segundo [Willett 1988], o agrupamento deve ser realizado baseado na similaridade entre os documentos, sendo que os documentos mais similares entre si passam a pertencer a um mesmo grupo.

Partindo desse pressuposto, esta proposta utiliza alguns conceitos da área de Recuperação de Informação e Inteligência Artificial, dentre os quais se destaca o modelo vetorial, análise de ligações e agrupamento, seguindo as linhas

de estudos dos trabalhos [Kleinberg 1999], [Calado, Ribeiro-Neto, Ziviani, Moura & da Silva 2003], [Macqueem 1967] e [Salton, Yang & Wong 1975].

Dentre os vários modelos da área de Recuperação de Informação o modelo vetorial clássico, definido em [Salton et al. 1975], utiliza somente informações baseadas no conteúdo dos documentos para verificar o grau de similaridade de um documento com determinada consulta ou com outro documento.

O objetivo deste trabalho é propor um método de agrupamento que combina informações de conteúdo dos documentos com informações extraídas da estrutura de ligações da *Web*. Essa estrutura de ligações é uma rica fonte de informações que geralmente carrega consigo a opinião do autor do documento a respeito dos outros documentos que são referenciados por ele. Observando essa propriedade do ambiente da *Web*, Kleinberg propõe um algoritmo chamado HITS, descrito na seção 4.3, que analisa a estrutura de ligações da *Web*, calculando valores que indicam a importância de um documento nesse ambiente (página *Web*) em relação a determinado assunto. Para a identificação desses valores são utilizadas as informações do quanto um documento é referenciado por outros documentos, assim ele é definido como uma boa autoridade, e o quanto um documento referencia a outros documentos que são boas autoridades, define um bom *hub*.

A principal contribuição deste trabalho é a proposta e avaliação de um algoritmo que agrega a informação de autoridade ao processo de agrupamento de documentos da *Web*.

1.2 Trabalhos Relacionados

Na *Web* a quantidade de documentos tem crescido muito e, a tarefa de recuperação da informação relevante tem se tornado cada vez mais difícil. Diante

dessa realidade, para melhorar a qualidade da resposta trazida por sistemas de recuperação de informação ao usuário, algoritmos de análise de ligações, como o HITS proposto em [Kleinberg 1999], que usa as informações da estrutura de ligação da *Web* para medir o grau de importância de um documento, tem sido estudados. O HITS define os graus de autoridade e *hub* dos documentos da *Web* de acordo com o a estrutura de ligações (*links*) onde os documentos são referenciados por e referenciam a outros.

Em [Calado et al. 2003] é descrito um estudo comparativo realizado entre o uso da informação de ligação local e global em estratégias de Recuperação de Informação aplicadas a um ambiente de hiperligações. A informação local é derivada de um conjunto de documentos retornados como resposta a uma consulta corrente definida pelo usuário, enquanto que a informação de ligação global considera todos os documentos da coleção. Nesse estudo, constatou-se que quando na análise local, onde são considerados os documentos do conjunto resposta, existe uma melhor qualidade em relação à análise global. No trabalho aqui proposto a análise de ligações é realizada localmente, considerando um grupo de documentos.

Em um ambiente onde o número de objetos a serem analisados é grande ou não se tem pessoas para fazer a investigação e descobrir quais deles são semelhantes, é comum utilizar um processo de agrupamento não supervisionado. Um estudo desse tipo de agrupamento foi desenvolvido por McQueen, que propôs o algoritmo k-médias, que escolhe aleatoriamente k elementos para iniciar o processo de agrupamento, e depois segue fazendo a identificação das características dos elementos participantes do agrupamento que influenciarão na formação dos grupos [Macqueem 1967].

A proposta de agrupamento aqui apresentada, tem o intuito de agregar de maneira diferenciada as informações da estrutura de ligações ao processo de

agrupamento. Esse estudo propõe uma técnica de agrupamento não supervisionado de documentos adaptando o algoritmo k-médias de tal maneira que os grupos formados serão representados pelas autoridades naquele assunto.

1.3 Organização da dissertação

Este trabalho está estruturado como se segue:

Capítulo dois - Recuperação de Informação: apresenta uma visão geral sobre Recuperação de Informação, suas principais definições e etapas.

Capítulo três - Agrupamento não supervisionado: traz uma visão geral sobre agrupamento automático de informações, suas classificações, e, mais detalhadamente o algoritmo k-médias.

Capítulo quatro - Análise de ligações: são vistas as características e definições da análise de ligações da *Web*. A partir dos conceitos de Recuperação de Informação e agrupamento.

Capítulo cinco - O agrupamento por Autoridade Local: apresenta a proposta de agrupamento utilizando autoridade e conteúdo, e a especificação do algoritmo *AAL* .

Capítulo seis - Avaliação de AAL: descreve os experimentos e uma análise dos resultados adquiridos com a implementação da técnica proposta.

E o Capítulo sete - Conclusões e trabalhos futuros: apresenta conclusões sobre este trabalho e sugestões de trabalhos futuros.

Capítulo 2

Recuperação de Informação

Este capítulo apresenta conceitos básicos de Recuperação de Informação necessários para a compreensão da proposta de agrupamento por similaridade e ligações da *Web*. Isto porque o cálculo da similaridade entre documentos usado neste trabalho tem sua origem na área de Recuperação de Informação (RI). Então são apresentados os fundamentos de RI, especificamente do modelo vetorial que é usado para calcular a similaridade entre os documentos baseada no conteúdo. Além disso, são apresentadas medidas de avaliação da qualidade de resultados retornados pelos sistemas de Recuperação de Informação que serão apresentadas para a avaliação dos agrupamentos (Capítulo 6).

2.1 Introdução à Recuperação de Informação

O processo de recuperação de informação consiste em identificar, em uma coleção de documentos, quais desses documentos atendem (supostamente) à necessidade de informação do usuário. Considerando que a informação é todo dado que é compreendido e faz com que quem a receba mude o seu estado

atual de conhecimento, no contexto de RI, informação relevante é aquela que é compreendida e satisfaz a necessidade do usuário.

Em [Mizzaro 1997] relevância ainda não tem um conceito bem compreendido, apesar de existir uma bibliografia considerável a respeito. Para Stefano Mizzaro a informação relevante está diretamente relacionada com o usuário, com a sua necessidade de informação (o contexto que é expresso na sua consulta) com o momento que isso ocorre. Além disso, determinado documento pode não ser relevante a uma consulta em determinado momento e ser relevante em outro [Mizzaro 1997].

Portanto, conclui-se que a relevância é um conceito estritamente relacionado ao usuário. Em meio a essa relação entre o usuário e o documento, no paradigma de RI verifica-se a existência de um terceiro elemento, o sistema, que auxilia o usuário a recuperar informação relevante, definindo assim um modelo genérico que possui três elementos básicos que se interagem: usuário, sistema e documentos.

Nesse modelo, o elemento central é o sistema que é a interface entre o usuário e os documentos. O usuário é uma pessoa que possui a necessidade de informação e se dispõe a buscá-la, e os documentos são registros que possuem potencial de transmitir informação. Para um documento ser recuperado e transmitir uma informação que seja relevante, essa deve estar de acordo com o interesse do usuário no momento da busca. Diferentemente dos sistemas de banco de dados que trabalham com dados estruturados e recuperam a resposta exata, os sistemas da área de Recuperação de Informação (SRI) tendem a recuperar uma aproximação do que seria considerado relevante e manuseiam dados extraídos dos próprios documentos.

No ambiente digital, nas últimas décadas o número de documentos vem se multiplicando tanto na tipologia quanto na complexidade. Isso se deu

principalmente com o advento da Internet, que é a maior biblioteca digital, o que torna uma tarefa difícil para o usuário encontrar manualmente o que realmente precisa. Nesse caso a melhor opção é fazer o acesso de maneira automatizada.

Para realizar o acesso automatizado a uma base de dados composta por documentos textuais é necessário definir uma representação para cada um desses documentos. Para tanto, o pré-processamento e indexação é realizado através de algoritmos da área de RI utilizando as próprias palavras presentes no conteúdo do documento.

2.2 Indexação e pré-processamento

O processo de indexação de documentos na Internet começa com a aquisição de documentos, que geralmente é realizada através de um *spider*, *crawler* ou robô de busca, que é um programa que parte de um conjunto inicial de *links* fazendo a varredura da *Web* e uma cópia dos documentos visitados, acrescentando sempre em sua lista cada novo *link* encontrado. Após realizada a aquisição dos documentos da *Web*, eles precisam ser representados para que sejam encontrados através de um sistema de recuperação de informação.

A representação de um documento tem como objetivo descrevê-lo através do seu conteúdo. Na área de RI geralmente são usados métodos automáticos de pré-processamento e indexação para extrair as características que representarão os documentos. Na fase de pré-processamento, para se definir quais os termos que caracterizarão o documento, são usados "filtros"que retiraram palavras de pouca relevância para o documento (*stop words*), as quais não conseguem diferenciá-lo de outro documento; os símbolos de pontuação e caracteres especiais (análise léxica); e em alguns casos, a extração de radicais

(*stemming*).

Mesmo depois de filtrados, o número de termos resultantes ainda pode ser alto. Esse número de características (termos) pode ser reduzido pela seleção dos N termos mais relevantes, que são identificados por meio da atribuição de pesos dos termos nos documentos. Estes pesos representam o grau de importância das características ou termos dos documentos. Pode-se também definir um limiar (*threshold*) do peso do termo no documento para definir seus representantes.

Depois de atribuídos os pesos às características dos documentos, para que seja otimizado o processo de busca, é necessário montar um índice. O usuário entregará ao sistema uma consulta usando palavras e como resultado lhe será entregue uma lista de documentos ordenada de acordo com a presença das palavras da consulta nos documentos. Então é interessante que o sistema tenha uma estrutura de acesso otimizado, contando com a mesma relação definida anteriormente (palavra $\Rightarrow$ documento).

A estrutura de armazenamento, chamada arquivo invertido é utilizada para facilitar o acesso aos documentos no momento de resposta do sistema à consulta do usuário.

Um arquivo invertido geralmente é composto um vocabulário, que é um vetor contendo todas as diferentes palavras de uma coleção, e para cada palavra do vocabulário, uma lista de todos os documentos onde a palavra ocorre [Baeza-Yates & Ribeiro-Neto 1999]. Cada lista é chamada lista invertida do termo. Assim, localizando a lista invertida do termo consultado obtém-se a lista de todos os documentos em que ele aparece. A seguir são descritos o arquivo invertido de um documento de uma coleção.

A estrutura do arquivo pode representar um único documento. Nesse caso, o vocabulário é o conjunto de todas as diferentes palavras do texto do

documento, e o arquivo invertido, também chamado de arquivo de ocorrências, é composto pelas posições nas quais a palavra ocorre no texto. Um exemplo de arquivo invertido de um documento pode ser visto na Figura 2.1, onde as ocorrências são o deslocamento em bits a partir do início do texto.

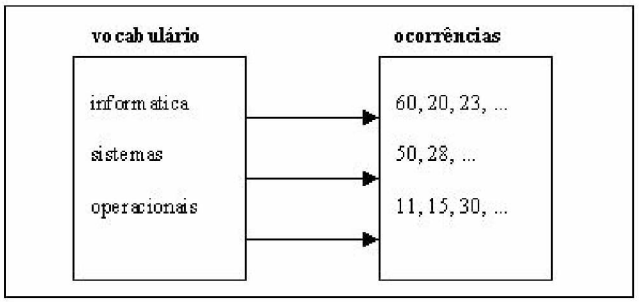

Figura 2.1: Arquivo invertido com listas de ocorrência do termo no documento.

Já a lista invertida de uma coleção de documentos tem como entrada, um vocabulário que contém todas as diferentes palavras da coleção. A partir desse vocabulário, tem-se acesso a um arquivo que lista o local onde cada palavra ocorre (identificador do documento). No caso de uma coleção, esse local geralmente é um documento no qual o termo ocorre. Outra estrutura é a lista de termos no documento. Este arquivo é aqui chamado de vetor de termos dos documentos e será utilizado para a comparação entre documentos. Conhecendo o documento (seu identificador), tem-se acesso ao respectivo vetor contendo os pesos dos termos contidos nesse documento. Na Figura 2.2 tem-se esboçado um exemplo do que seria um arquivo invertido e de um arquivo com o vetor de termos dos documentos de uma coleção. Essa figura mostra de maneira simplificada uma lista invertida de uma coleção de documentos, contendo ponteiros que fazem a ligação entre as estruturas de dados (vocabulário, lista de inversão e os vetores dos documentos). Onde, ao ser encontrado o termo da busca no dicionário, identifica-se a lista de documen-

tos onde ela ocorre (lista de inversão) os quais serão acessados. Devido à rapidez de acesso e facilidade para a identificação dos documentos relevantes a determinado termo, o arquivo invertido é uma das estruturas de armazenamento mais utilizadas pelos sistemas da área de Recuperação de Informação [Kowalski 1997].

Após serem indexados os documentos, o sistema pode disponibilizar ao usuário toda a base de dados para serem recuperados os documentos que (supostamente) são relevantes para a sua consulta.

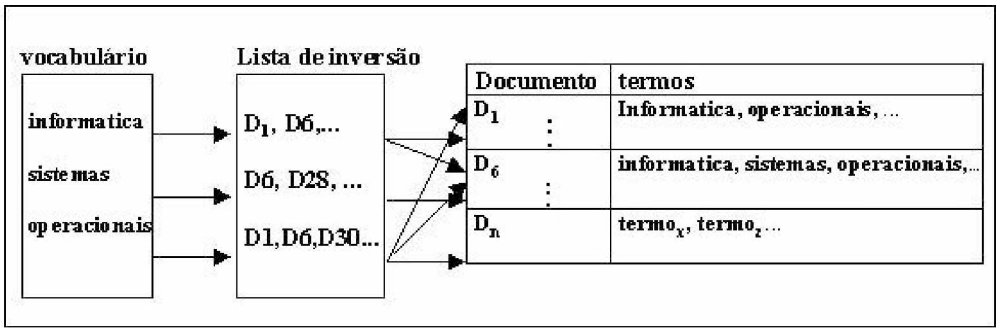

Figura 2.2: Arquivo invertido e vetor de termos de uma coleção de documentos.

A necessidade de informação do usuário é descrita através de uma consulta seguindo um formalismo definido pelo sistema, geralmente palavras-chave. Definida e submetida à consulta do usuário, o sistema faz o casamento (*matching*), isto é, o Sistema de Recuperação de Informação (SRI) faz a identificação dos documentos relevantes a uma consulta comparando as características dessa consulta com as características de cada um dos documentos presentes na base de dados.

Essa análise de semelhança (similaridade) entre características geralmente depende do modelo conceitual adotado pelo SRI, e é feita por uma classe de funções conhecidas por funções de similaridade [Kowalski 1997]. Essas fórmulas podem ser utilizadas não só para a identificação do grau de simi-

laridade entre uma consulta e um documento (documentos relevantes), mas também para a identificação da similaridade entre documentos, o que é extremamente útil para a área de agrupamento de documentos.

Existem vários modelos que permitem definir o conjunto de documentos considerados relevantes à determinada consulta, os quais são modelos conceituais ou abordagens genéricas para a recuperação de informações. Dentre eles tem-se os clássicos: booleano, vetorial e probabilístico. A seguir descreve-se o modelo vetorial, que apresenta bom desempenho e qualidade em RI e será utilizado para implementação dos agrupamentos de documentos.

2.3 Modelo Vetorial

Geralmente os métodos utilizados na recuperação de informações em textos têm como base o uso da palavra, que representa a unidade básica de acesso à informação. A partir dessa unidade foram desenvolvidos vários modelos com o objetivo de facilitar o acesso à informação e melhorar o resultado de uma busca ou consulta.

Dentre os vários modelos, o Modelo Espaço Vetorial (*Vector Space Model - VSM*) ou simplesmente Modelo Vetorial [Baeza-Yates & Ribeiro-Neto 1999], é muito utilizado por sua simplicidade, eficiência e eficácia.

Ele foi desenvolvido por Gerard Salton com o objetivo inicial de ser usado no sistema de recuperação de informação SMART [Salton 1971].

No Modelo Vetorial os documentos que possuem uma maior quantidade de termos da consulta, tendem a ser mais relevantes ao usuário. Quando é feita uma pesquisa (consulta) usando palavras-chave ou termos, os documentos que possuem os termos pesquisados (consulta) são dispostos em ordem decrescente em relação à ocorrência desses no documento.

Para cada documento são definidos os seus descritores (termos mais relevantes, palavras-chave) através dos métodos de pré-processamento e indexação comuns da área de Recuperação de Informação. Na representação do documento é feita a atribuição de pesos aos termos de indexação presentes na consulta e aqueles presentes nos documentos. Essa ponderação, atribuída a cada termo de indexação, permite calcular o grau de similaridade entre um documento e uma consulta. Para tanto é feita a comparação entre a representação vetorial dos documentos e da consulta do usuário conforme mostrado na próxima seção.

2.3.1 Representação vetorial

No modelo vetorial um documento é representado por um vetor em que cada elemento representa o peso, ou a relevância, do respectivo termo de indexação para o documento.

Cada elemento do vetor é considerado uma coordenada dimensional. Assim, os documentos podem ser colocados em um espaço euclidiano de t dimensões (onde t é o número de termos) e a posição do documento em cada dimensão é dada pelo seu peso.

Os pesos, quando normalizados, possuem valores que variam de 0 a 1, onde os pesos mais próximos de um (1) indicam os termos mais importantes, e os pesos menores caracterizam os termos menos importantes. Os termos que os documentos não possuem são considerados nos vetores, mas possuem valor zero (0).

A posição do documento em um espaço multidimensional é descrita pelo vetor, sendo que, cada termo de indexação representará um eixo ou uma dimensão. A Figura 2.3 mostra graficamente a representação de um documento d_1 com termos de indexação t_1 e t_2 com seus respectivos pesos 0.4 e 0.5.

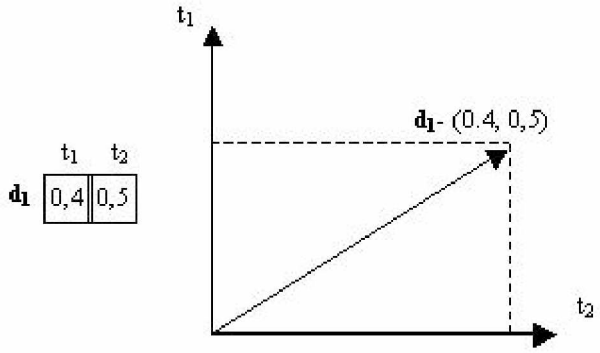

Figura 2.3: Representação vetorial de um documento em um espaço com dois termos de indexação.

Na Figura 2.4 os documentos d_1 e d_2 estão representados em um mesmo espaço vetorial com três dimensões. Os números representam os pesos de seus respectivos termos.

2.3.2 Cálculo de relevância

Conforme apresentado na seção anterior, cada termo do documento possui um peso associado, o qual expressa o seu grau de importância para o documento em que está inserido, mas, vale ressaltar que nem todas as palavras de um documento possuem a mesma importância, sendo que as palavras que ocorrem com maior freqüência em determinado documento (com exceção das *stopwords*[1]) tendem a ser mais importantes para ele, ou seja, costumam representar melhor o seu conteúdo. utilizadas com maior freqüência (com exceção das *stopwords*) costumam ter um significado mais importante. Pode-se con-

[1]Conjunto pré-definido de palavras que geralmente são eliminadas por serem de grande ocorrência e não diferenciarem um documento do outro, sendo então consideradas irrelevantes. Por exemplo, artigos, preposições, conjunções e advérbios.

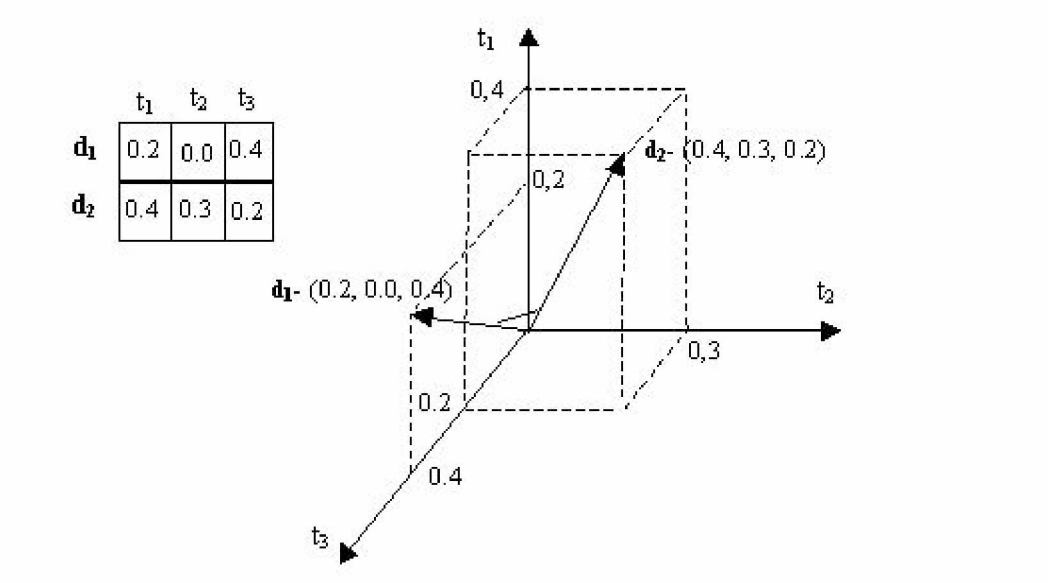

Figura 2.4: Representação vetorial com três termos de indexação.

siderar também que algumas palavras, por estarem em locais especiais do texto, como o título, possuem um maior grau de relevância para o documento em questão, já que o autor do documento deve tê-las colocado lá por considerá-las como sendo muito relevantes e descritivas para a sua idéia. Os substantivos e complementos também podem ser considerados elementos mais significativos na representação do um documento, por representarem melhor o assunto tratado por aquele texto (documento), sendo portanto mais significativos na representação do documento. Então o cálculo de relevância de uma palavra pode basear-se na freqüência das palavras, na análise estrutural do documento ou na posição sintática de uma palavra. Entretanto nem sempre as informações de estrutura estão disponíveis. Além disso o cálculo baseado em estrutura é complexo e nem sempre se justifica. Então neste trabalho para realizar os cálculos de importância do termo será considerada a sua freqüência. A técnica que será utilizada baseia-se na freqüência da pala-

vra no texto e na coleção. Nesse caso, existem várias fórmulas que podem ser utilizadas para calcular a importância de uma palavra baseando-se em sua freqüência [Van Rijsbergen 1979], [Salton & Buckley 1987*a*], algumas dessas fórmulas são apresentadas a seguir.

A freqüência do termo i no documento j ($freq_{ij}$) sendo também conhecida por fator tf [Baeza-Yates & Ribeiro-Neto 1999] ou *term frequency* (tf_{ij}), mede o número vezes que um termo k_i aparece em um documento d_j.

Além disso, a freqüência inversa de documentos (*inverse document frequency* - *idf*), quantifica a relevância do termo k_i como um fator discriminante do termo em relação a todos os documentos da coleção, aumentando assim a importância dos termos que aparecem em poucos documentos e diminuindo a importância de termos que aparecem em muitos documentos [Robertson & Walker 1997].

Assim, pode-se definir que dada uma coleção que possui N documentos e sendo n_i a quantidade de documentos que possuem o termo t_i, então o inverso da freqüência do termo na coleção, ou *idf* (*inverse document frequency*) é dado por:

$$idf_i = log\frac{N}{n_i} \tag{2.1}$$

Tanto documentos (d_j) quanto consultas (q) são vetores compostos pelos pesos dos termos do tipo $d_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$ e $q = (w_{1q}, w_{2q}, \ldots, w_{tq})$. Para calcular estes pesos, uma abordagem comum é balancear a importância dos termos intradocumento (tf) com a importância dos termos interdocumentos (idf).

Este valor é usado para calcular o peso, utilizando a seguinte fórmula:

$$w_{ij} = tf_{ij} \cdot idf_i \tag{2.2}$$

ou seja, é o produto da freqüência do termo no documento pelo inverso da freqüência do termo na coleção.

Uma alternativa de normalização dos pesos é considerar

$$tf_{ij} = \frac{freq_{ij}}{max_{\forall i} freq_{ij}} \tag{2.3}$$

O Modelo Vetorial propõe avaliar o grau de similaridade de um documento d_j com relação a uma consulta q como sendo a correlação entre os vetores d_j e q.

As principais vantagens do modelo vetorial são a sua simplicidade, a facilidade que ele provê de se calcular similaridades com eficiência e o fato de que o modelo se comporta bem com coleções genéricas

2.3.3 Similaridade

A similaridade entre dois objetos (documento-consulta ou documento-documento) é calculada considerando a informação das características que representam os objetos e seus respectivos graus de importância, sendo então possível identificar quais documentos podem satisfazer a consulta do usuário. Na maioria dos casos, são considerados relevantes à consulta os documentos que possuem mais características em comum com ela.

A similaridade mede numericamente os graus de semelhança entre dois objetos. Esse valor é o meio de distinguir, entre os objetos candidatos, quais são similares ou não em relação ao que o usuário está buscando.

No modelo vetorial, a consulta do usuário também é representada por um vetor. Dessa forma, os vetores dos documentos podem ser comparados com o vetor da consulta e o grau de similaridade entre cada um deles pode ser identificado. Os documentos mais similares à consulta são considerados relevantes para o usuário e retornados como resposta para ela. Uma das formas de se calcular a similaridade entre os vetores é observar o ângulo

entre estes vetores. Então a similaridade pode ser quantificada pelo cosseno do ângulo entre estes dois vetores, como mostrado na Figura 2.5:

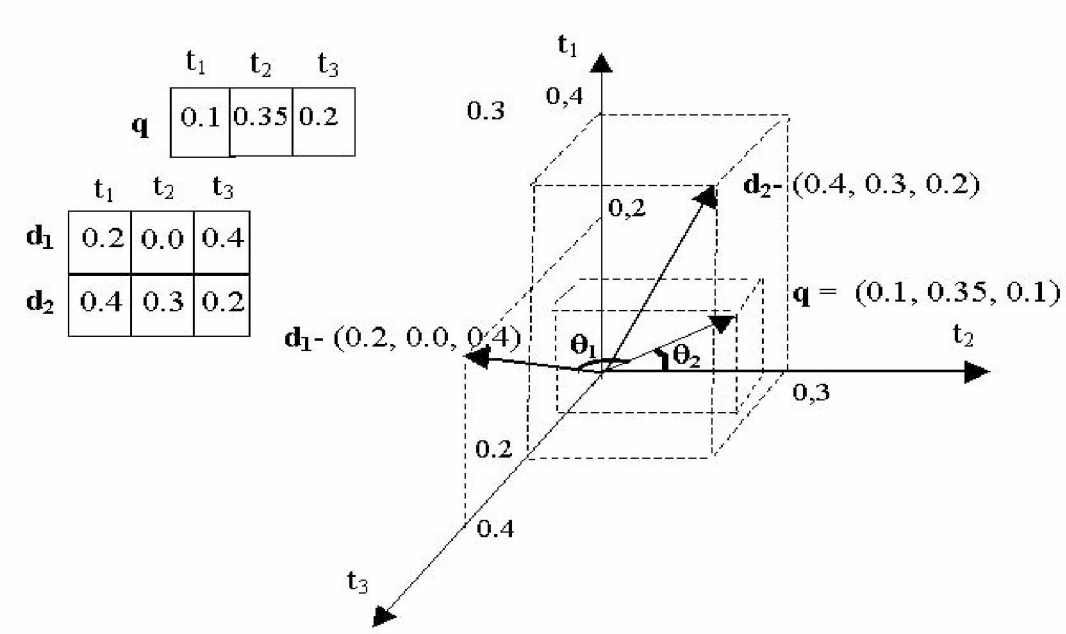

Figura 2.5: Representação de uma consulta q e dois documentos (d_1, d_2) e respectivos ângulos (Θ_1 e Θ_2) em um espaço vetorial.

Esse cálculo usando o cosseno é como Salton faz no seu modelo original, ele utiliza uma função, a *cosine vector similarity* [Salton & Buckley 1987*b*], que calcula o produto dos vetores de documentos através da fórmula apresentada a seguir:

$$sim_v(d_j, q) = \frac{\sum_{i=1}^{t} w_{ij} w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \tag{2.4}$$

nessa fórmula, d_j é o vetor de termos do documento; q representa o vetor de termos da consulta; w_{iq} são os pesos dos termos da consulta e w_{ij} são os pesos dos termos do documento.

Depois dos graus de similaridade terem sido calculados, é possível montar

uma lista (um *ranking*) de todos os documentos em ordem decrescente de graus de relevância em relação à consulta, dessa forma os documentos mais relevantes são mostrados primeiro [Harman, Fox, Baeza-Yates & Lee 1992].

A qualidade dessa lista de documentos relevantes retornada (*ranking*), pode ser avaliada com as medidas mostradas a seguir.

2.4 Avaliação de resultados

Depois de implementado, um sistema de recuperação de informação precisa ser avaliado, e, entre outros fatores, pode-se considerar a capacidade do sistema em retornar documentos que são considerados relevantes à necessidade de informação do usuário.

A avaliação pode ser realizada observando a qualidade do resultado retornado pelo sistema em relação a determinada consulta, e para isso é necessário conhecer o resultado ideal, ou seja, a lista de documentos que após a avaliação humana foram considerados como realmente relevantes àquela consulta.

Com esse intuito, coleções de referência são geradas,ou seja, essa coleção é composta por documentos que são conhecidos e analisados manualmente, de onde são identificados conjuntos (listas) de documentos considerados relevantes a determinadas consultas pré-definidas. Dessa forma tem-se: uma coleção de documentos conhecidos, as consultas e seus respectivos conjuntos de relevantes (conjunto ideal).

No intuito de avaliar um sistema de recuperação de informação em relação a uma coleção de referência, considera-se conhecido o conjunto de consultas e, para cada consulta seu respectivo conjunto ideal de relevantes. Então é feita a submissão dessas consultas, e, para cada uma delas, é realizada a comparação dos documentos retornados pelo SRI (ranking dos documentos retornados)

com o conjunto de documentos que realmente são relevantes (conjunto ideal).

Como apresentado na Figura 2.6, a submissão de uma consulta dessa base ao sistema em avaliação tem-se: A como um conjunto resposta retornado (ranking), R sendo o conjunto de documentos realmente relevantes à consulta, $R_a = R \cap A$ que são os documentos retornados que pertencem ao conjunto de relevantes, e aqueles retornados que não pertencem ao conjunto de relevantes.

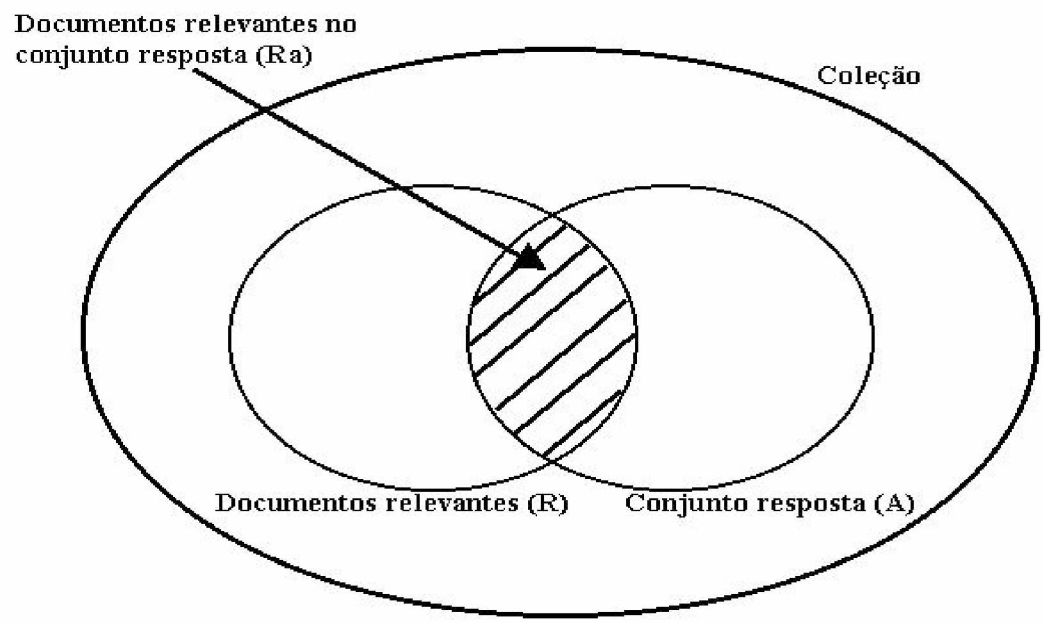

Figura 2.6: Elementos participantes do ambiente de avaliação de uma consulta em uma coleção de referência.

Observando a relação desses itens é possível obter a avaliação de sistemas com base na performance de recuperação. Aqui serão abordadas duas das medidas de avaliação mais utilizadas, a revocação (*recall*) e precisão (*precision*) e, a média harmônica (*harmonic mean*) [Baeza-Yates & Ribeiro-Neto 1999].

2.4.1 Revocação

A Revocação indica a capacidade de cobertura do sistema em relação ao conjunto de relevantes ideal conhecido.

Desta forma, em uma coleção de referência e dada uma consulta tem-se que o conjunto de documentos retornados pelo SRI em avaliação é comparado ao conjunto de relevantes ideal pertinente àquela consulta, que foi previamente definido, possibilitando medir a capacidade do sistema em trazer documentos que realmente são relevantes. A porcentagem de documentos relevantes presentes no conjunto de documentos retornados como resposta à consulta ($|R_a|$), em relação ao total de documentos presentes no conjunto ideal ($|R|$), define o índice de revocação do sistema em razão àquela consulta.

Dessa maneira, a revocação mede a capacidade do sistema em trazer os documentos que pertencem ao conjunto resposta ideal.

Definição: Revocação (Rev) é a fração dos documentos relevantes que foram recuperados, ou seja quanto do conjunto de documentos relevantes o sistema foi capaz de retornar.

$$Rev = \frac{|R_a|}{|R|} \tag{2.5}$$

Onde $|R_a|$ é a quantidade de documentos relevantes no conjunto $R_a = R \cap A$ e $|R|$ e a quantidade de documentos do conjunto ideal R.

2.4.2 Precisão

Uma outra medida de avaliação pode verificar a capacidade de um sistema em retornar somente documentos relevantes no *Ranking*, ou seja, do total de documentos retornados, quanto são considerados relevantes a uma consulta.

A relação existente entre o número de documentos que realmente são

relevantes e os que estão presentes no conjunto resposta retornado a uma consulta ($|R_a|$), e o total de documentos desse conjunto ($|A|$), define o conceito de precisão do sistema.

Assim pode ser verificada a qualidade da resposta do sistema em relação ao total dos documentos retornados, ou seja o quanto desse conjunto é considerado relevante, ou seja, o quanto o sistema consegue satisfazer a necessidade de informação do usuário sem trazer documentos não relevantes.

Definição: Precisão (P) é a fração do conjunto de documentos retornados que pertencem ao conjunto de relevantes, e mede a capacidade do sistema em deixar documentos não relevantes fora do resultado.

$$P = \frac{|R_a|}{|A|} \tag{2.6}$$

2.4.3 Média Harmônica

Uma combinação entre os valores de revocação e precisão do sistema, define a média harmônica. Essa medida pode ser utilizada com o intuito de realizar a avaliação dando a mesma importância à revocação e precisão.

A Média Harmônica (F) é a combinação valores de revocação e precisão. Onde: Rev_j é o valor de revocação do j-ésimo documento no ranking, e P_j é a precisão j-ésimo documento. [Baeza-Yates & Ribeiro-Neto 1999].

$$F_j = \frac{2}{\frac{1}{Rev_j} + \frac{1}{P_j}} \tag{2.7}$$

Conforme será apresentado no Capítulo 6, serão usados os conceitos de revocação, precisão e média harmônica de maneira relativa para a avaliação da qualidade do agrupamento gerado pelo AAL (Agrupamento por Autoridade Local).

Nesta proposta os grupos a serem avaliados devem ser gerados a partir de um processo não supervisionado. Para uma completa compreensão do processo, no próximo capítulo serão apresentados conceitos básicos sobre agrupamento não supervisionado (*clustering*) e, especialmente sobre o k-médias (*k-means*) que é o algoritmo base da proposta aqui apresentada.

Capítulo 3

Agrupamento não supervisionado

Neste capítulo serão apresentadas algumas técnicas de agrupamento e em especial o método k-médias que será utilizado como algoritmo base para a proposta de agrupamento apresentada nesta dissertação.

3.1 O processo de agrupamento não supervisionado

O homem tende a separar objetos em classes de elementos semelhantes, pois agrupar elementos similares facilita a localização de informações (por exemplo a disposição dos livros em bibliotecas tradicionais).

O processo de agrupamento se refere à separação automática de elementos similares em grupos sem a pré-definição das características desses grupos ou seja, os grupos são deduzidos (definidos) ao longo do processo baseando-se nas características dos próprios elementos candidatos ao agrupamento. De maneira intuitiva, na Figura 3.1 pode-se identificar dois grupos naturais exis-

tentes no conjunto de dados. No caso em que for calculada a distância entre os elementos existentes no conjunto dos dados candidato ao agrupamento, obtém-se os dois grupos (A e B) como resultado.

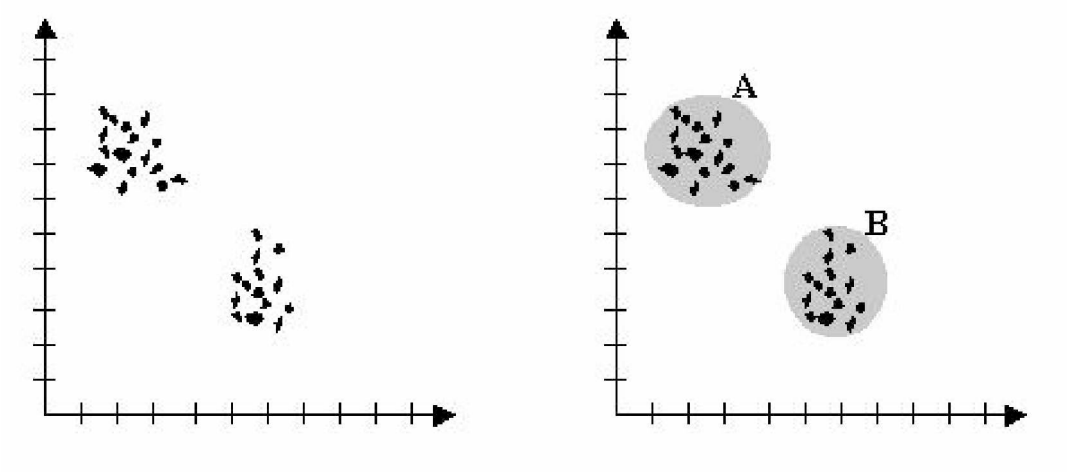

Figura 3.1: Exemplo de agrupamento.

O agrupamento identifica: um conjunto finito de grupos ou categorias que descrevam os dados; estes grupos são identificados a partir das características próprias dos dados; através do princípio de maximizar a semelhança intra-grupos e minimizar a semelhança intergrupos.

Enquanto o agrupamento é a geração de grupos a partir das características dos objetos (elementos) a serem agrupados, a classificação possui as classes pré-definidas e os elementos são alocados aos grupos (ou classes) com os quais mais se assemelham.

O processo de agrupamento não precisa necessariamente definir as classes, bastando separar os elementos em grupos. A posterior identificação das classes através de suas características é chamada de "análise do grupo"(*cluster analysis*), conforme [Willett 1988].

Segundo [Willett 1988], a separação dos elementos é feita com base numa

avaliação de similaridades entre os elementos, procurando colocar os elementos mais similares no mesmo grupo. Para tanto, devem ser escolhidas as características que irão representar cada objeto.

A análise de agrupamento é interessante quando não se tem um especialista para reconhecer a coleção, definir as classes e alocar os elementos, ou a coleção é muito grande e multivariada [Baeza-Yates & Ribeiro-Neto 1999].

3.1.1 Grupo

Em [Everitt 1980] são encontradas algumas definições de grupo (*cluster*):

Definição 1 - Um grupo (*cluster*) é um conjunto de entidades que se assemelham, e entidades pertencentes a grupos diferentes não são semelhantes.

Definição 2 - Grupo (*cluster*) pode ser definido como regiões conectadas de um espaço multidimensional contendo uma alta densidade relativa de pontos, separados de outras regiões por uma região contendo uma baixa densidade relativa de pontos.

Para [Willett 1988], grupo é uma coleção de objetos onde um é similar em relação a outro dentro do mesmo grupo e dissimilar em relação a objetos pertencentes a outro grupo. Ou seja, grupo é um conjunto de objetos, os quais possuem algumas características em comum.

O grupo possui um título (centro) capaz de representar todos os objetos nele contidos.

O agrupamento proposto tem como intenção formar grupos de documentos que possuem assuntos em comum.

Cada um desses grupos possui um representante que carrega consigo a missão de referenciar todos os elementos pertencentes àquele grupo. Esse representante será apresentado a seguir.

3.1.2 Representante do grupo

No processo de agrupamento de um conjunto de elementos, é possível utilizar-se de um elemento central capaz de representar todos os elementos deste conjunto. Este elemento é um identificador mais genérico capaz de representar o grupo como um todo, ou seja, todos os objetos nele contidos.

Em alguns métodos de agrupamento, para avaliar o grau de semelhança existente entre um elemento e o grupo, a comparação pode ser feita entre o candidato ao agrupamento e um representante do grupo. Isto facilita o processamento, pois evita a comparação com todos os elementos do grupo. O representante contém apenas um sumário das características comuns dos elementos do grupo, não representando fielmente todas as características do grupo, nem cada elemento individualmente, e sim o grupo em sua totalidade. Para que esse representante seja gerado o mais comum é a extração de características comuns nos elementos do grupo. Geralmente ele é criado durante o processo de agrupamento (pois, se fosse criado antes, o processo seria de classificação) e também pode ir sendo alterado e refinado durante a formação dos grupos e inclusão de elementos.

Douglass Cutting desenvolveu estudo e apresentou testes que confirmam que a utilização de um centro (chamado no caso de *profile* ou *cluster digest*), além de tornar mais rápido o processo de análise, corresponde aos resultados obtidos com a utilização de todos os elementos do grupo [Cutting, Karger & Pedersen 1993]. Mas, os resultados podem ser falhos se o conjunto de características selecionadas não conseguir representar corretamente o grupo.

Em alguns casos esse representante pode ser um centróide (*centroid*), que segundo [Willett 1988] é um representante (abstrato) da classe (grupo), sendo representado com as mesmas técnicas empregadas para os elementos participantes do agrupamento.

Podem ser extraídas as características que ocorrem em mais de um elemento do grupo. Esse elemento é o vetor central que contém todas as características dos demais vetores dos elementos pertencentes ao mesmo grupo.

O centróide é um conjunto de características centrais do grupo (*cluster*), que consegue representar todos os elementos que pertencem a ele (e somente estes).

No caso em que as características dos elementos forem representadas por pesos, pode-se também associar um peso às características do centróide. Para tanto, pode ser usada a média dos pesos de cada termo dos documentos pertencentes ao grupo representado por este centróide.

Na proposta de agrupamento definida neste trabalho, é feita a comparação do uso do centróide como representante do grupo com um representante que é um elemento participante do próprio grupo obtido, utilizando o conceito de análise de ligações. O representante será a maior autoridade local do grupo, cujo conceito será apresentado no Capítulo 6.

Além da definição do representante do grupo, existem outros fatores fundamentais para o processo de agrupamento, que serão mostrados na seção seguinte.

3.2 Componentes de uma tarefa de agrupamento

Em documentos não estruturados, caso tratado por essa dissertação, não há um local predeterminado que indique os atributos, ou seja, as características que melhor representam o documento. Portanto, é necessário estabelecer um método para identificar essas características que são capazes de caracterizar determinado objeto.

Em geral, o processo de agrupamento é composto por três etapas básicas: identificação e seleção de características, cálculo de similaridades e a etapa de agrupamento.

3.2.1 Identificação e seleção de características

Segundo [Willett 1988], todos os algoritmos de agrupamento estão de alguma maneira baseados em informações de medidas de similaridade entre os pares de elementos candidatos ao agrupamento. Mas, esse cálculo de similaridades só é possível quando primeiramente as características que serão usadas para representar os elementos forem identificadas e selecionadas.

Nessa etapa são identificadas as características que possuem um grau de importância para os objetos e depois é feita a seleção daquelas que possuem maior grau de discriminação. As características podem ser:

i) Originais - É a seleção do subconjunto de características originais mais importantes na descrição de um objeto, que serão usadas no processo de agrupamento.

ii) Transformadas - são realizadas operações nas características de entrada com o intuito de produzir novas características que serão importantes para o processo de agrupamento.

No caso de documentos que é o foco dessa dissertação, o conjunto de características que representam um documento geralmente é formado pelas palavras (termos) que o formam.

Além da definição das características, que em geral são extraídas por operações de pré-processamento e indexação (seção 2.2), têm-se outras fases importantes para o processo de agrupamento que é o cálculo da similaridade e o algoritmo de agrupamento.

3.2.2 Similaridade

Num processo de agrupamento existem vários meios de se avaliar a similaridade entre objetos, que podem ser elementos candidatos ao agrupamento ou centros dos grupos.

Em sua maioria os meios de cálculo de similaridade propostos procuram comparar vetores de características (elementos).

Os graus de similaridade entre os objetos (dados, documentos) são identificados baseando-se nas características identificadas na etapa anterior (seleção de características), utilizando por exemplo uma função de distância, que deve ser escolhida cuidadosamente. Em [Aamodt & Plaza 1994] a avaliação da similaridade subdivide-se em:

- Similaridade sintática - é a mais superficial, onde os atributos são comparados em termos de sua semelhança sintática.
- Similaridade semântica - propõe-se uma avaliação mais profunda, tentando abranger o significado dos casos e comparando-os.

A similaridade no caso da proposta aqui apresentada, trata-se de uma similaridade sintática, onde é verificada a ocorrência dos termos nos documentos definindo assim um grau de semelhança a ser calculado através da função cosseno entre vetores de termos de documentos, conforme apresentado na seção 2.3.3. Considerando esses conceitos tem-se subsídios para efetuar o agrupamento dos documentos.

3.2.3 Algoritmo de agrupamento

Nesta etapa é realizada a escolha do algoritmo de agrupamento, do qual dependerá a qualidade do resultado dos grupos que serão formados. O algoritmo vai definir a condição (regra) necessária para definir os grupos.

Um exemplo dessa condição é a imposição de um limite mínimo de similaridade de um elemento e outro, ou entre ele e o representante do grupo. Logo, se esse elemento satisfaz a restrição do grupo imposta, ele passa a fazer parte daquele grupo.

Geralmente os tipos de técnicas de agrupamento que podem ser definidos nessa etapa são hierárquicos e não-hierárquicos. Esses dois tipos serão detalhados na próxima seção.

3.3 Tipos de técnicas de agrupamento

Willett classifica os métodos de agrupamentos em hierárquicos e não-hierárquicos (particionais), conforme a existência ou não de relações hierárquicas entre os grupos [Willett 1988]. Em ambos os casos, o controle do processo de criação dos grupos pode utilizar os seguintes critérios de parada:

- número total de grupos;
- limite (mínimo ou máximo) do número de elementos em cada grupo;
- minimizar o erro quadrático;
- limite mínimo de similaridade entre o elemento e o representante do grupo;
- limite mínimo de similaridade entre os elementos;
- observando-se a presença obrigatória de características nos elementos.

3.3.1 Métodos Hierárquicos

No agrupamento hierárquico os grupos identificados são recursivamente analisados, fazendo com que as relações entre os grupos também sejam identifi-

cadas.

Esse tipo de algoritmo pode iniciar considerando cada objeto em um grupo distinto e durante o processo seguir fazendo a união de dois grupos que são considerados (após os cálculos de similaridade) mais próximos até que se tenha um critério de parada seja satisfeito [Jain, Mutty & Flynn 1999]. Nesse caso tem-se um agrupamento do tipo aglomerativo (*bottom-up*), construindo uma hierarquia de grupos conforme a Figura 3.2, onde os números indicam a ordem em que os grupos forem criados no processo aglomerativo.

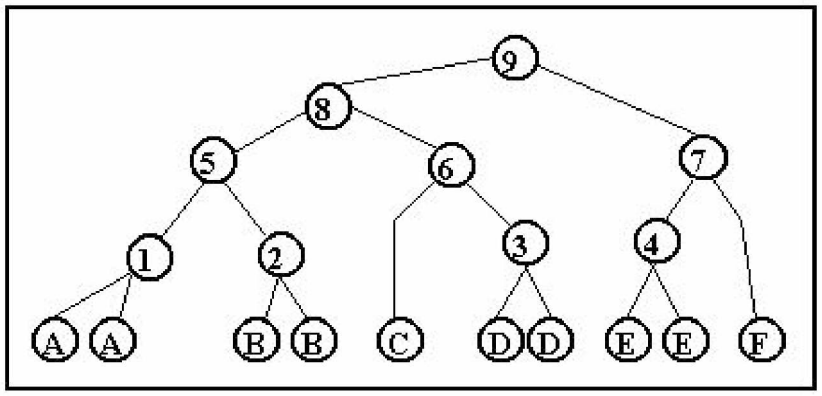

Figura 3.2: Resultado de agrupamento hierárquico aglomerativo.

No caso do exemplo mostrado na Figura 3.2, tem-se que o processo se inicia considerando que cada elemento define um grupo e, como é mostrado na Figura, os grupos mais comuns vão sofrendo fusões (aos pares) até que se tenha todos os elementos em um único grupo, como ocorre no nó raíz (9) da árvore. ‘

Um outro tipo de agrupamento hierárquico é o divisivo (*top-down*), onde inicialmente a coleção de objetos é considerada como um único grupo e, durante o processo, ocorre a decomposição em grupos menores até que um critério de parada seja satisfeito [Jain et al. 1999].

Nesses métodos são geradas árvores, seguindo a ordem inversa de formação em relação ao método aglomerativo. Ou seja, o processo é iniciado na raiz (nó 1), onde todos os elementos pertencem ao mesmo grupo, e a partir daí vão sendo realizadas divisões, que dependendo do critério de parada, pode acontecer de cada elemento definir um grupo. Como pode ser visto na Figura 3.3.

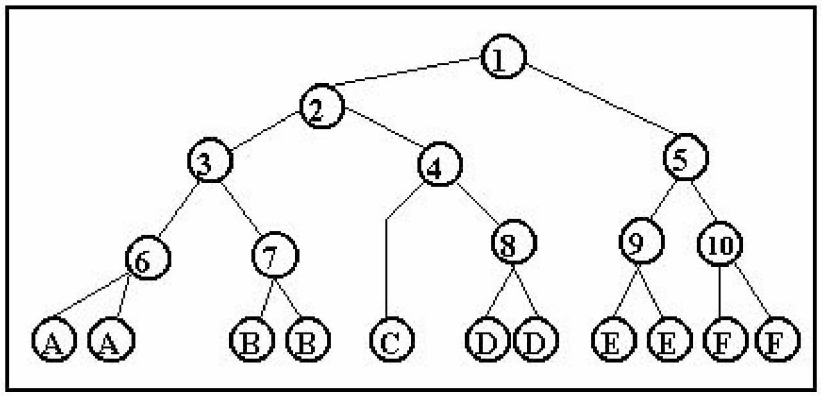

Figura 3.3: Resultado de agrupamento hierárquico divisivo.

Além dos métodos que definem essa relação hierárquica entre os grupos formados, existem também os métodos não-hierárquicos ou particionais que produzem grupos sem que exista essa relação hierárquica entre eles.

3.3.2 Métodos particionais (não-hierárquicos)

Os métodos não-hierárquicos ou particionais obtém os conjuntos agrupando os objetos constituindo grupos disjuntos e não hierárquicos, ou seja, o resultado é a geração de grupos (partições) que não possuem relação de hierarquia entre si.

Estes métodos procuram definir diretamente partições formadas pelos N elementos da coleção a serem agrupados em grupos distintos, de modo que

satisfaçam as premissas básicas: coesão interna e isolamento dos grupos. A coesão interna caracteriza grupos formados por elementos próximos uns dos outros, e o isolamento define grupos formados por elementos distantes dos elementos de outros grupos.

A maior vantagem desse método é poder atuar sobre conjuntos com elevado número de objetos, pois tais métodos em geral têm complexidade $O(N)$, onde N é o número de objetos do conjunto de dados [Halkidi, Batistakis & Varzigiannis 2001].

Essa técnica consiste em dividir os elementos em um número pré-definido de k grupos distintos.

Apesar de constituir grupos distintos, o particionamento não-hierárquico permite a possibilidade de colocar ou não determinado elemento em mais de um grupo. Quando os elementos são atribuídos a um único grupo diz-se que o processo é disjunto. Caso um elemento possa ser atribuído a mais de um grupo por possuir forte relação com mais de uma partição, diz-se que o processo não é disjunto.

Geralmente os algoritmos adotam restrições que impedem que um elemento pertença a mais de um grupo, atribuindo o objeto ao grupo de maior relação (similaridade).

Na Figura 3.4 os grupos são representados pelos grandes círculos que envolvem os objetos considerados semelhantes.

Como pode ser visto na Figura 3.4, os grupos não possuem ligações entre si, sendo totalmente isolados, assim, os documentos são totalmente separados.

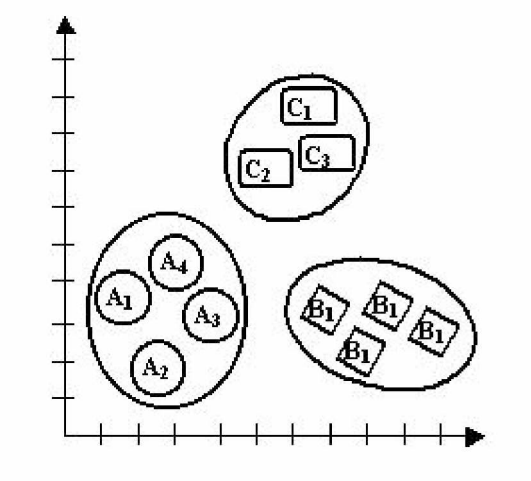

Figura 3.4: Resultado de um agrupamento por partição disjunta.

3.4 O algoritmo k-médias

O algoritmo k-médias (*k-means*) [Macqueem 1967], é muito conhecido e freqüentemente utilizado na análise de agrupamentos [Han & Kamber 2001]. Essa técnica exige a predefinição de critérios que produzam medidas sobre a qualidade da partição.

O algoritmo k-médias, é uma proposta de agrupamento baseada no centróide, ou seja, cada grupo possui um representante, que neste caso é o centróide do grupo. Esse centróide é um vetor médio calculado a partir dos valores das medidas dos elementos pertencentes àquele grupo, o qual é o ponto de comparação do grupo por ele representado. Assim, todos os elementos que estiverem próximos de um determinado centróide, passam a pertencer àquele grupo.

Os grupos são formados para otimizar um critério de particionamento objetivo, freqüentemente chamado de função de similaridade. Então os objetos dentro de um mesmo grupo são similares entre si, e dissimilares em relação

aos objetos pertencentes aos outros grupos [Han & Kamber 2001].

Dado uma base de N objetos e k partições (grupos a serem formados), onde $k \leq N$, existirão k centróides, representando esses grupos. Uma dificuldade do algoritmo é definir o k ótimo, ou seja, o número de grupos naturais da coleção.

Segundo [Han & Kamber 2001], o algoritmo toma o parâmetro de entrada k, e particiona um conjunto de N objetos em k grupos, de modo que a similaridade intragrupo é alta e a similaridade intergrupos é baixa. Sendo $X = \{x_1, x_2, \ldots, x_n\}$ a coleção de objetos, $G = \{g_1, g_2, \ldots, g_k\}$ o conjunto dos k grupos a serem formados, e $C = \{c_1, c_2, \ldots, c_k\}$ o conjunto dos k centróides representantes dos k grupos, um objeto x_i é atribuído ao grupo g_j que ele é mais similar, baseado na distância entre esse elemento e o centróide c_j. Depois que todos os elementos tenham sido alocados aos grupos de maior similaridade é feita a redefinição dos centróides dos grupos. Isso ocorre até que o critério de convergência seja satisfeito, o que pode ser quando os elementos não mais alternam entre os grupos, ou o erro quadrático, mostrado na equação 3.4 a seguir, pára de decrescer.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|d_i^{(j)} - c_j\|^2 \tag{3.1}$$

Onde $d_i^{(j)} j$ é o i-ésimo documento pertencente ao j-ésimo grupo e c_j é o centróide do j-ésimo grupo. Assim, em k grupos, n é o número total de objetos e $\|x_i^{(j)} - c_j\|$ é a medida de distância entre um objeto e o centróide do grupo em que ele está inserido. A função obtém a soma das distâncias dos n objetos aos seus respectivos centróides.

Abaixo estão informalmente apresentados os passos envolvidos no algoritmo k-médias:

Passo 1 O usuário/especialista deve especificar o número k de partições

Passo 2 O algoritmo escolhe, aleatoriamente, k pontos que serão os centros dos grupos iniciais - centróides.

Passo 3 Determinar, para cada elemento da coleção de dados, o grupo ao qual ele pertence. Isso é feito realizando o cálculo da similaridade entre o elemento e o centróide de cada grupo.

Passo 4 O elemento será alocado ao grupo do qual ele estiver mais próximo do centro.

Passo 5 É calculado um novo centróide para cada *cluster* (redefinição do centro que vai representar o grupo), ou seja, os pontos iniciais não são os centros definitivos dos grupos, eles são apenas uma tentativa inicial.

Passo 6 O processo se repete até que haja a convergência do processo. Que pode ser quando os centros dos *clusters* se estabilizem, isto é, o mesmo ponto é escolhido como centro durante algumas iterações.

Um exemplo é um conjunto de objetos candidatos ao agrupamento, como na Figura 3.5. Considerando que o usuário indicou três grupos ($k = 3$).

De acordo com o algoritmo k-médias descrito anteriormente, serão escolhidos aleatoriamente 3 elementos para serem os centróides iniciais. Onde os centróides serão marcados por "+".

Cada objeto será distribuído para um grupo ao qual ele se encontra mais próximo, isso será definido baseando-se na medida de similaridade em relação ao centróide. Mostrado na parte (a) da Figura 3.5.

Em seguida será calculado o novo centróide de cada um dos grupos, baseado nos elementos pertencentes àquele grupo. Agora é feita a redistribuição

de todos os elementos em relação aos novos centróides, o que pode ainda ser observado na parte (b) da figura.

Esse processo ocorre até que os grupos não se modifiquem mais (c).

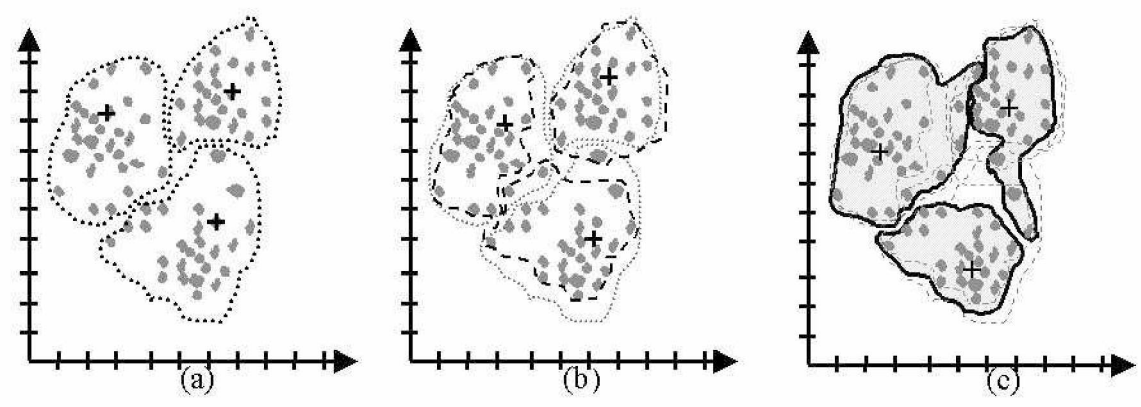

Figura 3.5: Exemplo de agrupamento baseado no algoritmo k-médias.

Algumas características do k-médias podem ser observadas:

- Algoritmo é significativamente sensível à localização inicial dos centróides.
- Pode acontecer que um grupo seja vazio, isto leva a que esse grupo não possa ser atualizado.
- Os resultados obtidos dependem do valor de k que foi definido.

A proposta de agrupamento dessa dissertação é inspirada no algoritmo k-médias aqui descrito e no conceito de autoridade que será apresentado no próximo Capítulo.

O agrupamento de documentos textuais tem como objetivo principal identificar documentos que possuam características comuns e separá-los agrupando-os em classes de documentos que contenham assuntos similares. O que pode ser exemplificado na figura 3.6:

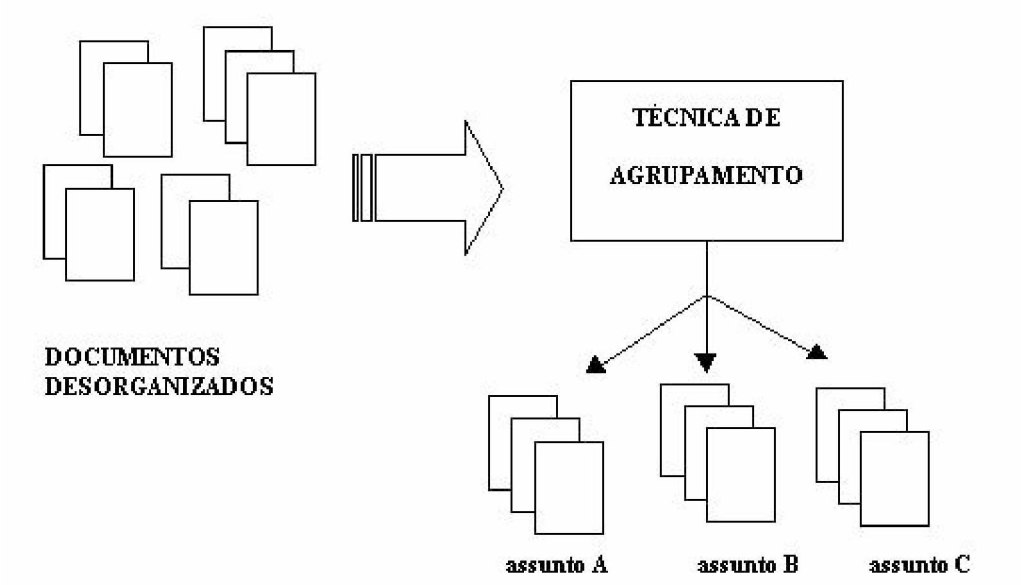

Figura 3.6: Exemplo de agrupamento de documentos textuais.

O processo de agrupamento baseia-se na hipótese de que documentos semelhantes tendem a permanecer em um mesmo grupo, pois possuem atributos em comum. Estes objetos pertencentes ao mesmo grupo, tendem a ser relevantes ao mesmo assunto. Essa hipótese, segundo [Van Rijsbergen 1979], é conhecida por hipótese de agrupamento (*cluster hypothesis*).

Capítulo 4

Análise de ligações

Este capítulo tem como objetivo apresentar uma visão geral de análise de ligação que fundamenta as modificações propostas nesta dissertação para o processo de agrupamento de documentos. Aqui são apresentados os principais passos do funcionamento do algoritmo clássico HITS [Kleinberg 1999] que é uma ferramenta para identificação da maior autoridade presente no grupo.

4.1 O processo de análise de ligação

Uma das informações mais ricas a serem extraídas de um ambiente de hiperligações (*hyperlinks*) é a sua estrutura de ligações [Calado et al. 2003]. A análise dessa estrutura, conhecida por análise de ligação é o processo de avaliar as ligações existentes entre documentos, ou seja, suas inter-referências (geralmente mencionadas como *links*). Esta análise contribui de maneira efetiva no processo de recuperação de informação no ambiente da *Web*, visto que resgata o conhecimento embutido nas ligações de documentos determinado subjetivamente pela atitude humana.

O estudo do comportamento das citações existentes entre documentos já tinha sido realizado na biblioteconomia antes do surgimento da *WWW*. Em alguns estudos o objetivo era encontrar, entre os artigos científicos publicados em jornais, quais eram os mais influentes em determinada área e a existência de co-citações entre eles, ou seja, a freqüência com que eles são citados por outros documentos.

Kleinberg [Kleinberg 1999], referenciando [Garfield 1972], cita uma das mais conhecidas medidas da área, que é o fator de impacto de Garfield, que define o número de citações a um documento no intervalo de dois anos.

As citações entre artigos científicos diferem das ligações entre páginas da *Web* por serem estáticas e unidirecionais, ou seja, após a publicação de um artigo não é possível haver modificações em suas referências, o que é diferente no ambiente *Web*, pois constantemente a estrutura de uma página é modificada (também as suas ligações) após a sua publicação inicial.

Segundo [Kleinberg 1999], a estrutura de ligações é uma informação que geralmente traduz o julgamento humano sobre os documentos.

As ligações existentes em uma página *Web* podem ter a função de conectar páginas de um mesmo site, de dar acesso a outros sites de assuntos que podem ser relevantes ou de propaganda (que não refletem a opinião do autor sobre a ligação criada).

A estrutura de ligações da *Web* pode ser analisada com o intuito de extrair informações relevantes que podem ser utilizadas para vários fins, como por exemplo: (i) associação de pesos de importância a páginas *Web* [Kleinberg 1999], [Amento, Terveen & Hill 2000]; (ii) identificação de páginas que se auto-referenciam e possuem um assunto específico, formando assim uma comunidade na *Web* [Flake, Lawrence & Giles 2000]; (iii) classificar documentos na *Web* [Calado 2004].

A associação de pesos de importância a páginas da *Web*, gera a definição de autoridades, ou seja, um documento é autoridade quando ele é referenciado por muitos outros. E, da mesma forma é possível identificar *hubs* que são páginas que referenciam muitas outras.

O algoritmo HITS para calcular o grau de boas autoridades e *hubs*, usa as informações da estrutura de ligações que cercam um documento, definindo o que no HITS é chamado de conjunto raiz de documentos.

Esse conjunto é então expandido com os documentos vizinhos (documentos que apontam para ele ou são referenciados por documentos do conjunto) formando assim um conjunto base de documentos. Por essa razão o algoritmo calcula os valores de hubs e autoridades locais. No contexto deste trabalho, tanto o conjunto raiz quanto o conjunto base são locais. Esse cálculo também pode ser aplicado a uma coleção, nesse caso o algoritmo calcula valores de autoridade e *hub* global para cada documento.

Para que o cálculo do índice de autoridade de uma página *Web* seja realizado através do algoritmo HITS, o processo se inicia definindo um grafo direcionado, como é mostrado a seguir.

4.2 Representação das ligações através de Grafos *Web*

Para o HITS, uma coleção de documentos *Web* com ligações entre si é interpretada como um grafo direcionado onde os documentos correspondem aos nós e as ligações (*links*) entre eles correspondem às arestas[1].

Cada documento da *Web* contém um determinado número de ligações de saída (*forward links* ou *outlinks*), e um determinado número de ligações de

[1] (d,p) representa a aresta referente a ligação (*link*) de d para p

entrada (*backlinks* ou *inlinks*).

O número de ligações de saída corresponde ao número de ligações no corpo do documento, e as de entrada são ligações de outros documentos que referenciam o documento em análise. No caso das ligações de entrada, existem problemas para efetuar os seus cálculos, devido às proporções da *Web* e sua alta taxa de atualização.

4.3 O algoritmo HITS

O algoritmo *Hyperlinked Inducted Topic Search* (HITS), proposto em [Kleinberg 1999] é utilizado para encontrar, em meio a um subconjunto de documentos relacionados a uma consulta, aqueles melhor referenciados (autoridade) e os documentos que possuem o melhor conjunto de apontadores (*hubs*) referenciando outros documentos (*links*). Como pode ser visto na Figura 4.1 Um *hub* é uma página que contém ligações para muitas outras páginas e, uma autoridade é uma página que é apontada por muitas outras.

Os valores de *hub* e autoridade são calculados baseado na relação de reforço mútuo. Essa relação estabelece que um bom *hub* é aquele que possui ligações para boas autoridades e boas autoridades são aquelas referenciadas por bons *hubs*.

Sendo um conjunto de documentos resposta a uma consulta, o valor de *hub* e autoridade local de cada documento pode ser calculado usando a estrutura de ligações dos documentos desse conjunto. A partir dessas informações é definido o conjunto base adicionando os documentos vizinhos aos já pertencentes ao conjunto raiz inicial.

A seguir apresentamos os passos do algoritmo HITS.

Passo 1 Construção do Grafo Base: O grafo base é formado pelo conjunto

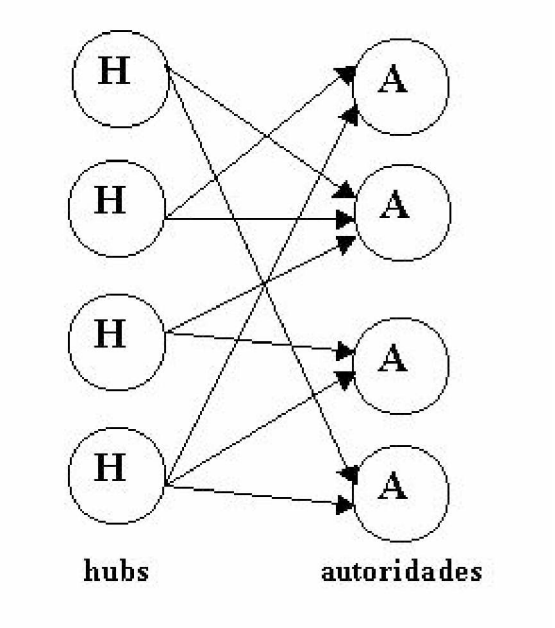

Figura 4.1: *Hubs* e autoridade.

resposta a uma consulta e sua vizinhança

1. Uma consulta do usuário é submetida;
2. Dentre os elementos pertencentes ao resultado da busca que foi retornado, os melhores dessa relação formam o conjunto raiz.
3. A expansão do conjunto raíz ocorre, agregando os documentos que referenciam (apontam) e os que são referenciados (apontados) por documentos contidos no conjunto raiz, formando assim o GRAFO BASE.
4. Eliminar arcos que conectam arcos direcionados que conectam nós em um mesmo domínio.
5. O resultado é um grafo direcionado contendo N nós.

Passo 2 Cálculo dos Pesos de *hub* e Autoridade:

1. inicialmente associa-se a cada nó (página) do grafo base um peso = 1 tanto para o valor de *hub* como para autoridade.

2. associa-se a cada nó (página) do grafo base um peso de *hub* e um peso de autoridade, que é calculado de acordo com a relação de reforço mútuo através de sucessivas iterações das seguintes equações de 4.1 e 4.2:

$$a(p) = \sum_{d|(d,p)\in E} h(d) \tag{4.1}$$

$$h(p) = \sum_{d|(p,d)\in E} a(d) \tag{4.2}$$

   O par ordenado (d,p) representa uma aresta partindo do nó d e chegando ao nó p, a qual pertence ao conjunto E de todas as arestas do grafo.

3. o cálculo dos pesos de *hub* e autoridade repete-se até o momento em que os valores de autoridade e *hub* não mais variam acima de um valor pré-definido entre sucessivas iterações. Neste momento diz-se que esta computação converge.

Passo 3 Filtro de *hubs* e Autoridades: após a convergência uma lista contendo os documentos maiores autoridade e os maiores *hubs* é retornada.

O algoritmo HITS foi apresentado com objetivo de que identificar a maior "autoridade local"existente nos conjuntos em formação durante o processo de agrupamento.

Capítulo 5

O Agrupamento por Autoridade Local

Neste capítulo será apresentada a proposta de agrupamento por autoridade e conteúdo. Para tanto, são usados os conceitos já apresentados de Recuperação de Informação, agrupamento e análise de ligações. No próximo capítulo será mostrado por meio de experimentos que a análise de ligações incorporada à estratégia agrupamento automático baseado na similaridade textual pode melhorar a qualidade da definição dos grupos de documentos formados.

5.1 Definição do Problema

A área de Recuperação de Informação reconhece a cerca de 40 anos os chamados dados não-estruturados como documentos que, como já foi mostrado, são dados que possuem um grande potencial de transmitir informação. No caso da Internet, apesar destes possuírem diferentes informações e estruturas, comumente sua forma é textual, o que dificulta a identificação de suas

características importantes.

Com isso, vêm acontecendo estudos e testes de algoritmos que lidam com dados não-estruturados,por exemplo, estudos de aplicação de técnicas de agrupamento automático em documentos textuais.

Segundo [Feldman & Hirsh 1997], o problema é que a maioria das técnicas de agrupamento foi desenvolvida para atuar sobre dados estruturados, ou seja, dados convencionais, armazenados em Sistemas de Gerência de Bancos de Dados, mais fáceis de serem tratados por meios computacionais.

No caso de documentos não-estruturados pertencentes a um ambiente que está em constante expansão como a Internet, ocorre o fato do usuário ter que manusear uma grande quantidade de informações, gerando sobrecarga de informações sobre ele.

5.2 Proposta

A Internet apresenta um crescimento em escala exponencial e, portanto, o volume de informações sobre determinado assunto torna-se cada vez maior. Isso dificulta o trabalho do usuário que em busca de determinada informação necessita selecionar dentre toda a coleção o grupo de documentos que merecem ser analisados. Observando esse fato e aproveitando das informações que podem ser extraídas da estrutura da *Web*, torna-se interessante a proposta de agrupamento por conteúdo e autoridade apresentada neste trabalho.

Na Recuperação de Informação (RI) o agrupamento tem sido utilizado há algum tempo com o intuito de organizar elementos potencializando assim a recuperação de documentos similares. Segundo [Van Rijsbergen 1979] e partindo do princípio da hipótese de agrupamento (*Cluster hypothesis*) documentos semelhantes tendem a ficar em um mesmo grupo.

A proposta deste trabalho é a análise de agrupamento (*clustering*) voltada para documentos textuais da *Web*, utilizando os conceitos de similaridade por conteúdo com análise de ligações (autoridade). Os grupos em formação serão representados pelas autoridades naquele assunto.

Tanto os documentos candidatos ao agrupamento, como os representantes dos grupos possuem a mesma representação, que geralmente são utilizados vetores de características/termos, o que viabiliza os cálculos.

Seguindo a idéia de referenciar um grupo pelo seu representante, na intenção de identicar o grupo com o qual o objeto mais se assemelha, o cálculo da similaridade dos objetos com cada um dos grupos pode ser realizado comparando apenas esse objeto com cada um dos representantes dos grupos ao invés de compará-lo com todos os elementos de cada um dos grupos.

Em um ambiente como a *Web*, a análise da estrutura de ligações traz informações bastante úteis, pois carrega consigo o julgamento do autor sobre o documento.

Observando esse ambiente de ligações (*Web*), com o propósito de gerar grupos definidos por assuntos neste trabalho, foi elaborado o AAL, um algoritmo de análise de agrupamentos que associa informações de conteúdo extraídas do modelo vetorial e de autoridade local obtidas a partir do algoritmo HITS proposto por Kleinberg.

Para validar este algoritmo foi feita a implementação e realizados testes do algoritmo aplicado a uma amostra da base real da *Web* brasileira de 1999, a WBR99 [Calado 2003][1]. Os experimentos serão discutidos no capítulo seguinte.

[1] A WBR99 foi cedida aos laboratórios da Facom exclusivamente para uso acadêmico

5.3 Especificação do algoritmo *AAL*

O algoritmo *AAL* (Agrupamento por Autoridade Local) proposto aqui, agrega informações de conteúdo e análise de ligações.

Objetivando a formação de grupos distintos de acordo com o assunto que tratam, não possuindo assim relação hierárquica entre si, a inspiração dessa proposta está em gerar partições como no k-médias, porém definindo grupos que utilizarão a informação de autoridade local para expressarem assuntos específicos.

Como visto no modelo vetorial, a medida de similaridade usando o cosseno do ângulo entre os vetores fornece o grau de similaridade entre eles. Essa medida pode ser usada na avaliação do grau de semelhança entre um documento e os representantes dos grupos para especificar a qual grupo ele irá pertencer. Assim é identificada uma semelhança do documento em relação ao assunto representado naquele grupo. Para tanto, é necessário definir o representante que, por se tratar de um agrupamento de documentos da *Web*, a similaridade entre cada documento e o elemento representante do grupo é calculada considerando as informações de conteúdo (termos/características) dos documentos.

A análise de ligações é utilizada de maneira recursiva na definição dos representantes dos grupos, onde esse serão as autoridades locais.

Intuitivamente, os grupos serão obtidos conforme a execução dos seguintes passos:

Seja uma coleção D com n documentos $D = \{d_1, d_2, \ldots, d_n\}$:

Passo 1 De acordo com a quantidade de documentos e a pluralidade de assuntos abordados pelos mesmos, define-se a quantidade de grupos a serem formados. Seja k a quantidade de grupos a serem formados.

Passo 2 Aleatoriamente seleciona-se k documentos para serem representantes dos grupos. Sejam $c_1, c_2, \ldots, c_k$ os representantes dos respectivos grupos $G = \{g_1, g_2, \ldots, g_k\}$ a serem formados.

Passo 3 Calcula-se a similaridade entre cada documento d_j da base com cada representante c_i.

Passo 4 Então, é feito o particionamento da coleção em k-grupos, alocando cada documento d_j ao grupo g_i de maior similaridade entre o documento d_j e o representante c_i, ou seja

$$\forall j,\ g_i = g_i \cup d_j\ se \forall\, l \neq i\ sim(d_j, c_i) > sim(d_j, c_l) \tag{5.1}$$

Passo 5 Redefine-se os representantes dos grupos que serão as respectivas autoridades locais de cada grupo g_i tendo como base o algoritmo HITS [Kleinberg 1999].

Passo 6 Compara-se o conjunto de representantes com o conjunto anterior. Se houver mudança, voltamos ao passo 3, caso contrário, finalizamos o processo de agrupamento proposto

Observe que o algoritmo proposto é uma versão do k-médias discutido na seção 3.4, alterando-se o passo 5, que define o representante do grupo.

Capítulo 6

Avaliação do AAL

Com o objetivo de avaliar o desempenho do AAL, neste capítulo serão abordados os resultados obtidos a partir de alguns testes realizados em comparação ao tradicional algoritmo de agrupamento k-médias. O desempenho aqui avaliado trata principalmente da qualidade dos grupos gerados pelo algoritmo proposto.

6.1 Recursos da avaliação

Objetivando-se a análise desempenho da técnica que agrega informações de ligações (AAL) ao processo de agrupamento não supervisionado, que utiliza informações de conteúdo dos documentos (k-médias), nesta seção são comparados os grupos formados como resultado da implementação dessas duas técnicas. São realizados alguns experimentos cujos resultados são comparados com grupos ótimos de documentos de mesmo assunto.

Na avaliação do desempenho da técnica de agrupamento proposta, é utilizada uma adaptação das tradicionais medidas de precisão e revocação, bem como a média harmônica, que são técnicas comumente utilizadas na área de

Recuperação de Informação.

Para a viabilização dos testes são implementados tanto o algoritmo de agrupamento proposto, o *AAL*, como o tradicional k-médias. Isto é feito para avaliar o impacto na precisão e revocação relativas causado pela fusão do paradigma de análise de ligações (HITS) com o da Recuperação de Informações (modelo vetorial) no processo de agrupamento.

São realizados experimentos utilizando as propostas dos algoritmos *AAL* e k-médias apresentados anteriormente. O número de documentos, o número de grupos e os k primeiros elementos escolhidos aleatoriamente, são os mesmos em cada teste para ambos os algoritmos.

A coleção de documentos WBR99 é utilizada na realização dos testes. Esta coleção teve seus documentos recolhidos da base de dados do engenho de busca TodoBR em Novembro de 1999, e cedida à Faculdade de Computação da Universidade Federal de Uberlândia pela Akwan Information Tecnologies para uso exclusivo de pesquisa em problemas de Recuperação de Informação. A coleção tem um tamanho de 20G, contendo cerca de 6 milhões de páginas; uma lista invertida de todas as páginas; uma lista de todas as palavras de cada página; uma lista de todos os links entre páginas e o texto das páginas, depois de removidas todas as tags HTML.

Com o objetivo de trabalhar com uma amostra de execução viável no ambiente computacional disponível (seção 6.3) e que possuísse as características mais próximas da coleção original, buscou-se coletar os documentos ao longo de toda a coleção original. Com esse intuito,selecionou-se aleatoriamente um dos 100 primeiros documentos e, varrendo a coleção com saltos de 100 documentos, selecionou-se cerca de 60.000 documentos para a massa de testes. Este conjunto de 60.000 documentos é aqui denominado coleção_de_teste.

6.2 Projeto e Implementação do *AAL*

Na implementação do algoritmo *AAL* foi utilizada a linguagem de programação C++. E, na agregação da informação de ligações ao processo de agrupamento, considerou-se o fator *In degree* de um documento (site), o que corresponde ao número de sites que o referenciam.

Dessa forma, no cálculo da autoridade local leva-se em conta o estudo realizado por Amento et al, que dentre várias medidas de estimativa de qualidade dos documentos da *Web* considerando a estrutura de ligações ou o conteúdo, ele observou pequenas diferenças entre as mesmas. Nesses testes foi identificado que três medidas baseadas em ligações e uma baseada em conteúdo são eficazes na tarefa de identificação da qualidade dos documentos. Após os vários testes efetuados, Amento et al concluiu que a medida *In-degree* é, no mínimo, igual quando comparada com os mais sofisticados algoritmos de Cálculo de Autoridade e *PageRank* [Amento et al. 2000].

Nesse artigo, Amento et al também afirma que não existe diferença significativa entre *In degree* e autoridade, sendo esses particularmente similares.

Sendo assim, devido a eficácia da medida *In-degree*, observada no referido artigo, e objetivando uma versão de menor custo computacional do algoritmo *AAL*, nesta dissertação utiliza-se a medida *In-degree* na definição dos representantes dos grupos em substituição ao algoritmo HITS citado no Capítulo 5.

6.2.1 Estruturas de Dados utilizadas

Dentre as estruturas de dados utilizadas no sistema, ressalta-se: a coleção de documentos (um conjunto de documentos), os termos, árvores e listas encadeadas. Dado um conjunto de documentos, para cada documento é

determinada uma lista de termos (lista de palavras existente no documento) sendo que para cada termo é definido um conjunto de 2 campos (com 4 bytes cada), onde o primeiro representa o termo (palavra) e o segundo representa quantas vezes o termo aparece no documento.

Para cada conjunto de documentos é definida uma árvore, sendo que cada elemento da árvore (nó) contém as informações sobre um documento, ou seja, o endereço da lista de termos do documento no arquivo que contém todos os documentos.

A Figura 6.1 ilustra as estruturas utilizadas para do desenvolvimento do programa (implementação) do *AAL*.

Para acessar as estruturas, são desenvolvidas rotinas em C++ que percorrem toda a árvore fazendo chamadas recursivas a cada novo nó encontrado na árvore. Para calcular a similaridade entre dois documentos, são carregados os termos de cada lista em uma lista ligada de palavras e, a partir das duas listas carregadas realizou-se o cálculo da similaridade entre os documentos.

A árvore é utilizada para otimizar o processo do cálculo da similaridade entre os documentos. Para obter melhoria de desempenho no acesso aos dados, ao calcular a similaridade entre duas listas de termos, verificou-se a existência de determinado termo na outra lista, reduzindo o número de comparações, independente da quantidade de termos na lista de documentos. A árvore é balanceada quando está sendo carregada, isto significa que tem a mesma quantidade de termos nos dois lados da árvore (esquerdo e direito); quando inicia o processo de verificar se um termo existe na lista, logo na primeira comparação já é eliminada a metade da lista.

A cada novo termo adicionado à lista, é criada uma estrutura contendo as informações do termo, e o ponteiro (endereço de memória) desta estrutura é incluído na árvore. O valor numérico do termo é comparado com o nó

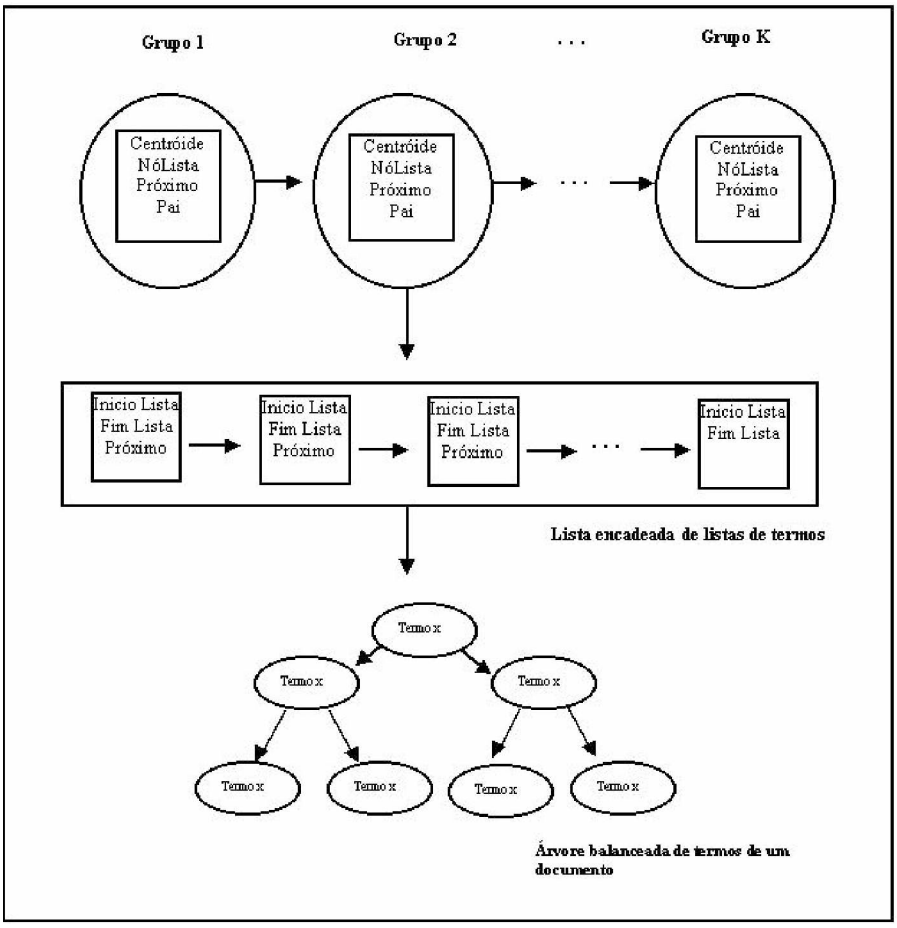

Figura 6.1: Estruturas utilizadas na implementação do *AAL*.

principal da árvore, se for maior segue para a direita, caso contrário, para a esquerda; esta comparação é feita até encontrar um local na árvore de forma que ao percorrer a árvore da esquerda para a direita os valores numéricos (termos) estejam em ordem crescente.

6.2.2 Funcionamento do programa

O programa desenvolvido trata-se da implementação de um processo de agrupamento não supervisionado conforme o método proposto pelo algoritmo

AAL, que define como representante do grupo a maior autoridade local (seção 5.3). E uma outra versão que é o tradicional k-médias (seção 3.5), onde o representante do grupo é definido pelo cálculo do centróide. Ele processa uma coleção de documentos, gerando como saída: os k grupos dados como entrada pelo usuário; um arquivo texto (.txt) de cada grupo contendo as URLs de todos os documentos nele contidos; e, um arquivo que descreve as etapas percorridas durante o processo de agrupamento.

6.3 Testes Realizados

Os testes realizados tiveram como objetivo a avaliação da qualidade dos grupos processados (gerados), ou seja, a capacidade do algoritmo em agrupar documentos que são semelhantes. Isso é feito observando os resultados retornados pela implementação do algoritmo aqui proposto, o *AAL*, e os retornados pelo k-médias.

O equipamento utilizado foi um micro computador com processador Pentium IV de 2GHz de *clock*, HD IDE de 80GB e 512MB de RAM.

O processo de avaliação do algoritmo de agrupamento (*AAL*) utilizando a coleção_de_teste de 60.000 documentos selecionada a partir da coleção completa é realizado considerando os seguintes elementos básicos de observação: Partição ótima e resultado retornado pelo agrupamento.

A partição ótima é gerada a partir da seleção manual de documentos que compartilham do mesmo assunto e estão presentes na coleção_de_teste. Dado um assunto, a partição ótima aqui é obtida por meio de uma seleção de URLs que sugerem o assunto. Selecionadas as URLs é feita a análise manual do conteúdo e decidida a pertinência na partição ótima. Então a partição ótima embora correta, não é completa, pois nem todas as URLs da coleção_de_teste

são observadas.

O resultado retornado pelo agrupamento será representado pelo grupo que possuir uma maior quantidade de documentos pertencentes à partição ótima. Os documentos formadores desse grupo constituirão o conjunto resposta a ser comparado com a partição ótima.

Cada teste consistiu em comparar o conjunto resposta retornado pelo agrupamento com a respectiva partição ótima. Sendo que, como o intuito é comparar o algoritmo de agrupamento proposto nesse trabalho (AAL) e o tradicional k-médias, cada algoritmo é executado com os mesmos k primeiros elementos e tiveram os seus resultados comparados com a mesma partição ótima, conforme descrito a seguir.

6.4 Roteiro de um teste

- Define-se um assunto e sua respectiva partição ótima, denotada por R.
- Define-se o número de grupos a serem formados pelos algoritmos, denotado por k. Em todos os testes utiliza-se k=12, baseado no número de classes usadas em um conhecido diretório de Web brasileira, o Cadê (www.cade.com.br).
- Executa-se o algoritmo de agrupamento AAL e o k-médias tradicional, cada um retornando os 12 grupos resultantes.
- Obtem-se qual o grupo dentre os 12 retornados será o representante do conjunto resposta durante a avaliação, denotado por A.
- Calcula-se as medidas de avaliação de A a saber: revocação, precisão e média hârmonica adaptadas para avaliar grupos de documentos da forma mostrada a seguir.

Definição das medias de avaliação[1]

As medidas de revocação, precisão e média harmônica são aqui adaptadas para a avaliação de grupos de documentos.

Sejam:

A= o conjunto resposta da execução do algoritmo de particionamento.

R= a partição ótima.

$R_a = R \bigcap A$.

$|R|$ o número de documentos na partição ótima.

$|R_a|$ o número de documentos pertences à Interseção entre R e A.

Então a revocação é aqui definida como sendo a relação entre o número de documentos da partição ótima que o algoritmo manteve no conjunto resposta pelo número total de documentos na partição ótima. Então dessa maneira tem-se:

$$Rev = \frac{|R_a|}{|R|} \tag{6.1}$$

A precisão define a capacidade do algoritmo em gerar grupos que contenham apenas documentos do mesmo assunto. Na avaliação aqui proposta, tem-se a precisão relativa, que é estabelecida a relação entre documentos o número de documentos da partição ótima que o algoritmo manteve no conjunto resposta pelo número total de documentos pertencentes ao conjunto resposta.

$$P = \frac{|R_a|}{|A|} \tag{6.2}$$

A média harmônica de cada teste relaciona a revocação e a precisão e

[1]Uma adaptação semelhante foi realizada em [Calado 2004] para avaliação de algoritmo de classificação.

Parâmetro	Descrição
K	Quantidade de grupos processada
N	Número de documentos existentes na coleção de teste
X	Número de documentos da partição ótima
Rev	Valor de revocação
P	Valor de precisão
F	Valor da média harmônica

Tabela 6.1: Parâmetros utilizados na avaliação

define valores como sendo:

$$F = \frac{2}{\frac{1}{Rev} + \frac{1}{P}} \tag{6.3}$$

Dessa forma, os testes realizados resultam em valores revocação (Rev), precisão (P) e média harmônica (F) relativos pois os mesmos são referentes a um grupo (partição ótima), que é definido mas não foram analisados todos os documentos da coleção_de_teste, podendo existir outros documentos referentes ao mesmo assunto mas que não fazem parte da partição ótima em questão.

A tabela 6.1 resume os parâmetros utilizados na avaliação, bem como suas devidas descrições.

6.5 Resultados dos testes

Para avaliar a qualidade dos grupos formados, definiu-se manualmente partições ótimas[2] para cada teste. Essas partições serão o ponto de observação

[2]As URLs dos documentos utilizadas para definir as partições ótimas estão disponíveis em www.facom.ufu.br/ ilmerio/ anacarolina

Teste 1	R_a	Rev	P	F
AAL	52	23,0%	0,28%	0,56%
k-médias	49	21,7%	0,26%	0,53%

Tabela 6.2: Resultados do primeiro teste

durante a análise dos resultados.

Os testes foram realizados com o intuito de avaliar a eficiência do AAL, para isso é utilizado o processo de comparação entre métodos agrupamento: AAL e o k-médias em termos da revocação e precisão relativas, e média harmônica.

Cada teste realizado com o AAL teve a duração de aproximadamente 5 dias, e o k-médias com aproximadamente 3,5 dias, com algumas variações que ocorrem pela característica de sensibilidade dos algoritmos em relação à escolha dos k primeiros elementos para serem os representantes dos grupos iniciais. O número de interações do algoritmo também sofre essa influência sendo que o AAL executou em média 46 interações, e o k-médias teve seu processamento realizado em média com 38 interações.

Para o primeiro teste (teste 1) definiu-se uma partição contendo 226 documentos sobre linux, e usando os documentos dessa partição para selecionar o grupo que possuía o maior número de documentos que também pertecem à partição ótima, tem-se os resultados apresentados na tabela 6.2. Para esse teste são utilizados k=12 grupos, o tamanho da coleção de testes N=60000 e o tamanho da partição ótima X=226.

Analisando os resultados do teste 1, o valor de revocação do AAL é de 23% enquanto do k-médias é de 21.7%, pode-se perceber que o AAL apresentou uma capacidade ligeiramente maior em agrupar os documentos de assuntos similares, em relação ao tradicional k-médias. Analogamente a precisão e a

Teste 2	R_a	Rev	P	F
AAL	69	$53,1\%$	$0,56\%$	$1,12\%$
k-médias	60	$46,2\%$	$0,49\%$	$0,97\%$

Tabela 6.3: Resultados do segundo teste

Teste 3	R_a	Rev	P	F
AAL	48	$51,6\%$	$0,46\%$	$0,9\%$
k-médias	45	$48,4\%$	$0,43\%$	$0,85\%$

Tabela 6.4: Resultados do terceiro teste

média harmônica são ligeramente melhores para o AAL.

Para o segundo teste (teste 2) é definida uma partição contendo x=130 documentos sobre saúde. Da mesma forma (k=12, N=60000).

Observando os resultados do teste 2, o valor de revocação do AAL é de $53,1\%$ e do k-médias é de $46,2\%$, onde se pode perceber que o AAL apresentou como resultado a capacidade de agrupar os documentos de assuntos similares, melhor 15% do que o k-médias. Analogamente a precisão e a média harmônica são melhores para o AAL.

Para o terceiro teste (teste 3) é definida uma partição contendo X=93 documentos sobre hotel. Para esse teste são utilizados (k=12, N=60000).

No teste 3 obteve-se a revocação do AAL em $51,6\%$ e do k-médias em $48,4\%$. Assim pode-se perceber que o AAL apresentou como resultado uma capacidade de agrupar os documentos de assuntos similares, melhor em 6,93

Para o quarto teste (teste 4) é definida uma partição contendo X=41 documentos sobre o banco Itaú. Para esse teste são utilizados (k=12, N=60000).

No teste 4, o valor de revocação do AAL é de 50% e do k-médias, é de $57,7\%$, pode-se perceber que o AAL apresentou como resultado uma capa-

Teste 4	R_a	Rev	P	F
AAL	20	50%	0,13%	0,26%
k-médias	24	57,7%	0,15%	0,3%

Tabela 6.5: Resultados do quarto teste

cidade de agrupar os documentos de assuntos similares 13,3% menor que o k-médias. Analogamente, houve perdas na precisão e média harmônica. Sobre esse teste realizado, é importante ressaltar que o grupo que foi selecionado como o grupo de análise, ou seja o que mais possuía documentos da partição sobre o Banco Itaú, era relativamente pequeno em relação aos outros grupos. Daí uma menor revocação.

Na maioria dos testes realizados os valores dos parâmetros encontrados são bem parecidos, e o AAL não apresenta bons resultados quando o grupo que está sendo avaliado é muito pequeno (teste 4). Nesse caso, trata-se de um grupo pequeno, o que justifica a pouca eficiência do método já que em uma estrutura de ligações quando o número de documentos é pequeno não existe ligações suficientes para definir uma autoridade localmente. Entretanto, o método sobressai em casos em que os grupos possuem um maior número de documentos, por apresentarem estrutura suficiente para definir uma boa autoridade local.

Observando os valores encontrados, calculou-se o ganho oferecido pelo AAL em relação ao k-médias em cada um dos testes realizados. A partir deles calculou-se um ganho de 2,16% de revocação em relação ao k-médias (tabela 6.6), 7,52% de ganho de precisão (tabela 6.7) e 7,17% de ganho na média harmônica (tabela 6.8). Os resultados indicam que quando se trata de grupos que possuem um número considerável de documentos, capaz de construir uma estrutura de ligações, as informações extraídas dessa estrutura

Revocação	AAL	k-médias	ganho
Teste 1	$23,0\%$	$21,7\%$	6%
Teste 2	$53,1\%$	$46,2\%$	15%
Teste 3	$51,6\%$	$48,4\%$	$6,7\%$
Teste 4	50%	$57,7\%$	$-13,3\%$
Média	$41,1\%$	$40,1\%$	$2,16\%$

Tabela 6.6: Relação entre os valores de revocação dos algoritmos: AAL e k-médias

Precisão	AAL	k-médias	ganho
Teste 1	$0,28\%$	$0,26\%$	$7,69\%$
Teste 2	$0,56\%$	$0,49\%$	$14,29\%$
Teste 3	$0,46\%$	$0,43\%$	7%
Teste 4	$0,13\%$	$0,15\%$	$-13,33\%$
Média	$0,36\%$	$0,33\%$	$7,52\%$

Tabela 6.7: Relação entre os valores de precisão dos algoritmos: AAL e k-médias

melhoram a qualidade do algoritmo k-médias.

Média Harmônica	*AAL*	k-médias	ganho
Teste 1	$0,56\%$	$0,53\%$	$5,66\%$
Teste 2	$1,12\%$	$0,97\%$	$15,46\%$
Teste 3	$0,9\%$	$0,85\%$	$5,88\%$
Teste 4	$0,26\%$	$0,3\%$	$-13,33\%$
Média	$0,71\%$	$0,66\%$	$7,17\%$

Tabela 6.8: Relação entre os valores de média harmônica dos algoritmos: *AAL* e k-médias

Capítulo 7

Conclusões e Trabalhos Futuros

Neste trabalho, um estudo de aplicação de informação de ligações foi combinado com técnicas tradicionais de RI, para a melhora de um algoritmo de agrupamento não supervisionado. O efeito dessa combinação foi avaliado experimentalmente.

Objetivando agregar qualidade ao agrupamento de documentos da *Web* de maneira não supervisionada, foi estudada uma forma de melhorar a qualidade do resultado de um processo utilizando as informações de ligações entre os documentos. Para tanto, o algoritmo tradicional de agrupamento, o k-médias, foi estudado e na fase de redefinição do representante do grupo, no caso o centróide, foi substituído pela autoridade local do grupo naquele instante.

O algoritmo AAL aqui apresentado é uma abordagem alternativa para o processo de identificação de documentos semelhantes em um ambiente que contenha ligações entre os documentos. Esta solução consiste na utilização das informações contidas na estrutura de ligações da *Web*, no caso a autoridade local, em conjunto com as informações textuais (conteúdo) no processo de agrupamento não supervisionado de documentos.

Os resultados experimentais apontam para uma melhoria da qualidade do agrupamento da ordem de $7,52\%$ em precisão, $2,16\%$ em revocação e $7,16\%$ em média harmônica. Observa-se entretanto que, para grupos pequenos a melhoria não foi constatada, inclusive com uma perda na qualidade. Isto é explicado por não constituir uma estrutura de ligações em pequenos grupos de documentos.

A solução apresenta limitações, uma delas é que mantém as restrições do método original de agrupamento (k-médias) utilizado para a proposta, que é a definição do número ótimo de partições, ou seja, as partições originais da coleção. Como trabalho futuro pretende-se aprimorar a técnica para que através da análise da estrutura de ligações da *Web*, se possa definir o k ótimo. E para a seleção dos k primeiros elementos que representam os grupos iniciais, um estudo futuro é utilizar o conceito de distância (dissimilaridade), ou seja escolher os k elementos mais diferentes, naturalmente com bons valores de autoridade. A idéia é que a escolha de autoridade retira a influência de aleatoriedade do algoritmo podendo trazer melhorias.

Referências Bibliográficas

Aamodt, A. & Plaza, E. [1994], 'Case-based reasoning: Foundational issues, methodological variations, and system approaches', *AICom - Artificial Intelligence Communications, IOS Press* **7**(1), 39–59.

Amento, B., Terveen, L. & Hill, W. [2000], Does authority mean quality? Predicting expert quality ratings of web documents, *in* 'Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', ACM.

Baeza-Yates, R. A. & Ribeiro-Neto, B. A. [1999], *Modern Information Retrieval*, ACM Press / Addison-Wesley.

Calado, P. [2003], The wbr99 collection, Relatório Técnico, Universidade Federal de Minas Gerais - Ciência da Computação.

Calado, P. [2004], Using Link Structure for Information Retrieval in The World Wide Web, Tese de Doutorado, Universidade Federal de Minas Gerais - Ciência da Computação.

Calado, P., Ribeiro-Neto, B. A., Ziviani, N., Moura, E. & da Silva, I. R. [2003], 'Local versus global link information in the web', *ACM Transations On Information Systens* **21**(1), 42–63.

Cutting, D. R., Karger, D. & Pedersen, J. [1993], Constant interaction-time scatter/gather browsing of very large document collections, *in* 'Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', pp. 126–135.

Everitt, B. [1980], *Cluster Analysis*, New York Academic Press.

Feldman, R. & Hirsh, H. [1997], 'Exploiting background information in knowledge discovery from text', *Journal of Intelligent Information Systems* **9**(1), 83–97.

Flake, G. W., Lawrence, S. & Giles, C. L. [2000], Efficient identification of web communities, *in* 'Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD-2000)', pp. 150–160.

Garfield, E. [1972], 'Citation analysis as a tool in journal evaluation', *Science 178* pp. 471–479.

Halkidi, M., Batistakis, Y. & Varzigiannis, M. [2001], 'On clustering validation techniques', *Journal Intelligent Information Systems* **17**(2-3), 107–145.

Halkidi, M., Nguyen, B. & Varzigiannis, M. [2003], 'Thesus: Organizing web document colletions based on link semantics', *VLDB Journal* **12**(4), 320–332.

Han, J. & Kamber, M. [2001], *Data Mining - Concepts and Techniques*, Academic Press, EUA.

Harman, D., Fox, E. A., Baeza-Yates, R. A. & Lee, W. C. [1992], Inverted files, *in* 'Information Retrieval: Data Structures & Algorithms', pp. 28–43.

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

Kleinberg, J. M. [1999], 'Authoritative sources in a hyperlinked environment', *Journal of the ACM* **46**(5), 604–632.

Kowalski, G. [1997], *Information Retrieval Systems. Theory and Implementation*, Kluwer Academic Publishers.

Macqueem, J. [1967], 'Some methods for classification and analysis of a multivariate observation', *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1**(3), 281–297.

Mizzaro, S. [1997], 'Relevance: The whole history', *Journal of the American Society for Information Science* **48**(9), 810–832.

Robertson, S. E. & Walker, S. [1997], 'On relevance weights with little relevance information', *Annual International ACM-SIGIR Conference On Research And Development In Information Retrieval (SIGIR'97)* pp. 16–24.

Salton, G. [1971], *The SMART Retrieval System - experiments in automatic document processing*, Prentice-Hall.

Salton, G. & Buckley, C. [1987*a*], Improving retrieval performance by relevance feedback, Relatório Técnico, Department of Computer Science, Cornell University.

Salton, G. & Buckley, C. [1987*b*], Term weighting approaches in automatic text retrieval, Relatório Técnico, Department of computer science, Cornell University.

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

Van Rijsbergen, C. J. [1979], *Information Retrieval*, 2ª edn, Dept. of Computer Science, University of Glasgow.

Willett, P. [1988], 'Recent trends in hierarchic document clustering: a critical review', *Information Processing & Management* **24**(5), 577–597.

Rua S3, 80 - Apto. 1504 - Ed. Bela Vista Hills - Setor Bela Vista
74.823-440 Goiânia (GO) - Brasil
anacarolina@goi.universo.edu.br
www.goi.universo.edu.br/~anacarolina

Dissertação de Mestrado em Ciência da Computação - versão α
documento escrito em LaTeX 2_ε
UFU - Universidade Federal de Uberlândia - FACOM