JOÃO GUILHERME RODRIGUES GALLO

OTIMIZAÇÃO DO *RANKING* DE DOCUMENTOS EM MÁQUINAS DE BUSCA NA WEB A PARTIR DA MINERAÇÃO DE DADOS SENSÍVEL A CONTEXTOS

Dissertação apresentada ao Curso de Pós-Graduação em Ciência da Computação da Universidade do Federal de Minas Gerais, como requisito parcial à obtenção do grau de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Wagner Meira Jr

Belo Horizonte

2009

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Otimização do Ranking de Documentos em Máquinas de Busca na Web a Partir da Mineração de Dados Sensível a Contextos

JOÃO GUILHERME RODRIGUES GALLO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. WAGNER MEIRA JÚNIOR - Orientador
Departamento de Ciência da Computação - UFMG

DR. EDUARDO ALVES DO VALLE JUNIOR
Bolsista de Pós-Doutorado - DCC - UFMG

PROF. FABIANO CUPERTINO BOTELHO
Centro Federal de Educação Tecnológica - CEFET-MG

PROF. MARCOS ANDRÉ GONÇALVES
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 28 de maio de 2009.

Ficha catalográfica elaborada pela Biblioteca do ICEx - UFMG

Gallo, João Guilherme Rodrigues

G172o Otimização do *ranking* de documentos em máquinas de busca na web a partir da mineração de dados sensível a contextos / João Guilherme Rodrigues Gallo. Belo Horizonte, 2009.

71 f. : il.; 29 cm.

Dissertação (mestrado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientador: Wagner Meira Júnior

1. Computação - Teses. 2. Mineração de dados (Computação) - Teses. 3. Recuperação da informação – Teses. I. Orientador. II. Título.

CDU 519.6*73(043)

Aos meus pais pela participação constante e à minha noiva Gabriela pelo altruísmo de mais do que se preocupar em entender os motivos dos meus projetos, me apoiar incondicionalmente, abrindo mão do que fosse necessário, para eu chegasse até aqui.

AGRADECIMENTOS

A colaboração de diversas pessoas, cada uma à sua maneira, foi essencial durante a realização deste trabalho. Agradeço principalmente ao professor, orientador e amigo Wagner Meira Jr. pela inestimável orientação, paciência, preocupação e pela oportunidade de tamanho crescimento pessoal e acadêmico. Um agradecimento especial também ao professor Nívio Ziviani pela ativa colaboração no desenvolvimento desse trabalho.

Agradeço também aos integrantes dos laboratórios LATIN e E-SPEED, colegas de sala, professores e funcionários do Departamento de Ciência da Computação da Universidade Federal de Minas Gerais que criaram o ambiente necessário para que todos os meus objetivos fossem alcançados.

Agradeço aos amigos-sócios do Buzzero.com pela compreensão nas ausências, apoio irrestrito e, em suma, por serem sócios também nessa jornada.

Por último, e com especial atenção, agradeço a todos os voluntários que gastaram parte de seu precioso tempo em solidariedade ao meu projeto. Dentre esses agradeço, em especial, aos meus pais, Edson e Elisa e à minha noiva Gabriela pelo esforço hercúleo na avaliação dos documentos retornados nos testes.

“Uma das mais cativantes ironias do pensamento moderno é o fato de que o método científico, do qual ingenuamente se esperou no passado que pudesse banir o mistério do mundo, deixa-o cada dia mais inexplicável” (Carl Becker, 1932).

RESUMO

Neste trabalho apresentamos uma nova abordagem de ordenação de documentos de mecanismos de busca na web a partir da mineração de dados sensível a contextos. Sua originalidade apresenta-se, especialmente, na aplicação conjunta de estratégias anteriormente adotadas de maneira isolada: Primeiro, os padrões de correlação entre os termos são considerados e processados de maneira eficiente. Segundo, é utilizada uma técnica de mineração de dados chamada de regras de associação para a ponderação dos termos e criação de conjuntos de termos semanticamente relacionados. Terceiro, a identificação do conceito buscado pelo usuário a partir da correlação semântica entre os termos de todas as consultas realizadas por um usuário em uma sessão de buscas. Por último, todo o processo é realizado sem a solicitação explícita de informação extra ao usuário. Resultados experimentais mostram que nosso modelo aumenta a precisão média na coleção avaliada para todos os tipos de consulta, sem que o custo computacional do processo seja excessivo. Nossos resultados sugerem que o nosso modelo apresenta ganhos consideráveis também para coleções genéricas de textos disponíveis na *Web*.

Palavras-chave: Recuperação da Informação. Mineração de Dados. Sessões de Busca, Contextos.

ABSTRACT

This work presents a new approach to rank documents in web search engines based on context sensitive data mining. It's novelty lies specially on the use of a set of previously used strategies which have not been put together yet: First, the correlation patterns among the terms are processed efficiently. Second, a data mining technique called association rules is used for creating semantic correlation of the terms. Third, the identification of the concept searched by the user is done based on terms submitted on a retrieval session. Finally, all the process is done without the explicit demand of extra information from the user. Experimental results show that our approach increases the average precision of the search results on the evaluated collections for all kinds of searches without a substantial increase of the process computational. Our results suggest that our model presents considerable gains for generic Web text collections.

Keywords: Information Retrieval. Data Mining. Search Sessions, Contexts.

LISTA DE ILUSTRAÇÕES

LISTA DE GRÁFICOS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A grande e vertiginosamente crescente disponibilidade de informações a baixíssimo custo na *Web* é um conceito unânime nos dias de hoje. As diversas formas de livre publicação de conteúdo e o acesso de uma parcela cada vez maior da humanidade transformaram radicalmente a forma de se pesquisar a respeito de qualquer assunto em praticamente todas as áreas do conhecimento humano. Recentemente, por exemplo, foi veiculado na mídia que o Google já conhecia mais de um trilhão de páginas[1], o que dá uma idéia do trabalho a ser realizado em uma sessão de buscas dessa ferramenta.

Encontrar o conteúdo desejado no meio de toda essa informação disponível tem se mostrado uma tarefa árdua, especialmente devido à grande quantidade de conteúdos irrelevantes retornados a cada busca. Os usuários de sistemas abertos de busca na *Web*, em geral, possuem pouca familiaridade com a utilização de sistemas e com conceitos de recuperação de informação, acabando por fornecer chaves de busca pequenas e com pouco poder discriminatório do conceito de seu interesse. Além disso, muitas vezes os termos submetidos a uma máquina de busca podem representar mais de um conceito.

Apesar de, geralmente, o usuário de um mecanismo de buscas realizar suas pesquisas por meio de sequências de consultas curtas, maior parte dos modelos utilizados pelos mecanismos de busca atuais levam em consideração cada consulta submetida a uma coleção de documentos isoladamente. Muita informação relevante a respeito do real interesse do usuário é desconsiderada e isso pode ser determinante na qualidade dos resultados retornados. Uma máquina de busca ideal deve ser capaz de coletar tanta informação adicional a respeito do interesse do usuário quanto possível. O contexto no qual a consulta se insere é, sem dúvidas, uma das informações mais importantes. Além disso, conforme descrito em (1), a recuperação sensível a contextos é considerada um dos grandes desafios da área de Recuperação de Informação.

Há vários tipos de contextos que podem ser explorados na otimização do *ranking* de documentos: destacam-se os geográficos, temporais, linguísticos e o perfil de navegação do usuário. Uma das estratégia tradicional para se obter informações complementares a respeito do contexto de interesse do usuário é a técnica de *Relevance Feedback* (2), uma

[1] Fonte: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html - acessado em 30 de setembro de 2008

forma reconhecidamente eficiente de se aumentar a precisão da tarefa de recuperação a partir da análise da qualitativa dos resultados pelo usuário.

Em muitos casos, o *Relevance Feedback* é obtido de maneira explícita por meio da seleção de categorias ou tópicos identificados para a consulta submetida ou pela escolha de subconjuntos de documentos considerados relevantes. Essa interação explícita com o usuário é problemática devido à relutância do mesmo em fornecer tais informações por não perceber os benefícios oferecidos. Para solucionar esse problema há a possibilidade de se aproveitar as informações fornecidas pelo usuário na própria utilização do sistema, o que é chamado de *Implicit Feedback*. Nessa técnica utiliza-se, por exemplo, os documentos retornados nas consultas anteriores e as informações a respeito de sua relevância fornecida pelo usuário ao selecionar ou não tais documentos na lista de resposta do mecanismo de busca.

A aplicação dessa estratégia se apóia no próprio comportamento usual do usuário ao realizar consultas complexas na máquina de busca. Como o usuário geralmente submete uma consulta, avalia os resultados e seleciona os documentos que lhe parecem relevantes ao tema consultado para então refinar sua busca realizando o mesmo processo repetidas vezes é possível coletar, a cada consulta submetida, informações a respeito do refinamento da chave de busca e dos documentos que aparentaram ser relevantes na lista retornada.

A maior parte dos trabalhos existentes utiliza unicamente nos documentos selecionados durante a busca para alimentar os algoritmos de *Implicit Feedback*. Nossa abordagem leva em consideração a correlação entre os termos da coleção, entre os termos dos documentos selecionados e entre as consultas submetidas durante toda a sessão de busca. Com isso buscamos ser capazes de identificar o conceito procurado pelo usuário ao invés de nos concentrarmos apenas nos termos.

Um desafio extra ao se realizar esse tipo de trabalho é a ausência de coleções de referência para a análise quantitativa dos resultados obtidos. Em geral as coleções são analisadas com foco nas abordagens tradicionais nas quais cada consulta é tomada de forma isolada e por isso mesmo não são adequadas a avaliar uma proposta como a nossa onde o foco é justamente identificar o conceito buscado e não apenas qualquer conceito do termo buscado. A solução desse problema se deu pela adequação de duas coleções de referência à nossa necessidade por meio de um experimento com usuários voluntários.

1.1 CARACTERIZAÇÃO DO PROBLEMA

Um mesmo termo pode representar vários conceitos ou entidades dependendo do contexto onde ele se enquadra. Uma busca com a chave “CRISTO SALVADOR” pode ter conjuntos resposta de documentos relevantes bem diferente se o usuário for um turista em busca de roteiros no Brasil ou um devoto em busca de textos religiosos.

A caracterização desse usuário, no entanto, pode ser uma tarefa por si só um tanto complexa também. De uma maneira geral há dois tipos de informações contextuais que podem ser utilizadas para se caracterizar implicitamente o usuário: *long-term context*, que analisa todo o histórico do usuário, e *short-term context*, que analisa apenas as informações imediatas do mesmo. A escolha das informações a serem utilizadas é um fator determinante do sucesso da abordagem.

A Mineração de Dados pode ser útil na identificação da correlação entre os termos em uma coleção de documentos. No entanto os termos podem também apresentar correlações distintas de acordo com a semântica dos mesmos. No exemplo citado acima o termo “CRISTO” pode ser relacionar com “RIO”, “COPACABANA” e “IPANEMA” em um determinado contexto e com “APÓSTOLO”, “BÍBLIA” e “DEUS” em outro. Por esse motivo a mineração de dados sensível a contextos mostra-se mais adequada a essa tarefa.

Esse trabalho busca, portanto, propor uma abordagem capaz de identificar o conceito buscado pelo usuário e, com isso, alterar o *ranking* dos documentos retornados de modo que os documentos onde o termo buscado tenha o significado buscado sejam melhor classificados enquanto aqueles onde o significado aparenta ser outro sejam penalizados.

1.2 SOLUÇÃO PROPOSTA

A abordagem proposta neste trabalho busca identificar o conceito semântico pesquisado pelo usuário através da definição da sessão de buscas como um contexto no qual as consultas estão correlacionadas. Isso pode ser realizado por utilizarmos diversas

informações passadas implicitamente pelo usuário, ou seja, sem que exista uma interação formal entre o usuário e o sistema na determinação da relevância dos documentos retornados.

Nas abordagens tradicionais de recuperação de informação, tais como os modelos booleanos, probabilísticos ou de espaços vetoriais, não são identificadas relações semânticas entre os termos da coleção. Além disso, cada consulta é considerada de forma isolada, sem que sejam levadas em consideração as demais consultas previamente realizadas ou os documentos nela selecionados. Nesses modelos, uma série de informações relevantes disponíveis capazes de auxiliar na identificação do conceito buscado pelo usuário, como por exemplo os contextos temporais e semânticos, são desconsideradas, o que acaba por reduzir a precisão dos documentos retornados nas consultas.

Existem diversas formas de se extrair e utilizar essas informações descartadas pelos modelos clássicos. A expansão dos termos da consulta e a recuperação da informação sensível a contextos, utilizadas em nossa abordagem, são algumas das que se mostraram mais eficientes e computacionalmente viáveis.

A análise do contexto da busca pode ser feita por meio da análise de relevância dos documentos retornados. Essa análise, conhecida como *Relevance Feedback*, torna possível inferir o provável interesse do usuário e, com isso, melhorar o *ranking* dos documentos que aparentam tratar desse conceito. Essas informações contextuais podem ser divididas em dois grupos de acordo com a sua natureza, as persistentes (*long-term* context) e as instantâneas (*short-term context*).

O primeiro tipo, *long-term context*, considera informações históricas coletadas ao longo do tempo ou por meio de cadastros. De uma maneira geral esse tipo de caracterização leva em consideração o interesse padrão do usuário, tipos de documentos ou fontes escolhidos, grau de escolaridade e o histórico global de consultas. Esse tipo de caracterização é especialmente interessante para se modelar o usuário e propor resultados que tenham sido considerados relevantes por outros usuários com modelo semelhante. Por outro lado, existe uma restrição muito grande com relação à identificação e monitoramento das ações online por parte dos usuários. Sendo assim, obrigá-lo a se identificar antes de iniciar uma sessão de buscas pode ser uma má idéia. Além disso há o fato de o mesmo usuário poder, dependendo da ocasião, ter perfis totalmente diferentes uma vez que seus objetivos ao realizar consultas no ambiente profissional podem ser ortogonais ao foco de suas buscas pessoais. Dessa forma, o uso do *long-term context* no problema tratado nesse trabalho não se mostra muito eficiente.

O outro tipo, *short-term context*, considera informações instantâneas como, por exemplo, oriundas de uma sessão isolada. Nele pode ser levada em conta toda a informação

disponibilizada na interação do usuário com o sistema durante a busca por um determinado tema. Essa informação tende a estar relacionada diretamente com o foco da busca do usuário e pode ser coletada de maneira totalmente transparente para o usuário que se mantém anônimo. Como a sessão de buscas pode ser facilmente determinada pelas configurações dos servidores da máquina de buscas, consideramos que o *short-term context* mostra-se como um caminho adequado à coleta dos dados a serem utilizados na identificação do significado semântico do termo utilizado na consulta.

A identificação dos conceitos propriamente dita é feita por meio de técnicas de mineração de dados, mais especificamente as regras de associação. Buscamos regras capazes de correlacionar termos a partir da sua co-ocorrência nas consultas realizadas, nos textos explicativos dos documentos retornados nas consultas, os *snippets*, e nos documentos da coleção.

A otimização do *ranking* dos documentos retornados é obtida por meio da expansão dos termos da consulta a partir de termos extraídos das regras de associação e das consultas anteriores. Esses termos funcionam como desambiguadores de conceitos que buscam melhorar o *ranking* dos documentos nos quais o termo buscado apresenta o significado semântico procurado pelo usuário.

1.3 ESTRUTURA DA DISSERTAÇÃO

Essa dissertação está organizada da seguinte forma: inicialmente foi apresentado, em linhas gerais, o problema a ser tratado, o cenário no qual ele se insere e a solução proposta. Nos capítulos 2 e 3 é feita uma revisão bibliográfica das duas áreas da Ciência da Computação nas quais essa dissertação se baseia. O capítulo 2 apresenta conceitos e algoritmos clássicos da área de Recuperação de Informação. No capítulo 3 os conceitos de mineração de regras de associação bem como estratégias e algoritmos são apresentados. O quarto capítulo apresenta a nossa estratégia propriamente dita. Nele explicitamos a aplicação dos conceitos apresentados nos dois capítulos anteriores. Os experimentos, bem como seus resultados são apresentados no capítulo 5. Por fim, no capítulo 6, são apresentadas as conclusões e trabalhos futuros.

2 RECUPERAÇÃO DE INFORMAÇÃO

A Recuperação da Informação (RI) trata da representação, armazenamento, organização e acesso à informação. Seu objetivo é dar aos usuários uma ferramenta através da qual seja possível acessar facilmente a informação desejada. (3)

De uma maneira geral, os mecanismos de busca na Web atuais utilizam índices muito similares àqueles que vêm sendo usados em bibliotecas por centenas de anos. A principal diferença está no tamanho da biblioteca na qual se faz a busca e na heterogeneidade estrutural dos documentos.

A primeira tarefa realizada ao se implementar um mecanismo de busca na *Web* é a coleta. Nela um conjunto de robôs conhecidos como coletores, ou *crawlers*, percorrem o grafo de documentos, a partir de um conjunto de documentos semente, seguindo os *hyperlinks* entre os documentos. À medida que o coletor visita as páginas o conteúdo das mesmas é armazenado para que sejam utilizados posteriormente.

A partir dos documentos coletados é elaborado então um índice dos termos e documentos, a fim de que seja possível a pesquisa eficiente em grandes volumes de dados. A estrutura de dados mais comum para essa tarefa é a lista invertida onde se registra, a partir de uma lista lexicograficamente ordenada de todos os termos presentes no vocabulário da coleção, sub-listas de todos os documentos nos quais o termo chave aparece.

Por último, tem-se o processador de consultas que, muitas vezes, é confundido com a própria máquina de busca. É no processador de consultas que são implementados os algoritmos capazes de comparar as chaves de busca com os documentos indexados e determinar aqueles que satisfazem às condições impostas pelo usuário. De acordo com o modelo adotado pelo mecanismo de buscas, diferentes estratégias são adotadas para determinar o subconjunto resposta da coleção que mais se aproxime do conjunto ideal que contenha todos os documentos relevantes e nenhum outro mais.

A precisão do mecanismo de buscas é dada pela razão entre o número de documentos relevantes e o número total de documentos retornados. A revocação, por sua vez, é determinada pelo percentual dos documentos relevantes que tenham sido retornados. Essas duas métricas são capazes de determinar a qualidade do mecanismo uma vez que identificam tanto a probabilidade de um documento retornado ser relevante quanto a de que se um documento é relevante ele tenha sido retornado. O desafio então é conseguir selecionar o

maior percentual possível dos documentos relevantes sem que sejam retornados também muitos documentos não relevantes.

De uma maneira geral o processo de recuperação da informação pode ser resumido da seguinte forma: O usuário elabora uma consulta que é usada como entrada do algoritmo de recuperação da informação. A consulta é processada pela máquina de buscas que seleciona os documentos a serem retornados ao usuário. Finalmente, os documentos são ordenados em função da sua relevância e são, então, examinados pelo usuário.

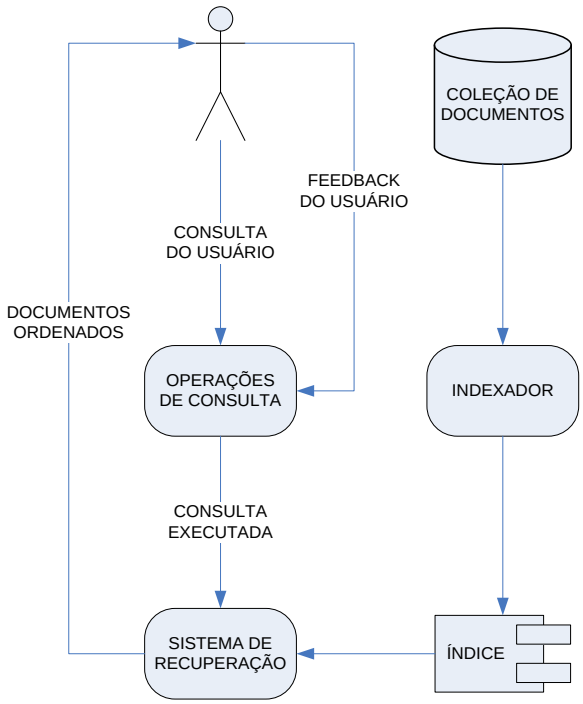

Figura 2.1 - Arquitetura de um Sistema de Recuperação da Informação

É justamente no momento em que o usuário analisa os documentos e seleciona aqueles que julga relevantes que se pode extrair informações extra a respeito do contexto da consulta capazes de melhor caracterizar o significado semântico dos termos submetidos pelo usuário.

2.1 MODELOS CLÁSSICOS DE RECUPERAÇÃO DA INFORMAÇÃO

Um modelo de recuperação da informação, RI, determina a forma como os documentos e consultas são representados e como a relevância de um documento para o

usuário é definida. Os três principais modelos existentes são: os Booleanos, os probabilísticos e os de espaços vetoriais.

Apesar de cada um desses modelos usar uma representação específica para documentos e consultas, o arcabouço no qual todos se baseiam é o mesmo: todos tratam cada documento ou consulta como uma "sacola de termos" (*bag of terms*) ou "sacola de palavras" (*bag of words*). A sequência e a posição dos termos nas frases são ignorados, as consultas são consideradas independentes umas das outras, ou seja, uma consulta não implica nenhuma interferência nas consultas seguintes. Cada termo é associado a um peso específico que indica sua capacidade discriminatória de documentos.

Definição 2.1 Dada uma coleção de documentos D, seja $V = \{t_1, t_2, ..., t_{|V|}\}$ o conjunto de termos distintos na coleção onde t_i é um termo. O conjunto V é, geralmente, chamado de vocabulário da coleção, e $|V|$ é o seu tamanho. Um peso $w_{ij} > 0$ é associado a cada termo t_i de um documento $d \in D$. Para um termo que não esteja presente no documento d, $w_{ij} = 0$. Cada documento d_j é então representado na forma de um vetor de termos $d_j = (w_{1j}, w_{2j}, ..., w_{|V|j})$, onde cada peso w_{ij} corresponde ao termo $t_i \in V$ e quantifica a importância do termo no documento. A sequência dos termos no vetor não tem significado.

Com essa representação vetorial, uma coleção de documentos pode ser representada apenas com uma tabela relacional ou matriz onde cada termo é um atributo e cada peso é o valor desse atributo. Cada modelo de recuperação da informação calcula *wij* de uma maneira específica.

2.1.1 Modelos Booleanos

Os primeiros sistemas de RI eram sistema booleanos e, ainda hoje, diversos sistemas comerciais baseiam-se nesse modelo. Esses modelos costumam ser citados frequentemente devido, em grande parte, à simplicidade de implementação e à facilidade dos usuários em lidar com os conceitos relacionados à teoria de conjuntos nos quais ele se baseia.

Em linhas gerais, as consultas consistem em combinações lógicas de termos utilizando os conectores "E" (AND), "OU" (OR) e "NÃO" (NOT). Cada termo da consulta estabelece um conjunto de documentos nos quais há sua ocorrência e a lista retornada ao

usuário surge das conjunções, disjunções e negações desses conjuntos. A função de *ranking* nesse caso é binária, ou seja, se o documento preenche os requisitos da consulta ele é selecionado, caso contrário não. Os documentos são apenas selecionados e não ordenados. A elaboração de consultas por um usuário que não conheça previamente o conteúdo dos documentos pode se tornar uma tarefa um tanto complexa.

O modelo Booleano estendido (4) acrescenta pesos aos termos e medidas de distância ao modelo Booleano. Inicialmente, os termos recebem pesos entre 0 e 1. Em seguida as conectividades semânticas ganham uma nova semântica e passam a ser modeladas pela medida de similaridade na distância não euclidiana em um espaço *t*-dimensional, onde *t* é o número de termos do vocabulário da coleção.

Esse modelo foi generalizado posteriormente no modelo *p-norm* onde os conectivos "OU" e "E" contém um parâmetro *p*. A partir da variação de *p* de 1 a ∞, a função de *ranking p-norm* varia entre os *rankings* do modelo de espaços vetoriais e Booleano. Teoricamente *p* pode ser configurada para cada conectivo, o que permitiria uma configuração *ad hoc* a cada consulta submetida.

Apesar do apelo conceitual o modelo Booleano estendido não se tornou popular. Uma das principais razões é o fato de ele não ser claramente formulado para o usuário, uma vez que as consultas mantêm a forma de uma fórmula Booleana, mas com uma semântica alterada.

2.1.2 Modelos Probabilísticos

Propostos pela primeira vez em (5), esses modelos passaram a ser conhecidos como modelo BIR (*Binary Independence Retrieval* – Recuperação de Independência Binária). Os modelos probabilísticos procuram resolver o problema de RI por um arcabouço probabilístico.

A idéia fundamental é que dada uma consulta de um usuário, existe um único conjunto de documentos que contém exatamente os documentos relevantes àquela consulta. Esse conjunto de documentos é conhecido como o conjunto-resposta ideal. Dada a descrição desse conjunto-resposta ideal, não haveria problemas em recuperar seus documentos; dessa

forma o processo de consulta se transforma no processo de determinar as propriedades de um conjunto-resposta ideal.

O problema é que não se sabe quais são essas características exatamente. Sabe-se, apenas, que há termos cuja semântica deveria ser usada para caracterizar essas propriedades. Já que essas propriedades não são conhecidas no momento da consulta é necessário um esforço para se descobrir, inicialmente, quais seriam esses termos. Essa descoberta inicial nos permite gerar uma descrição probabilística preliminar da resposta ideal que seria utilizada para retornar o primeiro conjunto de documentos. A partir daí o usuário examina os documentos retornados e decide quais são relevantes. O sistema utiliza então essas informações para refinar a descrição do conjunto-resposta ideal. Pela repetição desse processo espera-se chegar cada vez mais próxima da descrição real do conjunto-resposta ideal.

Definição 2.2 – Princípio Probabilístico: Dada uma consulta q e um documento d_j na coleção, o modelo probabilístico busca estimar a probabilidade de o usuário achar d_j relevante. O modelo assume que essa probabilidade de relevância depende apenas da consulta e da representação do documento. O modelo assume que existe um subconjunto de todos os documentos que o usuário iria preferir como resposta à consulta q. Tal conjunto ideal é chamado de R e deve maximizar a probabilidade de relevância para o usuário. Documentos no conjunto R devem ser relevantes à consulta e os documentos fora desse conjunto não.

Definição 2.3 Para o modelo probabilístico as variáveis de peso dos termos são todas binárias, i.e., $w_{ij} \in \{0,1\}$, $w_{iq} \in \{0,1\}$. Uma consulta q é um subconjunto dos termos. Seja R o conjunto de documentos ditos relevantes. Seja $\bar{R}$ o complemento de R, ou seja o conjunto de documentos não relevantes. Seja $P(R|\overrightarrow{d_j})$ a probabilidade de que o documento d_j seja um documento relevante à consulta q e $P(\bar{R}|\overrightarrow{d_j})$ a probabilidade de que d_j não ser relevante à q. A similaridade $sim(d_j,q)$ do documento d_j à consulta é definida pela razão

$$sim(d_j, q) = \frac{P(R|d_j)}{P(\bar{R}|d_j)}$$

Pela regra de Bayes tem-se:

$$sim(d_j, q) = \frac{P(\overrightarrow{d_j}|R) \times P(R)}{P(\overrightarrow{d_j}|\bar{R}) \times P(\bar{R})}$$

Nesse caso, $P(\overrightarrow{d_j}|R)$ determina a probabilidade de se escolher aleatoriamente o documento d_j entre os documentos relevantes *R*, já P(*R*) é a probabilidade de um documento de *R* ser escolhido aleatoriamente na coleção inteira.

Assumindo-se a independência dos termos, pode-se evoluir a fórmula para a seguinte forma:

$$sim(d_j, q) \sim \frac{\left(\prod_{g_i(\vec{d}_j)=1} P(k_i|R)\right) \times \left(\prod_{g_i(\vec{d}_j)=0} P(k_i|R)\right)}{\left(\prod_{g_i(\vec{d}_j)=1} P(k_i|\bar{R})\right) \times \left(\prod_{g_i(\vec{d}_j)=0} P(\bar{k}_i|\bar{R})\right)}$$

P(k_i|*R*) é a probabilidade de que o termo k_i esteja presente em um documento aleatoriamente selecionado no conjunto *R*. Ignorando os fatores constantes na coleção e fazendo o logaritmo da fórmula chega-se a forma final do *ranking* do modelo probabilístico.

$$sim(d_j, q) \sim \sum_{i=1}^{t} w_{i,q} \times w_{i.j} \times \left(\log \frac{P(k_i|R)}{1 - P(k_i|R)} + \log \frac{1 - P(k_i|\bar{R})}{P(k_i|\bar{R})}\right)$$

Como o conjunto *R* não é conhecido inicialmente, as probabilidades iniciais de $P(k_i|R)$ e $P(k_i|\bar{R})$ devem ser definidas arbitrariamente. Uma das formulas mais eficientes (3) é a seguinte:

$$P(k_i|R) = \frac{V_i + \frac{n_i}{N}}{V + 1}$$

$$P(k_i|R) = \frac{n_i - V_i + \frac{n_i}{N}}{N - V + 1}$$

Onde *V* é o subconjunto dos *r* primeiros documentos retornados pelo modelo probabilístico, V_i são os documentos de *V* que contêm o termo k_i, N é o número de documentos da coleção e n_i é o número de documentos da coleção que contém o termo k_i.

Teoricamente, a maior vantagem do modelo probabilístico é ordenar os documentos retornados em função da probabilidade de serem relevantes, no entanto, a separação inicial dos conjuntos de documentos relevantes e não relevantes mostra-se muito difícil. Além disso, o modelo não leva em consideração os pesos dos termos em função da sua raridade na coleção e considera que os termos são mutuamente independentes em relação à sua co-ocorrência.

2.1.2.1 Modelos Linguísticos Estatísticos

Originalmente propostos em (6), os modelos linguísticos estatísticos baseiam-se na probabilidade e têm suas bases na teoria estatística. A idéia básica dessa abordagem é simples: em primeiro lugar um modelo de linguagem é estimado para cada um dos documentos que, então, são ranqueados pela semelhança com uma consulta pelo modelo de linguagem dado. Idéias semelhantes foram utilizadas anteriormente no processamento de linguagem natural.

Seja a consulta q uma sequência de termos, $q = q_1, q_2, ..., q_m$ e a coleção de documentos D um conjunto de documentos, $D = \{d_1, d_2, ..., d_n\}$. Na abordagem do modelo de linguagem, consideramos a probabilidade de uma consulta q ser gerada por um modelo probabilístico baseado em um documento d_j, i.e., $\Pr(q|d_j)$. Para ranquear o documento na tarefa de recuperação estamos interessados em estimar a probabilidade posterior $\Pr(d_j|q)$. Pela regra de Bayes temos:

$$PR(d_j|q) = \frac{\Pr(q|d_j)\,\Pr(d_j)}{\Pr(q)}$$

Para o *ranking*, não é necessário calcular $\Pr(q)$, já que ele é o mesmo para todos os documentos. $\Pr(d_j)$ geralmente é considerado uniforme e também não afeta o *ranking*. Sendo assim só é necessário calcular $\Pr(q|d_j)$.

O modelo de linguagem utilizado na maioria dos trabalhos leva em consideração apenas palavras, ou seja, o modelo assume que cada palavra é gerada de maneira independente, o que é, essencialmente, uma distribuição multinomial das palavras. O caso geral é o modelo do *n-grama*, onde o enésimo termo depende dos n-1 termos anteriores.

Sendo assim tem-se:

$$\Pr(q = q_1, q_2, \dots, q_m|d_j) = \prod_{i=1}^{m} \Pr(q_i|d_j) = \prod_{i=1}^{|V|} \Pr(t_i|d_j)^{f_{iq}}$$

Onde f_{iq} é o número de vezes que um termo t_i ocorre em q, e $\sum_{i=1}^{|V|} \Pr(t_i|d_j) = 1$. O problema da recuperação é então reduzido à estimativa de $\Pr(t_i|d_j)$, que pode ser dado pela frequência relativa,

$$\Pr(t_i|d_j) = \frac{f_{ij}}{|d_j|}$$

Onde f_{ij} é o número de vezes que o documento t_i ocorre no documento d_j e $|d_j|$ é o número total de palavras em d_j.

O problema, nesse caso, é que a estimativa de um termo que não está presente em d_j tem a probabilidade 0, o que subestima a probabilidade desse termo no documento. Seguindo a solução do modelo de classificação Bayesiana ingênua de textos a função de probabilidade é alterada para:

$$\Pr(t_i|d_j) = \frac{\lambda + f_{ij}}{\lambda|V| + |d_j|}$$

Ao definir-se um valor pequeno, maior do que zero, para a probabilidade desses termos ocorrerem em tais documentos, a probabilidade de ocorrência do termo no documento deixa de ser subestimada.

2.1.3 Modelos de Espaços Vetoriais

Os modelos de espaços vetoriais (Vector Space Model – VSM) se baseiam na similaridade entre a consulta e os documentos. Os documentos não são selecionados pelo casamento exato dos termos da consulta com os do documento mas por uma estimativa de relevância.

Nesse modelo, as representações da consulta e dos documentos são feitas por meios de vetores em um espaço Euclidiano *t*-dimensional, onde *t* é número de termos presentes no vocabulário da coleção (7). Um algoritmo que implementa o modelo de espaços vetoriais determina a similaridade entre a consulta e um documento pela sua distância vetorial. A similaridade determina então a probabilidade de relevância de um documento.

Os pesos dos termos podem ser calculados de diversas maneiras (8) (9) (10), sendo que a forma considerada mais eficiente atualmente é dada por:

$t_{fi,j}$ – número de vezes que um termo *i* ocorre em um documento *j*;

idf_i – inverso do número de documentos em que um termo *i* ocorre.

A partir disso é possível dar maior peso a termos raros na coleção, que possuem um maior poder discriminatório e avaliar a recorrência desses termos no documento, aumentando a importância dos documentos onde o termo ocorre mais frequentemente. O peso w_{ij} de um termo *i* em um documento *j* é então dado por:

$$w_{ij} = t_{f_{ij}} \times idf_i = t_{f_{ij}} \times \log\frac{N}{df_i}$$

onde *N* é o número de documentos na coleção e idf_i é o inverso do número de documentos da coleção que possuem o termo *i*.

O peso de um termo *i* em uma consulta *q* é determinado de maneira semelhante:

$$w_{iq} = f\left(t_{f_{iq}}\right) \times idf_i = (1 + \log t_{f_{ij}}) \times \log\left(1 + \frac{N}{df_i}\right)$$

onde *N* é o número de documentos na coleção, $t_{f_{iq}}$é o número de ocorrências do termo *i* na consulta *q* e idf_i é o inverso do número de documentos da coleção que possuem o termo *i*.

Uma das métricas de *ranking* mais aceitas para o modelo de espaços vetoriais é a medida do cosseno. Nela a medida de similaridade é definida pelo produto escalar entre os vetores dos documentos $\overrightarrow{d_j}$, $1 \leq j \leq N$, e o vetor da consulta $\vec{q}$. Essa medida equivale ao cosseno entre o vetor da consulta e qualquer vetor de documento e é dada por:

$$sim(q, d_j) = \frac{\overrightarrow{d_j} \cdot \vec{q}}{\left|\overrightarrow{d_j}\right| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}}$$

Onde os fatores $\left|\overrightarrow{d_j}\right|$ e $|\vec{q}|$ correspondem às normas dos vetores do documento e da consulta respectivamente.

As principais vantagens do modelo vetorial são:

i. seu esquema de definição de pesos aos termos, capaz de melhorar o desempenho da recuperação
ii. seu esquema de casamento parcial que permite buscar documentos que se aproximem das condições de consulta
iii. a fórmula de *ranking* do cosseno ordena os documentos de acordo com o seu grau de similaridade com a consulta.

Apesar de sua simplicidade, o modelo de espaços vetoriais apresenta resultados que dificilmente podem ser otimizados sem a utilização da expansão dos termos da consulta ou do *Relevance Feedback*.

2.1.3.1 Modelo de Espaços Vetoriais Generalizado

Apesar do grande sucesso do VSM, ele considera que os termos são mutuamente independentes, o que é claramente uma simplificação que não traduz a realidade mas que teve

que ser feita por conveniência matemática e simplificação da implementação. A fim de capturar a correlação entre os termos, foi proposto o Modelo de Espaços Vetoriais Generalizado (Generalized Vector Space Model – GVSM) (11) (12). Nesse modelo, além dos termos atômicos (palavras) são também consideradas combinações entre eles (expressões), que podem ser adicionadas ao vocabulário de forma direta, ou seja, modeladas como termos atômicos.

Ao se inserir expressões ao vocabulário, minimiza-se o problema da consideração de independência mútua da co-ocorrência dos termos já que a dependência entre eles fica representada pelas próprias expressões, chamadas mintermos.

A medida do cosseno é também usada como fórmula de *ranking* do modelo de espaços vetoriais generalizado. Ela define a medida de similaridade para cada documento contendo qualquer termo da consulta definida pelo produto escalar entre o vetor do conjunto de documentos $\vec{d}_j$, $1 \leq j \leq N$, e o vetor da consulta $\vec{q}$.

$$sim(q, d_j) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} \sum_{r=1}^{2^t} c_{i,r}^{j} \times c_{i,r}^{q}}{\sqrt{\sum_{i=1}^{t} \sum_{r=1}^{2^t} \left(c_{i,r}^{j}\right)^2} \times \sqrt{\sum_{i=1}^{t} \sum_{r=1}^{2^t} \left(c_{i,r}^{q}\right)^2}}$$

onde $c_{i,r}^{j}$ é a soma dos pesos de todos os termos k_i contidos em um documento d_j para cada mintermo m_r e analogamente para a consulta q.

O modelo vetorial generalizado é computacionalmente mais complexo do que o tradicional e torna-se inviável em coleções de tamanho moderadamente grande devido a explosão exponencial do número de mintermos. Apesar de não garantir uma melhoria efetiva nas buscas em coleções genéricas, sua contribuição teórica deve ser levada em consideração.

2.2 MODELOS ORIENTADOS A CONJUNTOS

O problema de recuperação da informação originalmente era modelado pela teoria de conjuntos do modelo Booleano, os modelos atuais, no entanto, tendem a ser orientados a itens. O modelo original das abordagens orientadas a conjuntos atuais foi proposto em (13) e

propõe que uma abordagem orientada a conjuntos seria capaz de otimizar a eficácia da tarefa de RI pelo uso das relações estruturais presentes na coleção já que essa orientação é uma generalização da orientação a itens.

Os modelos orientados a conjuntos provêem um arcabouço para a representação dos conceitos derivados das co-ocorrências de termos que podem ser usados para melhor identificar os documentos na coleção. Exemplos dessa abordagem podem ser encontrados nos modelos de conjuntos nebulosos e no modelo Booleano estendido. A principal limitação está na explosão exponencial do número de correlações possíveis entre os termos da coleção, o que torna sua computação inviável no caso de coleções maiores. Dessa forma, a maior contribuição desses modelos está na definição dos conceitos a serem explorados pelas abordagens baseadas na correlação de termos (3).

Esses conceitos são fundamentais no desenvolvimento da nossa abordagem uma vez que toda a estratégia por nós proposta se baseia na correlação entre os termos submetidos pelo usuário e coletados pela análise de ações do usuário para identificar o conceito buscado pelo usuário.

As questões relacionadas à complexidade computacional desses modelos, no caso, foram resolvidas tomando como referência o modelo utilizado pelo *Set Based Vector Model* (14), capaz de realizar de modelar essas correlações de uma maneira mais rápida e simples do que as abordagens clássicas.

3 MINERAÇÃO DE DADOS

As correlações entre os termos, tanto da coleção como um todo quanto da sessão de buscas corrente, são determinadas por meio de técnicas de Mineração de Dados. Também conhecida como KDD (*Knowledge Discovery in Databases*), a Mineração de Dados é comumente definida como o processo de descoberta de padrões úteis ou conhecimento a partir de fontes de dados tais como: bancos de dados, textos, imagens, a *Web*, etc. Tais padrões devem ser válidos, potencialmente úteis e compreensíveis.

A origem do KDD se deu a partir da informatização de sistemas e a redução do custo de armazenamento de dados que possibilitaram às empresas acumular grande quantidade de dados. A extração de informações úteis e conhecimento desses dados, no entanto, se mostrou uma tarefa altamente complexa. Os métodos de análise de dados tradicionais, geralmente, não podem ser aplicados devido ao tamanho das bases de dados e à natureza dos mesmos, o que levou ao desenvolvimento de novos métodos de análise para essa nova realidade.

A Mineração de Dados aglutina os métodos tradicionais de análise de dados e algoritmos otimizados para o processamento de grandes volumes de dados. Ela possibilita também a análise de novos tipos de dados e de dados antigos sob uma nova perspectiva. Isso permite descobrir novos padrões úteis que de outra forma permaneceriam desconhecidos e até mesmo predizer o comportamento dos dados futuros. (15)

Uma tarefa de mineração de dados geralmente se inicia com a compreensão do domínio de aplicação pelo analista de dados que então identifica fontes de dados adequadas e o conhecimento a ser buscado. Tais dados são pré-processados a fim de que se adéquem à execução dos algoritmos de mineração de dados. Uma vez executado o algoritmo a saída é novamente processada para que se torne inteligível ao usuário final.

Figura 3.1- Processo de Descobrimento de Conhecimento em Bancos de Dados (KDD)

Originalmente a mineração de dados utilizava fontes estruturadas em bancos de dados relacionais, planilhas de cálculo e arquivos texto tabulares. Com o crescimento da *Web* e dos

documentos de texto, a mineração da *Web* (*Web Mining*) bem como de textos (*Text Mining*) têm se tornado cada vez mais importantes e populares.

3.1 REGRAS DE ASSOCIAÇÃO

As regras de associação buscam identificar padrões regulares nos dados. Elas são, provavelmente, o modelo de mineração de dados mais importante e estudado. Seu objetivo principal é a identificação das relações de co-ocorrência, chamadas associações, entre os itens da base dados.

Originalmente proposta em (16), a aplicação clássica das regras de associação é a análise de "cestas de compras", que busca descobrir relações entre os itens comprados por um consumidor. Por exemplo, seja a tabela 3.1 o registro de vendas de um supermercado.

TID	Items
01	{Carne, Frango, Leite}
02	{Carne, Queijo}
03	{Queijo, Botas}
04	{Carne, Frango, Queijo}
05	{Carne, Frango, Roupas, Queijo, Leite}
06	{Frango, Roupas, Leite}
07	{Frango, Leite, Roupas}

Tabela 3.1 - Exemplo de transações de vendas

É possível extrair dessa lista que, 43% das pessoas compram roupas, frango e leite e que 100% daquelas que compram roupas também compram frango e queijo.

Roupas → Frango, Leite [Suporte= 43%, Confiança = 100%]

No contexto do nosso trabalho, esse tipo de relacionamento pode ser usado para definir a correlação existente entre os termos existentes nos documentos ou entre aqueles usados como chave de busca nas pesquisas e respostas selecionadas, ou seja, consideradas relevantes pelo usuário.

3.1.1 Definição do Problema de Mineração de Regras de Associação

Formalmente, o problema de mineração de regras de associação pode ser definido da seguinte forma:

Definição 3.1 *Seja* $I = \{i_1, i_2, ..., i_m\}$ *um conjunto de itens. Seja* $T = (t_1, t_2, ..., t_n)$ *um conjunto de transações (a base de dados), onde cada transação* t_i *é um conjunto de itens tal que* $t_i \subseteq I$. *Uma regra de associação é uma implicação da forma*

$$X \rightarrow Y, \text{ onde } X \subset I,\ Y \subset I \text{ e } X \cap Y = \varnothing$$

X (ou Y) é um conjunto de itens, chamado ***itemset****.*

As transações podem ser representadas por meio de uma tabela binária onde cada linha corresponde a uma transação e cada coluna corresponde a um item da lista de itens possíveis em uma transação. A representação binária identifica com 1 a presença do item na transação e com 0 a sua ausência. Como a presença de um determinado item tem um significado muito mais importante do que sua ausência os registros da tabela são chamados variáveis binárias assimétricas.

TID	Carne	Frango	Leite	Queijo	Botas	Roupas
01	1	1	1	0	0	0
02	1	0	0	1	0	0
03	0	0	0	1	1	0
04	1	1	0	1	0	0
05	1	1	1	1	0	1
06	0	1	1	0	0	1
07	0	1	1	0	0	1

Tabela 3.2 - Representação binária das transações de exemplo

Diz-se que uma transação $t_i \in T$ contém um itemset X se X for um subconjunto de t_i. A contagem de suporte de X em T (denotado por X.Count) é o número de transações em T que contém X. A força de uma regra é medida pelo seu suporte e sua confiança.

Definição 3.2 *O suporte de uma regra* $X \rightarrow Y$, *é o percentual das transações em* T *que contém* $X \cup Y$ *e pode ser vista como uma estimativa da probabilidade* $Pr(X \cup Y)$. *O suporte da*

regra determina então o quão frequente a regra é aplicável nas transações T. Seja então n o número de transações em T. O suporte da regra $X \rightarrow Y$ será dado por:

$$Suporte = \frac{(X \cap Y).Count}{n}$$

Definição 3.3 *A confiança de uma regra $X \rightarrow Y$, é o percentual entre as transações em T que contém X, que* contém *Y e pode ser vista como uma estimativa da probabilidade Pr(X|Y). A confiança da regra determina o seu poder preditivo. Uma regra com um poder preditivo reduzido dificilmente terá grande utilidade. A confiança da regra $X \rightarrow Y$ é dado por:*

$$Confiança = \frac{(X \cap Y).Count}{X.Count}$$

A partir das definições tem-se que a regra retirada da tabela 3.1 significa que em 43% das compras realizadas o consumidor comprou Roupas, Frango e Queijo e que todos os consumidores que compraram Roupas compraram também Frango e Queijo.

Os conceitos de suporte e confiança são muito importantes pois uma regra com um suporte muito baixo pode ter ocorrido meramente ao acaso e que, geralmente, podem ser descartadas. A confiança por sua vez determina a confiabilidade da inferência proposta pela regra de associação, ou seja, a expectativa que se pode ter de que uma vez configurado o antecedente a ocorrência do consequente será verdadeira.

As regras de associação indicam a relação de co-ocorrência entre itens e não a relação causal dessa co-ocorrência, o que demanda o conhecimento das razões pelas quais a ocorrência dos antecedentes implicam ocorrência dos consequentes.

Com todos esses elementos em mente, pode-se definir o problema de mineração de regras de associação da seguinte forma:

Definição 3.4 Descoberta de Regras de Associação: *Dado um conjunto de transações T, encontrar todas as regras que tenham um suporte ≥ minsup e confiança ≥ minconf, onde minsup e minconf são os limites pré-determinados para o suporte e a confiança.*

Uma abordagem ingênua, por força bruta, nos levaria a calcular o suporte e a confiança de todas as regras possíveis, o que tem um custo computacional claramente proibitivo devido à sua natureza exponencial. O número de regras a ser minerado em uma base com d itens seria dado pela equação:

$$R = 3^d - 2^{d+1} + 1$$

Na prática, normalmente mais de 80% das regras geradas seriam descartadas para um suporte mínimo de 20% e uma confiança mínima de 50%, ou seja, a maior parte do

esforço computacional seria perdido. Para evitar esse desperdício a maioria dos algoritmos da área desacoplam os requisitos de suporte e confiança subdividindo a tarefa original em duas sub-tarefas principais:

a) Geração de Conjuntos de Itens Frequentes (*Frequent Itemsets*): busca identificar os *itemsets* que possuem um suporte igual ou superior ao *minsuport*.
b) Geração de Regras: busca identificar, dentre os *frequent itemsets*, quais possuem uma confiança igual ou superior à *minconf*.

3.1.2 Geração de Conjuntos de Itens Frequentes (*Frequent Itemsets)*

Podemos ilustrar o trabalho de geração dos itemsets através de uma estrutura em treliça como a da figura 3.2.

Essa estrutura contém todos os possíveis itemsets de uma base com os itens {A, B, C, D, E}. Normalmente, uma base com k itens pode gerar até $2^k - 1$ conjuntos frequentes, excluindo-se o conjunto vazio. Como k é um valor muito grande (especialmente quando se trata do vocabulário de diversos idiomas como na *Internet*), o número de *itemsets* gerados seria absurdamente grande.

Para fazer a contagem do suporte de cada *itemset* na estrutura da treliça seria necessário verificar a existência de cada *itemset* candidato em cada transação da base o que teria uma complexidade $O(N \times M \times w)$, onde N é o número de transações, $M = 2^k - 1$ é o número de *itemsets* candidatos e w é o número máximo de itens em uma transação.

A redução da complexidade computacional da geração dos *itemsets* frequentes pode basear-se em duas estratégias:

a) Reduzir o número de *itemsets* candidatos
b) Reduzir o número de comparações

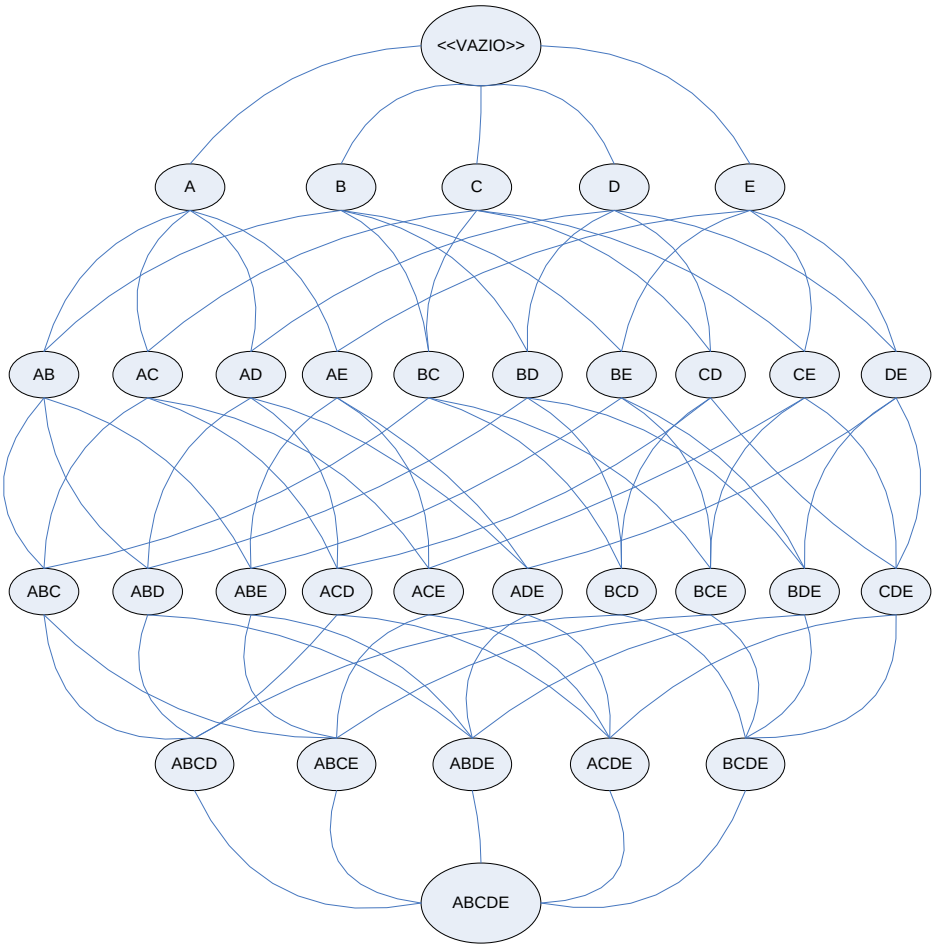

Figura 3.2 - Treliça de Itemsets

3.1.2.1 O Princípio *Apriori*

A fim de reduzir o número de *itemsets*-candidatos pode-se realizar uma poda na treliça baseando-se no princípio *Apriori*.

Teorema 3.1 Princípio *Apriori*: *Se um itemset é frequente, então todos os subconjuntos de seus itens também são frequentes.*

Com esse princípio em mente é possível realizar a poda na contagem do suporte tanto partindo do conjunto universo da coleção quanto do conjunto vazio. Partindo do universo, ao se encontrar um *itemset* frequente, pode-se assumir que todos os seus subconjuntos também são frequentes. Caso parta do conjunto vazio, ao se encontrar um *itemset* não frequente, pode-se assumir que todos os conjuntos dos quais aquele *itemset* é um subconjunto não é frequente. Em ambos casos torna-se desnecessário realizar a contagem do suporte para os *itemsets* que se sabe de antemão serem frequentes ou infrequentes.

O *Apriori* foi o primeiro algoritmo de mineração de regras de associação a utilizar a poda baseada no suporte a fim de controlar sistematicamente o crescimento exponencial dos *itemsets*-candidatos. Inicialmente cada item é considerado um candidato (1-itemset). Uma vez contado o suporte para cada um deles aqueles que apresentaram um suporte inferior ao *minsuport* são cortados e apenas os restantes são recombinados uma a um a fim de gerar *itemsets* da ordem seguinte (2-itemset). Esse processo se repete até que se chegue ao conjunto universo ou que não haja mais nenhum candidato.

O pseudocódigo do algoritmo de geração dos *itemsets* frequentes pode ser visto a seguir. Nele C_k representa o conjunto de *k*-itemsets candidatos e F_k o conjunto de *k-itemsets* frequentes.

Algoritmo 3.1 – Geração de *Itemsets* Frequentes pelo algoritmo *Apriori*

```
1:  k = 1.
2:  F_k = { i | i ∈ I ∧ σ({i}) ≥ N × minsup }.        (Encontra todos os 1-itemsets)
3:  REPITA
4:    k = k + 1.
5:    C_k = apriori − gen(F_{k−1})                    (Gera os itemsets candidatos)
6:    PARA CADA transação t ∈ T FAÇA
7:      C_t = subset(C_k, t).                         (Identifica todos os candidatos que pertencem a t)
8:      PARA CADA itemset candidato c ∈ C_t FAÇA
9:        σ(c) = σ(c) + 1                             (Incrementa a contagem do suporte)
10:     FIM
11:   FIM
12:   F_k = { c | c ∈ C_k ∧ σ(c) ≥ N × minsup }.      (Extrai os k-itemsets frequentes)
13: ATÉ QUE F_k = 0
14: Resultado = ∪F_k.
```

A função *apriori-gen*, presente na linha 5 do pseudocódigo realiza a geração e poda de *itemsets*-candidatos. A geração é feita baseando-se nos *(k-1)-itemsets* frequentes encontrados na iteração anterior. Já a poda é feita por meio de estratégias baseadas no suporte.

Existem diversas formas de se gerar os itemsets candidatos. No entanto deve-se ter em mente os seguintes requisitos para que o processo seja o mais eficiente possível:

a) Deve-se evitar a geração de candidatos desnecessários, para isso pode-se utilizar do princípio *apriori*.
b) Deve-se assegurar que todos os candidatos foram gerados, para assegurar essa completude é necessário que os *itemsets*-candidatos incluam todo o conjunto de *itemsets* frequentes.
c) O mesmo *itemset* não deve ser gerado mais de uma vez

Na abordagem por força bruta para geração de candidatos todos os *k-itemsets* são considerados candidatos potenciais realizando-se a poda para remover os desnecessários. O excessivo número de *itemsets* a serem examinados na poda faz com que esse processo torne-se demasiadamente custoso; a complexidade total da abordagem acaba por ter o limite assintótico $O\left(\sum_{k=1}^{d} k \times \left(\frac{d}{k}\right)\right) = O(d \cdot 2^{d-1})$.

Uma abordagem alternativa, chamada de $F_{k-1} \times F_1$, combina cada *(k-1)-itemset* com um item frequente. A completude nesse caso é garantida pelo fato de todo *k-itemset* frequente ser composto por um *(k-1)-itemset* e um *itemset* frequente. Apesar de apresentar uma grande redução do custo de poda se comparado ao método de força bruta, um grande número de candidatos desnecessários ainda é gerado nessa abordagem. O limite assintótico nesse caso é $O(\sum_k k|F_{k-1}||F_1|)$.

O método utilizado pela função *apriori-gen* é o $F_{k-1} \times F_{k-1}$, onde são mesclados apenas os pares de *(k-1)-itemsets* que apresentem o mesmo prefixo de tamanho *k*-2. A poda nesse caso é feita em um número bastante reduzido de *itemsets*.

3.1.2.2 Representações Compactas de *Itemsets* Frequentes

O número de *itemsets* considerados frequentes pode ser muito grande, por isso é uma boa idéia encontrar aqueles a partir dos quais seja possível derivar todos os outros. Duas formas muito aceitas de se fazer isso são os chamados *itemsets* frequentes máximos (*maximal frequent itemsets*) e fechados (*closed frequent itemsets*). (17)

Um *k-itemset* é dito máximo quando não há nenhum *(k+1)-itemset* frequente que o contenha. Ele é uma representação compacta ótima pois pelo princípio *apriori* pode-se inferir que todos os seus *subsets* sejam frequentes e não é redundante pois não se pode inferir a sua frequência a partir de nenhum outro *itemset*. A representação pelos *itemsets* máximos, no entanto, implica perda de informação a respeito do suporte dos *subsets*, sendo apenas possível determinar que eles são frequentes.

Como a informação a respeito do suporte em muitos casos é relevante, define-se como *itemset* fechado os *itemsets* cujos suportes de todos os seus *supersets* sejam diferentes do seu. Dessa forma obtém-se uma representação mínima mantendo todas as informações disponíveis na derivação dos *subsets* frequentes.

3.1.3 Geração de Regras de Associação

Uma vez identificados os *itemsets* frequentes chega o momento da segunda subtarefa da mineração de regras de associação, a geração de regras. Cada *k-itemset* frequente, *Y*, pode gerar até 2^k-2 regras de associação, ignorando-se aí as regras com antecedentes ou consequentes nulos.

Uma regra de associação pode ser obtida particionando o *itemset* frequente em dois *subsets*, *X* e $Y - X$, tal que a regra $X \rightarrow Y - X$, satisfaça o limite de confiança mínimo, *minconf*.

A confiança não é monotônica como o suporte, ainda assim é possível realizar a poda pela confiança baseando-se no teorema a seguir.

Teorema 3.2 Se a regra $X \rightarrow Y - X$ não satisfaz o limite de confiança, então qualquer regra do tipo $X' \rightarrow Y - X'$, onde X' é um subconjunto de *X*, também não irá satisfazê-la.

O algoritmo *Apriori* utiliza uma abordagem por níveis (*level-wise*) para a geração das regras. Nela, cada nível corresponde ao número de itens do consequente da regra. Inicialmente, apenas as regras de alta confiança que possuem um único item como consequente são geradas. Essas regras são então utilizadas para gerar as novas regras

candidatas. A poda nesse caso é realizada quando uma regra com um determinado consequente possui uma confiança baixa então todas as regras que possuem um subconjunto dos antecedentes e aquele consequente como subconjunto dos consequentes também terão uma confiança baixa. Por exemplo: se {abc}→{d} tem uma baixa confiança, {ab}→{cd}, {ac}→{bd}, {bc}→{ad}, {a}→{bcd}, {b}→{acd}, {c}→{abd}, também terão a confiança baixa e por isso não precisam ser computadas.

4 RECUPERAÇÃO DE INFORMAÇÃO PELO CONCEITO SEMÂNTICO

Esse capítulo apresenta em detalhe a abordagem proposta nesse trabalho. Embasados pelas teorias de Recuperação da Informação e Mineração de Dados buscamos definir uma abordagem capaz de otimizar o processo de buscas por meio da identificação da semântica dos termos buscados pelo usuário e não apenas pelos termos em si.

A chave para a identificação da semântica dos termos é o pré-processamento da coleção de documentos na busca das correlações que permitam identificar outros termos capazes de especificar o conceito buscado com maior exatidão. Para essa tarefa nós nos baseamos no modelo utilizado pelo *Set Based Vector Model* (14), que é capaz de realizar essa tarefa de uma maneira mais rápida e simples do que as abordagens clássicas como o *Generalized Vector Space Model* (11), apresentado anteriormente.

Em linhas gerais, o que se faz é buscar regras de associação capazes de expandir e desambiguar os termos utilizados nas consultas da sessão de buscas corrente ou presentes nos *snippets* dos documentos selecionados pelo usuário, i.e., potencialmente relevantes para ele.

Além do uso de termos correlacionados, nossa abordagem utiliza também os próprios termos utilizados nas consultas submetidas anteriormente, definindo um fator de redução de peso constante à medida que a consulta na qual o termo estava presente, se distancia da atual. O peso dos termos das consultas anteriores é dado, então, não só pela sua frequência na consulta e raridade na coleção, mas também pela distância entre a consulta na qual ele foi submetido e a consulta corrente segundo a função:

$$w_{iq} = \frac{f\left(t_{f_{iq}}\right) \times idf_i}{k \times n}$$

Onde:

$f\left(t_{f_{iq}}\right)$ é a freqüência do termo i na consulta

idf_i é o inverso da freqüência do termo i na coleção

k é o índice da consulta na qual o termo foi submetido

n é o número de termos da consulta k

4.1 MINERAÇÃO DA WEB - WEB MINING

Máquinas de busca são a aplicação de maior relevância na *Web*. Apesar de serem oriundas dos sistemas de Recuperação de Informação, as características dos documentos da *Web* demandam um novo conjunto de técnicas e algoritmos. As páginas não são compostas apenas de texto puro, pois apresentam alguma hierarquia chegando a ser semi-estruturadas em diversos casos. Além disso os documentos estão interconectados por meio de *hyperlinks* o que deve ser levados em consideração quando se projeta um sistema de RI para esse cenário. A maior diferença entre esses cenários, no entanto, é o tamanho das coleções a serem pesquisadas. Por maior que fosse a coleção a ser indexada pelos métodos tradicionais, ela ainda seria muito pequena se comparada aos mais de 1 trilhão de documentos atualmente conhecidos pelo Google (18). Isso dá uma importância ainda maior à questão do desempenho, uma vez que os usuários desejam respostas rápidas e precisas.

A mineração da *Web* procura descobrir informações úteis ou conhecimento a partir da estrutura de hiperlinks da *Web*, conteúdo das páginas e informações de uso (19). O conhecimento de mineração de dados é crucial à mineração de informações disponíveis na *Web*. No entanto, devido à natureza heterogênea dos dados disponíveis, geralmente apresentando-se de forma semi-estruturada ou não estruturada, uma série de algoritmos foram desenvolvidos especificamente para as tarefas de mineração da *Web*.

Diversos aspectos da *Web* podem ser considerados para extrair informações importantes. A mineração da estrutura da *Web* (*Web Structure Mining*) busca o conhecimento por meio dos *hyperlinks* que representam a estrutura da *Web*. Esse processo, conhecido como linkanálise é muito utilizado por mecanismos de buscas. Um bom exemplo disso é o algoritmo de *PageRank* do Google. A mineração da utilização da *Web* (*Web Usage Mining*) busca identificar os padrões de utilização dos *websites* por meio da análise dos *logs* dos servidores. Nosso trabalho, no entanto, é focado na mineração de conteúdo da *Web* (*Web Content Mining*) e busca ser capaz de identificar os conceitos tratados nas páginas coletadas a fim de possibilitar respostas mais relevantes aos usuários de máquinas de busca.

O algoritmo de *ranking* é a parte mais importante de uma máquina de buscas e, por isso mesmo, os algoritmos utilizados em mecanismos de busca comerciais são tratados como segredos industriais. Em geral, as demonstrações didáticas baseiam-se nos algoritmos divulgados pelo Google em (20). Os métodos tradicionais de *ranking* em RI, tais como a

medida da similaridade do cosseno não são suficientes para a *Web*. A grande quantidade de documentos faz com que praticamente qualquer consulta tenha uma enorme quantidade de respostas consideradas relevantes.

A principal tarefa então deixa de ser a de encontrar os documentos relevantes e passa a ser ordenar os documentos relevantes de forma que aqueles de maior qualidade sejam posicionados à frente dos demais. Esse problema geralmente é abordado pelo estudo da autoridade da página, ou seja, as citações feitas àquela página por outras páginas na *Web*, i.e., *links* entre as páginas. Ainda assim é importante que se leve em consideração o conteúdo das páginas e o destaque dado ao termo pesquisado nela. A análise do conteúdo leva em consideração o tipo de ocorrência: título, texto âncora, corpo ou url, número de ocorrências e posição das ocorrências.

Ainda assim resta um problema de difícil tratamento devido ao alto custo computacional da maioria das abordagens: a análise semântica dos termos. Um mesmo termo pode apresentar diversos significados e, nesse caso, uma página de um domínio de grande autoridade com total foco em uma determinada chave de busca do usuário pode, na verdade, não ter relação nenhuma com o conceito buscado. Esse é o principal problema atacado por nossa abordagem.

4.2 DEFINIÇÃO DOS CONTEXTOS

O contexto considerado para a mineração das regras de associação é a sessão de buscas. Entende-se por sessão de buscas a sequência de consultas com foco em um determinado tema. Apesar de não ser possível determinar exatamente esse conceito sem uma interação explícita do usuário, ele pode ser aproximado satisfatoriamente utilizando-se o *timeout* da sessão definida no servidor.

Diversos trabalhos focados na análise de *logs* de acessos de servidores já foram realizados, usamos a caracterização feita em (21) onde chega-se ao valor de 1000 segundos como um limite satisfatório na definição da sessão. Esse conceito pode ser extrapolado para o nosso cenário definindo assim o limite a ser parametrizado no servidor para o *timeout* da sessão.

4.3 RELEVANCE FEEDBACK

A técnica de *Relevance Feedback* é a mais utilizada para reformulação de consultas. Na abordagem inicialmente proposta, uma série de respostas são apresentadas ao usuário, como numa consulta comum, e ele deve selecionar aquelas consideradas relevantes. A estratégia por trás desse método está em identificar termos e expressões que sejam capazes de caracterizar os tipos de documentos selecionados como relevantes pelo usuário para que mais documentos semelhantes a eles sejam retornados nas próximas consultas por meio da reformulação dos termos das consultas. O efeito esperado é que as próximas consultas retornem os documentos mais relevantes melhor ranqueados.

O *Relevance Feedback* mostra-se mais eficiente do que outras estratégias de reformulação pois sua atuação é transparente para o usuário, transforma a tarefa de busca em uma sequência de consultas mais simples e provê um processo controlado de se enfatizar os termos mais relevantes e tirar peso dos demais.

No *Relevance Feedback* Explícito, o próprio usuário avalia a qualidade dos resultados retornados por meio de uma interface binária ou multivalorada no sistema. Essa estratégia foi acrescentada recentemente ao serviço de buscas do Google com o nome de *SearchWiki* (22). Outra ferramenta comercial que adota essa estratégia é o plugin de navegadores Exalead[2]. A interação explícita com o usuário muitas vezes dificulta o uso do recurso por usuários menos experientes. Infelizmente a grande maioria dos usuários de sistemas de busca na *web* têm esse perfil restringindo-se à submissão de poucas palavras-chave que, na sua concepção, sejam capazes de classificar os documentos que lhe interessam. Isso faz com que o *Explicit Feedback* seja utilizado eficientemente apenas por um pequeno grupo de usuários da máquina de busca.

O *Pseudo Feedback* automatiza a parte manual do *Relevance Feedback* ao assumir que os primeiros *k* documentos dos n retornados, onde $k << n$, são relevantes. O efeito colateral dessa estratégia é claro e acaba por dar nome à própria estratégia pois uma vez que não é considerada qualquer ação do usuário, não se tem realmente um *feedback* do mesmo.

[2] http://www.exaled.com

O *Implicit Feedback* (23) (24) (25) (26) utiliza técnicas capazes de capturar a interação do usuário com o sistema de forma transparente para agregar uma série de informações relevantes ao modelo de recuperação adotado, quer seja na chave de consultas, quer seja na coleção de documentos (27) (28). Através dele a relevância dos documentos para o usuário pode ser inferida pelo padrão comportamental do usuário sem que seja necessária uma intervenção explícita por parte dele. Nesse modelo o usuário não tem que cooperar com o sistema a fim de que ele se aprimore mas, apenas, pela busca natural da informação desejada. Muitas vezes o usuário nem mesmo sabe que seu *feedback* será utilizado na otimização do mecanismo de busca.

As informações conseguidas por meio do *Relevance Feedback* podem ser aproveitadas de duas maneiras distintas:

- Expansão da consulta (*Query Expansion)* – adiciona-se novos termos oriundos dos documentos relevantes.
- Redefinição dos pesos dos termos (*Reweighting*) – redefine-se os pesos dos termos de acordo com o julgamento de relevância do usuário

Um bom exemplo de utilização dessa estratégia é a extensão de navegadores da Internet *Surf Canyon*[3], que gera listas secundárias de documentos em função da relevância de cada um dos documentos retornados, exigindo apenas que o usuário clique em um ícone ao lado do documento retornado na consulta. Apesar de, nesse caso, a interação do usuário com o sistema ser exclusivamente em benefício próprio, acreditamos que a instalação de extensões aos navegadores utilizados pelos usuários uma tarefa que, por si só, diminui sensivelmente a aplicabilidade desse tipo de solução.

Acreditamos também que, quanto menor a alteração no padrão comportamental do usuário em uma sessão de buscas, maior a chance de uma estratégia ser bem aproveitada pela comunidade em geral e que, mesmo se ao ser comparada com estratégias que exijam uma interação explícita do usuário com o sistema de *feedback*, as abordagens não intrusivas apresentem resultados inferiores sua viabilidade e escalabilidade acabam por fazer delas as melhores opções. Em nosso trabalho utilizamos o *Implicit Feedback* determinado pela sessão de buscas corrente do usuário.

[3] http://www.surfcanyon.com/

Um usuário que realiza, por exemplo, uma consulta com os termos {“Paris”, “Hilton”}, terá como resposta à sua consulta em um modelo tradicional tanto documentos que falem a respeito do hotel Hilton na cidade de Paris quanto da socialite americana Paris Hilton. É improvável, no entanto, que ambos grupos sejam o foco da busca do usuário. A identificação da entidade sobre a qual o usuário deseja informações pode ser obtida por meio da análise dos termos submetidos anteriormente e dos documentos selecionados como resposta às consultas anteriores da sessão corrente. Se o usuário pesquisou sobre roteiros turísticos na França, por exemplo, a maior probabilidade é de que ele se interesse pela cidade e a rede de hotéis e que os documentos que tratem desse tema sejam mais relevantes do que aqueles que falam da socialite.

4.4 EXPANSÃO DE CONSULTAS

A elaboração de consultas por usuários que não conhecem bem a coleção de documentos a ser pesquisada tem se mostrado uma tarefa árdua. Nos logs de máquinas de busca na *Web* pode-se observar que os usuários acabam por reformular suas consultas várias vezes até chegarem às respostas consideradas relevantes. Sendo assim, podemos considerar que a primeira consulta de uma sessão de buscas não possui nada além de si mesma para alimentar o sistema de RI mas que as consultas seguintes podem aproveitar os termos já utilizados nas consultas bem como os documentos examinados pelo usuário a fim de otimizar a formulação de consultas e conseguir um resultado satisfatório mais rapidamente.

As abordagens tradicionais de expansão de consultas baseiam-se, geralmente em um do seguintes modelos:

a) *Feedback* Explícito: onde o usuário seleciona explicitamente dentre os documentos previamente retornados, por meio de *check boxes* ou alguma interface semelhante, aqueles que considera relevante;
b) Análise do contexto local: onde são usadas técnicas de agrupamento (*clustering*) local para analisar a co-ocorrência de termos nos documentos melhor posicionados na busca original;

c) Análise global: onde toda a coleção é previamente examinada e os agrupamentos são determinados por *thesaurus* baseados em similaridade ou estatística.

Nossa abordagem aplica tanto a análise global quanto a análise de contexto local. Da análise global identificamos a correlação entre termos em toda a coleção, dessa forma conseguimos identificar, por exemplo, que o termo “JAGUAR” se relaciona com “FERRARI” e que, apenas quando essas relação ocorre nas consultas “JAGUAR” se relaciona com “AUTOMÓVEL” e não se relaciona com “ANIMAL”. A análise do contexto local permite realizar a mineração de dados em uma base instantânea proveniente apenas dos termos submetidos nas consultas da sessão corrente e dos títulos e *snippets* dos documentos retornados. Isso permite que regras que não seriam identificadas por não apresentarem um suporte acima do limite mínimo na coleção sejam identificadas na sessão de buscas.

Como optamos por não interagir explicitamente com o usuário, a análise do *feedback* do usuário foi feita exclusivamente considerando que os documentos clicados pelo usuário possíam características que o levavam a considerá-los relevantes não estabelecendo regras que definissem o documento como não relevante. Isso foi feito porque, na prática, o usuário não clica em tudo o que considera relevante, sendo assim o fato de um documento não ser clicado não foi considerado negativamente. Por outro lado, apesar de um usuário, eventualmente, clicar em conteúdos fora do seu foco de busca inicial, esse não um comportamento usual.

4.5 MODELAGEM DA CORRELAÇÃO ENTRE OS TERMOS

Como dissemos na sessão anterior, a aplicação da análise global para a expansão das consultas se dá pela identificação da correlação semântica entre os termos toda a coleção. A fim de que sejam estabelecidas as estruturas de dados necessárias à mineração das regras de associação entre os termos e às tarefas de recuperação da informação, todos os documentos da coleção devem ser processados *a priori*.

Definição 4.1 - *Seja $T = \{k_1, k_2, ..., k_n\}$ o vocabulário de uma coleção C com N documentos, composto de um único registro k_x para cada termo distinto presente em algum documento da coleção.*

Definição 4.2 - *Um n-termset, S, é um subconjunto de n termos do vocabulário.*

Definição 4.3 - Seja *$V = \{S_1, S_2, ..., S_n\}$ o vocabulário de itemsets da coleção C, ou seja, o conjunto dos 2^t termsets distintos que podem ocorrer em um documento da coleção.*

Assim como para os termos do vocabulário da coleção, para cada *termset* é definida uma lista invertida contendo suas ocorrências na coleção. Essas listas serão usadas também para agilizar a tarefa de mineração de regras de associação de termos uma vez que tornam mais fácil a contagem do suporte.

Um *termset* é dito frequente quando seu suporte é maior ou igual ao suporte mínimo, definido como parâmetro na execução da mineração de regras de associação. Na realidade esses são os *termsets* realmente importantes para a nossa abordagem já que somente eles serão capazes de gerar as regras de associação que buscamos.

4.6 GERAÇÃO DE TERMSETS

Os termsets são gerados de maneira análoga à geração de itemsets pelo algoritmo *Apriori* (29). As listas invertidas geradas para as tarefas de recuperação da informação pelos modelos tradicionais são aproveitadas para realizar a contagem do suporte dos termos do vocabulário de forma otimizada, uma vez que através dela percorre-se a treliça de itens transversalmente.

Uma vez identificados os termos frequentes é feita a interseção entre as listas invertidas dos termos dois a dois, a fim de se estabelecer os *2-termsets* frequentes. Esse processo acontece de maneira sucessiva, assim como na geração de *itemsets* frequentes, até que não existam mais *termsets* para serem combinados.

4.6.1 *Termsets p*róximos

Uma restrição extra que pode ser imposta na geração dos *termsets* frequentes é a noção de proximidade. Para isso, além de estabelecer um valor limite para a distância entre os termos nos documentos, deve-se adaptar as listas invertidas para que armazenem também a posição das ocorrências dos termos nos documentos.

Essa abordagem mostrou-se capaz de agregar ainda mais valor semântico aos *termsets*, no entanto não foi adotada devido ao acréscimo do custo computacional, tanto em espaço quanto em processamento (14).

4.7 REGRAS DE ASSOCIAÇÃO DE TERMSETS

O simples fato de um termset estar presente em um grande número de documentos, no entanto, não é suficiente para determinar a existência de correlação semântica entre os termos que o formam. Conforme apresentado no capítulo 3, é necessário que exista uma relação causal entre as ocorrências para que se estabeleça uma regra forte, por isso, uma vez identificados os termsets frequentes, eles são analisados a fim de se estabelecer em cada um deles os possíveis grupos de antecedentes e consequentes para a geração das regras de associação.

Uma consideração que deve ser feita é o fato de diversas regras poderem estar contidas em outras regras de *itemsets* maiores. A simples remoção dessas regras pode não ser uma boa abordagem, uma vez que a confiança da regra mais genérica pode não dar a mesma relevância à correlação presente na regra menos genérica. Isso nos leva à escolha de utilização dos *termsets* fechados (extensão dos itemsets fechados - *Closed Itemsets*) ou *termsets* máximos (extensão dos itemsets máximos - *Maximal Itemsets*).

Definição 4.4 - *Fechamento de um Termset S_i:* *é o conjunto de todos os termsets que co-ocorrem com S_i, no mesmo conjunto de documentos e preservam as restrições de proximidade.*

Definição 4.5 - ***Termset Fechado:*** *é o maior termset de um fechamento* S_i.

No nosso caso optamos por utilizar os *termsets* fechados, descartando apenas os *termsets* que não agregavam qualquer informação adicional, assumindo assim o aumento na complexidade de espaço.

A computação dos *termsets* fechados foi feita implementando o algoritmo *CHARM* (30). Esse algoritmo baseia-se na ordenação de todos os *termsets* frequentes da coleção e no teste de fechamento deles. Todo *termset* contido em outro termset do mesmo fechamento é então removido da lista.

Definição 4.6 - ***Termset Máximo:*** *é um termset frequente que não esteja contido em nenhum outro termset frequente.*

Os *termsets* máximos são a representação mínima de todos os *termsets* frequentes, mas apesar de serem ainda mais compactos do que os *termsets* fechados implicam na perda da informação a respeito do suporte dos *termsets* contidos nele. Sua aplicação mostra-se especialmente útil quando a redução na complexidade computacional mostra-se mais importante do que a identificação individual dos suportes dos *termsets*.

4.8 ELEMENTOS DE OTIMIZAÇÃO

A análise do contexto local se dá então pela busca novas correlações de termos específicas para a sessão de buscas em questão em tempo real. Apesar da elevada complexidade computacional comum às tarefas de mineração de dados, o tamanho reduzido da fonte de dados favorece a realização das tarefas em um tempo aceitável. As transações a serem consideradas, nesse caso, passam a ser as consultas submetidas durante a sessão bem como os *snippets* dos documentos retornados pelo mecanismo de busca.

A escolha pelo uso apenas dos *snippets* ao invés do conteúdo total dos documentos baseia-se no fato de que essa é a única informação disponível ao usuário até o momento da sua escolha pelo documento. Como na análise do *implicit feedback* isso é visto como a determinação de relevância por parte do usuário, se considerássemos todo o conteúdo do documento estaríamos considerando conteúdo não avaliado pelo usuário.

As regras geradas dessa forma podem ser de dois tipos, um para maximizar os termos antecedentes e acrescentar os consequentes para documentos relevantes e outro para minimizar o peso dos considerados não relevantes:

1. $A_1, A_2, ..., N_a \rightarrow C_1, C_2, ..., C_n$ [RELEVANTE]
2. $A_1, A_2, ..., N_a \rightarrow C_1, C_2, ..., C_n$ [NÃO RELEVANTE]

No nosso caso, no entanto, apenas as regras que geram resultados com documentos relevantes foram consideradas.

Outro fator a ser considerado é a posição do documento na lista de respostas. É sabido que a posição do documento na lista de respostas, por si só, pode influenciar diretamente na probabilidade de um documento ser escolhido (31) e isso deve ser levado em consideração na hora de se estabelecer as regras. Isso é especialmente importante na determinação de regras que tenham como consequência a determinação de documentos não relevantes. Como em nossa abordagem apenas as regras que geram resultados com documentos relevantes foram consideradas, na busca pela simplificação da implementação da solução esse fator também foi desconsiderado.

Por último, é necessário considerar um fator decrescente de influência dos elementos de otimização à medida que a sessão evolui, ou seja, os elementos levantados na primeira consulta devem ter mais impacto na segunda consulta do que na terceira e assim por diante. Isso se deve também ao padrão comportamental dos usuários de refinamentos e expansões sucessivas.

De forma geral, nossa proposta busca considerar toda a informação fornecida pelo usuário durante a sessão de buscas aceitando, no entanto, algumas simplificações no modelo. A partir da segunda consulta da sessão de buscas, as consultas submetidas pelo usuário são expandidas pelos termos utilizados nas consultas anteriores, pelos termos obtidos através das regras de associação extraídas previamente de todo o vocabulário da coleção, pelos termos obtidos pelas regras de associação dos termos submetidos anteriormente e dos *snippets* dos documentos selecionados. Isso se dá através dos seguintes passos:

1. Uma lista invertida dos termos aponta as regras nas quais eles aparecem como antecedentes com isso adiciona-se os termos consequentes à consulta.
2. A cada rodada da sessão são criadas transações compostas pelos termos da consulta e dos *snippets* e títulos dos documentos selecionados. Essas transações são mantidas enquanto a sessão de buscas permanecer ativa. É feita a mineração de regras de associação em tempo real dessas transações em busca de novos termos que possam ser introduzidos nas próximas consultas. É interessante observar que esse processo é

realizado sempre com foco na alteração das consultas seguintes. O peso dessas regras deve ser uma função não só de seu suporte e sua confiança mas também do tamanho do conjunto de transações disponível.

3. A cada nova consulta da mesma sessão os termos introduzidos pelo usuário são replicados *R* vezes, onde *R* é o índice da rodada corrente. Com isso conseguimos que ocorra um decaimento linear no peso dos termos das consultas anteriores de acordo com a sua distância em relação à consulta corrente.

Dessa forma buscamos considerar termos que apesar de não serem explicitamente definidos pelo usuário na consulta corrente sejam importantes para definir o conceito semântico pelo qual o usuário está interessado.

4.9 EXEMPLO MODELO DE APLICAÇÃO DA PROPOSTA

A fim de deixar mais claro o processo explicado nas sessões anteriores mostraremos passo a passo o funcionamento da metodologia proposta a partir de um exemplo real de uma sessão de buscas realizada por um dos voluntários.

O tema da busca foi “EMPREGO” e o objetivo do usuário era conseguir sites com anúncios e ofertas de emprego. A sequência de chaves de busca submetida foi:

1. EMPREGO
2. GERENTE
3. MARKETING BH
4. MARKETING

Os resultados das buscas pela abordagem tradicional e pela proposta são mostrados nas tabelas a seguir.

CHAVE	TÍTULO DO DOCUMENTO	URL DO DOCUMENTO RETORNADO
EMPREGO	Anvisa - Legislação - Resolução	http://www.anvs.gov.br/legis/portarias/10_85.htm
	Anvisa - Legislação - Resolução	http://www.anvisa.gov.br/legis/portarias/10_85.htm
	A priori - Empregos - página atual	http://www.apriori.com.br/dados/empregos.htm
	Documento 5326137	http://www.jurinforma.com.br/cgi-bin/majordomo/get/responsabilidade-civil/responsabilidade-civil.0003
	Documento 4182742	http://www.jurinformatica.com.br/cgi-bin/majordomo/get/responsabilidade-civil/responsabilidade-civil.0003
	Jus Navigandi - Doutrina - A Aids e seus impactos nas relações de trabalho	http://www.jusnavegandi.com.br/doutrina/aids2.html
	Cadê? Serviços\Recursos Humanos\	http://cade.uninet.com.br/servrh.htm
	Cadê? Serviços\Recursos Humanos\	http://ie.cade.com.br/servrh.htm
	Lokaliza - Sistema de Busca	http://www.lokaliza.com.br/servicos/servrecu.htm
	O SEGURO-DESEMPREGO E O COMBATE A POBREZA: RECOMENDAÇÕES PARA O CASO BRASILEIRO	http://www.ipea.gov.br/redepesq/produtos/anpec/encontro/trabalhos/economia_do_trabalho2/JPZChahad.html
gerente	Anvisa - Legislação - Resolução	http://www.anvs.gov.br/legis/portarias/10_85.htm
	Anvisa - Legislação - Resolução	http://www.anvisa.gov.br/legis/portarias/10_85.htm
	MAIN1 - PASTA DE SERVIÇOS	http://www.gbsnetwork.com.br/inc_prof2.php3
	Grupo Orientrade - Trabalhe Conosco	http://www.grupoorientrade.com.br/institucional/trabalhe_conosco.asp
	GOYA - Banco de Empregos	http://www.goya.com.br/curriculo.htm
	Documento 4182742	http://www.jurinformatica.com.br/cgi-bin/majordomo/get/responsabilidade-civil/responsabilidade-civil.0003
	Documento 5326137	http://www.jurinforma.com.br/cgi-bin/majordomo/get/responsabilidade-civil/responsabilidade-civil.0003

	Dimensão	http://www.dimensaoconsult.com.br/cadast.htm
	Grupo Better - recursos humanos	http://www.betterrh.com.br/curr.asp
	A priori - Empregos - página atual	http://www.apriori.com.br/dados/empregos.htm
marketing + BH	Anvisa - Legislação - Resolução	http://www.anvs.gov.br/legis/portarias/10_85.htm
	Anvisa - Legislação - Resolução	http://www.anvisa.gov.br/legis/portarias/10_85.htm
	Grupo Orientrade - Trabalhe Conosco	http://www.grupoorientrade.com.br/institucional/trabalhe_conosco.asp
	MAIN1 - PASTA DE SERVIÇOS	http://www.gbsnetwork.com.br/inc_prof2.php3
	SuSE Linux 6.3 (i386) - November 1999 - "tetex"	http://www.delet.ufrgs.br/hilfe/pak/paket_inhalt_tetex.html
	Documento 3316003	http://www.iagusp.usp.br/preprints/9801/adriano/adriano.tar.gz.uu
	GOYA - Banco de Empregos	http://www.goya.com.br/curriculo.htm
	Cadê? Serviços\Comunicação\Propaganda e Marketing\	http://cade.uninet.com.br/servcomktg.htm
	Documento 3807573	http://cev.ucb.br/pipermail/cevmkt-l/2000-August.txt
	Cadê? Serviços\Comunicação\Propaganda e Marketing\	http://ie.cade.com.br/servcomktg.htm
marketing	Anvisa - Legislação - Resolução	http://www.anvs.gov.br/legis/portarias/10_85.htm
	Anvisa - Legislação - Resolução	http://www.anvisa.gov.br/legis/portarias/10_85.htm
	Grupo Orientrade - Trabalhe Conosco	http://www.grupoorientrade.com.br/institucional/trabalhe_conosco.asp
	MAIN1 - PASTA DE SERVIÇOS	http://www.gbsnetwork.com.br/inc_prof2.php3
	SuSE Linux 6.3 (i386) - November 1999 - "tetex"	http://www.delet.ufrgs.br/hilfe/pak/paket_inhalt_tetex.html
	Documento 3316003	http://www.iagusp.usp.br/preprints/9801/adriano/adriano.tar.gz.uu
	GOYA - Banco de Empregos	http://www.goya.com.br/curriculo.htm
	Cadê? Serviços\Comunicação\Propaganda e Marketing\	http://cade.uninet.com.br/servcomktg.htm
	Documento 3807573	http://cev.ucb.br/pipermail/cevmkt-l/2000-August.txt
	Cadê? Serviços\Comunicação\Propaganda e Marketing\	http://ie.cade.com.br/servcomktg.htm

Tabela 4.1 - Resultados de Busca Sensível a Sessão

CHAVE	TÍTULO DO DOCUMENTO	URL DO DOCUMENTO RETORNADO
EMPREGO	Anvisa - Legislação - Resolução	http://www.anvs.gov.br/legis/portarias/10_85.htm
	Anvisa - Legislação - Resolução	http://www.anvisa.gov.br/legis/portarias/10_85.htm
	A priori - Empregos - página atual	http://www.apriori.com.br/dados/empregos.htm
	Documento 5326137	http://www.jurinforma.com.br/cgi-bin/majordomo/get/responsabilidade-civil/responsabilidade-civil.0003
	Documento 4182742	http://www.jurinformatica.com.br/cgi-bin/majordomo/get/responsabilidade-civil/responsabilidade-civil.0003
	Jus Navigandi - Doutrina - A Aids e seus impactos nas relações de trabalho	http://www.jusnavegandi.com.br/doutrina/aids2.html
	Cadê? Serviços\Recursos Humanos\	http://cade.uninet.com.br/servrh.htm
	Cadê? Serviços\Recursos Humanos\	http://ie.cade.com.br/servrh.htm
	Lokaliza - Sistema de Busca	http://www.lokaliza.com.br/servicos/servrecu.htm
	O SEGURO-DESEMPREGO E O COMBATE A POBREZA: RECOMENDAÇÕES PARA O CASO BRASILEIRO	http://www.ipea.gov.br/redepesq/produtos/anpec/encontro/trabalhos/economia_do_trabalho2/JPZChahad.html

gerente	Grupo Orientrade - Trabalhe Conosco	http://www.grupoorientrade.com.br/institucional/trabalhe_conosco.asp
	MAIN1 - PASTA DE SERVIÇOS	http://www.gbsnetwork.com.br/inc_prof2.php3
	GOYA - Banco de Empregos	http://www.goya.com.br/curriculo.htm
	Dimensão	http://www.dimensaoconsult.com.br/cadast.htm
	Grupo Better - recursos humanos	http://www.betterrh.com.br/curr.asp
	Oportunidades	http://www.dossier-rh.com.br/scripts/oportunidade.asp
	Brasil Vitae - Currículo (cadastrar)	http://www.brasilvitae.com.br/portugues/p_pr_a02.asp
	TRH Curriculum	http://www.trh.com.br/cgi-bin/ccurric.pl
	Reportagens radiofônicas 1998	http://www.cnpm.embrapa.br/reporte/impfala98.html
	Reportagens radiofônicas 1998	http://ipe.nma.embrapa.br/reporte/impfala98.html
marketing + BH	SuSE Linux 6.3 (i386) - November 1999 - "tetex"	http://www.delet.ufrgs.br/hilfe/pak/paket_inhalt_tetex.html
	Documento 3316003	http://www.iagusp.usp.br/preprints/9801/adriano/adriano.tar.gz.uu
	Cadê? Serviços\Comunicação\Propaganda e Marketing\	http://cade.uninet.com.br/servcomktg.htm
	Cadê? Serviços\Comunicação\Propaganda e Marketing\	http://ie.cade.com.br/servcomktg.htm
	Documento 3807573	http://cev.ucb.br/pipermail/cevmkt-l/2000-August.txt
	Cadê? Internet\Serviços\Webmarketing e Divulgação de Sites\	http://cade.uninet.com.br/intsrvwmkt.htm
	Documento 3405191	http://humor.re7.com.br/humor/9f000000.nws
	Cadê? Serviços\Comunicação\Propaganda e Marketing\Agências de Publicidade\	http://cade.uninet.com.br/servcomktgag.htm
	S.I.M. - Sistemas de Informação de Marketing	http://www.fauze.com.br/artigo14.htm
	Documento 3796973	http://cev.ucb.br/pipermail/cevmkt-l/2000-September.txt
marketing	Cadê? Serviços\Comunicação\Propaganda e Marketing\	http://cade.uninet.com.br/servcomktg.htm
	Cadê? Serviços\Comunicação\Propaganda e Marketing\	http://ie.cade.com.br/servcomktg.htm
	Documento 3807573	http://cev.ucb.br/pipermail/cevmkt-l/2000-August.txt
	Cadê? Internet\Serviços\Webmarketing e Divulgação de Sites\	http://cade.uninet.com.br/intsrvwmkt.htm
	Cadê? Serviços\Comunicação\Propaganda e Marketing\Agências de Publicidade\	http://cade.uninet.com.br/servcomktgag.htm
	S.I.M. - Sistemas de Informação de Marketing	http://www.fauze.com.br/artigo14.htm
	Documento 3796973	http://cev.ucb.br/pipermail/cevmkt-l/2000-September.txt
	Cadê? Internet\Serviços\Webmarketing e Divulgação de Sites\	http://ie.cade.com.br/intsrvwmkt.htm
	Cadê? Serviços\Consultoria\Comunicação e Marketing\	http://cade.uninet.com.br/servconscmk.htm
	Cadê? Serviços\Consultoria\Comunicação e Marketing\	http://ie.cade.com.br/servconscmk.htm

Tabela 4.2 - Resultados da Busca Tradicional

Inicialmente foi pesquisado o termo “EMPREGO”, o resultado da pesquisa pela nossa abordagem foi idêntico ao da abordagem tradicional, isso ocorre por que os termos, apesar de

obviamente co-relacionados a outros termos pelo conhecimento histórico não ocorriam com uma frequência suficiente para que fossem geradas regras de associação a partir deles. Esse é uma dificuldade comum conhecida como a “Maldição da Dimensionalidade” (*Curse of Dimensionality*) e é consequência da grande quantidade de termos distintos presentes na coleção.

A atuação do nosso algoritmo pode ser sentida a partir da segunda consulta submetida pelo usuário. Nesse caso o termo utilizado foi “GERENTE”, nesse caso a mineração baseada nos documentos selecionados na primeira consulta e nos termos nela submetidos. Enquanto a pesquisa tradicional usou o modelo de espaços vetoriais para o termo “GERENTE” nosso modelo realizou a pesquisa por “EMPREGO GERENTE GERENTE GERENTE CURRICULUM”. Como pode ser observado, considerando o objetivo da pesquisa do usuário, nossa abordagem levou a uma precisão de 70% nos 10 primeiros documentos enquanto a abordagem tradicional teve uma precisão de 50%.

Na terceira e quarta consultas isso fica ainda mais evidente, os termos usados pelo usuário foram “MARKETING BH” e “MARKETING” respectivamente os termos adaptados pelo nosso algoritmo foram “EMPREGO GERENTE GERENTE GERENTE MARKETING BH MARKETING BH MARKETING BH MARKETING BH MARKETING BH MARKETING BH CURRICULUM VAGAS VAGAS” e “EMPREGO GERENTE GERENTE GERENTE MARKETING BH MARKETING BH MARKETING BH MARKETING BH MARKETING BH MARKETING BH MARKETING MARKETING MARKETING MARKETING MARKETING MARKETING MARKETING MARKETING MARKETING MARKETING CURRICULUM VAGAS VAGAS" o resultado observado foi que enquanto a abordagem tradicional não conseguiu nenhum documento relevante entre os 10 primeiros em nenhuma das duas consultas a nossa acertou 3 em cada uma delas.

Esse exemplo demonstra também um efeito colateral de nossa proposta que é o crescimento excessivo do número de termos a serem considerados na consulta, o que pode tornar o processo de recuperação da informação excessivamente complexo em sessões mais longas. A nosso favor temos o fato de, em geral, as sessões de busca não terem mais de 5 ou 6 consultas.

5 RESULTADOS EXPERIMENTAIS

No intuito de avaliarmos quantitativamente os ganhos obtidos pela nossa abordagem ela foi implementada em uma máquina de buscas baseada no modelo de espaços vetoriais. A escolha desse modelo se deu pela sua simplicidade e o seu uso como base de comparação em diversos artigos

Foram feitos testes com usuários e os resultados foram comparados com aqueles obtidos pela mesma implementação do modelo de espaços vetoriais sem as otimizações propostas no capítulo 4.

Nos testes, os usuário recebiam um grupo de temas sobre os quais deveriam realizar pesquisas como se estivessem utilizando um mecanismo de buscas comercial. Foram consideradas para a análise de precisão e revocação os 50 primeiros documentos retornados em cada consulta realizada.

5.1 A COLEÇÃO DE DOCUMENTOS PARA O EXPERIMENTO

Um desafio extra ao se realizar um estudo como o nosso é projetar experimentos para validá-lo. Em geral, coleções que incluam não apenas a base de textos e tópicos de teste, mas também um histórico de consultas e de documentos clicados, *clickthrough*, para cada tópico não são fáceis de se encontrar.

Uma maneira de se conseguir isso seria extrair os tópicos e históricos de consultas e *clickthroughs* associados a eles do registro de atividades, *log*, de um sistema de recuperação de informação como uma máquina de buscas. O problema nesse caso é que não haveria análise de relevância para esses dados e fazer essa análise com os recursos disponíveis seria impossível.

Outra alternativa é aproveitar os dados disponíveis de uma coleção consolidada como a TREC, onde já existe a base de textos, descrição de tópicos e a análise de relevância, apesar de não possuir históricos de consultas e *clickthrough*. O nosso trabalho então seria gerar esses históricos a fim de realizar nossa análise.

No nosso caso, as coleções utilizadas foram a WT2g (32) e a WBR99. Os históricos foram gerados de maneira semelhante à realizada por (23) com a TREC AP.

A WT2g é uma coleção de 2GB que consiste em 528.155 documentos HTML com 737.833 termos distintos coletados das seguintes fontes: *The Financial Times* (1991–1994), *Federal Register* (1994), *Foreign Broadcast Information Service*, and *LA Times*. Ela foi utilizada na *Web Track* da TREC-8 e é, na realidade, uma sub-parte da coleção VLC de 100GB que, por sua vez, é uma sub-parte da coleta *Internet Archive* de 300GB completada em 1997. Foram utilizadas também as consultas de exemplo 401 a 450 para avaliar o algoritmo de *ranking*.

Como grande parte dos usuários apresentou dificuldade no entendimento pleno das expressões em inglês contidas nos documentos da WT2g, os experimentos usando essa coleção acabaram por ser desconsiderados por não serem capazes de ilustrar os efeitos do uso da nossa abordagem.

Já a WBR99 é uma coleção de aproximadamente 16GB composta por 5.939.061 documentos da Internet brasileira e 2.669.965 termos distintos. Ela também disponibiliza a análise de relevância para as 50 consultas mais frequentes não relacionadas a sexo do *log* do mecanismo de busca Todobr[4]. Essas consultas analisadas foram utilizadas para estabelecer a precisão do algoritmo nos experimentos executados

Tabela 5.1 - Características das Coleções de Referência dos Experimentos

CARACTERÍSTICAS	COLEÇÃO	
	WT2g	WBR99
Número de Documentos	528.155	5.939.061
Número de Termos Distintos	737.833	2.669.965
Número de Tópicos Disponíveis	450	100.000
Número de Tópicos Utilizados	50 (401-450)	50
Média de Termos por Tópico (Disjuntivo)	10,80	1,94
Média de Termos por Tópico (Conjuntivo)	4,38	1,94
Número médio de Documentos Relevantes por Tópico	94,56	35.40
Tamanho	2 GB	16 GB

[4] http://www.todobr.com.br

5.2 O PROJETO DOS EXPERIMENTOS

Nossa principal hipótese é que o uso de contextos de busca (i.e. histórico de consultas e informações de *clickthrough*) podem melhorar a acurácia dos mecanismos de busca. Em particular, o contexto é capaz de fornecer mais informações a respeito do real interesse do usuário do mecanismo de buscas. Acreditamos também que a mineração de regras de associação seja a forma mais adequada para se extrair essa informação extra dos contextos de busca.

Sendo assim a maior parte de nossos experimentos busca comparar o desempenho de uma máquina de busca implementada utilizando apenas a consulta corrente com o desempenho da mesma máquina de busca otimizada para considerar também as informações históricas de consultas e *clickthrough*.

A coleção WT2g foi indexada e foi implementado um mecanismo de busca e uma interface web para consulta a seus documentos. O índice fornecido juntamente com a coleção WBR99 foi utilizado para implementar o mecanismo de busca a seus documentos.

Os mecanismos foram implementados e disponibilizados a um grupo de 60 voluntários que, por sua vez, realizaram tarefas de busca tendo como objetivo conseguir informações a respeito de um dos tópicos pré-determinados. O grupo tinha uma composição bastante heterogênea na *expertise* de uso de mecanismos de busca contemplando desde pessoas acostumadas a realizar apenas consultas simples pela interface padrão do buscador quanto pessoas com conhecimento de ferramentas de busca avançada ou mesmo da estrutura de funcionamento dos mecanismos de busca. Em todos os casos foram registrados o histórico de consultas bem como o *clickthrough*.

No mecanismo da WT2g, para cada usuário foram definidos 10 tópicos e dadas as descrições dos tópicos definidas pela TREC. Para cada tópico, a primeira consulta é o título do tópico dado na descrição original da TREC. Já na WBR99 cada usuário recebeu 50 tópicos, com uma breve descrição, que foram utilizados na submissão da primeira consulta.

Nos dois casos, uma vez que o usuário submete a consulta o mecanismo de busca retorna uma lista ordenada com os resultados da busca do usuário. O usuário deve examinar os resultados e pode clicar em um ou mais documentos retornados pelo mecanismo para examinar o documento completo.

O usuário pode também alterar sua consulta para uma nova consulta. Em nosso experimento, apenas as primeiras 4 alterações de consulta para cada tópico são utilizadas, uma vez que a primeira consulta é o próprio nome do tema e produz um resultado idêntico nas duas abordagens devido à ausência de histórico no início da sessão de buscas. O usuário deve selecionar o tópico a partir de um menu de seleção antes de submeter a primeira consulta para que seja possível simular o início e término de uma sessão de buscas.

Na figura 5.1 uma tela da máquina de buscas desenvolvida para os testes mostra em cima, à esquerda a caixa de seleção de temas. Em destaque abaixo, da barra divisora, o tema das buscas e sua descrição. As respostas aparecem de maneira semelhante à das máquinas de busca mais comuns na *Web*.

As interações dos usuários com o sistema são armazenadas em um banco de dados relacional, incluindo as consultas submetidas e os documentos clicados. Para cada consulta, nós gravamos os termos da consulta e as páginas de resultado associadas a ela. Para cada documento clicado, nós gravamos o título e o *snippet* mostrado na página de resultados. O *snippet* do artigo é gerado em função da consulta e é gerado em tempo de execução de acordo com a consulta.

Os testes com a coleção TREC foram abortados por apresentaram um fator limitante em função da língua dos documentos e temas, o inglês. Como vários voluntários apresentaram dificuldade na interpretação dos textos presentes na coleção, muitas vezes a escolha de documentos retornados e os termos utilizados na consulta e com isso ela era feita mais pela capacidade de interpretação de textos em inglês do que pela análise do conteúdo em si. Sendo assim optamos por analisar apenas os resultados oriundos dos experimentos com os documentos e temas apresentados na WBR99, na sua maioria em português.

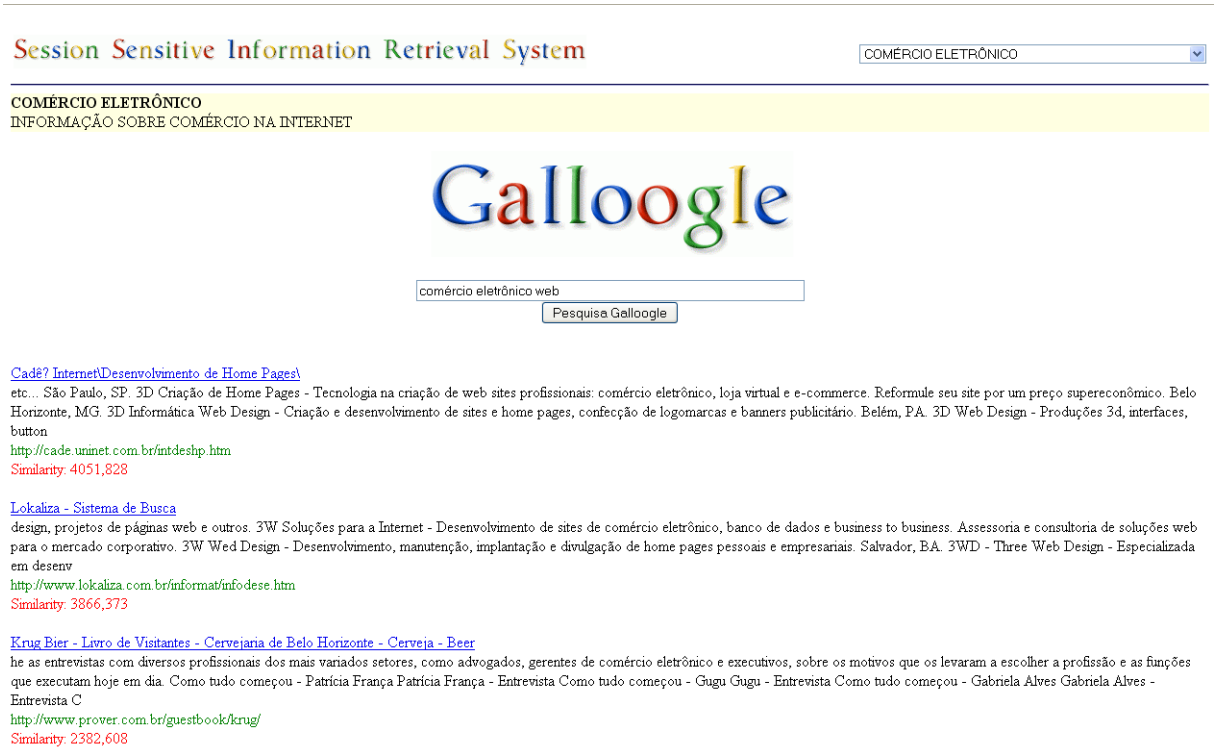

Figura 5.1 - Tela da Máquina de Busca de Testes

Os testes foram feitos por 59 pessoas que realizaram um total de 246 sessões de busca. Aproximadamente 46% das sessões de busca dos usuários tiveram entre 1 e 4 refinamentos dos termos de busca enquanto 13% consideraram a tarefa concluída logo na primeira consulta submetida.

Cerca de 11.000 documentos foram retornados tanto pelo nosso modelo quanto pelo modelo utilizado como *baseline*. Esse volume de dados nos permitiu uma análise estatística com resultados relevantes. É interessante ressaltar que os voluntários possuiam perfil profissionais, faixa etária e formação acadêmica heterogêneos.

A tabela 5.2 caracteriza essas sessões de busca quanto ao número de consultas realizadas pelo usuário e ao número médio de termos usados nas buscas realizadas. Ao ler essa tabela deve-se lembrar que a primeira consulta do experimento é gerada automaticamente pelo sistema utilizando os termos da definição do tema a ser pesquisado, logo o número médio de termos dessa consulta não é função de uma interação com usuário propriamente dita mas, principalmente, do título do tema escolhido por ele.

Através do exposto nessa tabela pode-se observar que a maior parte das sessões de busca tende a ser concluída até o quarto refinamento, ou seja, o usuário consegue modelar o seu objeto de pesquisa adequadamente até esse refinamento e, com isso, gerar resultados satisfatórios.

Outra caracterização interessante dos testes é o número médio de termos usados nas consultas em função da ordem das mesmas na sessão de buscas, mostrada na Tabela 5.3. Também, nesse caso, é necessário lembrar que a primeira consulta é gerada em função do tema escolhido.

Tabela 5.2 - Características das Sessões de Busca dos Experimentos

NÚMERO DE CONSULTAS	QUANTIDADE DE SESSÕES	NÚMERO MÉDIO DE TERMOS	VARIÂNCIA
1 Consulta	46	2,37	2,40
2 Consultas	51	1,68	0,43
3 Consultas	46	2,02	1,24
4 Consultas	35	1,81	0,45
5 Consultas	24	1,83	1,27
6 Consultas	11	1,91	4,59
Mais de 6 Consultas	129	2,46	1,88

Tabela 5.3 - Características das Consultas dos Experimentos

ORDEM DA CONSULTA	NÚMERO MÉDIO DE TERMOS	VARIÂNCIA
1ª Consulta	1,88	1,34
2ª Consulta	1,92	1,52
3ª Consulta	2,05	1,33
4ª Consulta	2,08	2,48
5ª Consulta	1,77	0,97
6ª Consulta	2,15	3,22
Acima da 6ª Consulta	2,39	3,51

Na tabela 5.3 ainda é possível observar uma tendência ao aumento do número de termos da consulta à medida que se avança nos refinamentos, o que pode indicar uma tentativa de aumentar o detalhamento do conceito buscado.

5.3 AVALIAÇÃO DOS RESULTADOS

A comparação entre os resultados da abordagem tradicional com a proposta deste trabalho foi feita inicialmente por meio de duas métricas de desempenho:

a) Média das Precisões Médias (*Mean Average Precision* – MAP): essa é a média padrão não interpolada e serve de boa medida da acurácia geral do *ranking*;
b) Precisão em 3, 10 e 50 documentos (PR@3docs, per@10docs e pr@50docs): apesar dessa medida não valer como média, ela é mais significativa do que a MAP pois reflete a utilidade para usuários que lêem apenas os primeiros documentos.

Posteriormente os resultados foram comparados, considerando-se também a ordem da consulta submetida, uma vez que se esperava um aumento cada vez maior na precisão do algoritmo à medida que novas características do conceito buscado fossem adicionados a cada busca. Para isso usou-se novamente a precisão média em 3, 10 e 50 documentos para cada um dos quatro refinamentos realizados.

5.3.1 Média das Precisões Médias

O gráfico 5.1 mostra o ganho ou perda percentual da média de documentos considerados relevantes em cada das 50 primeiras posições de resposta. No gráfico, o eixo das abscissas representa o desempenho do modelo de espaços vetoriais puro.

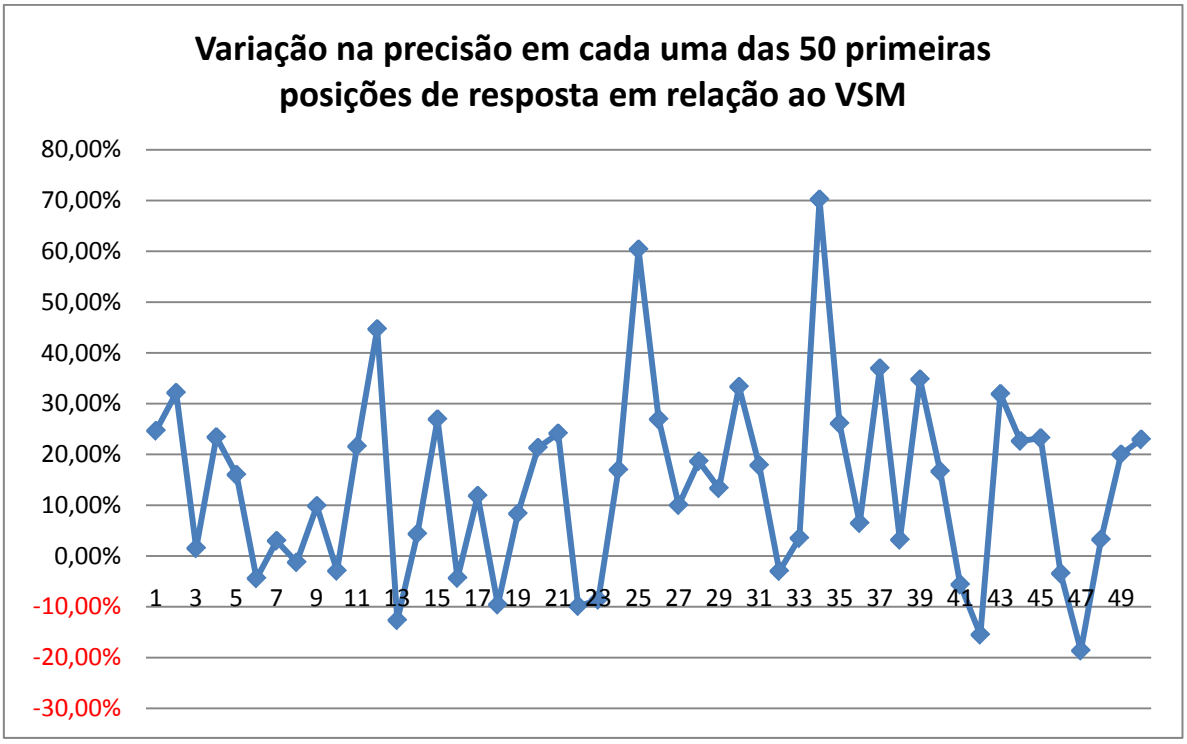

Gráfico 5.1 - Variação na precisão em cada uma das 50 primeiras posições de resposta

Como pode ser observado, há posições nas quais o desempenho médio da abordagem proposta é inferior ao do modelo clássico. Acreditamos que isso se deva ao fato dos documentos não relevantes ultrapassados por documentos relevantes com um *ranking* original um pouco inferior ainda continuarem mais bem ranqueados do que documentos relevantes cujo *ranking* original era muito inferior.

Por exemplo, supondo que os documentos que originalmente ocupavam as 5 primeiras posições não eram relevantes e passam a ocupar as posições de 6 a 10 tendo sido ultrapassados por outros 5 documentos cujo *ranking* foi aumentado pela nossa estratégia. O *ranking* desses documentos, determinado pelo modelo de espaços vetoriais, ainda é alto e por isso, mesmo com as otimizações, outros documentos que originalmente possuíam um *ranking* muito baixo não conseguem ultrapassá-los. A tendência é que esses documentos se concentrem em determinados pontos da lista de respostas e que com isso façam com que a precisão nesses pontos seja até mesmo pior do que o *baseline*.

De qualquer maneira, o ganho médio das precisões médias dos 50 primeiros documentos foi de 13,86% quando comparado à precisão do modelo de espaços vetoriais clássico.

5.3.2 Precisão em 3, 10 e 50 documentos (PR@3docs, per@10docs e pr@50docs)

O gráfico 5.2 mostra o ganho médio acumulado em cada uma das 50 primeiras posições de resposta quando comparado com a abordagem do Modelo de Espaços Vetorais, VSM, puro. Nele é possível observar o mesmo fenômeno do gráfico 5.1, o ganho médio em precisão nos primeiros documentos mostra-se muito maior e sofre uma queda considerável ainda nos 10 primeiros documentos, o que corresponde à primeira página de respostas padrão dos mecanismos de busca. É interessante observar, no entanto uma recuperação tendendo à estabilização na segunda metade do gráfico. Isso dá ainda mais força à suposição feita na análise do gráfico 5.1 pois indica que documentos não relevantes, muito bem ranqueados originalmente, se mantêm entre os documentos retornados; enquanto os documentos não relevantes que obtinham *rankings* apenas um pouco melhor do que outros considerados relevantes foram excluídos da lista dos 50 documentos melhor ranqueados.

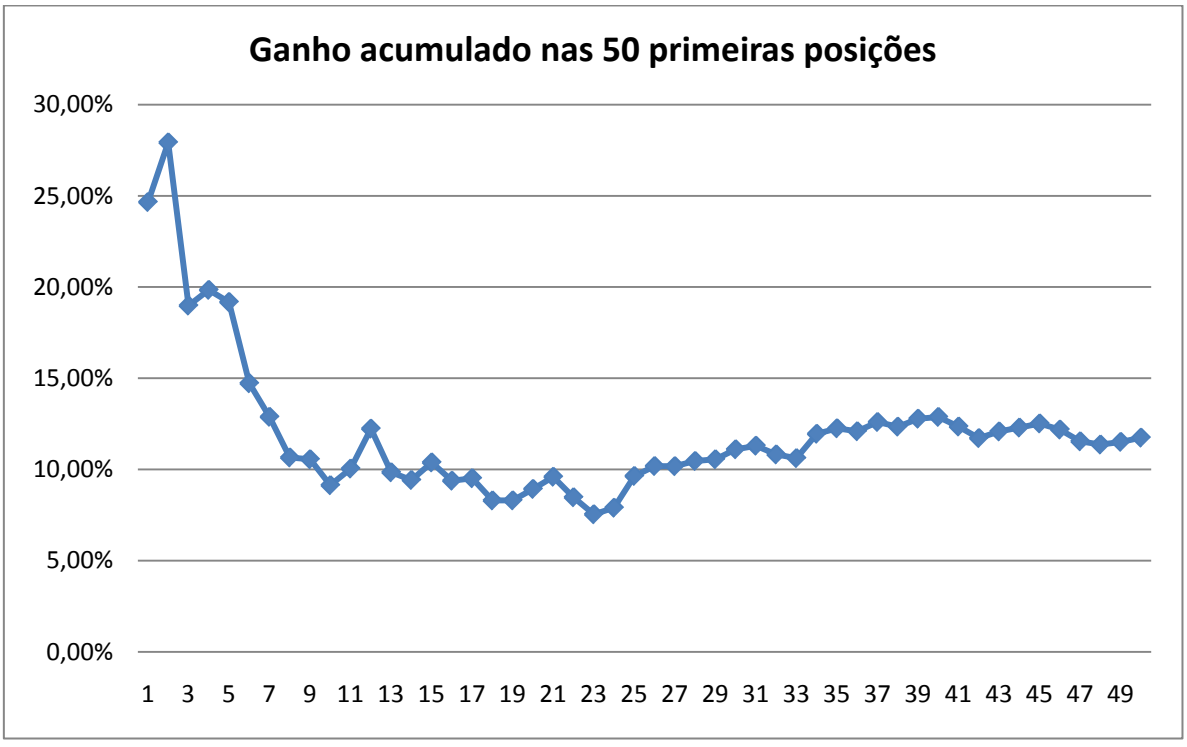

Gráfico 5.2 - Ganho acumulado nas 50 primeiras posições

Ao se considerar as métricas propostas tem-se os seguintes ganhos:

a) Pr@3Docs: 18,98%
b) Pr@10Docs: 9,13%
c) Pr@50Docs: 11,74%

Pelo que percebemos essa variação se dá pela reorganização dos documentos. Nas 3 primeiras posições, documentos relevantes que tinham um *ranking* ligeiramente mais baixo do que os mais bem posicionados conseguem ultrapassá-los. Daí o expressivo aumento de 18,98% nos 3 primeiros documentos.

Esses documentos que foram ultrapassados, no entanto, ainda apresentam um *ranking* elevado o suficiente para que não sejam ultrapassados por outros documentos relevantes que tiveram seus *rankings* aumentados pela nossa abordagem, daí a redução de precião apresentada ao se considerar os 10 primeiros documentos.

O efeito das nossas técnicas de otimização tendem a fazer mais efeito à medida que o *ranking* do documento aumenta pois a razão de sua participação no valor final do *ranking* do documento passa a ser mais relevante. Isso explica a recuperação apresentada na precisão em 50 documentos quando comparada com a precisão em 10 documentos.

Em suma, em todos os casos o desempenho da abordagem proposta mostrou-se superior, especialmente nos 3 primeiros resultados, notoriamente os mais considerados pelos usuários.

5.3.3 Precisão média em 3, 10 e 50 documentos em cada refinamento realizado

O gráfico 5.3 mostra o ganho/perda de precisão média até o sétimo refinamento submetido pelo usuário. Nele é possível observar um padrão semelhante nos ganhos médios até a sexta consulta submetida, o quinto refinamento. Ao contrário do que havia sido previamente esperado o ganho não aumentou à medida que os refinamentos foram submetidos, sendo até mesmo piores do que a abordagem clássica a partir do quarto refinamento. Mesmo com esse resultado negativo nossa abordagem mostra-se superior à clássica nos 3 primeiros refinamentos, o que contempla a maior parte das sessões de busca de acordo com a tabela 4 contempla a maioria das sessões de busca.

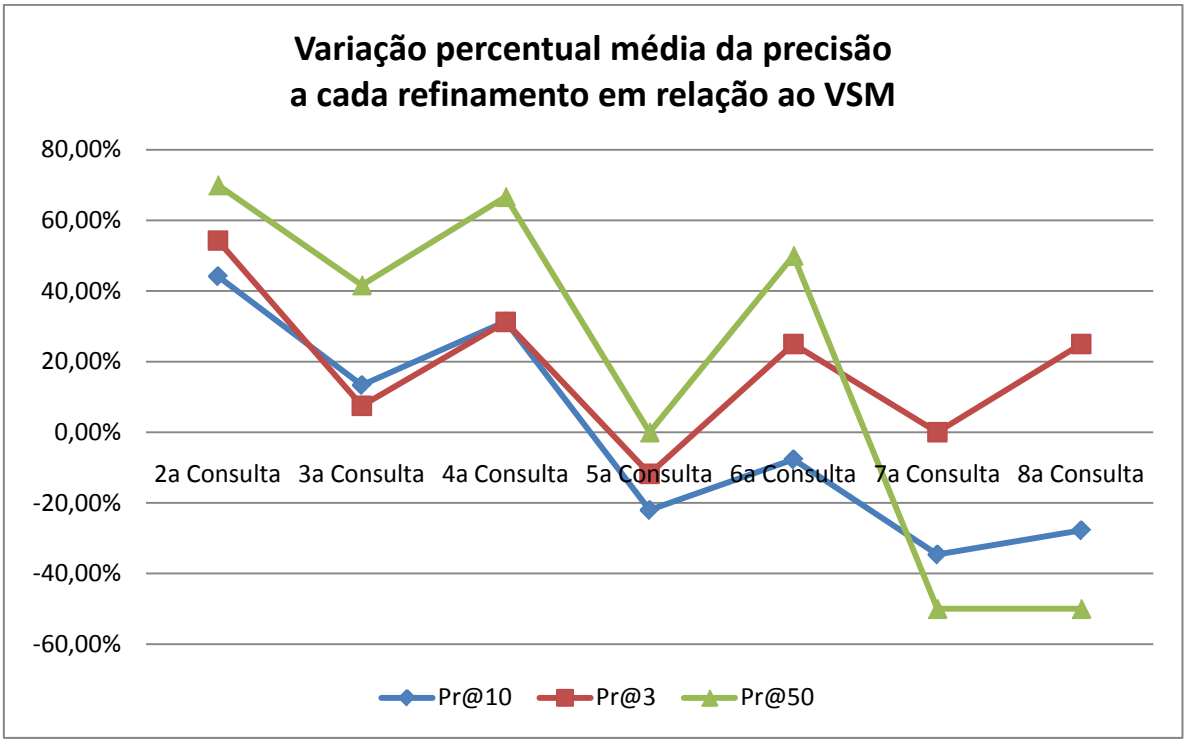

Gráfico 5.3 – Variação percentual média da precisão a cada refinamento em relação ao VSM

Acreditamos que a queda na precisão da nossa abordagem se dê pelo fato de o pelo uso acumulado dos termos extra na tentativa de se definir o contexto de busca na sessão acabe por se tornar um ruído nos dados utilizados na entrada do algoritmo de buscas. Esse resultado pode indicar que seja necessário rever o fator de peso dos termos nas consultas subsequentes na sessão de buscas.

5.4 ANÁLISE GERAL

De maneira geral o modelo proposto apresentou resultados superiores aos conseguidos através do modelo clássico, especialmente nas primeiras posições.

Documentos não relevantes continuam a aparecer no conjunto resposta devido ao elevado *ranking* obtido por eles no modelo no qual o nosso modelo se baseia. Isso que faz com que as otimizações impostas não sejam suficientes para elevar o ranking dos documentos relevantes com *rankings* muito baixos no modelo original.

À medida que o impacto percentual das otimizações no *ranking* se torna mais elevado, os documentos não relevantes tendem a perder ainda mais espaço para os relevantes.

Sessões muito longas podem levar nossa abordagem a piorar o desempenho do algoritmo de Recuperação da Informação devido ao excesso de informações de geradas por ela. De qualquer maneira, sessões com mais de 5 consultas não são muito usuais.

6 CONCLUSÕES E TRABALHOS FUTUROS

A crescente quantidade de informação disponível na *web* faz das máquinas de busca um dos principais temas de pesquisa atuais em Ciência da Computação. As características heterogêneas dos usuários, bem como as restrições ao treinamento deles, fazem com que as tentativas de otimização desses mecanismos sejam ainda mais desafiadoras.

Além dos modelos de recuperação de informação, apresentados no trabalho em suas abordagens clássicas, vários recursos podem ser implementados na busca pela melhoria da qualidade dos resultados retornados ao se submeter uma consulta em uma máquina de buscas. O uso da estrutura dos documentos, meta tags, a análise de *links* e a análise temporal são algumas das abordagens mais comuns para esse fim.

Nesse trabalho foi proposta uma nova abordagem para a otimização da classificação dos documentos recuperados por um dos algoritmos clássicos de recuperação de informação ou dele derivados. Essa abordagem leva em consideração o padrão comportamental do usuário de refinamento sucessivo de consultas ao utilizar uma máquina de buscas. Foram consideradas as correlações entre termos utilizados nas consultas e presentes nos documentos selecionados na sessão de buscas bem como os termos em si. Foram estabelecidos experimentalmente fatores decrescentes de influência na sessão de buscas.

Os resultados experimentais comparados com uma implementação VSM pura apontam para a eficiência da abordagem proposta com ganhos representativos, que chegaram próximo a 20% nas 3 primeiras posições do *ranking*, notoriamente as mais consideradas pelos usuários. O fato de as tarefas computacionalmente mais custosas serem realizadas *offline* permite que a abordagem seja implementada sem que haja um grande impacto no tempo de execução das tarefas de busca.

Na forma como foi proposta, nossa abordagem permite que sua implementação seja associada a praticamente todas as abordagens existentes com boas chances de potencializar seus ganhos.

Muito trabalho ainda deve ser realizado a fim de se definir as possibilidades de ganho de precisão e, especialmente, na análise da revocação conseguida pela nossa abordagem.

Como as implementações aqui realizadas baseiam-se no modelo de espaços vetoriais sem quaisquer otimizações, seria interessante implementar também outras estratégias

de otimização tais como a análise estrutural dos documentos e análise de elos a fim de que fossem analisados seus impactos individualmente e combinados.

A mineração de padrões frequentes nos *snippets* dos documentos selecionados pelo usuário pareceu ser uma boa fonte de informações a respeito do interesse do mesmo. O problema, nesse caso, é o fato de os *snippets* serem criados em tempo de execução e a complexidade computacional da tarefa de mineração mostrou-se proibitiva. A busca de alternativas para a realização dessa tarefa durante a sessão de buscas pode levar a ganhos consideráveis na precisão da recuperação da informação desejada.

Outro grande desafio é a análise da revocação semântica de um determinado termo na coleção, por exemplo, uma pessoa que busca por “CRISTO” consideraria relevantes conjuntos bem distintos de documentos se seu interesse fosse turístico ou religioso. As coleções conhecidas apresentam apenas a revocação considerando os termos como entidades unívocas o que é, claramente, uma simplificação grosseira.

Por fim, mas talvez a primeira extensão do trabalho a ser feita, é um estudo maior do peso a ser definido nos termos inseridos pelo sistema às chaves de busca uma vez que, experimentalmente, observou-se que ao contrário do esperado, o seu impacto ao longo da sessão de buscas chega até mesmo fazer com que o algoritmo tenha um desempenho inferior ao conseguido pelo modelo clássico de referência.

7 BIBLIOGRAFIA

1. *Challenges in information retrieval and language modeling.* **Allan, J. e al.** 2003. SIGIR. pp. 37(1):31-47.

2. **Rocchio, J.** Relevance feedback information retrieval. *The Smart Retrieval Systems Experiments in Automatic Document Processing.* 1971, pp. 313-323.

3. **Baeza-Yates, Ricardo e Ribeiro-Neto, Berthier.** *Modern Information Retrieval.* s.l. : Addison Wesley, 1999. 978-0201398298 .

4. **Salton, G. e McGill, M. J.** *Introduction to Modern Informatino Retrieval.* New York, NY : McGraw-Hill, 1983.

5. **Robertson, S. E. e Jones, K. Sparck.** Relevance weighting of search terms. *Journal of the American Society for Information Sciences.* 27, 1976, Vol. 3, pp. 129-146.

6. **Ponte, J. e Croft, W. B.** A Language Modeling Approach to Information Retrieval. *ACM SIGIR Conf. on Research and Development in Information Retrieval.* 1998, pp. 275-281.

7. *Computer evaluation of indexing and text processing.* **Salton, G. e Lesk, M. E.** 1, 1968, Journal of the ACM, Vol. 15, pp. 8-63.

8. **Salton, G. e Yang, C. S.** On the specification of term values in automatic indexing. *Journal of Documentation.* 1973, 29, pp. 351-372.

9. **Yu, C. T. e Salton, G.** Precision weighting - an effective automatic indexing method. *Journal of the ACM.* 23, 1976, Vol. 1, pp. 76-88.

10. **Sparck, J. K.** A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation.* 28, 1972, Vol. 1, pp. 59-75.

11. *Generalized vector spaces model in information retrieval.* **Wong, S. K. M, Ziarko, Wojciech e Wong, Patrick C. N.** Montreal, Quebec, Canada : ACM - New York, NY, USA, 1985. Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 18-25. ISBN: 0-89791-159-8.

12. **Wong, S. K. M., et al.** On modeling of information retrieval concepts in vector spaces. *ACM Transactions on Database Systems (TODS).* June de 1987, Vol. 12, 2, pp. 299-321.

13. *Set oriented Retrieval.* **Bookstein, A.** Grenoble, France : s.n., 1988. ACM-SIGIR Conference on Research and Development in Information Retrieval. pp. 583-596.

14. **Pôssas, B., et al.** Set-based vector model: An efficiente approach for correlation-based ranking. *ACM Trans. Inf. Syst.* 23, 2005, Vol. 4, pp. 397-429.

15. **Tan, Pang-Ning, Steinbach, Michael e Kumar, Vipin.** *Introduction to Data Mining.* Boston, MA, USA : Addison-Wesley Longman Publishing Co. Inc., 2005. 0-321-45052-7/978032142052790000.

16. *Mining Association Rules between Sets of Items in Large Databases.* **Agrawal, R., Imielinski, T. e Swami, A. N.** ACM SIGMOD : s.n., 1993. pp. 207-216.

17. **Tan, Pang-Ning, Steinbach, Michael e Kumar, Vipin.** *Introduction do Data Mining.* Boston, MA - USA : Pearson Addison Wesley, 2006. 0-321-42052-7.

18. **Alpert, Jesse e Hajaj, Nissan.** We knew the web was big... *Google Blog.* [Online] 25 de julho de 2008. [Citado em: 30 de stembro de 2008.] http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.

19. **Liu, Bing.** *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.* New York : Springer Berlin Heidelberg, 2007. 3-540-37881-2/978-3-540-37881-5.

20. **Brin, S. e Page, L.** The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks.* 30, 1998, Vols. 1-7, pp. 107-117.

21. **Arlitt, Martim.** Characterizing Web user sessions. *ACM SIGMETRICS Performance Evaluation Review.* 2, 200, Vol. 28, Setembro 2000, pp. 50-63.

22. **Dupont, Cedric.** SearchWiki: make search you own. *The Official Google Blog.* [Online] 20 de 11 de 2008. [Citado em: 10 de 01 de 2009.] http://googleblog.blogspot.com/2008/11/searchwiki-make-search-your-own.html.
23. *Context-sensitive information retrieval using implicit feedback.* **Shen, X., Tan, B. e Zhai, C. X.** New York, NY, USA : ACM, 2005. ACM SIGIR. pp. 43-50.
24. *Optimizing search engines using clickthrough data.* **Joachims, T.** 133-142 : ACM, 2002. ACM SIGKDD. pp. 133-142.
25. *Adaptative web search based on user profile constructed without any effor from the user.* **Sugiyama, K., Hatano, K. e M., Yoshikawa.** New York, NY, USA : ACM, 2004. WWW. pp. 675-684.
26. *Implicit feedback for inferring user preference: a bibliography.* **Kelly, D. e Teevan, J.** 2003. SIGIR. pp. 37(2)18-28.
27. *Query session based term suggestion for interactive web search.* **Chien, L., Huang, C. e Oyang, Y.** 2001. WWW 2001.
28. *Using terminological feedback for web search refinement: a log based study.* **Anick, P.** New York, NY, USA : ACM, 2003. ACM SIGIR. pp. 88-95.
29. *Fast algorithms for mining association rules.* **Agrawal, R. e Srikant, R.** 1994.
30. *Generating non-redundant association rules.* **Zaki, M. J.** Boston, MA, USA : ACM Press, 2000. 6th ACM SIGKDD International Conference In Data Mining. pp. 33-34.
31. *An experimental comparison of click position-bias models.* **Craswell, Nick, et al.** Palo Alto, California, USA : ACM, New York, NY, USA, 2008. Proceedings of the international conference on Web search and web data mining. pp. 87-94. ISBN:978-1-59593-927-9.
32. *Overview of the Eighth Text Retrieval Conference (TREC 8).* **VOORHEES, E. AND HARMAN, D.** Gaithersburg, MD : s.n., 1999. Proceedings of the 8th Text REtrieval Conference (TREC-8). pp. 1-23.
33. *A Contextualization Method of Browsing Events in Web-based Learning Content Management System for Personalized Learning.* **Wang, Feng-Hsu.** s.l. : IEEE Explore, 2007. IEEE International Conference on Advanced Learning Technologies. pp. 43-45. 10.1109/ICALT.2007.9.
34. *Evaluating implicit measures to improve web search.* **Fox, S., Karnawat, K., Mydland, M., Dumais, S., and White, T.** New York, NY, USA : ACM, 2005. ACM Transactions on Information Systems (TOIS). Vol. 3, pp. 147-168. ISSN:1046-8188.
35. *An experimental comparison of click position-bias models.* **Craswell, Nick, et al.** Palo Alto, California, USA : ACM, 2008. WSDM '08: Proceedings of the international conference on Web search and web data mining. pp. 87-94. 978-1-59593-927-9.
36. *Beyond PageRank: machine learning for static ranking.* **Richardson, Matthew and Prakash, Amit and Brill, Eric.** New York, NY, USA : ACM, 2006. WWW '06: Proceedings of the 15th international conference on World Wide Web. pp. 707-715. 1595933239.
37. *Comparing explicit and implicit feedback techniques for web retrieval.* **White, R.W., Jose, J. M. e Ruthven, I.** s.l. : NIST, 2002. TREC-10 interactive track report.
38. **C.T. Diop, A. Giacometti, D. Laurent, N. Spyratos.** Composition of Mining Contexts for Efficient Extraction of Association Rules. 2001.
39. **Theodorakis, Manos.** Contextualization: An Abstraction Mechanism for Information Modeling. *PhD thesis.* Crete, Greece : Department of Computer Science, University of Crete, 1999.
40. **Theodorakis, Manos, Spyratos, Nicolas e Constantopoulos, Panos.** Contextualization as an Independent Abstraction Mechanism for Conceptual Modeling. *Information Systems.* 2007, Vol. 32, pp. 24-60.
41. **Lin, Patrick Pantel and Dekang.** Discovering Word Senses from Text. *In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 2002, pp. 613-619.

42. *Fast generation of result snippets in web search.* **Turpin, , Andrew and Tsegay,, Yohannes and Hawking,, David and Williams,, Hugh E.** New York, NY, USA : ACM, 2007. SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 127-134. 978-1-59593-597-7 .
43. *Implicit user modeling for personalized search.* **Shen, Xuehua, Tan, Bin e Zhai, ChengXiang.** Bremen, Germany : ACM, 2005. CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management. pp. 824-831. 1-59593-140-6.
44. **Pan, Bing and Hembrooke, Helene and Joachims, Thorsten and Lorigo, Lori and Gay, Geri and Granka, Laura.** In Google We Trust: Users' Decisions on Rank, Position, and Relevance. *Journal of Computer-Mediated Communication.* February de 2007, Vol. 12.
45. **Lee, Yeong-Chyi, Hong, Tzung-Pei e Lin, Wen-Yang.** Mining association rules with multiple minimum supports using maximum constraints. *International Journal of Approximate Reasoning.* 1-2, July de 2005, Vol. 40, pp. 44-54.
46. **Ma, Bing Liu and Wynne Hsu and Yiming.** Mining Association Rules with Multiple Minimum Supports. *In Knowledge Discovery and Data Mining.* 337--341, 1999.
47. **Ng, Wilfred, Deng, Lin e Lee, Dik Lun.** Mining User preference using Spy voting for search engine personalization. *ACM Transactions on Internet Technology (TOIT).* October de 2007, Vol. 7, 14.
48. *Querying Contextualized Information Bases.* **Spyratos, Manos Theodorakis and Anastasia Analyti and Panos Constantopoulos and Nicolas.** 1999.
49. *Ranking Definitions with Supervised Learning Methods.* Chiba, Japan : ACM, 2005. Special interest tracks and posters of the 14th international conference on World Wide Web. pp. 811-819. 1-59593-051-5 .
50. *Text data mining: discovery of important keywords in the cyberspace.* **Arimura, H. Abe, J. Fujino, R. Sakamoto, H. Shimozono, S. Arikawa, S.** Kyoto, Japan : IEEE, 2000. Digital Libraries: Research and Practice, 2000 Kyoto, International Conference on. pp. 220-226. 0-7695-1022-1.
51. **Tu, Xin Li and Dan Roth and Yuancheng.** PhraseNet: Towards Context Sensitive Lexical Semantics. 2003.
52. *Overview of TREC-8 Web track.* **David Hawking, Ellen Voorhees, Nick Craswell, Peter Bailey.** Gaithersburg, Maryland USA : s.n., 1999. Proceedings of TREC-8. pp. 131-150.
53. *UCAIR: Capturing and Exploiting Context for Personalized Search.* **Xuehua Shen, Bin Tan, ChengXiang Zhai.** 2005.
54. *Why we search: visualizing and predicting user behavior.* **Adar, , Eytan and Weld,, Daniel S. and Bershad,, Brian N. and Gribble,, Steven S.** Banff, Alberta, Canada : ACM, 2007. Proceedings of the 16th international conference on World Wide Web. pp. 161-170. 978-1-59593-654-7 .
55. **Pôssas, B.** Um Novo Modelo de Ordenação de Documentos Baseado em Correlação entre Termos. *Tese de Doutorado.* Belo Horizonte : UFMG, 2005.
56. **Witten, I. e Frank, E.** *Data Mining: Pratical Machine Learning Tools and Techiniques.* San Francisco, CA, USA : Elsevier, 2005. 81-312-0050-7.
57. **Abiteboul, S, Buneman, P. e Suciu, Dan.** *Data on the Web: From Relations to Semistructured Data and XML.* San Francisco, CA - USA : Morgan Kauffman Publishers, 2000. 1-55860-622-X.
58. **Witten, I. H., Moffat, Alistair e Bell, Timothy C.** *Managing Gigabytes: Compressing and Indexing Documents and Images.* San Diego, CA, USA : Morgan Kaufmann Publishers, 1999. 978-1-55860-570-1.
59. **Jain, Raj.** *The Art of Computer System Performance Analysis.* New York, NY, USA : John Wiley and Sons, 1992. 0-471-50336-3.

60. **Barry, Wilkinson e Michael, Allen.** *Parallel Programming: Techniques and Applications Using Networkek Workstations and Parallel Computers.* Upper Saddle River, NJ, USA : Pearson Prentice Hall, 2005. 0-13-140563-2.
61. **Quinn, Michael J.** *Parallel Programming in C with MPI and OpenMP.* New Delhi, India : Tata McGraw-Hill, 2003. 978-0-07-058201-9.