UMA ANÁLISE DE FERRAMENTAS PARA MINERAÇÃO DE CONTEÚDO DE PÁGINAS WEB

Lúcia Helena de Magalhães

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA CIVIL.

Aprovada por:

_____

Prof. Nelson Francisco Favilla Ebecken, D.Sc.

_____

Profª Myrian Christina de Aragão Costa, D.Sc.

_____

Prof. Alexandre Gonçalves Evsukoff, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

JUNHO DE 2008

MAGALHÃES, LÚCIA HELENA

Uma análise de ferramentas para mineração de conteúdo de páginas Web [Rio de Janeiro] 2008.

XI, 76 p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia Civil, 2008)

Dissertação - Universidade Federal do Rio de Janeiro, COPPE.

1. Mineração Web, Recuperação da Informação; Extração da Informação

I . COPPE/UFRJ II. Título (série).

Dedico este trabalho aos meus pais
José Praxedes e Hilda, a meus irmãos e ao
meu esposo Helder que sempre me
apoiaram e incentivaram.

AGRADECIMENTOS

Dedico esta dissertação a todas as pessoas que contribuíram direta ou indiretamente para a realização do presente trabalho.

Agradeço a Deus por tudo que tenho e por ter me suprido a ótima saúde e coragem de que tanto precisei nesta gloriosa etapa da minha vida.

Aos meus pais, por me darem sempre total apoio em todos os momentos de minha vida.

À toda minha família, pelo amor e dedicação, em especial a minha querida irmã Teresinha que sempre esteve ao meu lado nos momentos mais difíceis.

Ao meu esposo Helder pela compreensão e carinho.

A todos docentes do programa de pós-graduação em Engenharia Civil.

Ao meu orientador, prof. Nelson Ebecken, pelo estímulo e orientação que muito agregaram em minha formação.

A todos os meus amigos e colegas, que sempre me ajudaram, torceram e acreditaram em mim.

Finalmente, gostaria de agradecer a todos os Funcionários e Professores dos cursos de Mestrado em Computação de Alto Desempenho que me auxiliaram nesta conquista.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

UMA ANÁLISE DE FERRAMENTAS PARA MINERAÇÃO DE CONTEÚDO DE PÁGINAS WEB

Lúcia Helena de Magalhães

Junho/2008

Orientador: Nelson Francisco Favilla Ebecken
Programa: Engenharia Civil

A Mineração de Conteúdo na Web é uma técnica para descobrir e analisar informações úteis da Internet. A proposta deste trabalho é apresentar uma análise de ferramentas para mineração de conteúdo de páginas web e analisar os resultados destas técnicas no processo de recuperação e extração da informação. Os modelos de recuperação de informação apresentam estratégias de pesquisa de documentos relevantes através da busca automática da informação. Assim, será feito uma análise comparativa das páginas devolvidas pelos mecanismos de busca, levando em consideração a precisão, organização e a qualidade dos documentos recuperados. Quanto às ferramentas de extração da informação, serão também analisados os resultados relacionados à automatização do processo, funcionalidades para exportação dos dados extraídos e a qualidade do extrato.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

AN ANALYSIS OF TOOLS FOR MINING OF CONTENT FROM WEB PAGES

Lúcia Helena de Magalhães

June/2008

Advisor: Nelson Francisco Favilla Ebecken
Department: Civil Engineering

Web Content Mining is a technique to discover and analyze useful information from the Internet. The proposal of this work is to examine some of the principle tools for web content mining and to analyze the results of these techniques in the recovery process and information extraction. All these information recovery models implement different strategies. In this way, it will be made a comparative analysis of the returned results taking in consideration the precision, organization and the quality of the recovered documents. The extraction information tools will also be analyzed through the results linked to the automation process, functionalities for export the extracted data and the quality of this extract.

SUMÁRIO

INDICE DE FIGURAS

INDICE DE TABELAS

1 INTRODUÇÃO

Com o aumento de popularidade da WEB, um grande volume de dados e informação foi gerado e publicado em inúmeras páginas web, nas quais se encontram dados das mais diversas áreas do conhecimento humano.

Este volume de informações disponível na Internet e as taxas diárias de crescimento tornam cada vez mais necessários mecanismos eficientes e eficazes para extração e mineração de conhecimentos úteis da web, pois a busca de informação relevante e necessária em tempo útil torna-se cada vez mais uma necessidade crítica.

> Cada vez é mais difícil encontrar informação relevante na crescente "selva" de informação não estruturada, disponível. Quando, no final dos anos 60, surge o projeto ARPANET, o embrião daquilo que viria a ser a Internet, ninguém imaginava o tão rápido crescimento, disseminação e divulgação desta rede de computadores. (CORDEIRO, 2003, p.10)

A web tem muitas características, as quais serão apresentadas a seguir, que faz da procura de conhecimento e informação relevante na internet uma grande tarefa (LIU, 2007).

- A quantidade de dados na web é grande e continua crescendo muito;
- Dados na Web são de fácil acesso.
- Existem vários tipos de dados na web, tais como, tabelas estruturadas, páginas web semi-estruturadas, textos não estruturados, multimídia, etc.
- Informação na web é heterogênea, devido a diversas autorias de páginas web. Muitas páginas possuem o mesmo ou similar conteúdo usando palavras ou formatos completamente diferentes, isso é um problema para integração da informação de múltiplas páginas.
- A informação na web é conectada. Existem *hiperlinks* entre páginas web de um site e para diferentes sites. Dentro de um site, os *hiperlinks* são usados como mecanismo de organização e em sites diferentes, *hiperlinks* representam implicitamente transferência de autoridade para a página a qual se faz referência. Isto é, páginas que são apontadas

por vários sites são normalmente páginas de alta qualidade porque muitas pessoas confiam nelas.

- Muitas informações na Web são redundantes. Muitas informações ou partes dessas aparecem em muitas páginas ou em sites com poucas alterações. Esta propriedade tem sido muito explorada nas tarefas de mineração de dados na web.
- Somente algumas informações na web são úteis. Informações de páginas da web tais como anúncios, política de privacidade, links de navegação, etc, podem ser removidas para um bom resultado na mineração e análise da informação.
- A web consiste em duas superfícies importantes. Uma visão superficial e uma visão profunda. A Web superficial é composta de paginas que podem ser navegadas usando um *browser* normal e é também pesquisável pelas populares máquinas de busca. Já a web profunda consiste de base de dados que somente podem ser acessados através de uma interface apropriada para a consulta.
- A web é útil. Muitos sites comerciais permitem que pessoas executem operações úteis, como por exemplo: pagar contas, comprar produtos, etc.
- A Web é dinâmica. Informações na web mudam constantemente. Manter o ritmo da mudança e monitorá-las são assuntos importantes para muitas aplicações.
- A Web é uma sociedade virtual. Ela permite interação entre pessoas, organizações e sistemas automatizados. Uma pessoa pode comunicar com outras pessoas em qualquer lugar do mundo facilmente e instantaneamente e, as pessoas podem também expressar sua opinião através de fóruns, blogs e sites de opinião.

Todas essas características da web fazem com que a descoberta de informação e conhecimento na internet seja uma tarefa interessante.

A presente pesquisa visa apresentar os conceitos de *Web Mining*, descrever suas subáreas, que são *Web Structure Mining*, *Web Usage Mining* e *Web Content Mining*, com destaque especial em *Web Mining Content*, que é o objetivo principal desta pesquisa. Além disso, analisar algumas ferramentas para Mineração e extração de conteúdo de páginas Web e discutir algumas máquinas de busca que são o mecanismo mais usado para Recuperação da Informação.

O primeiro capítulo deste estudo trará uma apresentação a respeito de *Web Mining*, salientando suas peculiaridades e seus conceitos básicos, além de apresentar as fases e as categorias do processo de mineração na Web.

O segundo capítulo trata da Recuperação da Informação, que é uma forma de ajudar o usuário a encontrar informação necessária em uma grande coleção de documentos da Web, destacando seus principais modelos.

O terceiro capítulo apresenta um estudo sobre as máquinas de busca, mecanismo usado para ajudar o usuário no processo de recuperação da informação, conceituando *crawling*, *indexing* e *searching,* que são as principais funções dos Motores de Busca. Além disso, apresentar uma análise de alguns sites de busca tais como: *Google*, *Medline* e *Web Crawler, Terrier*, etc.

O penúltimo capítulo comenta o processo de clusterização de páginas web, apresentando os objetivos e o conceito de cluster. Faz também um estudo de alguns mecanismos de pesquisa, tais como: *Snaket*, *Northern Light Search*, *Clusty*, *Kartoo*, *MetaCrawler* e *Grokker*, etc. Estes sites de busca organizam os resultados em categoria, ou seja, em cluster, ajudando assim, os usuários a encontrarem o que estão procurando. A formação de cluster é um tipo de Mineração de Conteúdo que visa organizar os resultados de forma a facilitar a navegação do utilizador.

No último capítulo serão abordadas as técnicas de *Web Content Mining* para extrair e representar os documentos de páginas Web a partir dos seus termos mais relevantes.

Assim, feitas essas abordagens do que vem a ser tratado no presente estudo, será efetuado um aprofundamento de suas questões.

2 WEB MINING

A Web é hoje a maior fonte de informação eletrônica de que dispomos. Todavia, por causa da sua natureza dinâmica, a tarefa de encontrar informações relevantes se torna muitas vezes uma experiência frustrante. Por isso, sente-se uma aspiração para que a Web realmente alcance todo o seu potencial e se torne uma ferramenta mais utilizável, compreensível e eficaz. Nesse contexto, a web aparece como uma possibilidade óbvia a ser explorada.

Quando se enfoca a mineração de informações no ambiente da Internet, utiliza-se a expressão Mineração de Dados na Web ou Web Mining. Segundo Cook e Holder (2000), Mineração de Dados na Web pode ser definida como a descoberta e análise de informação útil originada da internet.

De acordo com Kosala e Blockeel (2000) Web Mining é o uso de técnicas de Mineração de Dados para descobrir e extrair automaticamente informações a partir de documentos e serviços da Web.

Para Liu (2007), Web Mining visa descobrir conhecimento e informação útil de conteúdo de páginas, estruturas de hiperlinks e dados de usuário.

"Embora Web Mining use muitas técnicas de Mineração de Dados, ela não é puramente uma aplicação da tradicional Data Mining devido a heterogeneidade e a natureza dos dados da web que são semi-estruturado ou não estruturados". (LIU, 2007)

> A Mineração Web é uma área de pesquisa que visa integrar as tecnologias Web e a Mineração de Dados, focalizando o desenvolvimento de novas ferramentas e métodos para análise e descoberta de conhecimento de dados na Web. Ela pode ser definida como a descoberta e análise de informações úteis, novas e interessantes da Web, onde, a partir das informações descobertas, seja possível demonstrar características, comportamentos, tendências e padrões de navegação do usuário na Web (COOLEY et al., 1997).

Portanto, Web Mining refere-se ao processo completo de descoberta de informação e conhecimento útil, a partir de dados da Web.

2.1 Categorias da Web Mining

Web Mining, de acordo com Kosala e Blockeel (2000), pode ser dividido em três subáreas: *Web Structure Mining*, *Web Usage Mining* e *Web Content Mining*. A seguir, será apresentada uma breve descrição sobre cada uma dessas categorias, com maior ênfase em *Web Content Mining*, que é o objetivo desta pesquisa.

2.1.1 Web Structure Mining

Web Structure Mining procura descobrir um modelo sobre a estrutura de links da Web. O modelo é baseado na topologia de *hiperlinks*, com ou sem a descrição destes links. Este modelo pode ser usado para categorizar páginas Web e ser útil na geração de informações similares e relacionadas entre diferentes sites.

> A Mineração de estrutura da Web descobre conhecimento útil de hiperlinks, que representa a estrutura do site. Por exemplo, do link, nós podemos descobrir importantes páginas web, que, incidentemente, é a tecnologia fundamental para maquinas de busca. Podemos também descobrir comunidades virtuais que compartilham interesse comum. (LIU, 2007, p. 237)

Uma visão interessante em Web Mining é que a Internet possui mais informações que somente o conteúdo de suas páginas. A referência cruzada entre sites ou páginas de um mesmo site contém em si, um conhecimento implícito a respeito do documento propriamente dito.

Através da interconexão entre documentos, a Rede Mundial de Computadores pode revelar mais informações do que simplesmente as relacionadas ao conteúdo dos documentos. Por exemplo, muitos links apontando para um documento indicam sua popularidade, enquanto muitos links saindo de um documento indicam uma riqueza de tópicos cobertos pelo mesmo. Além disso, descobrindo a estrutura que os links formam,

pode-se estudar como o fluxo de informação afeta o projeto de um site, fornecendo dicas de como melhorá-lo.

O processo que tenta descobrir o modelo que está por trás dessa estrutura de links, ou seja, o processo de inferir conhecimento através da topologia, organização e estrutura de links da Web entre referências de páginas, é chamado de *Web Structure Mining*.

2.1.2 Web Usage Mining

Web Usage Mining refere-se à descoberta de padrões de acesso através da análise de interação do usuário com páginas Web. Os dados de uso da Web incluem basicamente os dados obtidos através dos registros de acesso aos servidores Web. A Mineração do Uso da Web está focada em técnicas que possam descrever e predizer o comportamento do usuário, enquanto esse estiver interagindo com o site.

> *Web Usage Mining* refere-se à descoberta de padrões de acesso de usuário na Web, que registra todo clique feito por cada usuário. Em *Web Usage Mining* se aplica muitos algoritmos de Data Mining para a descoberta de conhecimento. (LIU, 2007, p.7)

Servidores Web armazenam dados em relação ao acesso de suas páginas de forma permanente. Apesar de ser limitada a análise desses registros, eles podem explicar o comportamento de usuários que buscam informações sobre determinado assunto e auxiliar na estruturação de um site. Alguns programas são associados a esses registros para enriquecer a quantidade e o tipo de informação sobre o acesso a determinado site.

Cada servidor Web guarda, localmente, uma coleção de registros bem estruturados: os logs de acesso. Esses armazenam informações sobre a interação dos usuários cada vez que é feito um acesso ao site. *Web Usage Mining* utiliza-se desses dados para descobrir informações sobre os usuários da Web, tais como seus comportamentos e seus interesses.

Como a informação dos logs é bem estruturada, podem ser aplicadas técnicas típicas de Mineração de Dados a estes registros descobrindo assim, conhecimentos úteis.

Web Usage Mining é capaz de descobrir perfil do usuário que pode ser útil na personalização da interface, ou do conteúdo, de forma a ajudar o site a atingir seus objetivos. Sua aplicabilidade no Marketing é de extrema importância, pois saber quem freqüenta um determinado site pode ser de grande valia para este setor. Além disso, descobrindo-se o padrão de acesso dos usuários, pode-se programar um servidor Proxy para efetuar o download das próximas páginas que o usuário provavelmente irá visitar.

A utilização de *Web Mining* é muito variada e vai desde a descoberta de novos conhecimentos nos dados da Web até a melhoria na recuperação da informação. Além disso, existe a possibilidade de se aumentar a eficiência dos sites com a definição do comportamento dos usuários, por meio da análise dos *logs* de acesso.

Em empresas que possuam Intranet, essa análise pode facilitar e aperfeiçoar a infra-estrutura organizacional e formação de grupos de trabalho. Duas tendências para análise desses dados mostram-se claras: a determinação de padrões de acesso geral e a personalização de uso (REZENDE, 2003).

2.1.3 Web Content Mining

O processo de descoberta de informações úteis a partir do conteúdo de páginas Web é chamado de *Web Content Mining*.

A Mineração de Conteúdo descreve a descoberta de informações úteis de conteúdos, dados e documentos da Web, através da busca automática de informação de pesquisas *on-line* (PAL, 2000). Vale salientar que o conteúdo da Web não se constitui apenas de texto ou hipertexto, mas abrange uma ampla variação de tipos de dados, tais como áudio, vídeo, dados simbólicos, metadados[1] e vínculos de hipertexto.

Conforme Cooley (2000), a Mineração de Conteúdo pode ser descrita como sendo a busca automática dos recursos e recuperação das informações disponíveis na rede. Como exemplo desta abordagem, pode-se citar as ferramentas de busca, como Altavista, Yahoo, Google, entre outros.

[1] Metadados são "dados sobre dados". Metadados referem-se a estrutura descritiva da informação sobre outro dado, o qual é usado para ajudar na identificação, descrição, localização e gerenciamento de recursos da web.

Para Liu (2007) o processo de minerar, extrair e integrar dados úteis, informação e conhecimento de conteúdo de páginas web é chamado de *Web Mining Content.*

3 RECUPERAÇÃO DE INFORMAÇÃO

A web contém uma quantia de informação incrível. Encontrar a informação certa é uma tarefa de pesquisa desafiadora.

> Se todos os arquivos na Web fossem claramente marcados com palavras-chave e outros metadados que descrevessem perfeitamente o seu conteúdo e se os usuários fossem treinados em como fazer pesquisas, a recuperação de informação relevante não exigiria algoritmos sofisticados, elas poderiam ser encontradas através de simples consultas. (LINOFF e BERRY, 2001, p.43).

Para Martha (2005) Recuperação de Informação é uma ciência que estuda a criação de algoritmos para recuperar dados provenientes de textos livres, que constituem a maior parte de documentos em forma digital disponível nos dias atuais, sobretudo após a internet.

A recuperação de informação ajuda o usuário a encontrar informação necessária a partir de uma grande coleção de documentos de textos. Em relação à Web, o tópico diz respeito à obtenção de páginas que obedeçam a determinados critérios de alguns usuários.

Segundo enciclopédia livre, a Wikipédia, a Recuperação de Informação (RI) é uma área da computação que lida com o armazenamento de documentos e a recuperação automática de informação associada a eles. É uma ciência de pesquisa sobre busca por informações em documentos, busca pelos documentos propriamente ditos, busca por metadados que descrevam documentos e busca em banco de dados, sejam eles relacionais e isolados ou banco de dados interligados em rede de hipermídia, tais como a World Wide Web. A mídia pode estar disponível sob forma de textos, de sons, de imagens ou de dados.

> O conceito de recuperação de informação (*Information Retrieval*) surge neste contexto como um processo onde são devolvidos e ordenados por ordem de importância os documentos mais relevantes de acordo com uma consulta (*query*) especificada pelo utilizador. (CAMPOS, 2005, p.1).

A Recuperação de Informação trata da automatização do processo de recuperação de documentos relevantes e suas principais funções são: a representação, a indexação[2] e a busca por documentos que satisfaçam à necessidade do usuário.

> Recuperação de Informação (IR) é um campo de estudo que ajudar os usuários a encontrar informações que combinem com as suas necessidades. Tecnicamente, IR estuda a aquisição, organização, armazenamento, recuperação e distribuição de informações.
> (LIU, 2007, p.183)

Segundo Liu (2007) a IR é diferente da classificação de dados usando consulta SQL, porque os dados em base de dados são altamente estruturados e armazenados em tabelas relacionadas, enquanto informação em textos não é estruturada. Não há uma linguagem de consulta como SQL para recuperação de textos.

A indexação de páginas Web, que facilita o processo de recuperação, é bem mais complexa que o processo de indexação utilizado em bancos de dados tradicionais. A enorme quantidade de páginas na Web, seu dinamismo e atualizações freqüentes, fazem da indexação uma tarefa aparentemente impossível. E, na verdade, este é um dos grandes desafios dos serviços de busca atuais: indexar toda a Web.

Paginas web também são totalmente diferentes dos documentos de textos convencionais. Primeiro, as paginas web contém *hiperlinks* e âncoras, o que não existe em documentos tradicionais. Os *hiperlink*s são extremamente importantes para pesquisa e desempenham um papel central no algoritmo de classificação de busca. Texto de ancoras associados a *hiperlinks* também são cruciais, pois esses são freqüentemente uma descrição mais precisa da página para a qual o link aponta.

Segundo Liu (2007), as páginas web são semi-estruturadas. Essas têm diferentes campos, por exemplo, títulos, metadados, corpo, etc. A informação contida em certos campos, (como por exemplo, o campo titulo) é mais importante do que em outros campos. Além disso, o conteúdo das páginas é tipicamente organizado e apresentado em vários blocos estruturados. Alguns blocos são importantes e outros não. Descobrir efetivamente o bloco principal do conteúdo de páginas web é útil para pesquisas, porque termos que aparecem em tais blocos são efetivamente mais importantes.

[2] Indexação é essencialmente um processo de classificação onde é realizada uma análise conceitual do documento ou elemento de informação.

3.1 Modelos de Recuperação da Informação

Os modelos de recuperação de informação apresentam estratégias de busca de documentos relevantes para uma consulta. Tanto a consulta feita pelo usuário, quanto os documentos que compõem a coleção a ser pesquisada, são representados pelos seus termos.

3.1.1 Modelo Booleano

O Modelo Booleano é (LIU, 2007) um dos primeiros e mais simples modelos de Recuperação de Informação. Com pouca prática, é possível construir consultas complexas devido a flexibilidade de misturar operadores booleanos com a palavras-chave.

O Modelo Booleano, que considera a consulta como uma expressão booleana convencional, liga seus termos através de conectivos lógicos **AND**, **OR** e **NOT**. Por exemplo, a consulta (*web* AND *Mining*) diz que os documentos recuperados podem conter ambos os termos *web* e *Mining*. Outro exemplo de consulta seria (*web* OR *Mining*) onde pelo menos um destes termos deve estar em cada documento recuperado.

O Modelo Booleano (LIU, 2007) raramente é usado sozinho. Muitos mecanismos de busca suportam algumas formas de limitadores de recuperação booleana usando explícitos operadores de inclusão e exclusão. Por exemplo, poderia ser emitida a seguinte consulta para o Google: mineração - dados + "Recuperação de Informação", em que + (inclusão) e – (exclusão) são operadores Booleanos similares a AND e NOT respectivamente.

Nesse modelo os documentos recuperados são aqueles que contêm os termos que satisfazem a expressão lógica da consulta. O documento é considerado relevante ou não relevante a uma consulta e não existe resultado parcial. Talvez essa seja a maior desvantagem do método booleano. Porém, é bastante claro que a freqüência dos termos e a proximidade entre eles contribuem significativamente para a relevância do documento.

O modelo Booleano é muito mais utilizado para recuperação de dados do que para recuperação de informação. Possui a vantagem de ser facilmente programável e exato,

além da expressividade completa se o usuário souber exatamente o que quer. Em contrapartida, apresenta como desvantagens: a dificuldade do usuário especificar o que quer, pouco retorno, não existe ordenação, a formulação das consultas são difíceis para usuários inexperientes e o resultado pode ser nulo ou muito extenso.

3.1.2 Modelo Vetorial

O Modelo Vetorial talvez seja o mais conhecido e o mais usado na Recuperação da informação (LIU, 2007).

Nesse modelo, os documentos são representados como modelos no espaço e as consultas são representadas como documentos. Cada documento da coleção é considerado como um vetor multidimensional, onde cada dimensão do vetor representa uma palavra. O peso da consulta e do documento é calculado baseado no peso e direção dos respectivos vetores e a medida da distância de um vetor entre a consulta e o documento é usada para ordenar os documentos recuperados.

Segundo Cardoso (2008), o Modelo Vetorial representa documentos e consultas como vetores de termos. Os documentos devolvidos como resultados de uma consulta são representados similarmente, ou seja, o vetor resultado para uma consulta é montado através de um cálculo de similaridade. Aos termos das consultas e documentos são atribuídos pesos que especificam o tamanho e a direção de seu vetor de representação. Ao ângulo formado por esses vetores, conforme figura 1, dá-se o nome de q. O termo cos(q) determina a proximidade da ocorrência.

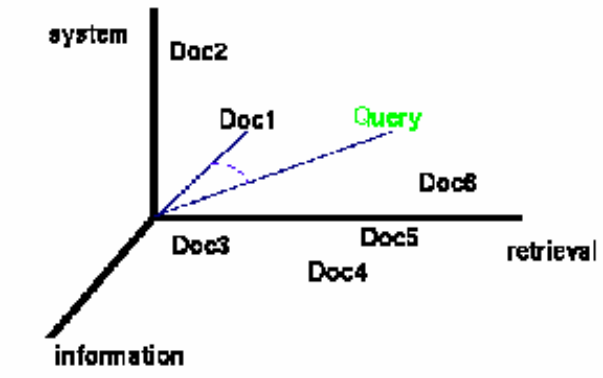


Figura 1: Modelo Vetorial (DUQUE, 2007)

Para Liu (2007), um documento no modelo de espaço de vetor é representado como um vetor de peso no qual cada peso de componente é computado baseado em alguma variação de **TF (*Term Frequency*)** ou **TF-IDF scheme (**IDF – Inverse *Document Frequency*). O peso w_{ij} do termo t_i em um documento $\boldsymbol{d}_j$ não está em {0,1} como no modelo booleano, mas pode ser qualquer outro número.

No método TF, o peso de um termo t_i no documento $\boldsymbol{d}_j$ é o número de vezes que t_i aparece no documento $\boldsymbol{d}_j$ e é denominado por f_{ij}**.** O peso f_{ij} é calculado pela Equação 1 (LIU, 2007).

A deficiência desta estratégia é que ela não considera a situação em que um termo aparece em muitos documentos da coleção. Em compensação, pode-se utilizar os valores TF para verificar a importância dos termos nos documentos, pois os termos mais freqüentes podem representar conceitos importantes associados ao tema principal do texto.

No modelo de ponderação TF-IDF, a pontuação de um termo é feita levando em consideração a freqüência do termo no documento e em todos os outros documentos da coleção. Sendo assim, considerando N o número total de documentos em que o termo t_i aparece pelo menos uma vez e f_{ij} a contagem da freqüência de termos t_i no documento $\boldsymbol{d}_j$. Então, a freqüência do termo (denotado por tf_{ij}) de t_i em d_j é dado por (LIU, 2007):

$$tf_{ij} = \frac{f_{ij}}{\max\{\{ f_{1,j}, f_{2,j} ..., f_{|V\ |j}, \}} \qquad (1)$$

O |v| é o tamanho do vocabulário da coleção e o máximo é computado em cima de todos os termos que aparecem no documento. Se o termo t_i não aparece no documento $\boldsymbol{d}_j$, $tf_{ij} = 0$.

A freqüência da cada palavra pelo inverso (denotada por idf_i) do termo t_i é dado pela equação 2 (LIU, 2007).

$$idf_i = \log \frac{N}{df_i} \qquad (2)$$

Onde *df* é o número de documentos em que a palavra *i* ocorre pelo menos uma vez e N representa o número de elementos da coleção de documentos. Se um termo aparece em um grande número de documentos na coleção, provavelmente ele não é importante. Por fim, o peso do termo TF-IDF é dado por Liu (2007), conforme a equação 3:

$$w_{ij} = tf_{ij} \times idf_i \qquad (3)$$

Por conseguinte, o modelo vetorial representa documentos e consultas como vetores de termos. Os termos são ocorrências únicas nos documentos e o vetor resultado é obtido através de um cálculo de similaridade. Tem como vantagens a simplicidade de implementação, a facilidade que ele provê de se computar similaridades com eficiência e o fato de que o modelo se comporta bem com coleções genéricas.

Outros métodos de ponderação substituem a parcela IDF por outras funções, como podem ser visto, por exemplo, em MATSUNAGA, 2007.

3.1.3 Modelo Probabilístico

O Modelo Probabilístico é o modelo clássico de recuperação de informação baseado na interpretação probabilística da relevância de um documento para uma dada interrogação e tem fundações na teoria estatística.

Esse modelo supõe que exista um conjunto ideal de documentos que satisfaz a cada uma das consultas ao sistema, e que esse conjunto de documentos pode ser recuperado.

Através de uma tentativa inicial com um conjunto de documentos (para a qual se podem utilizar técnicas de outros modelos, como o vetorial) e do *feedback* do usuário em sucessivas interações, busca-se aproximar cada vez mais desse conjunto ideal, por meio de análise dos documentos considerados pertinentes pelo usuário. O valor desse modelo está em considerar a interação contínua com o usuário como um caminho para refinar o resultado continuamente.

Segundo Cardoso (2008), o Modelo Probabilístico descreve documentos considerando pesos binários, que representa a presença ou ausência de termos. E o vetor resultado é gerado com base na probabilidade de que um documento seja relevante para uma consulta. Segundo esse autor, o modelo usa como principal ferramenta matemática o Teorema de Bayes e tem como vantagem o desempenho prático.

As desvantagens do Modelo Probabilístico (CARDOSO, 2008) são a dependência da precisão das estimativas de probabilidade e o fato de não considerar a freqüência do temo no documento e ignorar o problema de filtragem da informação[3].

[3] Selecionar documentos pertinentes com base no perfil do usuário.

4 MÁQUINAS DE BUSCA

As máquinas de busca são ferramentas essenciais para recuperar informação da Web. A quantidade de informações na Internet é tão grande e diversificada que é praticamente impossível encontrar tudo o que se precisa sem o uso de um mecanismo de busca. Sem as máquinas de busca a web seria quase inútil.

> Um sistema de busca é um conjunto organizado de computadores, índices, bases de dados e algoritmos, reunidos com a missão de analisar e indexar as páginas *web*, armazenar os resultados dessa análise e devolve-los posteriormente adequando-os a uma pesquisa que preencha os requisitos indicados pelo utilizador por ocasião de uma consulta. (CAMPOS, 2005, p. 8)

"As máquinas de busca são pequenas aplicações, disponíveis em determinados sites/portais, que numa descrição simples, possibilitam a obtenção de um conjunto de "links" (URL' s) de páginas contendo informação relacionada com aquilo que um utilizador procura" (CORDEIRO, 2003).

O motor de busca, conhecido também como programas de busca, mecanismo de procura, ferramenta de busca, são programas que tem três funções básicas: identificar páginas da web, indexar estas páginas em um banco de dados e devolver o resultado da consulta em um mecanismo de pesquisa com interface.

> [...] procuram ser de fácil utilização, através de uma interface gráfica amigável em um sistema de hipermídia, sendo que a solicitação de busca é concretizada em segundos e as respostas são apresentadas diretamente pelos links - pontos de acesso para as páginas e/ou por categorias de assunto ou, ainda, segundo critérios de parametrização fornecidos pelos usuários, podendo apresentar ainda, roteiros de ajuda e exemplos de estratégias de busca diversificados.
> (BUENO, 2000, p. 23).

Para a maioria dos usuários, o conteúdo das páginas é o que realmente os interessa. Encontrar o conteúdo certo é muito mais fácil com a ajuda de um motor de busca. A tarefa

que as máquinas de busca desempenham é o mais útil exemplo de *Web Mining Content* e o nome formal para o mecanismo que os motores de busca realizam é *Information Retrieval*.

Para Campos (2005) a organização dos resultados devolvidos pelos mecanismos de busca facilitará a navegação do utilizador, pois usando essas ferramentas, os processos serão automatizados e o usuário não mais necessitará de fazer uma procura exaustiva da página do seu interesse, o que significa um considerável ganho de tempo que o mesmo poderá dedicar a outras tarefas.

Segundo Campos (2005) as funções dos Motores de Busca são: *crawling*, *indexing* e *searching*.

4.1 *Crawling*

Uma forma das máquinas de busca descobrirem sites é através de *crawler*, ou seja, rastejador. Este processo é realizado por programas chamados *bots* ou *spiders* criados especialmente para este fim.

Para Menczer(2007) in Liu (2007) *Web Crawlers* são programas que automaticamente vasculham páginas web para colher informação que podem ser analisadas e mineradas em um local *on-line* ou *off-line*.

Esse robô interage diretamente com a Web. Possui como função descobrir novos documentos na internet de forma a torná-los consultáveis. Os *crawlers* automaticamente e recursivamente visitam páginas Web, lêem-nas, copiam-nas, e, seguem os *hiperlinks* contidos nelas (SELBERG apud ZANIER, 2006).

O *crawler* captura e transmite muitos sites simultaneamente de forma eficiente, tentando prever a similaridade entre o conteúdo do arquivo e a consulta do usuário.

> Cada *bot* na verdade é um agente comum que faz requisições aos servidores web. Esses baixam todo tipo de informação sobre o seu site, e na verdade vão além disso: INTERPRETAM o conteúdo (como palavras repetidas, nome das imagens, links, formatação do código HTML, etc), armazenando assim para podê-los visitar posteriormente e buscar novas modificações ou não". (REBITTE e BP, 2006, p.14).

Segundo Rebitte e BP (2006) todos os dados rastejados pelos *spiders* são gravados no banco de dados das Máquinas de Busca e mesmo havendo modificações no site, os

robôs voltam ao site em média de 3 em 3 dias e recuperam as modificações. A figura 2 apresenta o crawler do Google(*googlebot*) no processo de recuperação da informação.

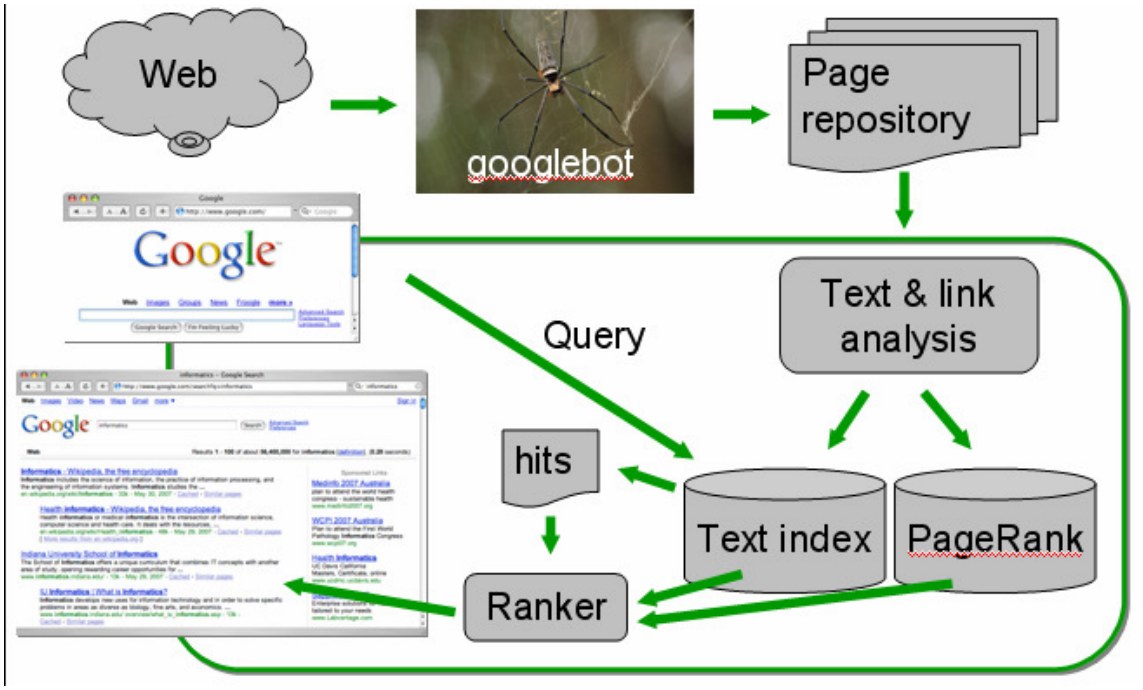


Figura 2: O crawler em uma máquina de busca (MENCZER, 2007)

Os mecanismos de busca (MARKOV, 2007) colecionam documentos da web e cria o índice de acordo com as palavras ou termos contidos nestes documentos. Porem, para indexar um conjunto de páginas web, é necessário que todos os documentos estejam disponíveis para o processamento em um repositório local. A coleta de todos os documentos é feita folheando a web exaustivamente e armazenando todas as páginas visitadas. Este processo é realizado pelos *crawlers*.

Um bom exemplo de um *crawler* é o WebSPHINX (http://www.cs.cmu.edu/~rcm/websphinx) (MARKOV, 2007) que é usado para visualizar e analisar a estrutura da web de forma gráfica. Abaixo ilustraremos o WebSPHINX, que é uma classe Java interativa e desenvolvida para ambiente web. Para realizar o teste, foi passado para o WebSPHINX a URL do site da UFRJ.

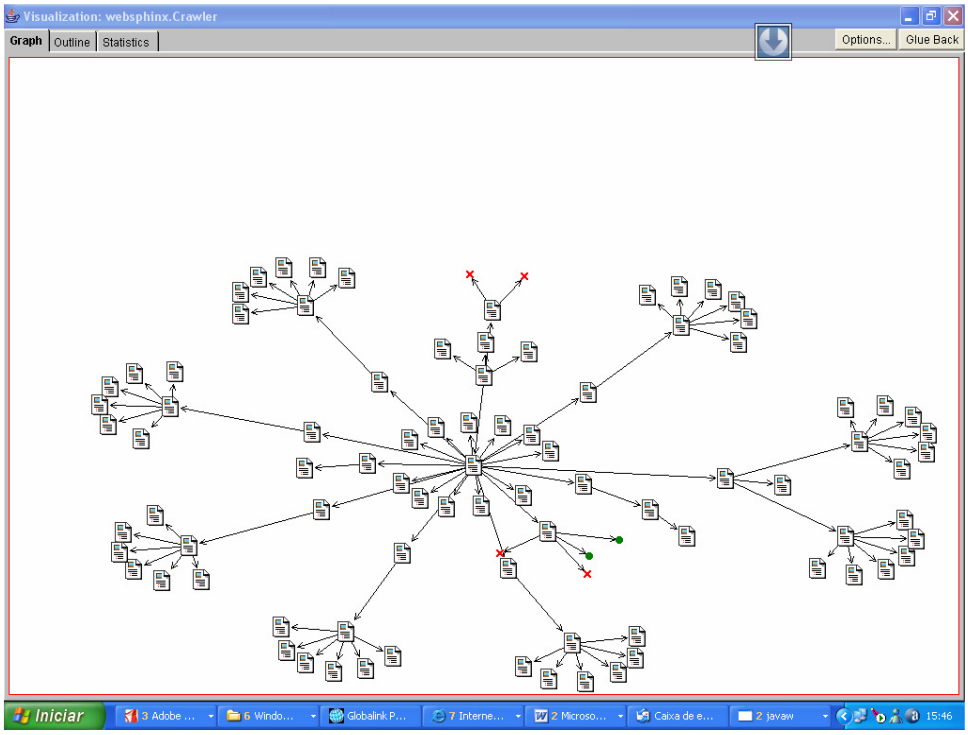


Figura 3 : Imagem gerada pelo Websphinx

A figura acima mostra o gráfico da estrutura dos links do site da UFRJ. Embora a visualização gráfica da estrutura da web seja uma característica agradável desse *crawler*, isso não é o mais importante. Na realidade, o papel básico de um *crawler* é colecionar informações sobre páginas web para as máquinas de busca. Isso pode ser conteúdo textual de páginas, títulos, cabeçalhos, estruturas de tags ou estruturas de links da web.

Essas informações são organizadas corretamente para um acesso eficiente e armazenadas em um repositório para ser usadas para indexação e pesquisa. Assim, o *crawler* não é somente uma implementação de um algoritmo de visualização de uma coleção de páginas web em forma de um gráfico, mas é também um analisador e *parser* HTML.

Além do que já foi apresentado acima, o rastreador WebSPHINX permite:

- Salvar páginas para o disco local para que as mesmas possam ser folheadas off-line.
- Concatenar páginas de modo que elas possam ser visualizadas ou impressas como um único documento.
- Extrair todos os textos correspondentes a um determinado padrão de um conjunto de páginas.

Outro exemplo de *crawler*, citado por Liu (2007), é o *MySpider*, um applet Java disponível em *myspiders.informatics.indiana.edu* que faz uso de dois algoritmos: O *best-N-first* e *InfoSpiders*.

O *MySpider* é interativo (LIU, 2007), o usuário submete uma consulta conforme faz em uma máquina de busca e o resultado é mostrado em uma janela. Contudo, ao contrário dos Motores de Busca, esta aplicação não tem nenhum índice para procurar os resultados. A Web é rastejada em tempo real. As páginas julgadas relevantes são exibidas em uma lista que é ordenada por critério de seleção do usuário: *score* e *recency*. O *score* é simplesmente a similaridade entre o conteúdo (coseno) e a consulta. A *recency* de uma página é estimada pelo cabeçalho de última modificação, se retornado pelo servidor.

A figura 4 mostra o applet *MySpiders* em ação. Neste exemplo, foi feito uma consulta por "*Web Mining*" usando o algorítimo *InfoSpiders*.

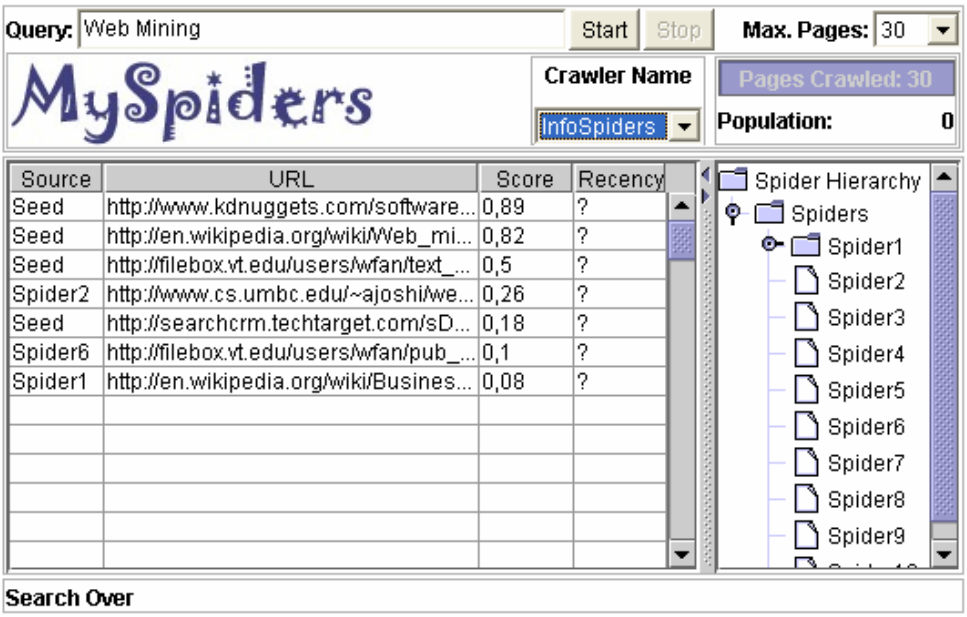


Figura 4: Imagem do resultado da consulta por "Web Mining"gerado pelo aplicativo MySpiders

4.2 Indexação

Segundo Salton (apud CORREA, 2003), a indexação é o processo em que as palavras contidas nos textos são armazenadas em uma estrutura de índice para viabilizar a pesquisa de documentos através das palavras que eles contêm.

"Indexar é produzir um índice menor, porém mais eficiente, para representar o conteúdo original que facilite a recuperação de informações". (HERSH et al. apud MARTHA, 2005).

Muitas das vezes conhecido como catálogo, o *index* é como um livro gigante que contém uma cópia de todas as páginas *web* que o *spider* encontrou. As mesmas poderão ser usadas para construir um índice invertido de palavras-chave para posterior classificação dos documentos em diretórios, ou para a construção de um grafo de *hiperlinks* de forma a desenvolver um *ranking* de links. (PAGE *et al* apud CAMPOS, 2005).

> Índices invertidos são criados para possibilitar melhorias significativas no desempenho e na funcionalidade da busca. As buscas usam os índices extraídos dos documentos-texto para comparação com a consulta do usuário. O arquivo de índices invertidos possui como edificadores as palavras-chave do texto e para cada palavra-chave, há um conjunto de referências que apontam para os documentos onde essas palavras- chave ocorrem. O processo de obtenção de palavras-chave que identificam o conteúdo do documento é chamado de Indexação Automática. (SALTON apud CORREA, 2003. p.20)

A estrutura do arquivo invertido é composta por dois elementos (BAEZA-YATES apud CORREA, 2003): o vocabulário e as ocorrências. O vocabulário é o conjunto de palavras retiradas do texto. As ocorrências são os documentos onde essas palavras aparecem.

A figura abaixo mostra a utilização de arquivos invertidos para o armazenamento dos termos que identificam os documentos. Os termos ou palavras-chave são extraídos dos textos e ficam armazenados juntamente com as referências para os respectivos documentos. Através dessas referências é possível recuperar os documentos desejados (CORREA, 2003).

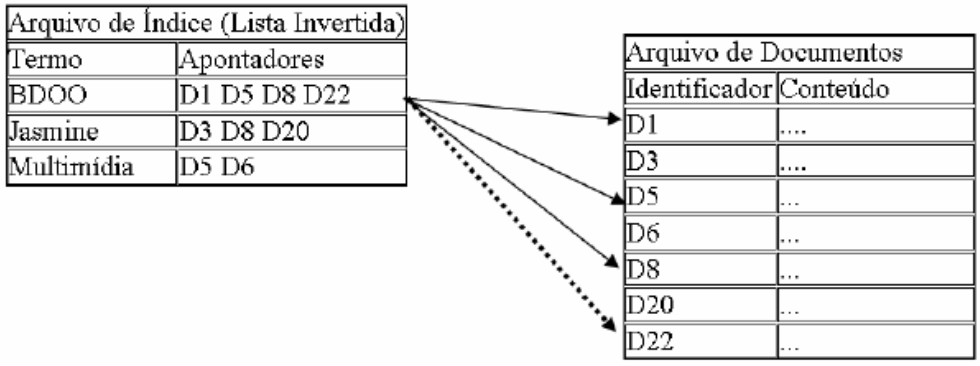


Figura 5: Estrutura de arquivo invertido (CORREA, 2003, p. 21)

Segundo Yates apud Loh in Santos (2000), existem três tipos de indexação: indexação tradicional, indexação do texto todo e indexação por tags.

Na indexação tradicional, uma pessoa determina os termos que caracterizam os documentos, os quais farão parte do índice de busca. Porém, este tipo de índice, além de ser uma atividade trabalhosa e que consome muito tempo, pode conter falhas devido a intervenção humana.

> A indexação do texto todo, todos os termos que compõem o documento fazem parte do índice. Neste segundo tipo de técnicas indexação (full-text) procura indexar todos os termos, sem estruturas hierárquicas entre eles. As ferramentas indexam automaticamente todas as palavras do documento, gerando índices volumosos. Para diminuir o problema do tamanho dos índices, existem técnicas como árvores, listas invertidas, etc. Também para diminuir o tamanho dos índices, mas principalmente para evitar a indexação de termos indesejados ou inconsistências no índice, podem ser aplicados filtros nesse processo. (SANTOS, 2000. p.67)

O terceiro tipo de indexação (por tags) procura indexar somente as partes relevantes do documento. Para tanto, as ferramentas automatizadas deverão analisar cada documento, procurando por marcas (tags) que identificam as partes mais significativas.

> Na indexação por tags apenas algumas partes do texto são escolhidos, automaticamente, para gerar as entradas no índice (somente aquelas consideradas mais caracterizadoras). Nesse tipo de indexação procura-se indexar somente as partes relevantes do documento. Para tanto, as ferramentas automatizadas deverão analisar cada documento, procurando por marcas (tags) que identificam estas partes mais importantes. (SANTOS, 2000. p.67).

A indexação é essencialmente um processo de classificação onde é realizada uma análise conceitual do documento ou elemento de informação. Por exemplo, nas técnicas baseadas no modelo do espaço vetorial, a indexação envolve a atribuição de elementos de informação a certas classes, onde uma classe é o conjunto de todos os elementos de informação para o qual um termo de indexação (ou palavra-chave), em particular, tem sido atribuído. Os elementos de informação podem fazer parte de várias classes. Algumas técnicas atribuem pesos aos termos de indexação de um elemento de informação de forma a refletir sua relativa relevância (GIRARDI, 1998). Nas técnicas baseadas no modelo estatístico, os termos de indexação são extraídos a partir de uma análise de freqüência das

palavras ou frases em cada documento e em toda a fonte de informação. Nas técnicas lingüísticas, os termos de indexação são extraídos utilizando técnicas de processamento da linguagem natural, por exemplo, análise morfológica, lexical, sintática e semântica (GIRARDI, 1995).

4.3 Searching

O motor de busca é a terceira parte do software. Este é o programa que percorre o índice invertido de forma a encontrar os milhares de páginas armazenadas no *index* para encontrar correspondências com a consulta do usuário.

Sendo as ferramentas de busca o mecanismo mais utilizado para recuperação das informações na Web, um estudo descritivo e exploratório será realizado para melhor identificar e analisar as características, funções, formas de catalogação, indexação e recuperação das informações via estas máquinas. Como universo de pesquisa, enfoca-se as ferramentas *Google, Northern Light Search, Snaket, Clusty, WebCrawler, Kartoo, MetaCrawler, Grokker e Copernic* com o intuito de estruturar um referencial teórico para o processo de descrição, armazenamento, recuperação e disseminação da informação na Internet. Serão analisadas as características funcionais das ferramentas de busca citadas e será feito um levantamento bibliográfico sobre o tema proposto.

4.3.1 Google

Google[4] é o nome da empresa que criou e mantém o maior site de busca da internet, o Google Search. O serviço foi criado a partir de um projeto de doutorado dos estudantes Larry Page e Sergey Brin da Universidade de Stanford em 1996. (WIKIPÉDIA, 2008)

O Google é composto por uma série de *crawlers*, os *GoogleBots*, distribuídos por várias máquinas e um servidor de URL que envia listas de URLs para os *crawlers* procurarem. Como os *crawlers* seguem os links de uma página para outra, o motor de busca consegue encontrar milhares de páginas (CAMPOS, 2005).

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

> O Google atualiza sua base de informações diariamente. Existe o crawler Googlebot, um "robô" do Google que busca por informações novas em tudo o que for site. Isso é realmente interessante porque cerca de aproximadamente 4 dias depois de uma matéria ser publicada em um site já é possível encontrá-la no Google. (CAMPOS, 2005)

As páginas encontradas pelos *GoogleBots* são guardadas em um repositório, ficando associado a cada página um ID chamado de *DocID*, o tamanho da mesma e o URL. A função de *indexing* é feita pelo *indexer*, que lê a página web de uma URL e constrói um conjunto de *hits* (ocorrências de palavras) do texto. O *hits* tem um *WordID,* posição da palavra no documento, um tamanho aproximado da fonte e a informação capitalizada. (KONCHADY, 2005). O index é ordenado alfabeticamente por termos e a lista de documentos que contém esses termos é armazenada numa estrutura intitulada por *Inverted Index*.

Para Liu (2007) o *Inverted Index* de uma coleção de documentos é basicamente a estrutura de dados que une cada termo diferente com a lista de todos os documentos que contém o termo. Assim, na recuperação, ele reduz o tempo para encontrar o documento que contem o termo da consulta.

Abaixo apresentaremos algumas dicas de como montar um site perfeito para que o Google (REBITTE e BP, 2006) recupere com eficiência as informações contidas nesse site:

- **Palavras-chave:** Definir as palavras-chave no cabeçalho das páginas do site conforme exemplo:

<meta name = "keywords" content = "Web Mining, Search engine, Information Retrieval">

- **Título**: O Título é um fator que contribui muito para um bom posicionamento, sendo assim, o título deve refletir o conteúdo da página. Exemplo:

<TITLE>Web Mining Content</TITLE>

- Os tags **H1** e **H2**: Estes tags são de grande importância, pois são usados para definir títulos e subtítulos respectivamente. Exemplo

<H1>Web Mining</H1>

<H2>Web Mining Content</H2>

- O atributo **ALT** do tag IMG(de imagem): Este atributo deve ser usado para a descrição da imagem e levado muito em consideração como organização para os robôs. Exemplo:

<img src=”figura.gif” name = ”descrição da figura”>

- **Hiperlinks**: Use sempre texto como hiperlinks, isso é muito importante para os robôs.

Web Mining

- **PageRank:** Trocar links entre sites pode ajudar na classificação, principalmente se o site aliado estiver uma boa qualificação no *PageRank*.

A figura 6 mostra o resultado de uma pesquisa, na qual foi fornecido para o Google as palavras “Web Mining Content”.

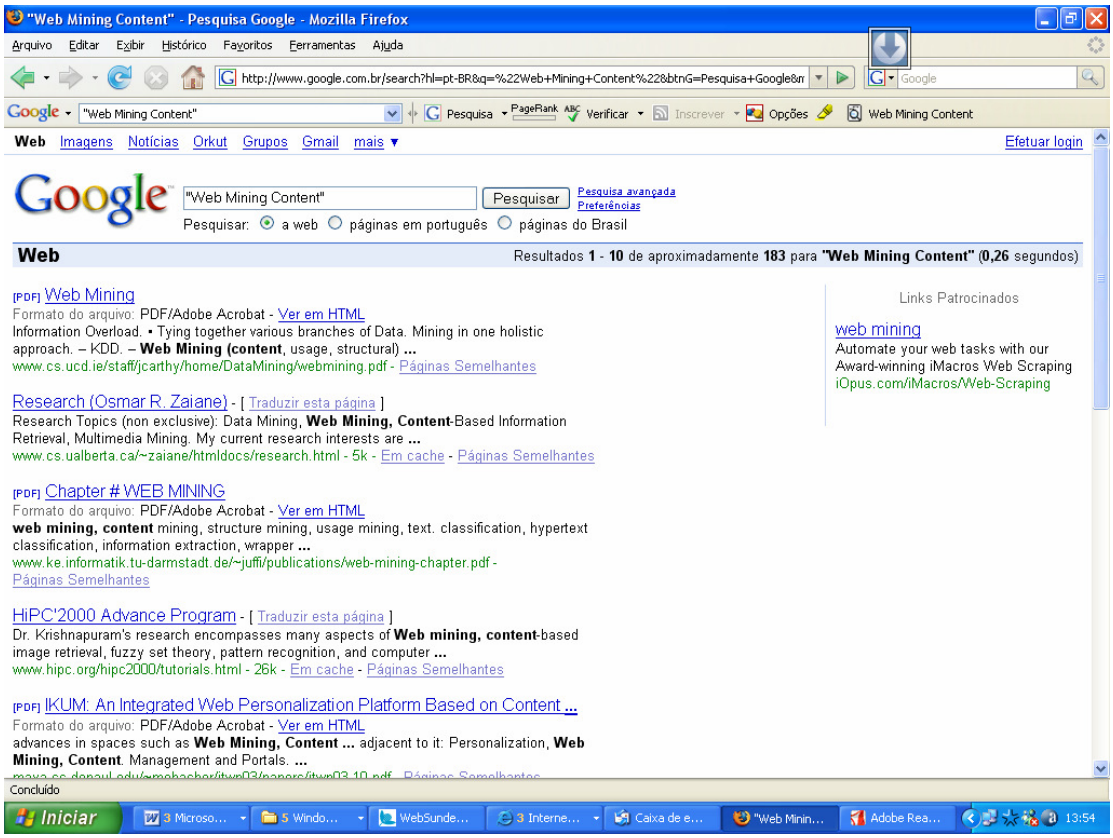


Figura 6: Imagem gerada pelo google em 11/01/08

4.3.1.1 Tecnologias de busca utilizadas pelo Google

Segundo Zanier (2006) as tecnologias de buscas utilizadas pelo Google são: o *crawler GoogleBot* e o algoritmo *PageRank*.

O robô do Google, *GoogleBot*, percorre a Web para explorar páginas e enviar suas URL’s ao índice de documentos.

> Existe um servidor URL que envia listas de URLs para serem percorridas por vários processos do *GoogleBot*. Os URLs encontrados são enviados ao servidor de URL, que checa se estes URLs já foram previamente percorridos. Caso negativo, o URL é adicionado ao índice de URLs que devem ser percorridos pelo *GoogleBot*. (ZANIR, 2006, p.39)

O *PageRank* é um algoritmo usado pelo motor de busca Google para ajudar a determinar a relevância ou importância de uma página.

> O PageRank interpreta um link da página A para a página B como um voto da página A para a página B. Mas, analisa também o valor da página que dá o voto. Os votos dados por páginas importantes pesam mais e ajudam a tornar outras páginas importantes (BRIN & PAGE, apud ZANIR, 2006, p.39).

4.3.2 Medline

O sistema Medline[5] (KONCHADY, 2006) é um sistema de busca que pesquisa documentos médicos da Biblioteca Nacional de Medicina e que foi desenvolvido em 1960 e avaliado em 1971.

> Medline foi baseado nos sistemas de organização de arquivos invertidos como muito de seus contemporâneos. Surpreendentemente, a organização de arquivos invertidos continua a ser a base de muitos das grandes máquinas de busca na web hoje. (KONCHADY, 2006. p.185)

A linguagem de busca no Medline (KONCHADY, 2006) foi uma simplificação da linguagem de consulta Booleana, em que os termos da consulta são combinados logicamente usando os operadores Boolean AND, OR, e NOT.

A figura 7 ilustra o resultado da busca pela palavra-chave “heart” usando o sistema Medline para a pesquisa.

[5] http://www.nlm.nih.gov/

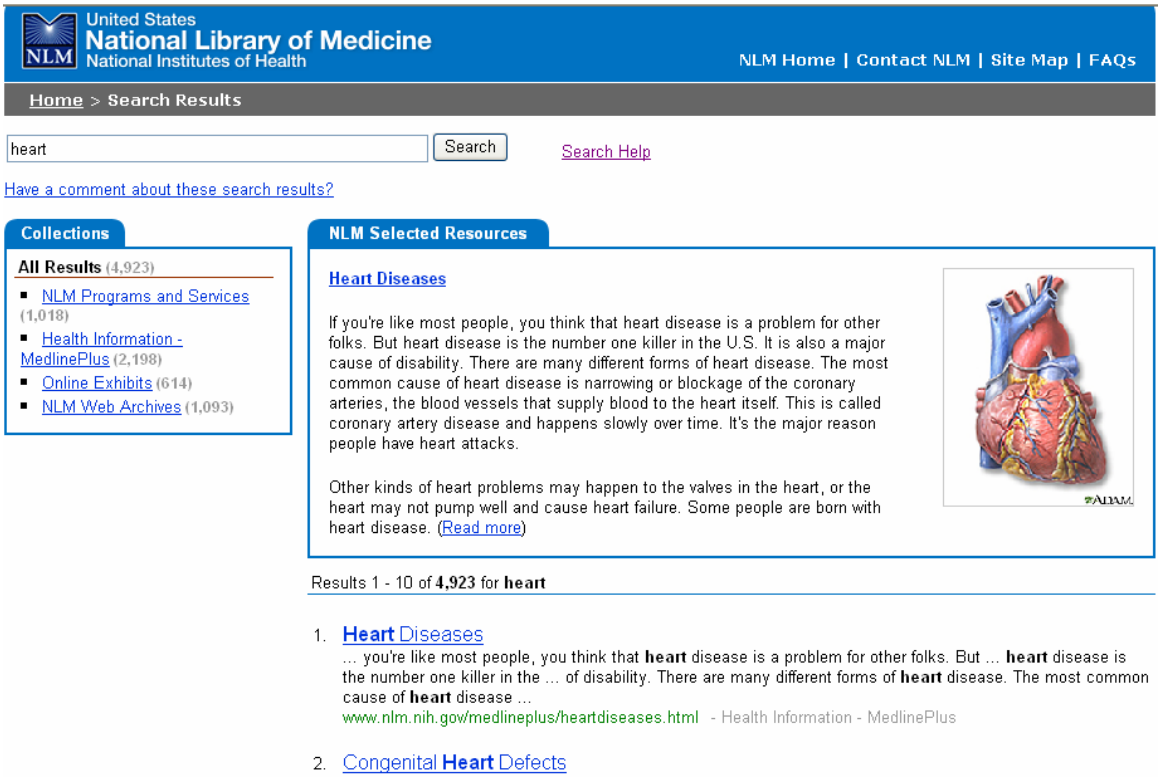


Figura 7: Imagem gerada pelo Medline usando a palavra-chave “heart” na consulta

4.3.3 WebCrawler

O WebCrawler[6] é um meta busca que procura simultaneamente em vários dos principais mecanismos de busca da Web, como Yahoo, Google e outros sites de busca, e traz os resultados de todos numa única página. O sistema é útil, mas tem suas desvantagens. O resultado demora um pouco mais e não é possível usar todas as funções de busca avançada disponíveis nos sites originais. Ainda assim, pode ajudar a poupar o tempo do usuário no processo de recuperação de informação.

A figura 8 ilustra o resultado de uma pesquisa gerado pelo WebCrawler usando a palavra-chave “Web Mining” na consulta.

[6] http://www.webcrawler.com/webcrawler/

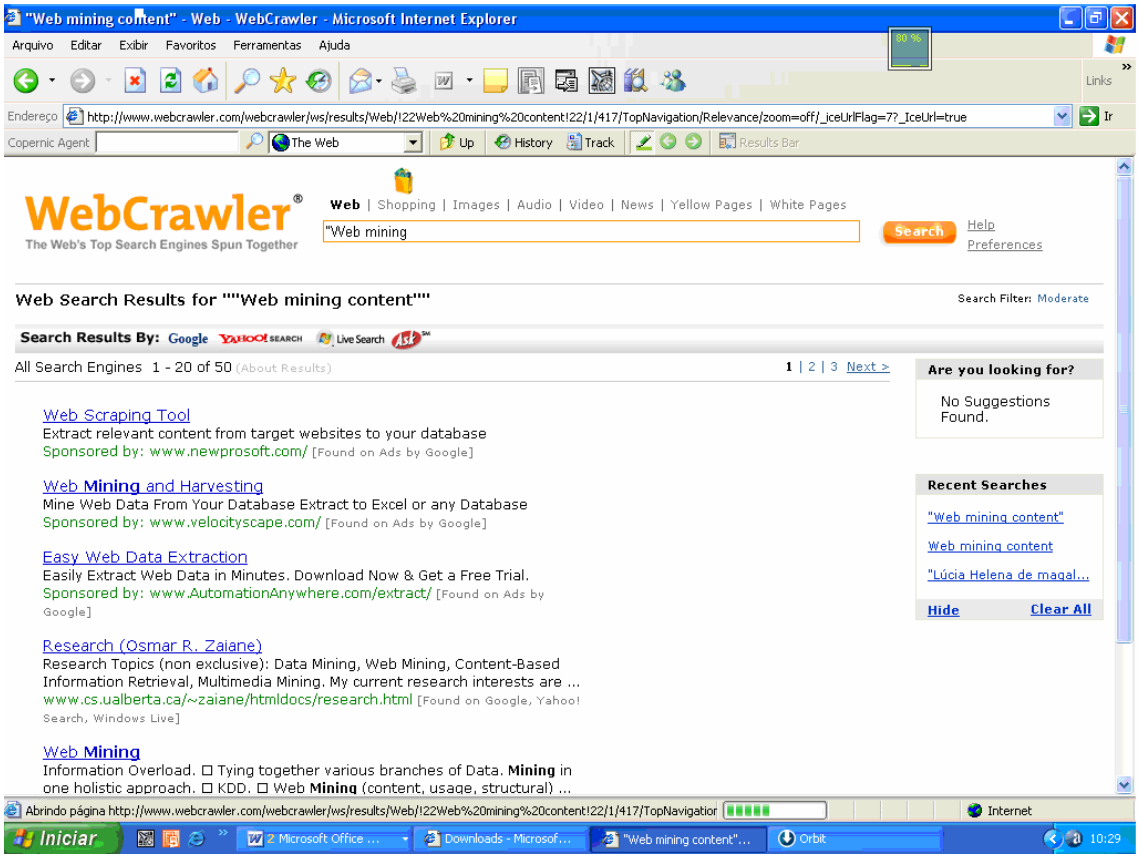


Figura 8: Imagem gerada pelo WebCrawler usando a palavra-chave “Web Mining” na consulta

4.3.4 Terrier

Terrier[7] é uma ferramenta de busca altamente flexível e eficiente, escrito em Java e que foi desenvolvido pelo Departamento de Ciências Computacionais, na Universidade de Glasgow. Possui versão de código aberto que oferece uma plataforma de investigação e experimentação para recuperação de texto de forma transparente.

A investigação constante posta em Terrier faz com que o mesmo se torne em um amplo campo de recuperação de informação, tornando Terrier uma ideal e forte plataforma para desenvolver e avaliar novos conceitos e idéias.

O Terrier pode ser usado em aplicações de busca em desktop e como máquina de busca na Web. Sua em função é recuperar documentos relevantes resposta à necessidade do usuário, frequentemente formulada por uma consulta. Para isso, a máquina de busca usa

[7] http://ir.dcs.gla.ac.uk/terrier

modelo de recuperação para estimar a relevância dos documentos que satisfaçam a necessidade do usuário.

As técnicas e modelos de recuperação implementados no Terrier para melhorar o ranking de documentos relevantes são: (OUNIS, et al, 2007)

- Um *framework* probabilístico para modelos de ponderação.
- Vários novos métodos de recuperação confeccionados especialmente para busca na Web;
- O Terrier possui um configurável *crawler* chamado de Librador, que tem sido usado para o desenvolvimento do Terrier em várias aplicações industriais.

As Máquinas de busca geralmente devolvem documentos pertinentes a consulta do usuário, calculando a relevância do conteúdo do documento. O calculo de relevância é feito usando modelos de ponderação. O Terrier implementa uma variedade de modelos, tais como o DFR(Divergence form Randomness) que é um tipo de modelo probabilístico que se baseia na freqüência estatística que um termo aparece em um documento ou em uma coleção de documentos.

Quanto à arquitetura da plataforma Terrier, destacaremos seus dois principais componentes: indexação (*indexing*) e recuperação (*retrieval*). A indexação descreve o processo durante o qual Terrier analisa um documento e representa a informação da coleção na forma de um índice que contém as estatísticas sobre a freqüência do termo em cada documento e na coleção inteira. Quanto à recuperação, o Terrier devolve os documentos pertinentes à consulta do usuário com base na relevância do documento. A relevância de um documento é calculada através do modelo de ponderação DFR.

A figura 9 mostra a imagem gerada pelo Terrier, que está atualmente disponível como máquina de busca no site do Departamento de Computar Ciência na Universidade de Glasgow (http://www.dcs.gla.ac.uk/search). Nessa pesquisa foi usado a palavra-chave “Web Mining”. A figura 10 ilustra o resultado do *Terrier Desktop Search* em uma pesquisa pela palavra “Terrier”.

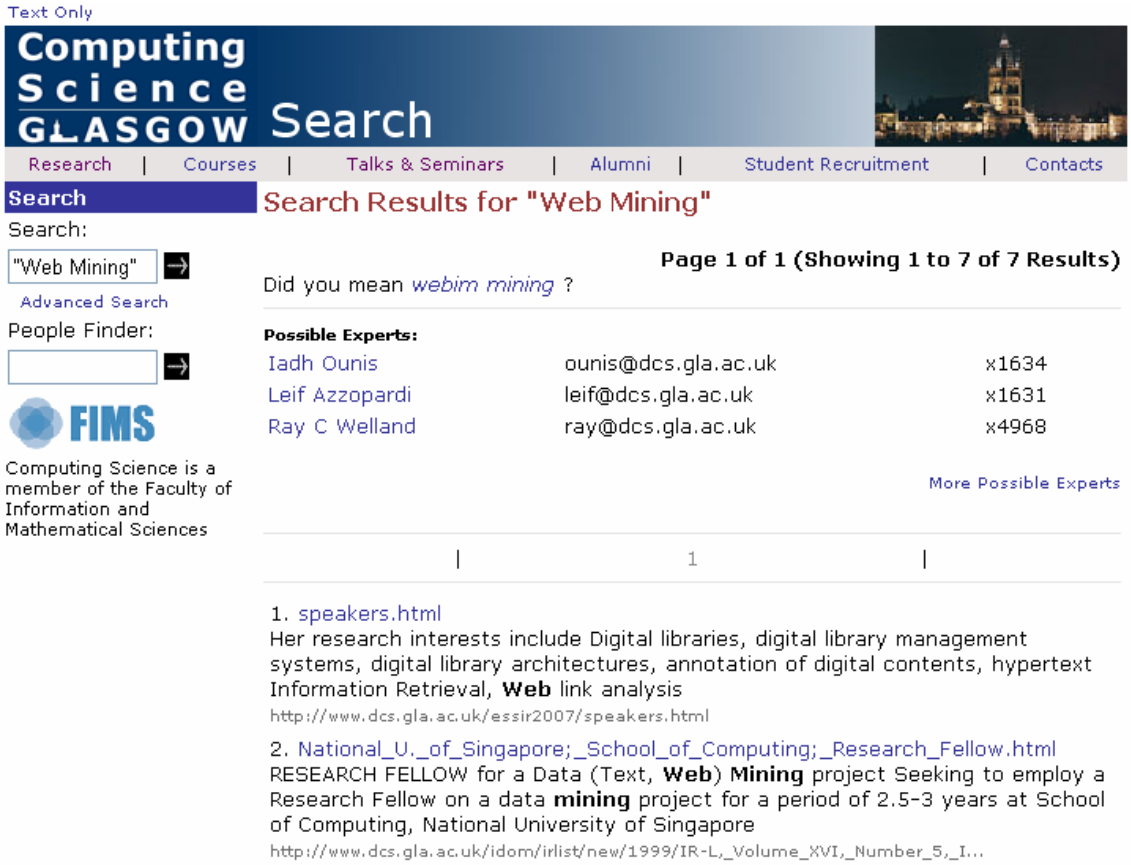


Figura 9: Imagem gerada pelo Terrier em 11/04/2008

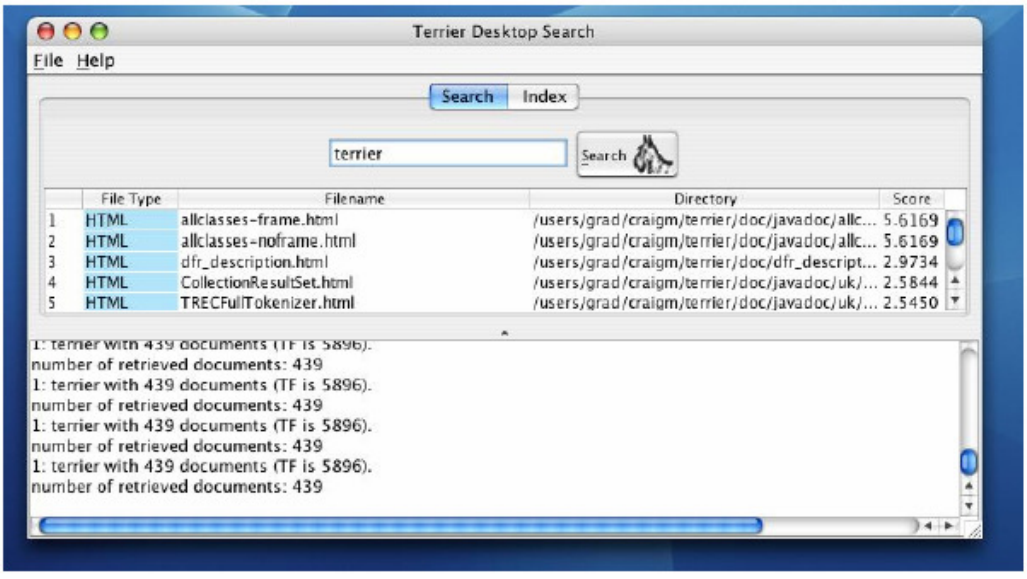


Figura 10: Imagem gerada pelo Terrier Desktop Search.

4.4 Análise dos resultados

Analisando as máquinas de busca Google, WebCrawler e Medline, nota-se que todas fornecem uma ajuda significativa aos usuários no processo de recuperação de informação, sendo que o Google e o Medline devolveram um número grande de páginas e praticamente todas as 10 primeiras páginas foram relevantes. A tabela 1 mostra o resultado comparativo das três ferramentas com maior precisão.

	Google	**Medline**	**WebCrawler**	**Terrier**
Número de páginas Retornadas	341.000	4.917	50	200
Facilidade de uso	bom	bom	bom	bom
Precisão das 10 primeiras páginas	90%	100%	50%	100%
Interface Gráfica	boa	ótimo	razoável	razoável
Organização do resultado	tópicos	tópicos	tópicos	tópicos
Tempo de resposta	0,27s	Não informado (muito rápido)	Não fornecido (lento)	Não informado (rápido)

Tabela 1: Comparação entre máquinas de busca – organização por tópicos

Através da análise das 10 primeiras páginas encontradas pelo WebCrawler, pode-se observar que 50% dos resultados realmente foram sobre "Web Mining", mas as páginas devolvidas eram sites de software dessa área, e o que realmente era de interesse na pesquisa seriam artigos e tutorias sobre o tema. Somente uma das páginas era um tutorial sobre Web Mining. Assim sendo, pode-se concluir que só uma página era realmente relevante, ou seja, 10%.

Ao fazer uma analise minuciosa das 10 primeiras páginas encontradas pelo meta busca WebCrawler, observou-se que o que realmente era de maior interesse na pesquisa, não estava nas primeiras páginas do ranking e sim nas páginas seguintes.

Já o Medline teve um excelente resultado, alem da quantidade de documentos recuperados, todos os 10 primeiros documentos foram relevantes. Assim, este sistema pode ser considerado como uma excelente opção para usuários que tem interesse em pesquisas na área médica.

Quanto ao Google, esta é a máquina de busca mais estudada e conhecida de todos. Pode-se observar através de artigos acadêmicos que o Google não só almeja o lucro, pois está sempre ligado ao meio acadêmico. E através da análise, conclui-se que o Google tem um ótimo desempenho, pois recupera uma grande quantidade de documentos e praticamente todas as 10 primeiras páginas retornadas foram relevantes. Além disso, o Google oferece

um sistema de pesquisa avançada que permite aplicar filtros restringindo o tipo de arquivo, o idioma, a região, entre outros.

Já o Terrier, apresentou um ótimo desempenho, sendo que 100% dos resultados foram relevantes quando o tema da pesquisa foi a palavra “Terrier”, o que não aconteceu quando a palavra-chave foi “Web Mining”, pois como o Terrier é uma máquina de busca do site do Departamento de Computar Ciência na Universidade de Glasgow, fica bastante restrito o assunto a ser pesquisado.

5 CLUSTERIZAÇÃO DE PÁGINAS WEB

Métodos de categorização (clustering) (BERENDT et al. 2004) dividem um conjunto de objetos em subconjuntos, chamados de *clusters*, tal que os objetos pertencentes a um determinado cluster são similares e os que não pertencem a este cluster são diferentes. Alguns métodos são hierárquicos e constroem *clusters* de *clusters*. Os métodos de agrupamento são baseados em similaridades dos objetos.

Para Zanier (2006) a clusterização na Web é definida como o agrupamento de documentos de acordo com uma medida de similaridade computada, utilizando-se associações entre palavras e frases.

A figura 11 mostra o processo de clusterização (CORREA, 2003).

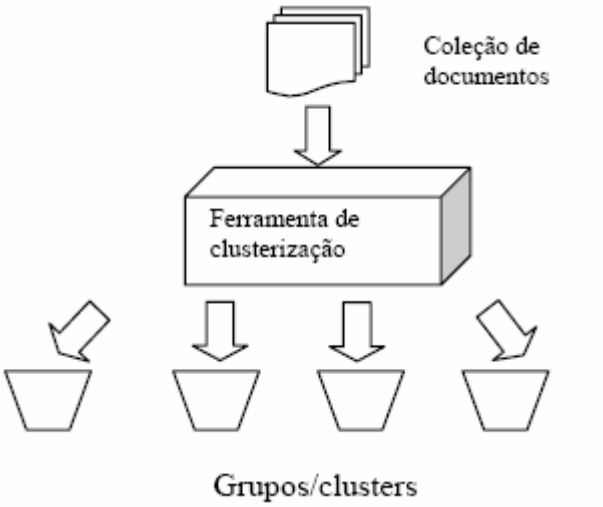


Figura 11: Processo de Clusterização

A idéia de distância também pode ser usada na tarefa não direcionada a Data Mining chamada *automatic cluster detection* (descoberta de agrupamento automático) (LINOFF e BERRY, 2001).

> O objetivo do *clustering* é formar grupos diferentes uns dos outros, mas não forçosamente disjuntos, contendo membros muito semelhantes entre eles. Ao contrário do processo de classificação que segmenta informação associando-lhe grupos já definidos, o *clustering* é uma forma de segmentar informação em grupos não previamente definidos. (CAMPOS, 2005 p. 34)

O objetivo do *cluster* é formar grupos de itens semelhantes. Em um banco de dados estruturado estes itens são registros que descrevem algo como um cliente ou um produto e na Web, os itens são documentos de hipertexto (LINOFF e BERRY, 2001).

Agrupar documentos (BERENDT et al., 2004) significa que um grande número de documentos é dividido em grupos com conteúdos similares. Isto é normalmente um processo intermediário para otimizar pesquisas e recuperar documentos. O agrupamento de documentos é caracterizado por características de documentos (palavra-chave ou combinação de palavras) e estas são exploradas para acelerar a recuperação ou para melhorar o desempenho de pesquisas por palavra-chave.

Há vários algoritmos que podem ser aplicados para agrupar dados textuais. Uma aproximação popular é agrupamento aglomerados. Cada página parte como único membro de seu próprio *cluster*. O algoritmo recursivo encontra os dois *clusters* mais semelhantes e os funde. Este processo continua até que todos os *clusters* sejam fundidos em um grande *cluster* que contenha todas as páginas. É evidente que um *cluster* que contém todas as páginas não é mais útil que uma coleção de *clusters* que contêm poucas páginas.

Há vários razões por querer encontrar dados agrupados, pois com um grupo de registros semelhantes, padrões podem ser visíveis, o que seria obscurecido em um grupo maior. (LINOFF e BERRY, 2001).

Na Internet, o *cluster* é um processo que poder ajudar os usuários a encontrar o que estão procurando. A formação de *cluster* é um tipo de Mineração de Conteúdo que organiza os resultados de forma a facilitar a navegação do utilizador pela lista de páginas devolvidas por um motor de busca, automatizando assim os processos, pois o usuário não mais necessitará de fazer uma procura exaustiva da página de seu interesse, o que significa um considerável ganho de tempo.

Serão apresentados a seguir alguns exemplos de máquinas de busca que organiza o resultado de uma pesquisa na web em pastas (*folders*) personalizados, um tipo de clusterização semi-automático.

5.1 Snaket

A p-rincipal dificuldade da máquina de busca é descobrir o que é relevante para o usuário. O conjunto de palavras-chave fornecido pelo pesquisador no momento da consulta pode resumir necessidades diferentes e talvez o resultado devolvido pelo motor de busca não seja realmente o que o usuário almeja.

Nos últimos tempos, tem surgido interesse comercial em ferramentas para recuperação de informação, que exponha os resultados de uma consulta em novas maneiras que possa ajudar o usuário em sua difícil tarefa de pesquisa. Como exemplo, podemos citar um software chamado *Snaket*, disponível em http://snaket.di.unipi.it/, que fornece os resultados de uma pesquisa em uma hierarquia de clusters que são nomeados por frases significativas retiradas do título e dos *snippets* de cada resultado, ou seja, do *abstract* e não do documento inteiro.

Assim, *snippets* que partilham as mesmas sentenças e falam de um mesmo tema devem, portanto, ser agrupados juntos num *cluster* e os *labels* (etiquetas) do cluster assume uma descrição mais geral seguido de uma descrição específica. Quando os *labels* dos vários *clusters* partilharem uma sentença mais abrangente, será encontrado o "pai" desses clusters, em termos hierárquicos.

Segundo Ferragina e Gulli (2008), o *Snaket* busca os *snippets* em 16 máquinas de pesquisa, na coleção de livros da Amazon, A9.com, nas notícias do Google News e nos blogs do *Blogline,* e constrói os *clusters* em resposta a consulta do usuário, organizando a informação recuperada em categorias. Nomeia os *clusters* com sentenças de comprimento variado e usa algumas funções de classificação que explora duas bases de conhecimento, devidamente construídas pela máquina no momento de pré-processamento para seleção das sentenças e para o processo de designação do *cluster*. Além disso, organiza os *clusters* em diferentes níveis hierárquicos.

Liu (2007) faz referência a dois motores de busca: *Vivisimo* e *Northern Light*, cujo objetivo desses pesquisadores é produzir uma taxonomia para ajudar na navegação e organizar os resultados da pesquisa em um pequeno número de *clusters* hierárquicos, categorizando assim, os resultados da busca.

Para exemplificar, foi feita uma busca por "Web Mining" no *Snaket* e pode-se visualizar na figura 12 que a ferramenta fornece o resultado da consulta em vários níveis

de detalhamento, de modo que o usuário não precisa perder tanto tempo vasculhando várias páginas do resultado, pois ele pode navegar através das pastas que são nomeadas inteligentemente de acordo com o conteúdo que elas possuem.

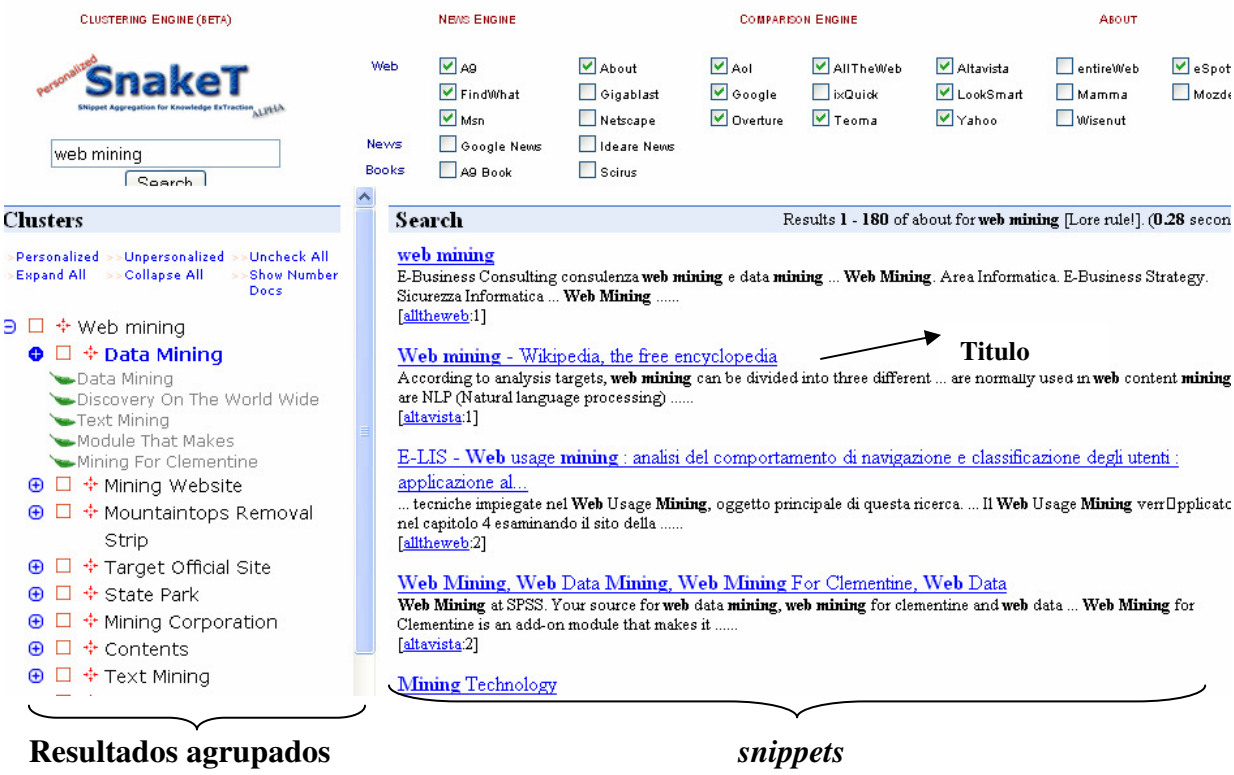


Figura 12: Imagem gerada pelo Snaket

5.2 Northern Light Search

A Northern Light[8] é uma máquina de busca que pesquisa base de documentos empresarias e organiza o resultado da pesquisa em grupos. Qualquer que seja a pesquisa, o resultado é apresentado em duas colunas. Os melhores documentos aparecem do lado direito, e as pastas que contem os resultados refinados da pesquisa, aparecem do lado esquerdo.

A figura 13 apresenta o resultado da busca pelas palavras “Web Mining” e “Web Data Mining” usando o sistema *Northern Light Search*. Os *folders* são interessantes, pois

[8] http://www.nlsearch.com

são rotulados com nomes inteligentes e significativos, o que facilita o processo de encontrar informações relevantes.

Figura 13: Imagem gerada pelo Northern Light Search

5.3 Clusty

O Clusty[9] é um meta busca da *Vivísimo Incorporated* que organiza os resultados da busca em *clusters*. Segundo Zanier (2006), o Clusty utiliza os resultados de outros mecanismos de busca, tais como o MSN Search, Asl.com, Gigablast, entre outros, para constituir sua base.

A figura 14 mostra a imagem gerada pelo Clusty em uma consulta usando a palavra-clave "Web Mining". Os resultados encontrados são agrupados em tópicos. Esse processo é realizado pelo algoritmo *Vivísimo Clustering Engine* (ZANIER, 2006) que subdivide os resultados em grupos (clusters) com base na semelhança do conteúdo.

[9] http://www.clusty.com

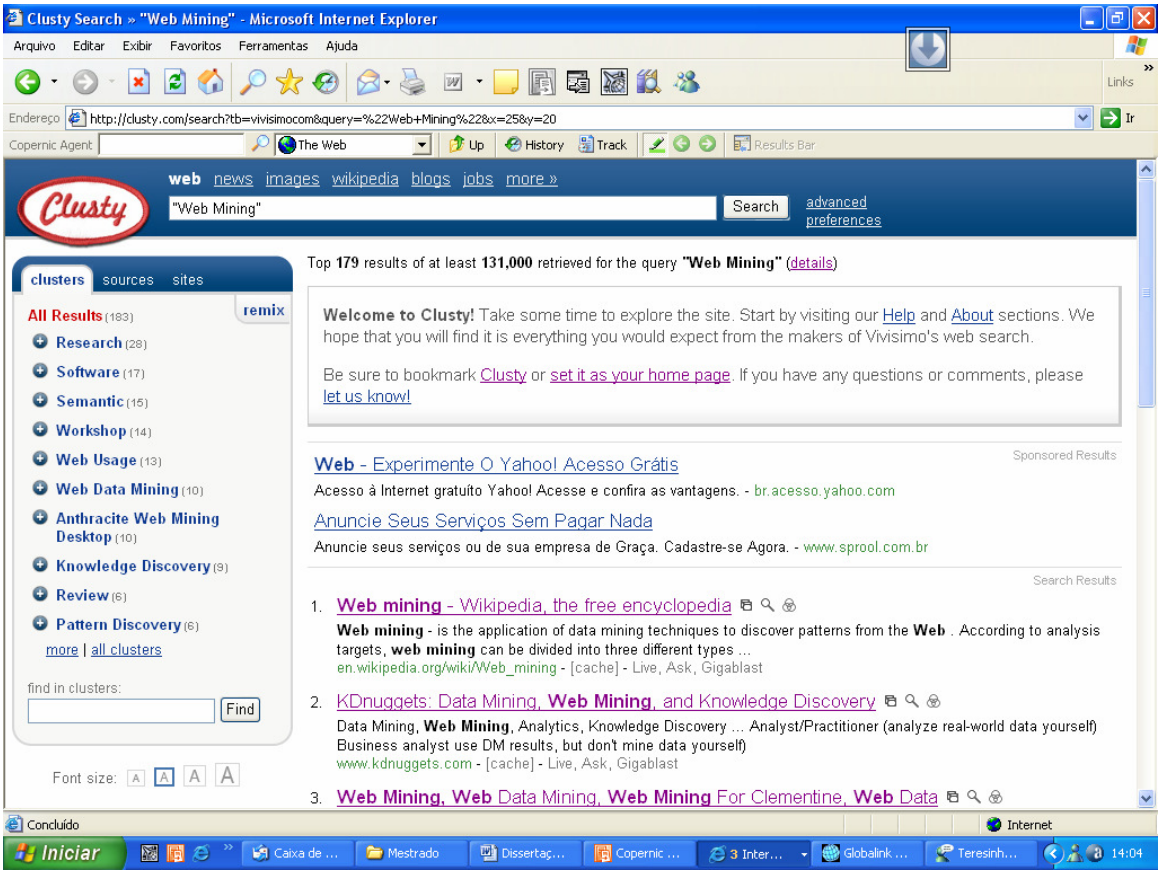


Figura 14: Imagem gerada pelo Clusty em 10/02/2008

5.4 Kartoo

Considerando mais uma das alternativas de obter conhecimento na Web, um serviço de busca muito interessante é o Kartoo[10] (MACEDO; FILHO, 2008), que apresenta não apenas páginas, mas também, o seu inter-relacionamento. Além de organizar os resultados em clusters, sua forma de apresentação em mapa conceitual interativo é muito mais significativa do que a simples listagem de páginas.

O Kartoo é um motor de busca que se distingue dos motores comuns pela sua forma de apresentação, que por sua vez pode ser explorada num sistema de pastas e sub-pastas. É um projeto francês, baseado na tecnologia *Flash*, *on-line* desde 2001, mas que no momento torna-se muito mais falado e divulgado.

[10] http://www.kartoo.com

Como motor de busca, tem menor potencialidade de um Google ou de um Yahoo, mas o seu conceito é interessante, permitindo um vasto leque de opções, incluindo a possibilidade de salvar o mapa para posterior utilização. Além disso, permite a impressão do mapa ou o seu envio por e-mail.

A figura 15 ilustra o resultado de uma pesquisa pelas palavras-chave "Web Mining" gerado pelo Kartoo. Observa-se que o conjunto de informações disponíveis é muito mais estruturado e rico que uma simples recuperação de páginas isoladas.

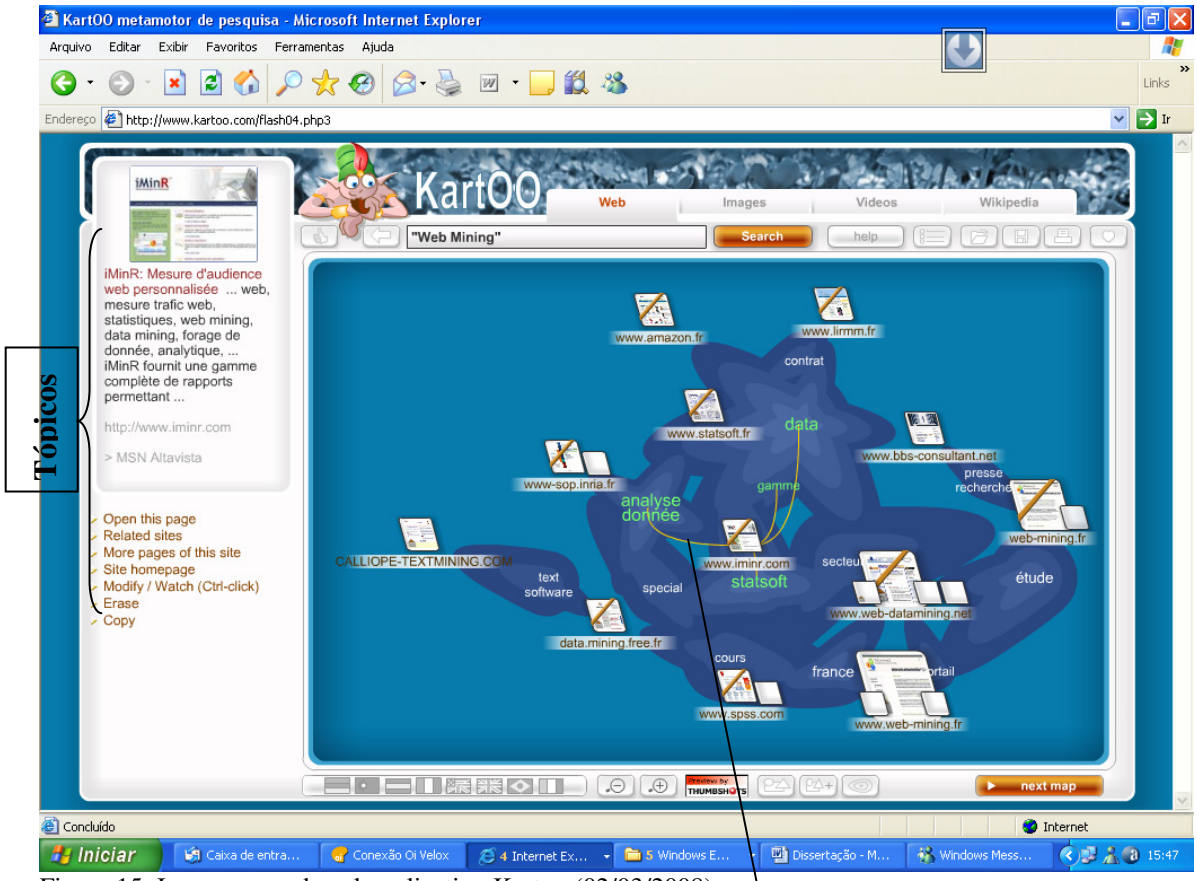


Figura 15: Imagem gerada pelo aplicativo Kartoo (02/03/2008).

5.5 MetaCrawler

O MetaCawler[11] é um exemplo (ZAMIR, ETZIONI, 1998) de um motor de busca que organiza os resultados em *clusters*.

[11] http://www.metacrawler.com/

A figura 16 mostra a imagem gerada pelo MetaCrawler em uma pesquisa pela palavra-chave “Web Mining”.

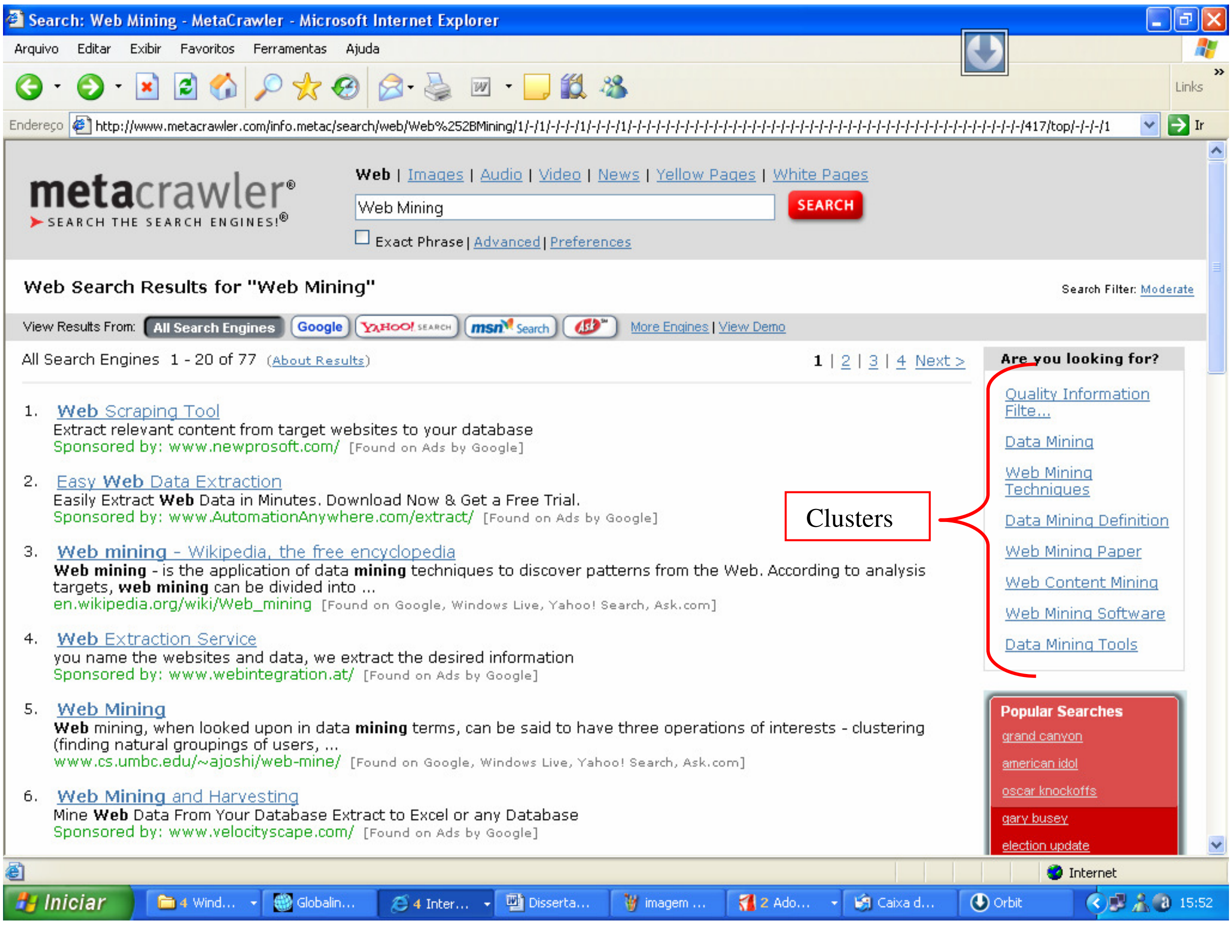


Figura 16: Imagem gerada pelo MetaCrawler

5.6 Grokker

Grokker[12] (MACEDO; FILHO, 2008) é uma nova forma de navegar na Internet de forma mais visual, que usa técnicas de *clustering* para organizar e agrupar o resultado das buscas.

A pesquisa através do Grokker, apresentada na figura 17, retornou onze círculos grandes de tamanhos diferentes, contendo em seu interior outros círculos com tamanhos variados. Cada um dos onze círculos grandes agrupa os sites com assuntos relacionados, cujo tamanho é proporcional ao número deles.

[12] http://www.grokker.com/

Portanto, essa ferramenta de busca informa ao pesquisador os *clusters* por assunto, dentro do tema da pesquisa, facilitando a busca por grupo de temas relacionados. Por exemplo: caso o pesquisador queira maiores informações sobre Mineração de Conteúdo Web, o *cluster*, denominado pelo Grokker de "*Web Content Mining*", será eleito para as análises. Ao clicar no círculo correspondente, esse se expande para permitir a melhor visualização de seu conteúdo, formado por pequenos círculos que representam os sites relacionados ao assunto em questão.

Conseqüentemente, passando-se o cursor do mouse sobre o circulo menor, tem-se uma breve descrição do site correspondente, tais como: endereço, número de documentos relacionados ao assunto da pesquisa (no caso *Web Mining Content*), etc, o que facilita de sobremaneira o trabalho do pesquisador, pois esse não precisa acessar a página web para saber seu conteúdo, ficando esta análise apenas para os sites que realmente mereçam atenção.

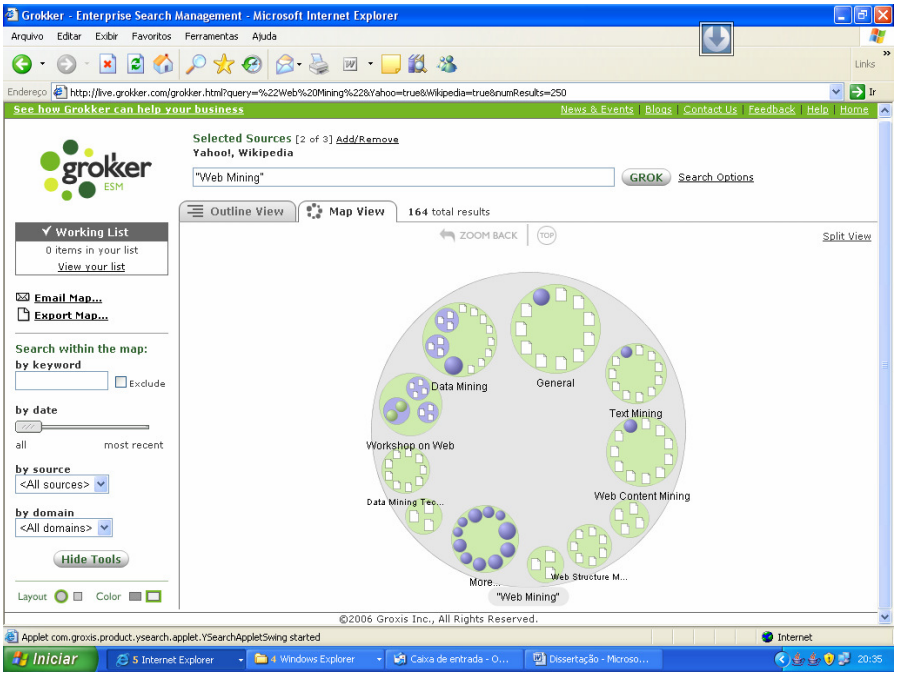


Figura 17: Imagem gerada pelo aplicativo Grokker (02/03/2008).

> Cabe lembrar que a interpretação dos resultados da pesquisa é que irão definir a maior ou menor utilização desse tipo de ferramenta, e a sua visualização como uma mola propulsora da descoberta de relações entre o objeto da pesquisa e temas até então não visualizados pelo pesquisador. (MACEDO; FILHO, 2008, p.12)

5.7 Copernic

Copernic[13] é um software proprietário da *Copernic Technologies*, subdividido nos seguintes programas: *Copernic Agent Basic*, *Copernic Agent Personal*, *Copernic Agent Professional*, *Copernic Summarizer* e *Copernic Desktop Search*.

Copernic Agent Basic é um software livre que reúne várias ferramentas de busca em um só programa. O software tem o poder de consultar uma variedade de fontes de informação simultaneamente. A partir de um único ponto de acesso, os usuários podem criar simples ou avançada consultas e obter resultados rapidamente com qualidade e relevância.

O programa oferece flexibilidade e interface amigável. Suas ferramentas são integradas a interface do brawser. Além de permitir ao usuário obter resultados facilmente, o *Copernic* remove resultados duplicados e elimina todos os links inválidos.

Dentre as principais funções do *Copernic* destaca-se:

- Detectar e remover automaticamente links com problemas;
- Analisar páginas web automaticamente;
- Salvar páginas para navegação off-line, filtrar e pesquisar nestas páginas;
- Detectar páginas duplicadas com diferentes endereços;
- Extrai a data da última modificação das páginas;
- Extrair os conceitos principais das páginas utilizando o Copernic Summarizer.

Infelizmente algumas funções não puderam ser testadas, pois o Copernic é um software proprietário e muitas destas funções estão disponíveis apenas na versão Copernic Agent Professional.

A figura 18 mostra o resultado de uma pesquisa passando para a ferramenta a palavra-chave “Web Mining”.

[13] http://www.copernic.com/

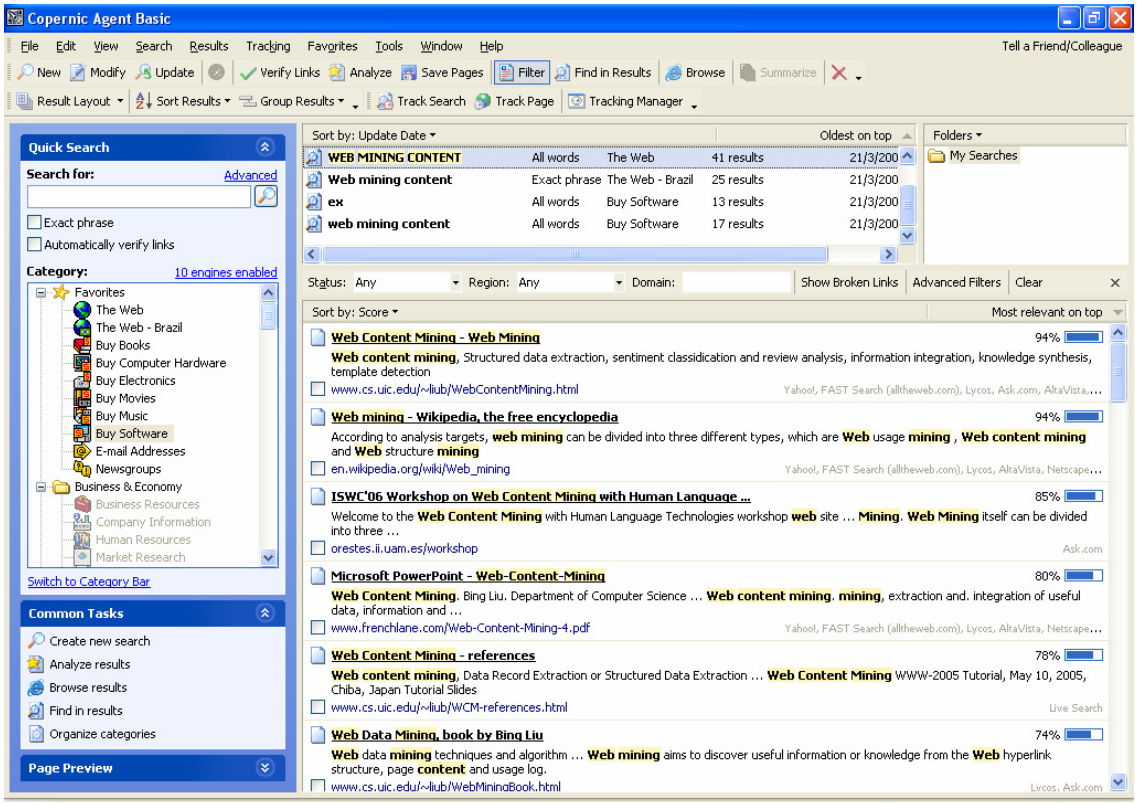


Figura 18: Imagem gerada pelo Copernic em 15/02/08

Compatibilidade do Copernic:

Windows 95, Windows 98, Windows ME, Windows 2000, Windows NT, Windows XP.

5.8 Análise dos resultados

A tabela 2 apresenta uma comparação entre os resultados fornecidos pelas máquinas de busca apresentadas, sendo fornecido a palavra "Web Mining" para a consulta.

	SNAKET	Northern	Clusty	Kartoo	MetaCrawler	Grokker	Copernic
Número de páginas Retornadas	180	314	179	28300	77	164	36
Facilidade de uso	sim	sim	sim	sim	sim	sim	sim
Precisão das 10 primeiras páginas	80%	50%	70%	70%	80%	100%	50%
Interface Gráfica	boa	boa	boa	ótima	boa	ótima	ótima
Organização do resultado	cluster	cluster	cluster	cluster	Cluster	cluster	cluster
Tempo de resposta	28s	Não informado (rápido)	Não informado (rápido)	Não informado (não muito rápido)	Não informado (rápido)	Não informado (não muito rápido)	Não informado (rápido)

Tabela 2: Avaliação dos resultados das máquinas de busca – organização por cluster

Analisando os resultados do *Snaket*, observou-se que 80% deles foram relevantes, apesar de duas páginas que foram retornadas eram páginas de empresa de software desta área. Já os dois últimos documentos eram páginas iguais, uma fornecida pelo site de busca *Alta Vista*[14] e a outra pelo *Alltheweb*[15]. Isso foi um ponto negativo encontrado pelo *Snaket*, pois já que o mesmo busca informações relevantes em outros sites de busca, o ideal seria que o mesmo usasse um filtro para evitar documentos repetidos. Em contrapartida, o *Snaket* teve um excelente desempenho e pouco tempo de resposta. Quanto aos clusters, o Snaket subdividiu o resultado em varias categorias: *Web Mining, Text Mining, Mining Corporation, Contents*, etc. Foram analisadas as páginas pertencentes aos clusters *Web Mining* e *Text Mining*, dos documentos analisados dessas duas categorias, todos foram relevantes, o que caracteriza ainda mais o bom desempenho do *Snaket*.

Os resultados encontrados são organizados em clusters que são nomeados com palavras significativas, o que facilita muito o processo de busca de informações relevantes pelo usuário.

O *Northern Light* não apresentou um bom resultado, isso se deve ao fato de pesquisar apenas em bases empresarias. Apesar de organizar os resultados em categorias, as pastas retornadas não se tratavam de tópicos relevantes em relação ao tema consultado. Quanto aos clusters, o *Northern Light* subdividiu o resultado em várias categorias: *Software engineering; Computer Insdustry, IT Technologies, Business Issues, Data Mining*, etc. Analisando cada clusters, a maioria das páginas retornadas em cada cluster foram relevantes, por exemplo, o cluster "Data Mining" retornou 22 documentos, sendo que 20 desses sites realmente são pertinentes ao nome do cluster, pois se tratavam de assuntos referentes a Data Mining.

Quanto ao software Clusty, é possível observar que os rótulos dos clusters são bastante expressivos, o que ajuda muito o usuário, minimizando o tempo de busca, já que o mesmo poderá consultar apenas os documentos pertencentes ao cluster que mais satisfaz o seu interesse.

Na análise dos resultados do Clusty, observou-se que 70% são relevantes, já que uma página praticamente não havia conteúdo importante referente ao assunto e duas páginas apareceram repetidas. Entre as páginas relevantes, duas dessas eram páginas de software desta área, que apesar de estar de acordo com a pesquisa, não contém definição

[14] http://www.altavista.com.br
[15] http://www.alltheweb.com/

ou descrição do tópico pesquisado. Quanto aos clusters, o Clusty subdividiu o resultado em várias categorias: *Research, Software, Semantic, Web Usage, Web Data Mining, Pattern Discovery, Web Structure Mining,* etc. Analisando os resultados pertencentes em todos os clusters, observou-se que o Clusty teve um ótimo desempenho. Tendo como exemplo, o cluster *Software*, dos 20 sites pertencentes a essa categoria, todos foram relevantes.

O Kartoo por sua vez, apresenta como principal característica a apresentação por mapa interativo, apresentação do inter-relacionamento entre os sites e a quantidade de documentos retornados. Quanto à relevância dos resultados, a análise das 10 primeiras páginas não apresentou um bom resultado. Apesar de 70% das páginas estarem de acordo com a pesquisa, 50% das páginas analisadas eram sobre softwares ou empresas que prestam serviços nesse campo. Já uma análise minuciosa dos clusters, estes apresentaram um bom resultado, analisando os clusters *Text and Web Mining* e *Data Mining Techniques*, cada cluster retornou 12 sites, sendo que todos os documentos da cada categoria foram relevantes.

Avaliando os resultados devolvidos pela meta-busca *MetaCrawler*, das 10 primeiras páginas analisadas, 80% dos documentos eram relevantes. Porém das 8 páginas que se tratavam de Web Mining, três dessas eram sobre softwares, o que não era de muito interesse na pesquisa. Quanto aos clusters, o WebCrawler subdividiu o resultado em várias categorias: *Data Mining, Web Content Mining, Web Mining Software, Web Mining Paper*, etc. Através de uma análise mais detalhadas dos clusters, notou-se que realmente os documentos pertencentes a cada categoria estava subdivido corretamente, processo que facilita muito aos usuários na busca por documentos relevantes.

O Grokker, além de apresentar uma interface gráfica muito interessante, com os rótulos dos clusters especificando claramente o tipo de conteúdo que contém, ele apresentou 100% dos resultados da pesquisa complacente, apesar de 30% dessas páginas serem sobre software, foi a que apresentou melhor resultado dentre as diversas máquinas de busca analisadas até o momento, pois trouxe conceitos, artigos, tutoriais e etc sobre o tema em questão. Quanto a subdivisão por categorias, o Grokker trouxe vários clusters etiquetados com nomes significativos. *Data Mining, Web Content, Business Intelligence, Web Usage, Web Semantic*, etc são alguns exemplos dos clusters retornados pelo Grokker. Ao analisar o cluster *Web Content*, dos 13 documentos pertencentes a essa categoria, 11 foram relevantes e o cluster Text Mining, dos 17 documentos pertencentes a essa categoria 16 foram relevantes.

Para concluir, o último meta-busca analisado foi o Copernic, que apesar de não ter apresentado um dos melhores desempenhos, pode ser considerado uma boa opção para os usuários, devido as diversas funcionalidades que o mesmo oferece. Um dos motivos de apenas 50% das dez páginas analisadas serem consideradas relevantes foi o fato de que o Copernic apresentou quatro páginas repetidas e estas foram consideradas não relevantes. Quanto aos clusters, não foi possível a análise dos mesmos, pois a opção de separar o resultado em categorias só é permitida na versão do *Copernic Agent Professional*.

6 EXTRAÇÃO DE INFORMAÇÃO

A extração de informação presta um grande serviço à mineração da Web. Denomina-se extração de informação à tarefa de identificar fragmentos específicos que constituem o núcleo semântico de um documento em particular e construir modelos de representação da informação (conhecimento) a partir do documento (PAL, 2002).

Com os documentos recuperados, a tarefa seguinte é transformar esses dados de forma que algoritmos de mineração e aprendizagem de máquina possam ser aplicados de forma efetiva para a obtenção ou extração de elementos pertencentes a determinados documentos.

> Enquanto que a "Information Retrieval" pretende extrair documentos importantes, entre um universo de documentos possíveis, a "*Information Extraction*" pretende identificar elementos relevantes no interior de determinados documentos, os quais já sabemos que contém a informação que nos interessa. Os elementos relevantes extraídos serão depois armazenados em alguma estrutura previamente definida, por exemplos numa tabela de uma Base de Dados. (CORDEIRO, 2003, p.17)

Para Santos (2002), os sistemas de Extração de Informação apresentam algumas fases, apesar de existirem muitas variações de sistema para sistema. Segue abaixo as principais funções executadas em um sistema de extração de informação. (CARDIE apud SANTOS, 2002).

a) Cada texto de entrada é primeiramente dividido em sentenças e palavras, numa etapa de *tokenização*[16] e *tagging* (colocação de etiquetas). Muitos sistemas etiquetam cada palavra com a respectiva classe gramatical e, possivelmente, classes semânticas neste ponto do processo;

b) A fase de análise da sentença compreende um ou mais estágios de análises sintáticas ou *parsing* que, juntas, identificam grupos de substantivos, grupos de verbos,

expressões preposicionais e outras estruturas simples. Em alguns sistemas, o *parser* também localiza, num nível superficial, sujeitos e objetos diretos e identifica conjunções e outras expressões complexas. Em algum ponto, antes, durante ou depois dos passos principais da análise sintática, o sistema de extração de informação também procura e etiqueta entidades semânticas relevantes ao tópico de extração;

c) A fase de extração é o primeiro componente do sistema que é específico para o domínio de aplicação. Durante a fase de extração, o sistema identifica relações entre entidades relevantes no texto;

d) O trabalho principal na fase de *merging* é a resolução co-referenciada ou resolução anafórica: O sistema examina cada entidade encontrada no texto e determina se tal entidade se refere a uma entidade já existente ou se ela é nova e deve ser adicionada ao nível de discurso do sistema que representa o texto;

e) As inferências no nível de discurso feitas durante a fase anterior, isto é, de *merging*, auxiliam a fase de geração de gabaritos, a qual determina o número de eventos distintos no texto, mapeia os itens individuais de informação extraídos de cada evento e produz gabaritos de saída. Inferências sobre o domínio de aplicação específico podem também ocorrer durante a geração de gabaritos.

Os métodos para extração da informação geralmente envolvem a escrita de código específico, popularmente chamados de *wrappers,* responsáveis pelo mapeamento de documentos para algum modelo de representação do conhecimento. O problema é que para cada documento da Web temos que escrever um código específico, tornando o trabalho manual. Como os documentos da Web não possuem uma semântica agregada às informações que contém, e nem mesmo um padrão de como apresentar essas informações ao usuário, temos que aprender acerca da estrutura individual de cada documento e escrever código para essa estrutura em particular.

Vários métodos foram desenvolvidos para a extração de informação tanto em documentos não-estruturados[17] quanto em semi-estruturados. (KUSHMERICK, 1997)

É importante salientar a diferença entre as fases de recuperação e extração de informação. As técnicas de extração de informação buscam derivar conhecimento de

[16] Processo que consiste em identificar tokens, ou palavras, em um texto (SANTOS, 2002)

[17] Documentos não-estruturados são dados que não possuem nenhuma estrutura, tais como um textos livres, uma imagem, um vídeo, já os documentos semi-estruturados possuem alguma estrutura como, por exemplo, as páginas de internet e documentos XML. Têm-se ainda os dados estruturados, tais como aqueles presentes nos bancos de dados relacionais.

documentos recuperados segundo a forma como um documento está estruturado e representado enquanto as técnicas de recuperação de informação visualizam o documento apenas como um conjunto de palavras (PAL, 2002).

Na área de extração de informação, a principal preocupação é saber extrair elementos textuais, que poderão ser palavras ou pequenas expressões, consideradas importantes, a partir de documentos. Os elementos relevantes extraídos serão posteriormente armazenados sob uma forma estruturada, permitindo assim, um rápido e eficiente acesso à informação.

Para Konchad (2005), os sistemas de extração geralmente convertem textos não estruturados em um formato que possa ser enviado para uma tabela em um banco de dados ou para uma planilha do Excel. Informações úteis como nome de pessoas, telefones, emails, data, preço de produtos, etc, mencionados no texto podem ser extraídos sem uma profunda compreensão do texto. Métodos simples de reconhecimento de padrão ou algoritmo de aprendizado de máquina são suficientes para extrair informações de textos.

6.1 Ferramentas para Extração de Dados na Web

6.1.1 HTML *Parsing*

Baseado nos exemplos de Loton (2002) será apresentado um *HTML parser,* conhecido também como *Wrapper,* baseado nas implementações do Java 2 SDK.

Tendo estabelecido o tipo de conteúdo que se deseja trabalhar, procura-se descobrir a estrutura e o arranjo dos elementos HTML ou XML.

Para simplificar, será assumido que o *HTML parser* será utilizado para todo conteúdo XML e HTML, pois o código genérico *WebParserWrapper* entende o tipo de conteúdo e invoca o *parser* correto para esse conteúdo. (LOTON, 2002).

A figura 19 mostra o exemplo de uma página que poderia ser analisada, apesar de simples, essa página contém bastante complexidade para demonstrar alguns princípios importantes.

Preços Promocionais

Computador	COMPUTADOR INTEL CELERON D331	799
	COMPUTADOR INTEL CELERON 420	899
	COMPUTADOR INTEL CELERON 420	1299
Notebook	NOTEBOOK POSITIVO MOBILE V52	1498
	NOTEBOOK HP COMPAQ C710 INTEL CELERON M 530	1599

Figura 19: Página HTML baseada no exemplo de Loton (2002).

A figura 20 mostra o código fonte da página HTML da figura 19 apenas para distinção das tags HTML, que servem para determinar a estrutura do conteúdo do texto. As tags <table>, <tr> e <td> são definidas como tags de estruturas enquanto a tag <b> (negrito), por exemplo, não é.

```
<html>
<body>
Preços Promocionais
<table BORDER=1 CELLSPACING=2 CELLPADDING=0 WIDTH="70%" bgcolor="#C0FFFF">
    <tr>
        <td VALIGN=TOP>Computador</td>
    <td>
            <table BORDER=1 WIDTH="100%" >
                <tr>
                    <td>COMPUTADOR INTEL CELERON D331 </td>
                    <td ALIGN=RIGHT><b>799</b></td>
                </tr>
                <tr>
                    <td>COMPUTADOR INTEL CELERON 420</td>
                    <td ALIGN=RIGHT><b>899</b></td>
                </tr>
                <tr>
                    <td>COMPUTADOR INTEL CELERON 420</td>
                    <td ALIGN=RIGHT><b>1299</b></td>
                </tr>
            </table>
        </td>
    </tr>
    <tr>
        <td VALIGN=TOP>Notebook</td>
            <td>
            <table BORDER=1 WIDTH="100%" >
                <tr>
                    <td>NOTEBOOK POSITIVO MOBILE V52</td>
                    <td ALIGN=RIGHT><b>1498</b></td>
                </tr>
                <tr>
                    <td>NOTEBOOK HP COMPAQ C710 INTEL CELERON M 530 </td>
                    <td ALIGN=RIGHT><b>1599</b></td>
                </tr>
            </table>
        </td>
    </tr>
</table>
</body>
</html>
```

Figura 20: Código fonte da Página HTML apresentada na figura 19.

O JDK tem embutido um conversor HTML. Assim, ao invés de implementar ou comprar um, pode-se usar as classes para conversão que estão dento do pacote *javax.swing*.

No final do processamento do *HTML parsing,* tem-se um vetor de elementos conforme mostrado na figura 21:

```
.text[0]            Preços Proporcionais
.table[0].@bgcolor[0]                   #COFFFF
.table[0].@cellpadding[0]               0
.table[0].@width[0] 70%
.table[0].@cellspacing[0]               2
.table[0].@border[0]                    1
.table[0].tr[0].td[0].@valign[0]        top
.table[0].tr[0].td[0].text[0]           Computador
.table[0].tr[0].td[1].table[0].@width[0]                    100%
.table[0].tr[0].td[1].table[0].@border[0]                   1
.table[0].tr[0].td[1].table[0].tr[0].td[0].text[0]          Computador Intel Celeron D33
.table[0].tr[0].td[1].table[0].tr[0].td[1].@align[0]        right
.table[0].tr[0].td[1].table[0].tr[0].td[1].text[0]          799
.table[0].tr[0].td[1].table[0].tr[1].td[0].text[0]          Computador Intel Celeron 420
.table[0].tr[0].td[1].table[0].tr[1].td[1].@align[0]        right
.table[0].tr[0].td[1].table[0].tr[1].td[1].text[0]          899
.table[0].tr[0].td[1].table[0].tr[2].td[0].text[0]          Computador Intel Celeron 420
.table[0].tr[0].td[1].table[0].tr[2].td[1].@align[0]        right
.table[0].tr[0].td[1].table[0].tr[2].td[1].text[0]          1299
.table[0].tr[1].td[0].@valign[0]        top
.table[0].tr[1].td[0].text[0]           Notebook
.table[0].tr[1].td[1].table[0].@width[0]                    100%
.table[0].tr[1].td[1].table[0].@border[0]                   1
.table[0].tr[1].td[1].table[0].tr[0].td[0].text[0]          Notebook positivo Mobile V52
.table[0].tr[1].td[1].table[0].tr[0].td[1].@align[0]        right
.table[0].tr[1].td[1].table[0].tr[0].td[1].text[0]          1498
.table[0].tr[1].td[1].table[0].tr[1].td[0].text[0]          Notebook HP Compaq C710 Intel M530
.table[0].tr[1].td[1].table[0].tr[1].td[1].@align[0]        right
.table[0].tr[1].td[1].table[0].tr[1].td[1].text[0]          1599
```

Figura 21: Resultado HTML *parsing*

6.1.2 FilterViewer

Para extrair alguma parte do conteúdo da página, será usado o *FilterViewer* (LOTON, T 2002) que foi implementado para mostrar o conteúdo da estrutura de marcação HTML ou XML. O *FilterViewer* invoca o arquivo *HTML Parsing* visto no item 6.1.1 para descobrir a estrutura e o arranjo dos elementos HTML.

A figura 22 mostra o resultado do FilterViewer. Foi fornecido para o campo URL o exemplo de página HTML apresentado na figura 19 e não foi usado filtro, portanto o campo FILTER foi preenchido com asterisco (*).

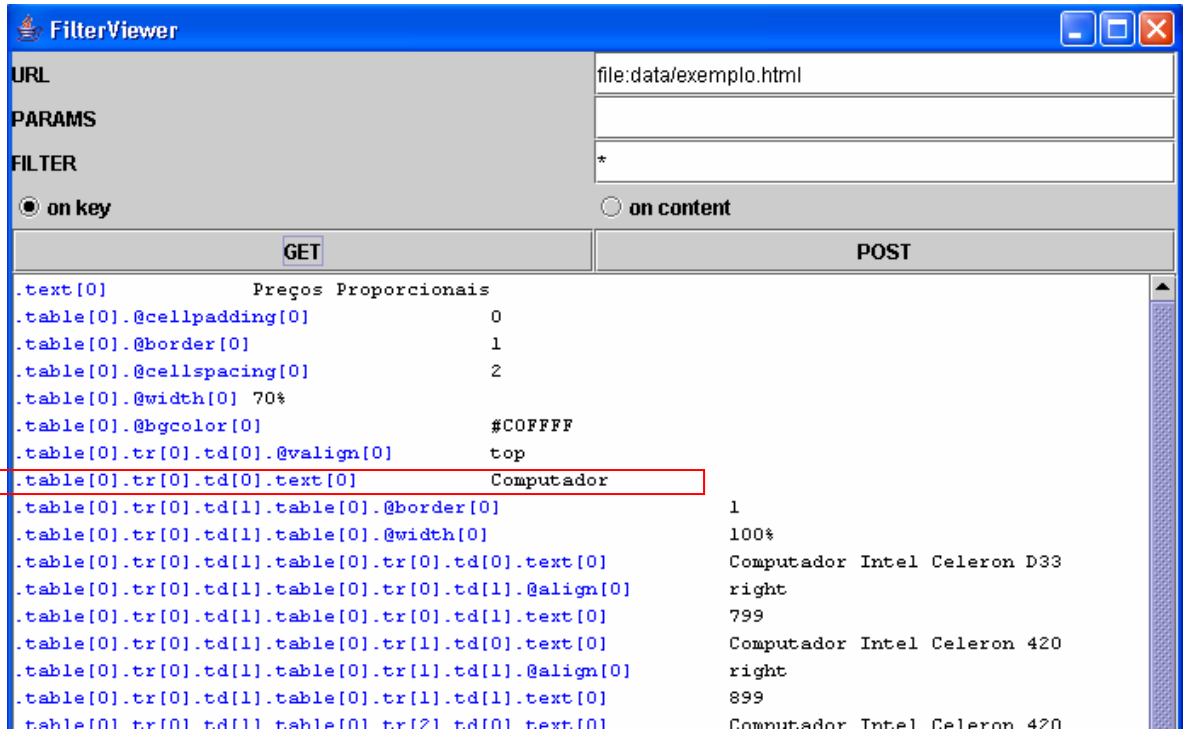


Figura 22: Imagem gerada usando o *FilterViewer* sem uso de filtro

Caso seja necessário extrair apenas a palavra “computador” do conteúdo do site, deve ser usado no filtro o vetor “.table[0].tr[0].td[0].text[0]”. Mas isso não seria tão interessante, talvez fosse mais importante extrair um relatório da coluna com os nomes dos computadores e da coluna dos preços, conforme exemplificado na figura 23.

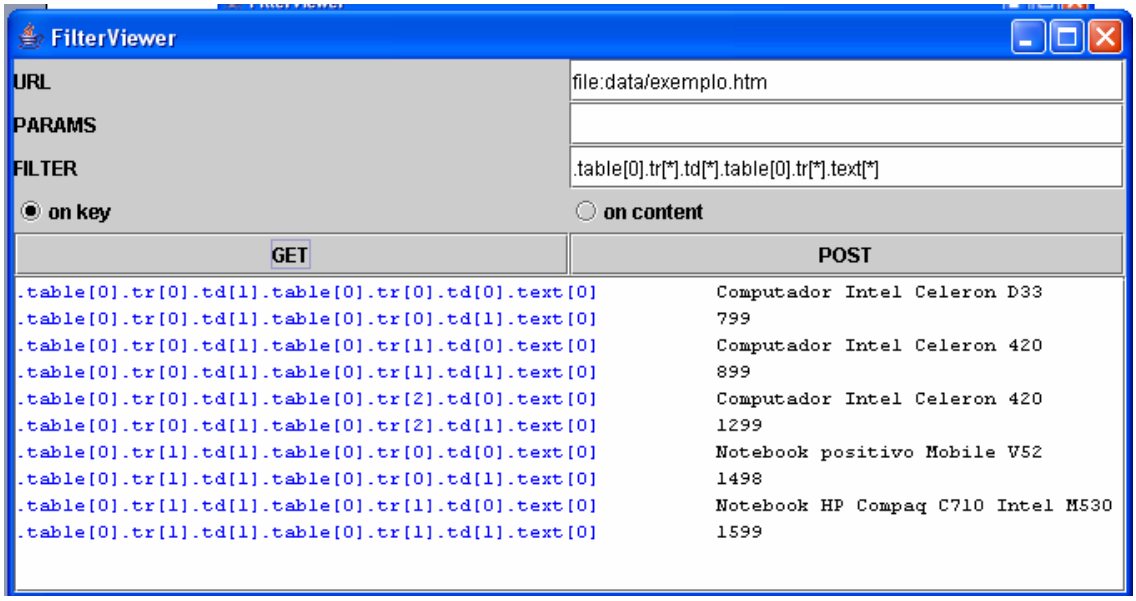


Figura 23: Imagem gerada usando o *FilterViewer* com filtro

6.1.3 SQLGUI

Conforme exemplificado anteriormente, quando se usa o *FilterViewer,* obtém-se um conjunto de elementos pertencentes às linhas e/ou colunas de uma tabela. Com o *SQLGUI* (LOTON, 2002) será permitido aplicar filtros para selecionar elementos em colunas fazendo restrições conforme é feito usando SQL em Banco de Dados Relacional.

A figura 24 ilustra o resultado gerado pelo *SQLGUI*. Observa-se que foram selecionados apenas os nomes dos computadores, não sendo apresentado os notebooks, isso se deve a restrição utilizada para selecionar somente os elementos da coluna que contenha a palavra "computador".

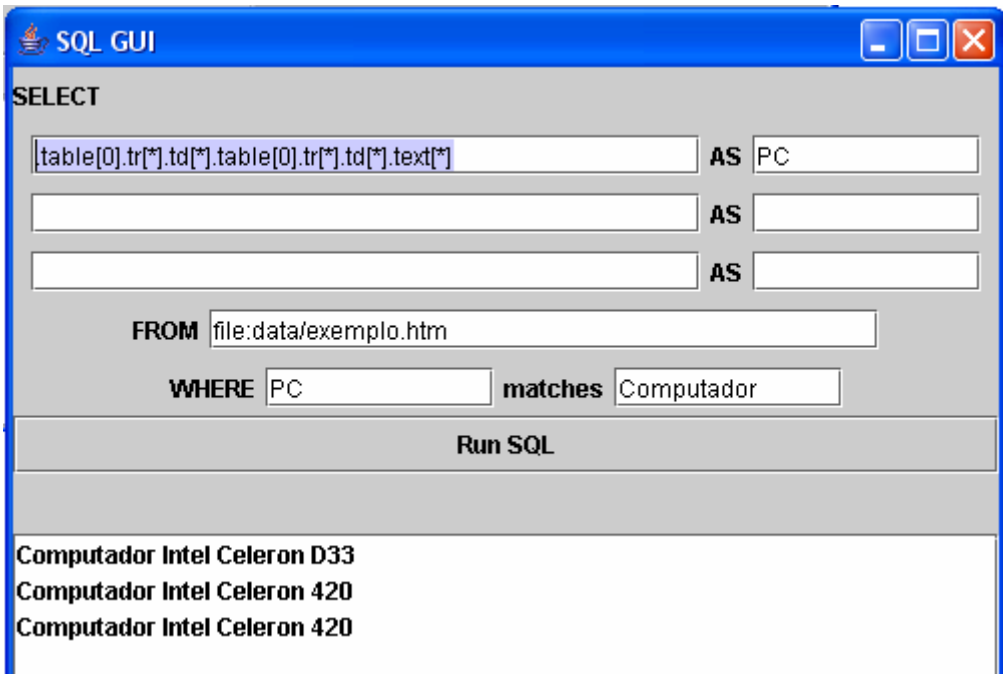


Figura 24: Figura gerada pelo *SQLGui* baseada no exemplo de LOTON (2002), página 99

6.1.4 Web Data Extractor 6.0

É um poderoso extrator de dados da web. Extrai URLs, títulos, textos, e-mails, telefones e fax de websites. Possui vários filtros para restringir a extração, como filtro de URLs, tamanho de arquivos, data modificada, etc e todos os dados são salvos diretamente no HD em arquivo texto.

É um software proprietário com Licença livre por 15 dias, disponível em: http://www.webextractor.com/ ou http://www.rafasoft.com/.

Abaixo, as figuras do *Web Data Extractor*, sendo que a primeira apresenta as várias opções de extração, a segunda, o resultado extraído e a última, as opções de filtro.

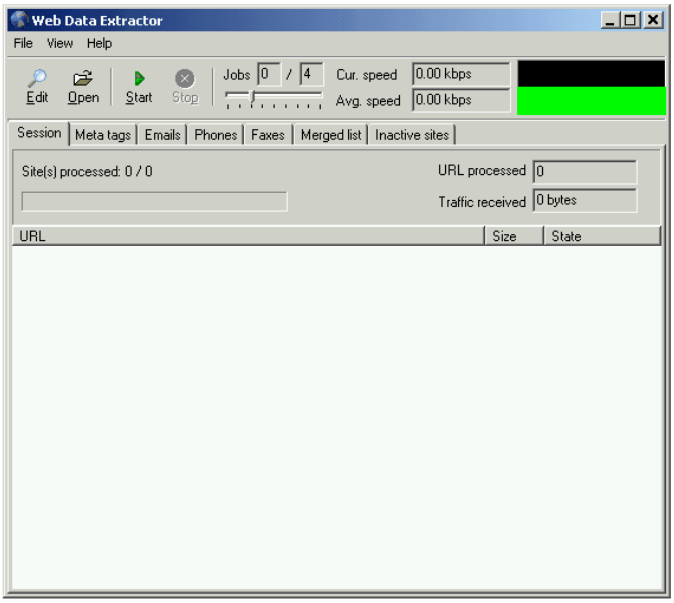


Figura 25: Figura do Web Extractor exemplificando as opções de extração

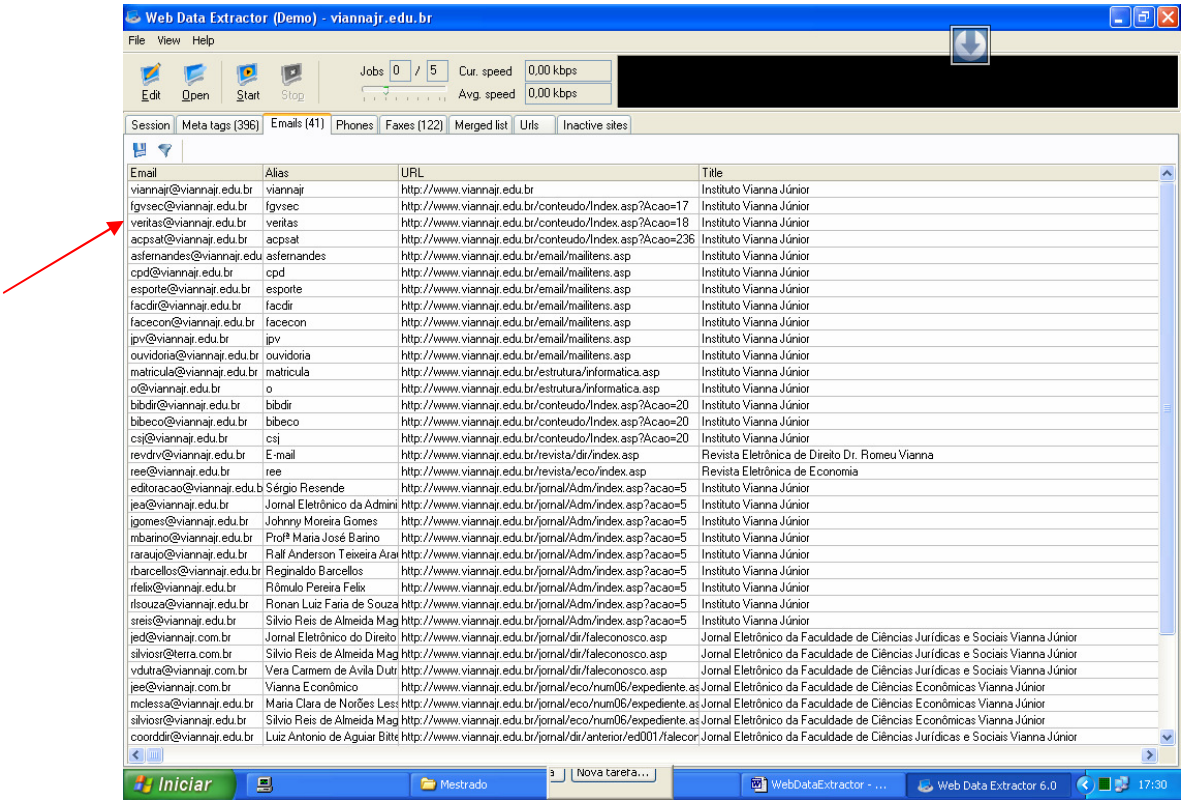


Figura 26: Resultado de busca por e-mail no site http://www.viannajr.edu.br usando o Web Data Extrator

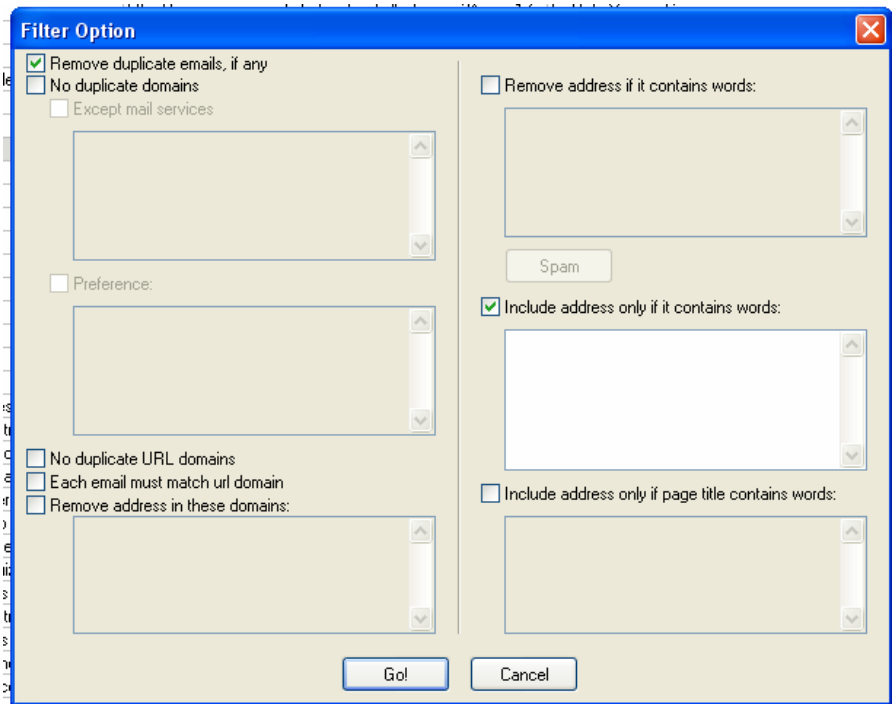


Figura 27: Opções de filtro do Web Data Extrator

6.1.5 Web Content Extractor

Web Content Extractor é o um software proprietário, usado para extração de dados da web e que permite criar padrões de dados e regras de rastreamento, ou seja, permite a especificação de regras para um determinado domínio, que irão identificar padrões nos textos, que determinarão as informações que se pretende extrair.

É de grande valia para o usuário, imagine, por exemplo, que o usuário tenha necessidade de coletar as informações sobre todos os produtos de um site (nomes, descrições, preços, etc), o sistema pode fazer isso de uma forma praticamente automática. Os dados extraídos podem ser exportados para uma variedade de formatos, incluindo Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL, MySQL ou scripts. Essa variedade de formatos de exportação permite processar e analisar os dados extraídos.

As figuras apresentadas a seguir mostram as telas do *Web Content Extractor*. A figura 28 mostra o processo de seleção de dados, a figura 29 mostra o processo de extração de padrões e a figura 30 mostra o resultado da extração exportado para o Excel.

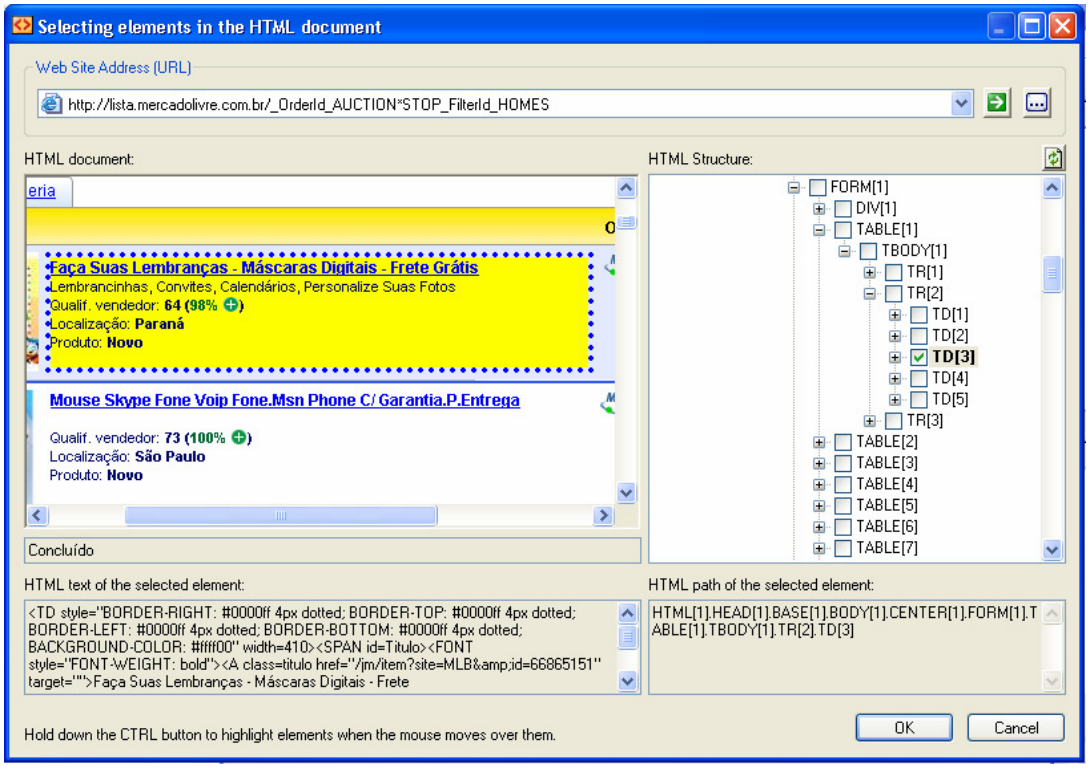


Figura 28: Figura gerada pelo *Web Content Extractor* na seleção dados

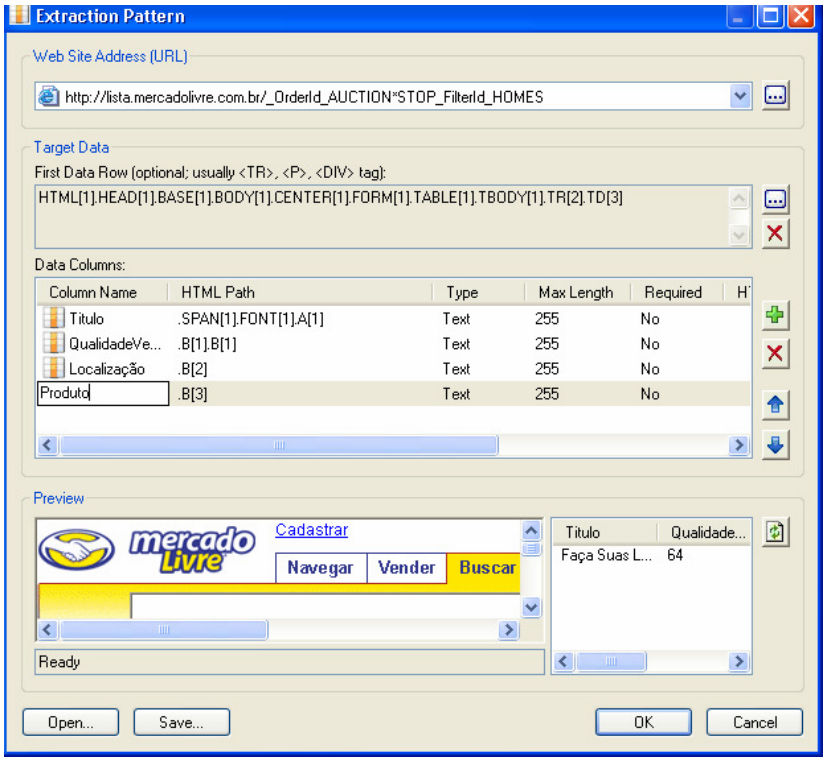
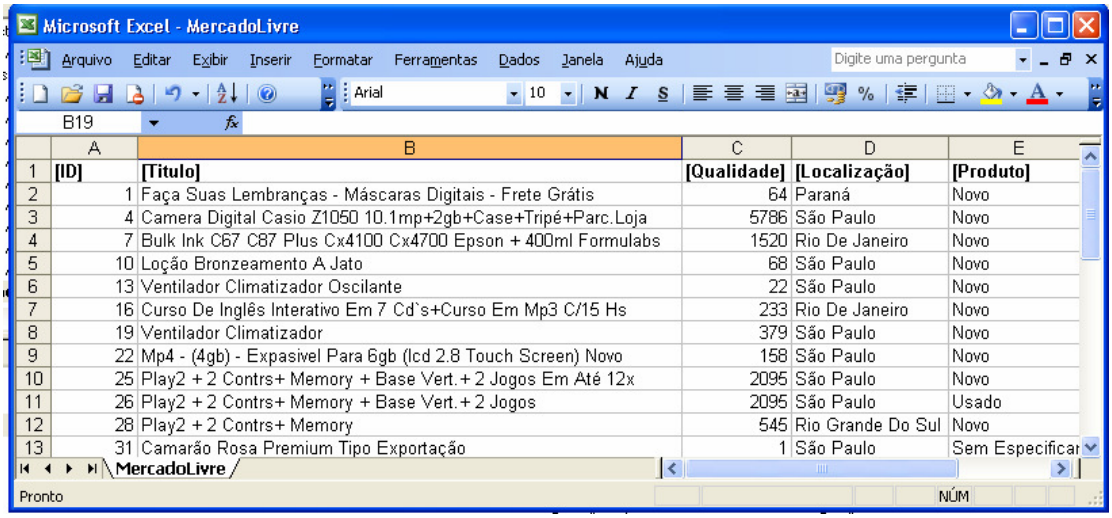


Figura 29: Extração padrão utilizando o *Web Content Extractor*



Microsoft Excel - MercadoLivre

Arquivo Editar Exibir Inserir Formatar Ferramentas Dados Janela Ajuda — Digite uma pergunta

Arial 10

B19

	A	B	C	D	E
1	[ID]	[Titulo]	[Qualidade]	[Localização]	[Produto]
2	1	Faça Suas Lembranças - Máscaras Digitais - Frete Grátis	64	Paraná	Novo
3	4	Camera Digital Casio Z1050 10.1mp+2gb+Case+Tripé+Parc.Loja	5786	São Paulo	Novo
4	7	Bulk Ink C67 C87 Plus Cx4100 Cx4700 Epson + 400ml Formulabs	1520	Rio De Janeiro	Novo
5	10	Loção Bronzeamento A Jato	68	São Paulo	Novo
6	13	Ventilador Climatizador Oscilante	22	São Paulo	Novo
7	16	Curso De Inglês Interativo Em 7 Cd`s+Curso Em Mp3 C/15 Hs	233	Rio De Janeiro	Novo
8	19	Ventilador Climatizador	379	São Paulo	Novo
9	22	Mp4 - (4gb) - Expasivel Para 6gb (lcd 2.8 Touch Screen) Novo	158	São Paulo	Novo
10	25	Play2 + 2 Contrs+ Memory + Base Vert.+ 2 Jogos Em Até 12x	2095	São Paulo	Novo
11	26	Play2 + 2 Contrs+ Memory + Base Vert.+ 2 Jogos	2095	São Paulo	Usado
12	28	Play2 + 2 Contrs+ Memory	545	Rio Grande Do Sul	Novo
13	31	Camarão Rosa Premium Tipo Exportação	1	São Paulo	Sem Especificar

MercadoLivre

Pronto NÚM

Figura 30: Resultado da extração exportado para o Excel utilizando o *Web Content Extractor*

6.1.6 Lixto

O sistema Lixto é um software baseado na tecnologia Java, que tem como função a extração automática de informação da Web e a geração de *wrapper*[18].

Lixto é totalmente visual. Os usuários interagem com o sistema para identificar informações de interesse e para criar filtros. Os w*rappers* gerados com o Lixto baseiam-se em exemplos selecionados pelo usuário. Primeiramente, o usuário abre uma página de exemplo e, em seguida, acrescenta padrões para extrair informações relevantes. Cada padrão caracteriza um tipo de informação e consiste de vários filtros. Cada filtro é composto de várias condições. Após especificar as condições e definir os filtros, o sistema pode também definir que intervalo de distancia os elementos podem ocorrer.

[18] Tecnologia utilizada para extrair informações relevantes de paginas HTML e converter para o formato XML.

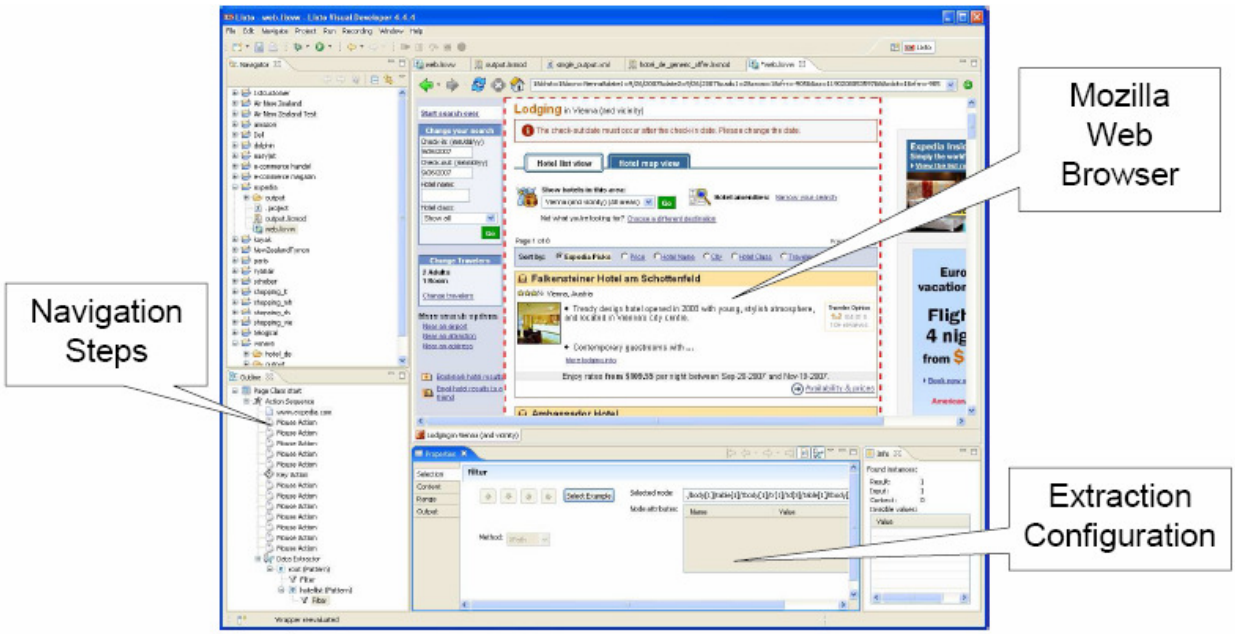


Figura 31: Lixto Visual (HERZOG, M.)

6.1.7 Automation Anywhere

Automation Anywhere é um software proprietário da *Tethys Solutions* que permite extrair dados da web facilmente. Como exemplos, podem ser extraídos catálogos de produtos, dados de estoque, tabelas de preços, etc.

Dentre as características da ferramenta destaca-se:

- *Automation Anywhere* permite extração de dados da web sem requerer nenhuma programação;
- Permite extrair dados da web e exportá-los para banco de dados, planilha do Excel ou outra aplicação.

A figura 32 ilustra as possíveis opções de exportações oferecidas pelo *Automation Anywhere* e a figura 33 mostra o resultado do extrato exportado para o Excel. Para a extração foi necessário identificar padrões nos textos, que determinaram as informações que se pretendia extrair. Os dados também foram extraídos do site do Mercado Livre, observa-se que tanto o *Automation Anywhere* quanto o *Web Content Extractor* obtiveram os mesmos resultados.

Figura 32: Automation Anywhere

Microsoft Excel - MercadoLivre

Arquivo Editar Exibir Inserir Formatar Ferramentas Dados Janela Ajuda

	A	B	C	D	E
1	[ID]	[Titulo]	[Qualidade]	[Localização]	[Produto]
2	1	Faça Suas Lembranças - Máscaras Digitais - Frete Grátis	64	Paraná	Novo
3	4	Camera Digital Casio Z1050 10.1mp+2gb+Case+Tripé+Parc.Loja	5786	São Paulo	Novo
4	7	Bulk Ink C67 C87 Plus Cx4100 Cx4700 Epson + 400ml Formulabs	1520	Rio De Janeiro	Novo
5	10	Loção Bronzeamento A Jato	68	São Paulo	Novo
6	13	Ventilador Climatizador Oscilante	22	São Paulo	Novo
7	16	Curso De Inglês Interativo Em 7 Cd`s+Curso Em Mp3 C/15 Hs	233	Rio De Janeiro	Novo
8	19	Ventilador Climatizador	379	São Paulo	Novo
9	22	Mp4 - (4gb) - Expasivel Para 6gb (lcd 2.8 Touch Screen) Novo	158	São Paulo	Novo
10	25	Play2 + 2 Contrs+ Memory + Base Vert.+ 2 Jogos Em Até 12x	2095	São Paulo	Novo
11	26	Play2 + 2 Contrs+ Memory + Base Vert.+ 2 Jogos	2095	São Paulo	Usado
12	28	Play2 + 2 Contrs+ Memory	545	Rio Grande Do Sul	Novo
13	31	Camarão Rosa Premium Tipo Exportação	1	São Paulo	Sem Especificar

MercadoLivre

Pronto NÚM

Figura 33: Resultado da extração exportado para o Excel utilizando o *Automation Anywhere*.

6.2 Ferramentas para Sumarização de Páginas Web

O processo de sumarização envolve métodos para encontrar uma descrição compacta para um subconjunto de dados. É muito utilizado na descoberta de conhecimento em textos (*Text Mining*), visando identificar palavras ou frases mais importantes do

documento ou conjunto de documentos. É útil para reduzir a quantidade de material em um documento, embora mantenha a mesma informação.

6.2.1 Copernic Summarizer

O *Coppernic Summarizer* é um software usado para sumarização de páginas Web e documentos de textos.

O *Copernic Summarizer* é um software comercial, sendo possível testar a versão trial por 30 dias. Essa ferramenta usa avançados algoritmos estatísticos e lingüísticos para resumir os conceitos-chave de um texto e extrair as mais relevantes frases para produzir uma versão condensada do texto original. Com a adição de uma única tecnologia proprietária chamada *WebEssence,* textos irrelevantes e outros conteúdos não essenciais encontrados em páginas da Web são ignorados e o resultado é um resumo notavelmente preciso.

Segundo Corrêa (2003) a tecnologia de inteligência artificial integrada ao produto, permite a compreensão do conteúdo do documento e a extração de conceitos e sentenças-chave. A ferramenta incorpora dois componentes: modelos estatísticos e processo de conhecimento intensivo. "O modelo estatístico pode ser aplicado para vários idiomas para aproximação do vocabulário específico. Ele inclui estimativas Bayesianas e sistemas de regras derivadas da análise de milhares de documentos". (CORREA, 2003, p. 17)

As principais características da ferramenta *Copernic Summarizer* são:

- O software pode analisar um texto de qualquer tamanho, sobre qualquer assunto;
- Produzir relatório de sumário para conceitos de texto pelo processamento de documentos;
- Permite ao usuário informar o tamanho do resumo em porcentagem ou em quantidade de palavras;
- Permite exportar os sumários em vários formatos: Word, HTML, XML, Rich Text, arquivo de texto;
- Além do resumo, a ferramenta determina os conceitos, que são palavras e combinações de palavras mais importantes no contexto.

- Permite resumir uma página da web a partir de um endereço URL**,** diretamente do software (tela principal) sem abrir o site, ou através do browser com o *Copernic Summarizer* integrado.

A figura 34 apresenta um resumo do site da UFRJ[19]. Apesar de a ferramenta pedir a entrada de textos em Inglês, Francês, Espanhol ou Alemão, o resultado de um site em português foi bastante satisfatório. A figura 35 mostra *o Copernic Summarizer* integrado ao browser.

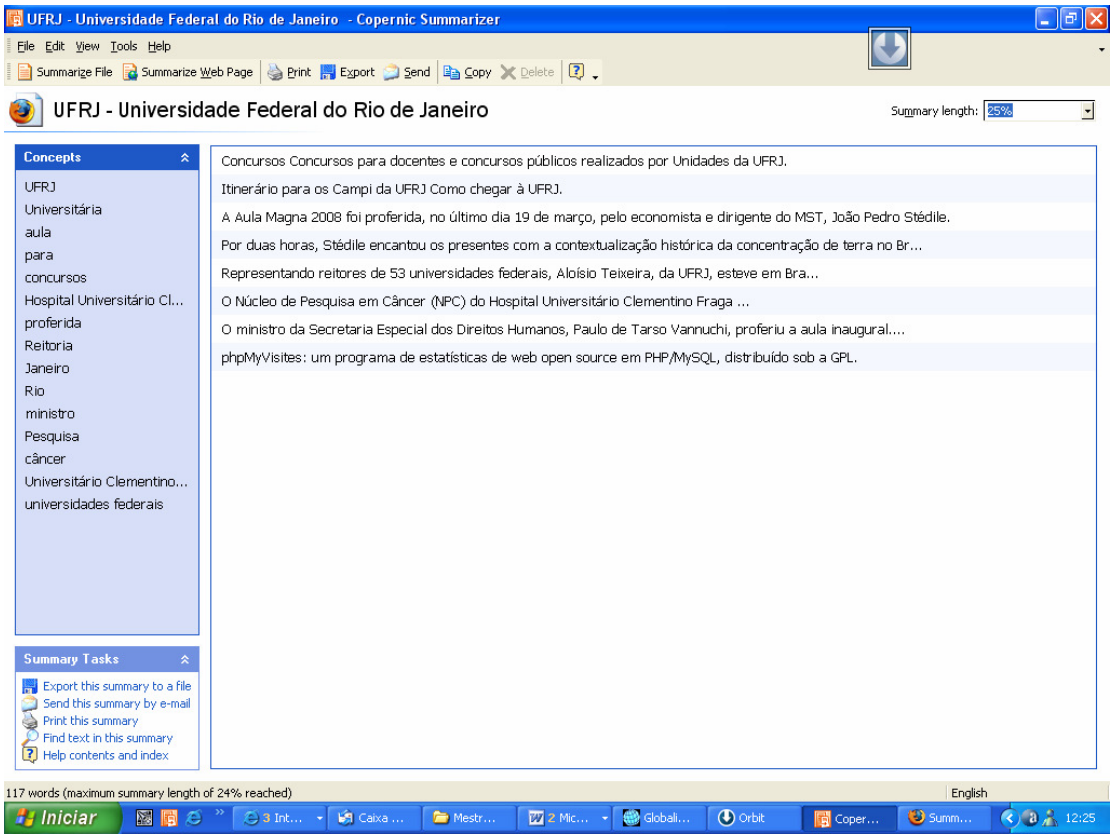


Figura 34: Figura gerada pelo *Copernic Summarizer* em 01/03/2008

[19] Htto://www.ufrj.br

Figura 35: Imagem gerada pelo *Copernic Summarizer* integrado ao browser.

6.2.2 TextAnalyst

O *TextAnalyst* é uma ferramenta usada na descoberta de conhecimento em textos, processo conhecido como Mineração de textos (Text Mining), baseado na tecnologia de Redes Neurais. O software pode ser utilizado também no processo de descoberta de conhecimento de textos da Web, sendo necessário salvar as páginas da web no formato de arquivo texto, pois a entrada para o *TextAnalyst* é um texto, do qual as informações relevantes são extraídas.

A ferramenta *TextAnalyst* é distribuída por *Megaputer Intelligence* (CORREA, 2003). A figura 36 apresenta a tela principal do *TextAnalyst* e mostra as três opções iniciais do software, que permite ao usuário analisar novos textos e criar a base de conhecimento, abrir uma base de conhecimento existente ou ainda utilizar o tutorial.

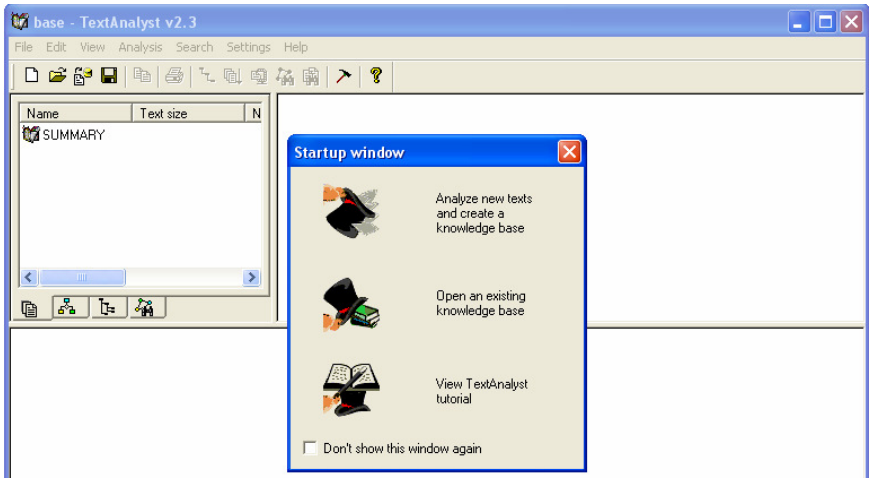


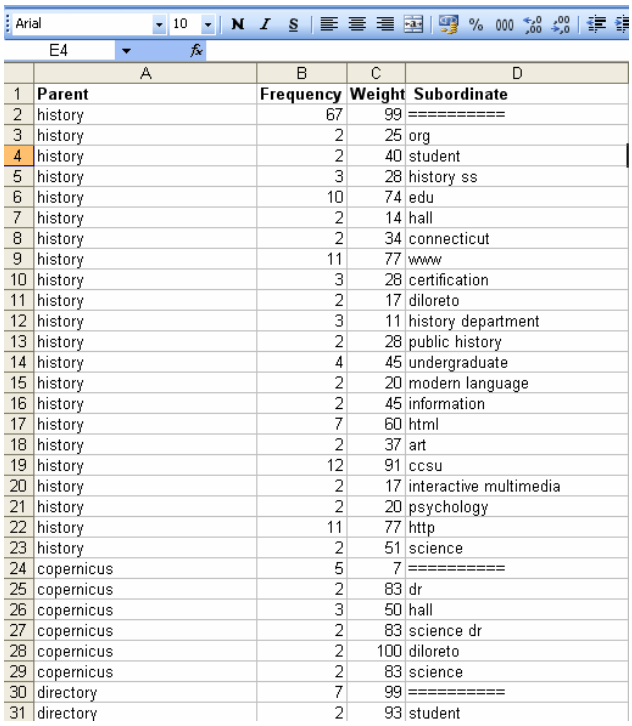
Figura 36: Tela principal do TexAnalyst

Na opção de analisar novos texto e criar uma base de conhecimento, esta ferramenta determina quais conceitos, ou seja, palavras e combinações de palavras que são mais importantes no contexto do texto investigado. Algoritmos matemáticos do TextAnalyst determinam a importância de cada conceito e o atribui um peso numérico semântico, que é a medida da probabilidade de que este conceito é importante no texto estudado. Simultaneamente, o TextAnalyst determina os pesos das relações entre os conceitos individuais no texto e hiperlinks e cria uma Rede Semântica sem utilizar conhecimento prévio sobre o assunto. A estrutura resultante, chamada Rede Semântica, é um conjunto dos mais importantes conceitos destilada a partir da análise do texto, juntamente com as relações semânticas entre esses conceitos no texto.

As principais características da ferramenta TextAnalyst são:

- Para cada conceito são obtidos dois valores que representam os pesos semânticos: dos conceitos em relação ao conceito “pai” e os pesos semânticos dos conceitos em relação ao documento.
- Permite, também, especificar determinadas palavras-chave, caso não tenha certeza se a ferramenta irá recuperá-las a partir do texto, possivelmente devido ao seu baixo peso semântico, as palavras-chave podem ser adicionadas pelo usuário através de um dicionário externo.
- Permite criar sumários dos textos de entrada baseando-se em textos semânticos.
- Os resultados podem ser exportados para um arquivo no formato HTML ou para um formato compatível com a planilha do Excel.

A figura 37 ilustra o resultado da extração dos conceitos mais importantes extraídos do site http://www.artsci.ccsu.edu/departments.htm e exportado para uma planilha do Excel. Para a extração dos conceitos mais relevantes, todas as páginas do site citado foram concatenadas usando o software Websphinx, que concatena todas as páginas, transformando-as em um único documento html. Após a concatenação, o arquivo html foi salvo no formato txt, já que o TextAnalyst suporta apenas os formato txt e rtf como entrada.



	A	B	C	D
1	**Parent**	**Frequency**	**Weight**	**Subordinate**
2	history	67	99	==========
3	history	2	25	org
4	history	2	40	student
5	history	3	28	history ss
6	history	10	74	edu
7	history	2	14	hall
8	history	2	34	connecticut
9	history	11	77	www
10	history	3	28	certification
11	history	2	17	diloreto
12	history	3	11	history department
13	history	2	28	public history
14	history	4	45	undergraduate
15	history	2	20	modern language
16	history	2	45	information
17	history	7	60	html
18	history	2	37	art
19	history	12	91	ccsu
20	history	2	17	interactive multimedia
21	history	2	20	psychology
22	history	11	77	http
23	history	2	51	science
24	copernicus	5	7	==========
25	copernicus	2	83	dr
26	copernicus	3	50	hall
27	copernicus	2	83	science dr
28	copernicus	2	100	diloreto
29	copernicus	2	83	science
30	directory	7	99	==========
31	directory	2	93	student

Figura 37: Resultado gerado pelo *TextAnalyst* exportado para o Excel

6.3 Análise dos resultados

	FilterViewer	**SQLGui**	**Web Data Extractor**	**Web Content Extractor**
Tarefas	- Extrair alguma parte do conteúdo da página; - mostrar o conteúdo da estrutura de marcação HTML ou XML.	- Extrair alguma parte do conteúdo da página; - Mostrar o conteúdo da estrutura de marcação HTML ou XML. - Restringir a consulta	- Extrair a URLs, títulos, textos, e-mails, telefones e fax de websites; - Restringir pesquisas através de filtros	- Permite criar padrão de dados e regras de rastreamento; _ extração de conteúdo de páginas de acordo com o padrão definido pelo usuário
Facilidade de utilização	Necessidade de conhecimento das tags HTML ou XML	Necessidade de conhecimento das tags HTML ou XML	sim	Sim
Automatização de processo	Não	Não	sim	Praticamente automático, mas o usuário precisa selecionar os dados para extração de padrão.
Interface visual	boa	Boa	boa	**boa**
Funcionalidades para exportação de dados	Não possui	Não possui	Apenas em arquivo texto	Permite extrair para Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL MySQL
Qualidade do extrato	Não possui (Só apresenta o resultado)	Não possui (Só apresenta o resultado)	Baixa qualidade (apenas textos)	Ótima

Tabela 3: Comparação entre as ferramentas de extração

	LIXTO	Automation Anywhere	Copernic Summarizer	Text Analyst
Tarefas	- Permite criar padrão de dados e regras de rastreamento; - Extração de conteúdo de páginas de acordo com o padrão definido pelo usuário	- Permite criar padrão de dados e regras de rastreamento; - Extração de conteúdo de páginas de acordo com o padrão definido pelo usuário	Sumarização de páginas Web e documentos de textos. - Extração dos principais conceitos	- Sumarização - Criação de base de conhecimento
Facilidade de utilização	Sim	Não	Sim	Sim
Automatização de processo	Praticamente automático, mas o usuário precisa selecionar os dados para extração de padrão.	Praticamente automático, mas o usuário precisa selecionar os dados para extração de padrão.	Sim	Necessidade de salvar as páginas HTML em formato txt
Interface visual	boa	boa	boa	boa
Funcionalidades para exportação de dados	Permite extrair para Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL MySQL	Permite extrair para Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL MySQL	Word, HTML, XML, Rich Text	Permite exporter para o Microsoft Excel
Qualidade do extrato	Ótima	Ótima	ótima	ótima

Tabela 4: Continuação da tabela 3

Ao Analisar as ferramentas de extração de informação, observa-se que no *FilterViewer* o processo não é automático, o usuário precisa informar o vetor HTML para filtrar as colunas que o usuário tem interesse. Tem também como desvantagem permitir a entrada apenas de uma URL.

O SQLGui tem como característica permitir a restrição da consulta, mas não permite a exportação dos resultados, além de exigir do usuário conhecimento das tags HTML ou XML. Já o Web Data Extractor permite a extração da informação de forma praticamente automática, além da alta performance nesse processo. Em contra partida, o mesmo já trás de forma pré-definida o que se pode extrair, não permitindo ao usuário extrair algo que seja de seu interesse, além das opções fornecidas pela ferramenta. A figura

Quanto ao Lixto, as descrições a respeito do mesmo foram feitas com base em artigos, pois ele é um software proprietário e não foi possível conseguir uma versão demo para teste. De acordo com as pesquisas realizadas sobre essa ferramenta, parece que ela tem um processo semelhante ao *Web Content Extractor* e ao *Automation Anywhere*, pois a extração do conteúdo das páginas é feita de acordo com o padrão definido pelo usuário. Esse software também permite a exportação dos dados extraídos para um formato estruturado.

O *Copernic Summarizer* e o *TextAnalyst* são ferramentas que são muito utilizadas em mineração de textos e que também podem ser utilizadas para sumarização de páginas web. Além de ter apresentado um ótimo desempenho nos testes realizados, permitem também a exportação dos dados para o excel, word, entre outros. Para uma análise mais detalhada dos softwares de sumarização, foi passado tanto para o *Copernic Summarizer* quando para o *TextAnalyst* o artigo "Ferramentas de Busca", disponível em http://www2.dbd.puc-rio.br/pergamum/tesesabertas/0024134_02_cap_03.pdf, para comparar o resumo produzido pelas duas ferramentas. A figura 38 apresenta o sumário do artigo produzido pelo *Copernic Summarizer* e a figura 39 apresenta o sumário produzido pelo *TextAnalyst.*

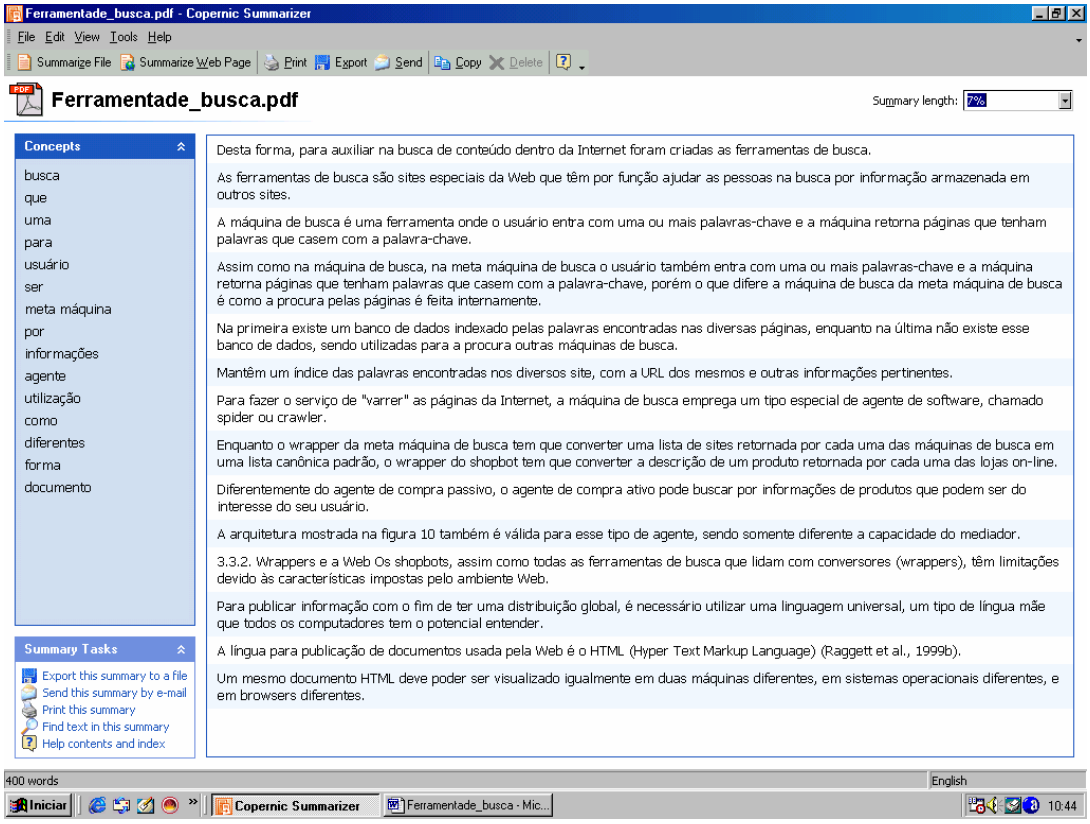


Figura 38: Sumário gerado pelo Copernic Summarizer

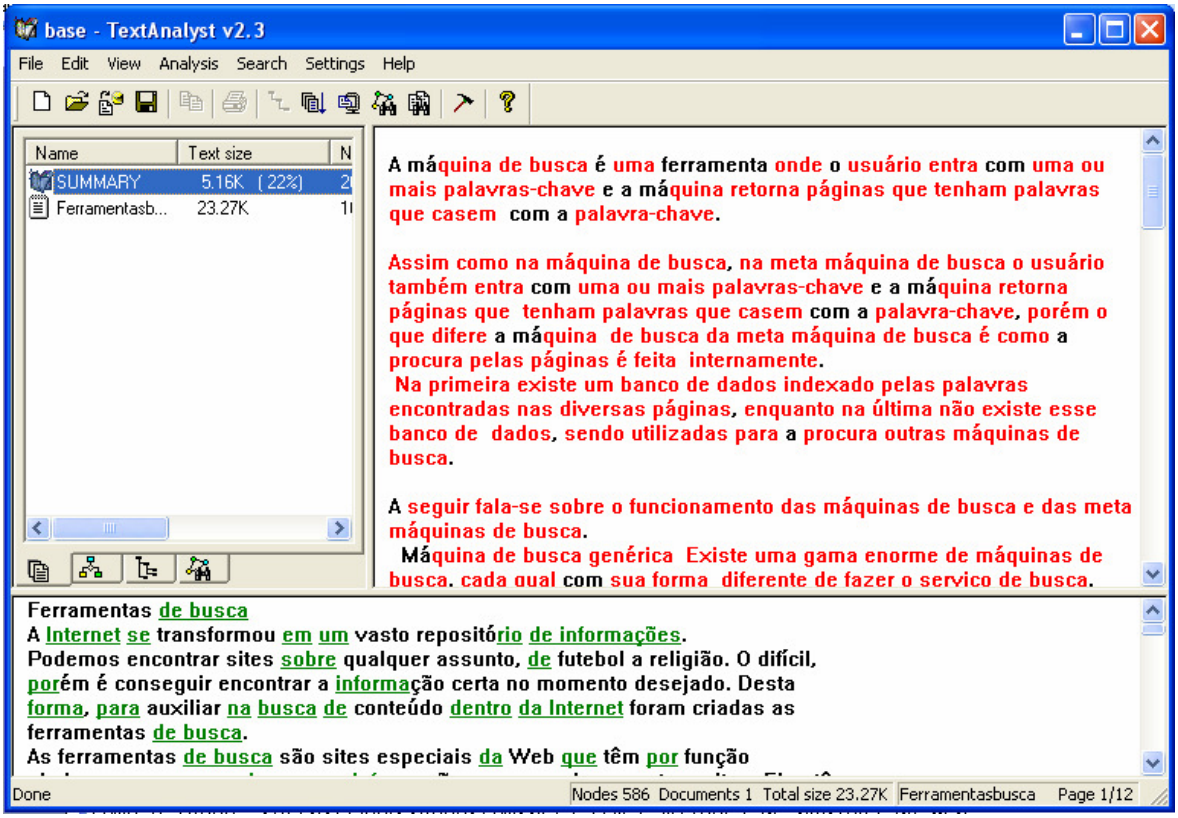


Figura 39: Sumário gerado pelo TextAnalist

O texto com palavras grifadas em vermelho na parte superior da figura 39 representa o sumário, o Texto com palavras grifadas em verde representa o texto original para análise.

O *TextAnalyst* apresentou um bom resultado, pois o sumário preservou a idéia central do texto e a apresentou de forma clara. As sentenças apresentaram sem problemas de coerência e coesão e o sumário preservou a essência do texto mantendo o texto no estilo de redação conforme o texto original.

Assim como o *TextAnalyst*, o *Copernic Summarizer* também foi bem sucedido na tarefa de sumarização. O sumário preservou a idéia central do texto, porém o texto veio estruturado em forma de tópicos não preservando o formato de redação do texto original. Contudo, o resultado obtido tanto no *TextAnalyst* quanto no *Copernic Summarizer* demonstrou o bom potencial dessas ferramentas na tarefa de sumarização de textos

CONCLUSÃO

Na elaboração desta dissertação procurou-se evidenciar que a área de descoberta do conhecimento tem passado por um avanço em muitos aspectos, onde um considerável crescimento foi observado em aplicações da Web, incluindo principalmente a utilização de técnicas de Mineração de Dados na Web.

Esta mineração é uma área de pesquisa que vem se desenvolvendo a partir da crescente interação entre as técnicas de Mineração de Dados empregadas ao contexto Web. Assim sendo, esse estudo apresentou uma visão geral da Mineração de Conteúdo da Web, além de apresentar as outras modalidades de mineração: Mineração de Estrutura e Mineração de Uso.

O que dificulta o estudo da Mineração de Conteúdo da Web é a delimitação de seu escopo, por se tratar de uma área multidisciplinar vinculada a outras áreas, como a recuperação de informação, a aprendizagem de máquina e os agentes de informação.

No Brasil, parece que os projetos ainda estão bastante tímidos e muitos ainda buscam posições de destaque através dos documentos indexados nos mecanismos de busca, sem a devida preocupação com a qualidade da informação, preocupando-se apenas com o ranking dos mesmos. Esse estado revela a necessidade de melhorias no tratamento da informação no cenário brasileiro de indexação eletrônica.

Inúmeras implementações foram detectadas na bibliografia de projetos disponíveis via rede, com relação à indexação por ranking, através de algoritmos de balanceamento de pesos e estatística. Porém, precisa ainda ser mais bem trabalhada a visão semântica, se o objetivo for o de aperfeiçoar e agregar qualidade à indexação eletrônica, almejando a relevância dos resultados nas buscas via Internet, evitando a recuperação de informações não relevantes.

Em relação à Recuperação de Informação realizada pelos serviços de busca, observou-se que nenhum mecanismo de busca consegue recuperar todas as páginas existentes, ou seja, as páginas recuperadas por essas ferramentas de buscas não são as mesmas. Cada serviço de busca recupera um conjunto de páginas, algumas são recuperadas por mais de um serviço, mas nenhum indexa todas as páginas da Web, pois as páginas

retornadas pelas máquinas de buscas são diferentes. Se o usuário utilizar apenas um serviço de busca não conseguirá recuperar muitas das páginas que poderiam ser interessantes para sua consulta.

Como na maioria dos problemas do mundo real, os agrupamentos na Web não têm fronteiras e muitas vezes se sobrepõem consideravelmente. Além disso, maus exemplos (outliers), bem como dados incompletos podem facilmente ocorrer no conjunto de dados, devido a uma ampla variedade de razões inerentes a navegação na web. Assim, Web Mining se depara com a presença significativa de ruído e outliers, (ou seja, exemplares ruins). Além do mais, os conjuntos de dados na Web são extremamente grandes, o que dificulta a descoberta de conhecimento.

Atualmente, a busca por informação na web tem crescentemente se tornado importante e popular. Isso se deve à riqueza de informação presente na web e, como se vive uma época de constantes mudanças, buscar novos conhecimentos e permanecer sempre atualizado é de extrema importância. Portanto, apesar da web permitir isso, muitas vezes não é fácil buscar a informação que se almeja, pois a internet não fornece uma estrutura de conceitos com tópicos e sub-tópicos conforme é estruturado um livro tradicional.

Por conseguinte, através das análises realizadas com as ferramentas de busca, pode-se dizer que muitos dos documentos recuperados, apesar de se tratar do assunto pesquisado, no caso, a pesquisa sempre foi realizada passando para a consulta a palavra-chave “web mining”, não satisfaria a necessidade de um usuário ainda inexperiente nesta área, pois os primeiros sites não trouxeram definição e descrição detalhada em tópicos sobre o assunto. Se um usuário deseja aprender algo sobre Web Mining, deve precisar conhecer a definição de Web Mining e qual é o seu objetivo. Em seguida, o usuário poderia desejar conhecer os subtópicos mais importantes como clusterização, sumarização, etc.

A maioria das ferramentas analisadas retornou o site http://www.kdnuggets.com/ dentre as primeiras posições, porém, este site não trás a mínima explicação do que é Web Mining. Contudo, esse site é mais satisfatório para pessoas que já são familiarizadas com o assunto e necessita de fontes adicionais, e não é designado para pessoas que desejam aprender sobre esse assunto.

Trabalhos relacionados como o sistema WebLearn (LIU, CHIN e NG, 2003) visam obter um conjunto de páginas relevantes de máquinas de busca e estrutura-las em tópicos e

sub-tópicos para facilitar o acesso a informação de maneira mais fácil conforme em livros tradicionais.

O processo de clusterização realizado por algumas máquinas de busca é efetivamente útil, pois a visualização dos resultados em grupos de documentos facilita muito a localização da informação desejada. Essa estratégia incorporada em alguns mecanismos de busca faz com que as consultas sejam realizadas de forma mais eficaz e mais produtiva, facilitando a vida do usuário. Portanto, os mecanismos de busca por clusterização obtiveram resultados mais satisfatórios e de melhor compreensão textual e visual para o usuário do que os mecanismos de busca tradicionais, pois a maioria dos documentos pertencentes aos clusters retornados pelas máquinas de buscas analisadas foi relevante. Além disso, os clusters foram rotulados corretamente, pois a rotulação do cluster é tão importante quanto a qualidade do conteúdo do cluster.

Quanto às ferramentas de extração, estas ainda estão em fase de estudos e análises, e precisam de muito amadurecimento, pois das ferramentas analisadas nenhuma permite a extração totalmente automática e não é de fácil uso para usuários inexperientes.

As ferramentas de sumarização de textos apresentaram resultados bastante semelhantes, tendo um bom desempenho na tarefa de sumarização, manteve a idéia central do texto, o que ressalta a utilidade desses sistemas de sumarização.

Contudo, este trabalho procurou, de uma forma concisa, caracterizar e definir o alcance da Mineração de Conteúdo da Web e oferecer indicações para o aprofundamento no estudo dos diferentes tópicos da área.

REFERENCIAS BIBLIOGRÁFICAS

BERENDT, B. et al. **Web Mining: From Web to Semantic Web.** Sringer. New York, 2004.

BERNERS – LEE, T., HENDLER, J., LASSILA, O. **The Semantic Web**. Scientifc American, May, 2001.

BORTOLETO, S. MOREIRA, L. BUENO, T. S. **Mineração de dados com aplicação de XML.** UNICENP, 2005.

BRACHMAN, R.; ANAND, T. "**The Process of Knowledge Discovery in Databases**: A Human-Centered Approach". In: Advances in Knowledge Discovery and Data Mining. AAAI Press, 1996.

BUENO, M. C. **Recuperação da informação: uso estratégico das ferramentas de busca**. 2000. 43 f. Trabalho de conclusão de curso apresentado ao Programa de Pós-Graduação a nível de Especialização do Departamento de Biblioteconomia - Faculdade de Filosofia e Ciência, Universidade Estadual Paulista, Marília.

CAMPOS, R. N.T., 2005. **Agrupamento Automático de Páginas Web Utilizando Técnicas de Web Content Mining.** Dissertação de M.Sc, Universidade da Beira Interior, Covilhã, 2005.

CARDOSO, O. **Recuperação da informação.** Disponível em: http://www.dcc.ufla.br/infocomp/artigos/v2.1/olinda.pdf. Acesso em: 31 jan. 2008.

COOK, Diane J. e HOLDER, Lawrence B. **Graph-Based Data Mining**.

COOLEY, R. W. et al. "**Web mining: Information and Pattern Discovery on the World Wide Web**". In: Proceedings of International Conference on Tools with Artificial Intelligence (ICTAI), 1997, 10p.

COOLEY, R. W. "**Web usage mining: Discovery and application of Interesting Patterns from Web data**". PhD thesis, Dept. of Computer Science, University of Minesota, 2000.

COOLEY, R., MOBASHER, B. e SRIVASTAVA, J. **"Data preparation for mining world wide web browsing patterns**". Knowledge and Information Systems, 1999.

COOLEY, R., MOBASHER, B. e SRIVASTAVA, J. **Web Mining: Information and Pattern Discovery on the World Wide Web**. In: 9th IEEE International Conference on Tool with Artificial Inteligence. Newport Beach, 1997. Anais. IEEE, 1997.

CORREA, A. C.G., 2003. **Mineração de documentos baseados em informação semântica no ambiente AMMO.** Dissertação M.Sc.Universidade Federal de São Carlos.

CORDEIRO, J.P.C., 2003. **Extracção de Elementos Relevantes em Texto/Páginas da World Wide Web.** Dissertação M.Sc. Faculdade de Ciências da Universidade do Porto.

CUTTING, D.D. KARGER, J. PEDERSON, J. & SCATTER, J. "**A cluster based approach to browsing large document collections**". Proceedings of the Fifteenth International Conference on Research and Development in Information Retrieaval, 1992.

DEGROOT, T. (1986) - **Probability and Stastistics.** Addison Wesley, MA.

DUQUE, C. G. **Recuperação da Informação e Máquinas de Busca.** Disponível em: http://www.bax.com.br/Bax/Disciplinas/BiblioDigi/ProgramaECI/ApresentacaoKlauss/maquinas.ppt Acesso em 10/12/2007.

ETZIONE, O., "**The World Wide Web Quagmire or gold mine**" Communications of the ACM, 1996.

FERRAGINA, P., GULLI, A. **The Anatomy of a Hierarchical Clustering Engine for Web-page, News and Book Snippets.** Disponível em http://citeseer.ist.psu.edu/708583.htm. Acesso em 11/08/07.

Ferramentas de Busca. Disponível em: http://www2.dbd.puc-io.br/pergamum/tesesabertas/0024134_02_cap_03.pdf Acesso em 10/06/2008

FORTES, D. **A nova internet.** Revista INFO. Abril/2002

GIRARDI, R. *"***Classification and Retrieval of Software through their Descriptions in Natural Language***",* Ph.D. dissertation, No. 2782, University of Geneva, December 1995.

GIRARDI, R. "**Main Approaches to Software Classification and Retrieval**". Em: Ingeniería del Software y reutilización: Aspectos Dinámicos y Generación Automática. Editores J. L. Barros y A. Domínguez. (Universidad de Vigo – Ourense, de l6 al 10 de julio de 1998). Julio, 1998.

GIRARDI, R. "**Main Approaches to Software Classification and Retrieval**". Em: Ingeniería del Software y reutilización: Aspectos Dinámicos y Generación Automática. Editores J. L. Barros y A. Domínguez. (Universidad de Vigo – Ourense, del 6 al 10 de julio de 1998). Julio, 1998. **IEEE Intelligent Systems,** Los Alamitos, v.15, n.2, p. 32-41, Mar/Abr 2000.

GRUBER, T.R. A **Translation Approach to Portable Ontologies**.**Knowledge Acquisitio**n, v.5, n.2, p.199-200,1995.

HECK, M., 2006, "Vivisimo Races of the Search Pack". **InfoWorld.** Disponível em: http://vivisimo.com/docs/infoworld.pdf Acesso em 25/01/2008.

HERZOG, M. **Optimizing Pricing Through Effective Online Market Intelligence.** Disponível em: http://www.lixto.com/images/File/Downloads/PPS_Lixto_Webinar_en.pdf acesso em 30/01/2007.

KONCHADY, M. **Text Mining Application Programming.** Charles River Media, Boston, 2006.

KOSALA, R.; BLOCKEEL H. "**Web Mining Research: A survey**". In: *SIGKDD Explorations*, vol. 2 -1, June 2000, pp. 1-15.

KUSHMERICK, N "**Wrapper Induction for Information Extraction**". Doctoral thesis. University of Washington, Department of Computer Science and Engineering, 1997.

LINOFF, G.S.; BERRY, M. J. A. **Mining the Web.** Transforming Customer Data into Customer Value. New York: Wiley, 2001.

LIU, B. **Web Data Mining.** Exploring Hiperlinks, Contents, and Usage Data. Chigago. Springer, 2007.

LIU, B. CHANG, K. C-C. "Editorial; Special Issue on Web Content Mining"**. SIGKDD Explorations.** v.6, n.2. pp. 1-3. Disponível em: http://kdd.org/explorations/issues/6-2-2004-12/editorial.pdf Acesso em 11/12/2007.

LIU, B.; CHIN, C.W; NG, H.T. **Mining Topic-Specific Concepts and Definitions on the Web.** Disponível em: http://www.cs.uic.edu/~liub/publications/WWW-2003.pdf. Acesso em 12/01/2008.

LOTON, T. **Web Content Mining with Java.** Techiniques for Exploiting the World's Biggest Information Resource. New York: Wiley, 2002.

MACEDO, F. Q., FILHO, F. T. **Inteligência Competitiva Aplicada à Pesquisa: Uma Abordagem aos Motores não Tradicionais de Busca na Internet para Análise de Governança Corporativa.** Disponível em: http://quoniam.univ-tln.fr/pdf/Travaux/Gouvernance_d'entreprise.pdf Acesso em 02/03/2008

MACHADO, L. S**.,** 2002. **Mineração do uso da web na educação a distância: propostas para a condução de um processo a partir de um estudo de caso.** Dissertação de M.Sc, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

MARINHO. L. B. GIRARDI, R. **Mineração na Web.** Disponível em: http://www.sbc.org.br/reic/edicoes/2003e2/tutoriais/MineracaoNaWeb.pdf. Acesso em 02/03/2008

MARKOV, Z.; LAROSE, D. T. **Data Mining the Web:** Uncovering Patterns in Web Content, Structure, and Usage. New Britain: Wiley, 2007.

MARTHA, A.S., 2005. **Recuperação de informação em campos de texto livre de prontuários eletrônicos do paciente baseada em semelhança semântica e ortográfica.** Dissertação de M.Sc, Universidade Federal de São Paulo, São Paulo.

MATSUNAGA, L. A., 2007. **Uma metodologia de categorização automática de textos para a distribuição dos projetos de lei às comissões permanentes da câmara legislativa do distrito federal.** Dissertação de M.Sc, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

MENCZER, F. **Web Crawling**. Indiana University School of Informatics in Web Data Mining by Bing Liu Springer, 2007.

OUNIS, et al. **Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web.** Disponível em http://www.upgrade-cepis.org/issues/2007/1/up8-1Yahoo!.pdf. Acesso em 05/03/2008

PAL S. K., Varum Talwar, Pabitra Mitra, "**Web Mining in Soft Computing Framework**: Relevant, State of the Art and Future Directions", 2002.

PAL, S. K.,TALWAR, V., MITRA, P., "**Web Mining in Soft Computing Framework**: Relevant, State of the Art and Future Directions", 2000.

REBITTE, L. BP, M. V. **Dominando Tableless. Seu site entre os primeiros nos sites de busca!.** Alta Books. Rio de Janeiro. 2006.

REZENDE, S. **Sistemas Inteligentes: fundamentos e aplicações**. São Paulo: Manole, 2003. 525 p.

SALTON, G. & Bucckley, C. **Term-weighting approaches in Automatic Retrieval, Information Processing & Management**, Vol. 24, No 5, 1988.

SALTON, G. **Automatic Text Processing: The Transformation, Analisys, and Retrieval of Information by Computer**, Addison Wesley, 1989.

SILVA, J. M. , SILVEIRA, E. S. **Apresentação de Trabalhos Acadêmicos. Normas Técnicas**. 4ª ed.

SILVIA, F.R. **Ferramentas de busca na Internet: um estudo do Google, Yahoo! e Metaminer. "**XXVI ENEBD – Encontro Nacional de Estudantes de Biblioteconomia". Universidade Estadual Paulista, 2003

SANTOS, M. A. M.R, 2002, **Extraindo regras de associação a partir de textos**. Dissertação de M.Sc, Pontifícia Universidade Católica do Paraná, Curitiba, Brasil.

SANTOS, R. G. 2000. **Utilização de Técnicas Data Mining na busca de conhecimento na Web.** Universidade Federal de Pelotas. Pelotas, RS.

SOWA J. F., **Ontology, metadata and semiotics, Conceptual Structures: Logical, Linguistic, and Computational Issues**, Lecture Notes in AI, Springer-Verlag, Berlin., 2000 , p 55-81

SPILIOPOULOU, M.; FAULSTICH, L. C. "WUM: **A tool for Web Utilization Analysis**". In: Lecture Notes in Computer Science (LNCS), vol. 1590, March 1999, pp. 184-203.

SPILIOPOULOU, M. et al**. "Improving the effectiveness of a web site with web usage mining**". In: Proceedings of the Workshop on Web Usage Analysis and User Profiling (WEBKDD '99), 1999, pp. 51-56.

SPILIOPOULOU, M. et al. "**A data miner analyzing the navigational behavior of Web users**". In: Proceedings of the Workshop on Machine Learning in User Modelling of the ACAI´99 Int. Conf., 1999, pp. 430- 437.

STUDER, R., BENJAMINS, R., FENSEL, D**. Knowledge Engineering and Methods Data and knowledge Engineering**, 1998. P. 161-197.

Terrier. A Practical Overview. Department of Computing Science, University of Glasgow, 2005. Disponivel em: http://ir.dcs.gla.ac.uk/terrier/publications/TerrierSessionSlides.pdf. Acesso em 05/03/2008.

ZAIANE, O. R., **WEB Mining**: Concepts, Practices and Research. In: Simpósio Brasileiro de Banco de Dados, Tutorial, XV SBBD, 2000, João Pessoa. Anais... João Pessoa: SBBD, 2000. p. 410-474.

ZAÏANE, O. "**Web Usage Mining for a better Web-based learning Environment**", In: Proceedings of Conference on Advanced Technology for Education, 2001, pp. 60-64.

ZAÏANE, O.; LUO, J. **Towards Evaluating Learners' Behaviour in a Web-based Distance Learning Environment**. In: Proceedings IEEE International Conference on Advanced Learning Technologies (ICALT 2001), 2001, 4p.

ZAÏANE. O. et al. "**Discovering web access patterns and trends by applying OLAP and data mining technology on web logs**". In: Advances in Digital Libraries, 1998, pp.19-29.

Zamir, O., Etizioni, O.: **Web document clustering: A feasibility demonstration. In: Research and Development in Information Retrieval.** (1998) 46-54.

ZANIER, A.M.A., 2006. **A Evolução dos mecanismos de busca on-line: A melhoria nos resultados obtidos.** Dissertação de M.Sc, Faculdade de Economia e Finanças IBMC. Porto Alegre. Rio de Janeiro, RJ.

W3C (WORLD WIDE WEB CONSORTIUM). Disponível em: <http://www.w3.org/ Acesso em: 05/12/ 2007.

WIKIPEDIA. **Recuperação da Informação.** Disponível em http://pt.wikipedia.org/wiki/Recupera%C3%A7%C3%A3o_de_informa%C3%A7%C3%A3o. Acesso em: 05/01/08.

WIKIPEDIA. **Google.** Disponível em http://pt.wikipedia.org/wiki/Recupera%C3%A7%C3%A3o_de_informa%C3%A7%C3%A3o. Acesso em: 05/01/08

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo