Visualização de Informação Aplicada a Resultados de Mecanismos de Busca

Cláudio Maximiliano Zaina

Dissertação de Mestrado

Instituto de Computação
Universidade Estadual de Campinas

Visualização de Informação Aplicada a Resultados de Mecanismos de Busca

Cláudio Maximiliano Zaina[1]
Julho de 2005

Banca Examinadora:

Profª Drª Maria Cecília Calani Baranauskas (Orientadora)
Instituto de Computação – UNICAMP

Prof. Dr. Sergio Roberto P. da Silva, Departamento de Informática
Universidade Estadual de Maringá – UEM

Profª Drª Anamaria Gomide
Instituto de Computação – UNICAMP

Profª Drª Ariadne B. R. Carvalho
Instituto de Computação – UNICAMP

[1] Apoio financeiro da CAPES.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP

Bibliotecário: Maria Júlia Milani Rodrigues – CRB8a / 2116

Zaina, Cláudio Maximiliano

Z13v Visualização de informação aplicada a resultados de mecanismos de busca / Cláudio Maximiliano Zaina – Campinas, [S.P. :s.n.], 2005.

Orientador : Maria Cecília Calani Baranauskas

Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Computação.

1. Interação homem-máquina. 2. Ferramentas de busca na web. 3. Visualização. I. Baranauskas, Maria Cecília Calani. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Título em inglês: Information visualization applied to search engine results.

Palavras-chave em inglês (Keywords): 1. Human-computer interaction. 2. Search agents. 3. Visualization.

Área de concentração: IHC

Titulação: Mestre em Ciência da Computação

Banca examinadora: Profa. Dra. Maria Cecília Calani Baranauskas (IC-UNICAMP)

Profa. Dra. Anamaria Gomide (IC-UNICAMP)

Profa. Dra. Ariadne B. R. Carvalho (IC-UNICAMP)

Prof. Dr. Sérgio Roberto P. da Silva (UEM)

Data da defesa: 22/07/2005

Termo de Aprovação

Tese defendida e aprovada em 22 de julho de 2005, pela Banca Examinadora composta pelos Professores Doutores:

_____

Prof. Dr. Sergio Roberto P. da Silva, Departamento de Informática
Universidade Estadual de Maringá – UEM

_____

Profª Drª Anamaria Gomide
Instituto de Computação – UNICAMP

_____

Profª Drª Ariadne B. R. Carvalho
Instituto de Computação – UNICAMP

_____

Profª Drª Maria Cecília Calani Baranauskas (Orientadora)
Instituto de Computação – UNICAMP

Dedico este trabalho a todos aqueles que se importam com o próximo e trabalham não só para o próprio crescimento, mas para o de todos.

Uma vida sem buscas não é
digna de ser vivida.

Sócrates

Agradecimentos

Gostaria de agradecer antes de tudo a meus pais, que sempre, com carinho, me apoiaram incondicionalmente nas minhas decisões, por mais absurdas que elas possam ter parecido.

Agradeço à Profª Drª Maria Cecília Calani Baranauskas, que me deu um significado completamente novo e maravilhoso à palavra "orientação": uma guia com carinho, elegância, humor, coerência e, sobretudo, inteligência. Essa combinação acaba por manter coeso um grupo que, de orientados, passam instantaneamente a admiradores.

Gostaria também de agradecer aos meus companheiros, aqueles que compartilham comigo o ponto de vista de "pós-graduandos": Eduardo Sartori, grande incentivador desta minha "empreitada acadêmica", Luís Augusto, a primeira mão amiga a me ser estendida dentro do Instituto de Computação, e Diogo, com seu humor e inteligência atordoantes. O grupo de orientação é um caso completamente à parte. O cuidado e carinho que uns temos com os outros é algo que me comove. Obrigado a todos, mas principalmente àqueles que são meus contemporâneos: Amanda, Carlos, Juliano, Raquel, Roberto e Sílvia (em ordem alfabética para não criar confusão!).

Obrigado a todos meus amigos que estão longe mas cuja lembrança traz luz e tranqüilidade quando me são mais necessários. E aqui estão Fátima Gomes Marin, que tem a capacidade de acreditar em mim, mesmo quando nem eu mesmo acredito e Marcelo Antônio Monteiro, um irmão que a vida me deu.

Não posso deixar de agradecer também ao pessoal da secretaria do Instituto de Computação: Vera, Daniel, Ademílson e Flávio, que sempre estão a postos de bom humor para resolver nossa papelada.

Muito obrigado a todos.

Resumo

A quantidade de informação disponível atualmente na Internet é tão vasta que localizar uma informação desejada em tal depósito torna-se uma tarefa quase impossível. Ferramentas conhecidas como mecanismos de busca retornam, a partir de palavras-chave fornecidas pelo usuário, uma lista de documentos que, supostamente, contêm a informação desejada, relacionada com as palavras-chave fornecidas. Entretanto, multiplicidade semântica das palavras-chave, jargões, entre outros problemas, fazem com que uma grande quantidade de documentos recuperados não tenham de modo algum relação com aquilo que se deseja obter. Para contornar este problema, novas abordagens têm sido introduzidas: categorização dos resultados dos mecanismos de busca, emprego de análise de aglomerados nos documentos etc.

Este trabalho utiliza técnicas de Visualização de Informação para propor uma representação que auxilie o usuário a identificar visualmente, dentre os itens tidos como resposta, quais deles realmente têm possibilidade de sanar suas necessidades. Após uma pesquisa que visou identificar o perfil e os hábitos dos usuários de mecanismos de busca, desenvolvemos um sistema que incorpora vários princípios de Visualização de Informação.

O sistema ReVEL – Representação Visual de Elementos de Lista – provê uma camada de representação gráfica sobre a lista de respostas obtida pelo usuário à consulta feita a mecanismos de busca. Testes de Usabilidade conduzidos com o sistema proposto mostraram sua efetividade comparativamente a sistemas similares e sugeriram novos elementos para continuidade de pesquisa.

Abstract

The amount of information available in the Internet is so vast that finding the desired information in such an unstructured repository easily becomes almost impossible. Search engines were developed and made available to help people locate documents in the Web. These tools collect keywords from the user and return a document list that, supposedly, brings the desired information. A problem arise when ambiguous words are used as keywords: their multiple meanings make the search engine return lots of documents unrelated to the given keywords. New techniques were used to distinguish useful from ordinary documents: categorization of the search results itens, cluster analisys in the documents etc.

This work applies Information Visualization techniques in order to aid the user identify, in a visual way, the needed documents among all the others from the query. We ran a small survey aiming to acquire some information about the search engine user and his/her behavior. Then, we developed a system that makes uses of Information Visualization principles expecting this to be an aid for the user.

The system named ReVEL – *Representação Visual de Elementos de Lista* (Visual Representation of List Elements) – provides a graphical representation layer over the result list returned to the user by the search engine. Usability tests realized with this proposed system show its effectiveness in comparation to similar systems. The results of these user tests also suggest directions for future work.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A Internet está rapidamente se transformando na fonte primária de informação hoje em dia. Praticamente qualquer assunto tem uma explicação ou referência na Internet. A quantidade de informação, porém, é tão vasta que surge um problema: como localizar justamente a informação desejada em tal depósito imenso e desestruturado? Diversos modos de localização apareceram, porém aquele que se estabeleceu foi o mecanismo de busca (*search engine*). Mecanismos de busca são sistemas que, dado um conjunto de palavras-chave fornecidas pelo usuário, retornam uma página contendo uma lista de documentos nos quais aquelas palavras-chave são encontradas. Cada elemento da lista – um documento disponível na Internet – vem acompanhado de uma breve descrição de seu conteúdo, com freqüência nos trechos nos quais algumas das palavras-chave ocorrem, assim como seu endereço – onde o documento pode ser encontrado.

A lista de respostas é supostamente ordenada pela relevância dos documentos com relação ao conjunto de palavras-chave. Infelizmente, devido à multiplicidade semântica, jargões, entre outros problemas, documentos irrelevantes à necessidade de informação do usuário, porém lícitos segundo as palavras especificadas, permeiam a lista diluindo informação útil em uma grande quantidade de informação inútil. Percorrer a lista, examinando os documentos um a um e avaliando sua utilidade facilmente transforma-se em uma tarefa improdutiva e tediosa.

Várias abordagens foram propostas com a finalidade de auxiliar o usuário na tarefa de busca. Alguns classificam os documentos da lista de resultados segundo categorias pré-estabelecidas (Chen e Dumais, 2000), outros criam categorias com base nos próprios documentos encontrados (Hearst e Pedersen, 1996). Há propostas que representam de forma gráfica os documentos, seja distribuindo-os pelo espaço (como a de Benford *et al.*, 1999, ou a de Leuski e Allan, 2000) ou ainda representando onde, dentro dos documentos, aparecem as palavras-chave (Hearst, 1995). Há ainda o Kartoo (Kartoo), que questiona diversos mecanismos de busca e resume para o usuário as respostas de todos eles na forma de um mapa.

Há um campo de estudo, cujos limites ainda são incertos hoje em dia, conhecido como Visualização de Informação. Visualização de Informação é o emprego de representações visuais, interativas, de dados abstratos, utilizando-se das capacidades naturais humanas de processamento visual com a finalidade de amplificar a cognição. Seu objetivo não são as figuras em si, mas a compreensão, a assimilação rápida de informação ou o monitoramento de vasta quantidade de dados (Card *et al.*, 1999). Trabalhos recentes em

Visualização de Informação mostram como o computador pode servir de intermediário no processo de uma assimilação rápida de informação. Extensos conjuntos de dados são reduzidos a uma forma gráfica de tal modo que a percepção humana possa detectar padrões que revelem a estrutura subjacente dos dados mais facilmente que qualquer análise direta dos dados (Robertson *et al.*, 1993).

Um fator que não deve passar despercebido é que os dados a serem representados têm uma característica em comum: são abstratos. Dados como diâmetro, pressão, distâncias, índices pluviométricos etc., têm uma representação imediata por sua relação com grandezas naturais. Como representar, porém, semelhança entre textos, confiança de pagamento, expectativa de retorno de investimento? Valores de cunho abstrato não têm um mapeamento unívoco para uma representação. A meta então é verificar e validar qual representação faz mais sentido para o ser humano.

O objetivo de visualização de texto é transformar informação textual em uma nova representação visual que revele padrões temáticos e relacionamentos entre documentos de um modo similar a como o mundo natural é percebido (Wise *et al.*, 1995).

1.1 Objetivos e Métodos

Este trabalho envolve investigar a problemática do acesso aos resultados de mecanismos de busca e identificar soluções para a representação da informação contida nesses resultados. É nossa proposta aplicar os conceitos de Visualização de Informação no *design* e desenvolvimento de um sistema que represente os documentos recuperados de uma lista de resposta a uma consulta feita a um mecanismo de busca. O *design* e desenvolvimento do sistema devem necessariamente observar as diretivas e recomendações pautadas pela disciplina de Interface Humano-Computador – IHC. Interface Humano-Computador é o estudo de como as pessoas projetam, implementam e usam sistemas computacionais interativos e como computadores afetam indivíduos, organizações e sociedade. Isto engloba não somente a facilidade de uso, mas também novas técnicas de interação que apóiem tarefas do usuário, provendo melhor acesso à informação, além de como informação é apresentada em sistemas computacionais (Myers *et al.* 1996).

Testes de Usabilidade são utilizados para avaliar preliminarmente a interface do sistema. Os Testes de Usabilidade visam avaliar os fatores que caracterizam a usabilidade de um sistema, ou seja: facilidade de aprendizado, facilidade de uso, eficiência de uso e produtividade, satisfação do usuário, flexibilidade, utilidade e segurança no uso (Nielsen, 1993). Adicionalmente às questões de usabilidade da interface propriamente dita estão questões mais diretamente relacionadas à representação proposta.

Neste trabalho é conduzida uma pesquisa inicial junto a usuários de ferramentas de busca para estudar o comportamento que estes apresentam ao relacionar-se com mecanismos de busca, especialmente quanto às suas expectativas de sucesso em localizar as informações desejadas. Essa pesquisa pode dar indicações adicionais de alguma semelhança entre características ou comportamentos relatados pelos usuários e aqueles descritos na literatura.

As principais contribuições esperadas deste trabalho envolvem um maior conhecimento sobre o usuário de mecanismos de busca, seu perfil, comportamento e dificuldades; um sistema que, utilizando princípios de Visualização de Informação, auxilie o usuário na tarefa de identificar, dentre os documentos tidos como úteis pelos mecanismos de busca, quais são realmente aqueles que trazem a informação por ele desejada.

Esta dissertação está organizada da seguinte forma: no Capítulo 2 apresentamos conceitos básicos de Visualização de Informação, suas características e propriedades. Ainda neste capítulo, elencamos algumas das soluções existentes para visualizar resultados de consultas. O Capítulo 3 contém a proposta, realização, resultados e conclusões de uma pesquisa realizada na Unicamp com o intuito de obter mais informações sobre as características sócio-culturais e os hábitos de utilização dos usuários com relação a mecanismos de busca. No Capítulo 4, apresentamos os detalhes de *design* e implementação do ReVEL, uma ferramenta que ilustra nossa proposta de visualização de resultados de consultas a mecanismos de busca. O Capítulo 5 traz as avaliações com os usuários, na forma de Testes de Usabilidade, de três modos de exibição de resultados de mecanismos de busca: a forma de lista comum – Google –, uma visualização disponível na Internet – Kartoo – e nossa proposta – ReVEL. No Capítulo 6, apresentamos uma discussão sobre o problema e sobre nossa abordagem, algumas conclusões obtidas ao longo deste trabalho e propostas para a continuação deste trabalho na forma de trabalhos futuros.

Capítulo 2

Visualização de Informação

A utilização de imagens para representar informações – com finalidades míticas, religiosas, estratégicas etc. – está presente na história do homem já há cerca de trinta ou quarenta mil anos, desde que foram traçadas linhas em argila até as incisões rupestres de cerca de quinze mil anos atrás (Lommel, 1978). O homem tem, ao longo do tempo, disposto do recurso visual para transmitir, perpetuar, informar, detalhar fatos. Atualmente pesquisas têm sido feitas no sentido de tornar a absorção destas informações cada vez mais fácil e natural.

2.1 Conceitos Gerais

Uma imagem vale mais que mil palavras? Larkin e Simon (1987) concluem em seu estudo clássico que diagramas possuem características que os tornam superiores a descrições verbais:

- Diagramas podem agrupar informações que são usadas juntas evitando, portanto, grandes quantidades de busca pelos elementos necessários à inferência utilizada na resolução do problema;
- Diagramas tipicamente usam localização – posicionamento – para agrupar informação sobre um dado elemento, evitando a necessidade de casar rótulos simbólicos;
- Diagramas podem facilitar um grande número de inferências da percepção, as quais são extremamente fáceis para humanos.

Entretanto os autores alertam que apenas os diagramas construídos de modo a tirar proveito destas características são úteis. Estes pontos por si só não garantem que todo diagrama é superior à sua descrição textual.

Um exemplo clássico de uma representação efetiva de informações em um diagrama é o excelente "*Carte figurative des pertes successives en hommes de l'Armée Français dans la campagne de Russe 1812-1813*" de Charles Joseph Minard, exibido na Figura 2.1. Este diagrama por si só, conta a história da perda humana do exército de Napoleão na marcha para Moscou e na sua retirada. Diversas variáveis estão representadas simultaneamente: tamanho, mobilização, localização e direção de movimento do exército além da temperatura

em diversos momentos durante a retirada de Moscou. Ficam evidentes o volume humano despendido ao longo do trajeto – exibido em tom claro na ida até Moscou e em preto na volta – e o manejo das tropas – ramificações ao longo do trajeto – com a clareza que nenhuma tabela numérica poderia representar.

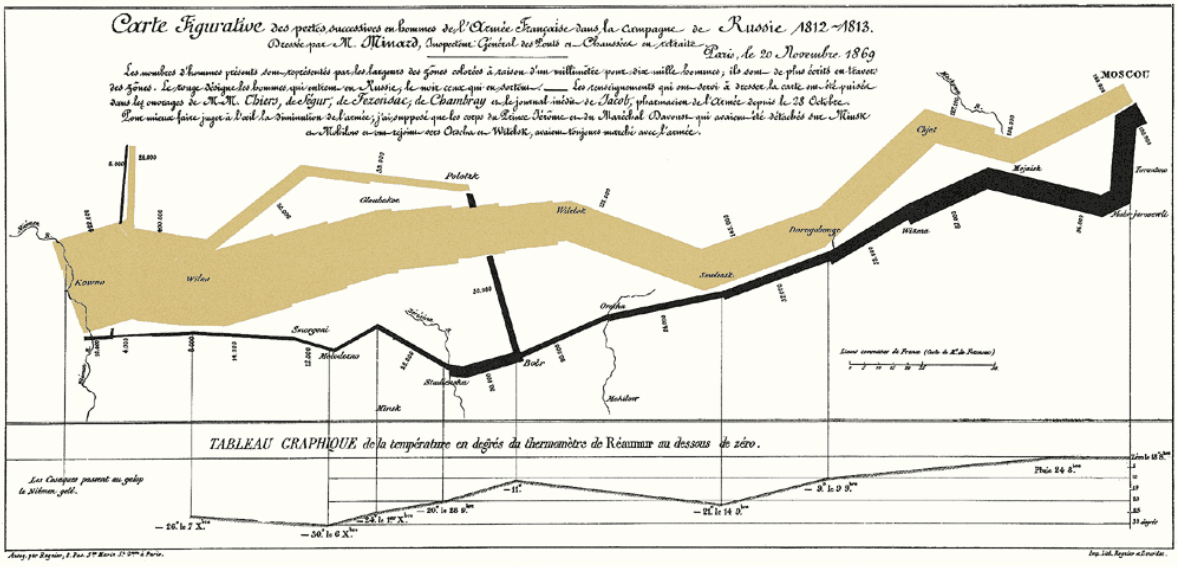


Figura 2.1: Diagrama da Marcha de Napoleão para Moscou
(extraído de www.math.yorku.ca/SCS/Gallery/minard/orig.gif)

Tukey (1977), com seu livro "Exploratory Data Analysis" começou um movimento dentro da Estatística pela utilização de diagramas que visassem uma compreensão estatística imediata dos dados. Mais recentemente, Edward Tufte publicou obras que rapidamente se transformaram em clássicos sobre boas práticas para representar informação. Ele definiu uma teoria para gráficos de dados em "*The Visual Display of Quantitative Information*" (Tufte, 1983) que se empenha em maximizar a quantidade e clareza das informações disponibilizadas. Seu livro "*Visual Explanations*" (Tufte, 2002) elenca quatro recomendações para uma boa apresentação de informação:

1. Coloque os dados em um contexto apropriado para expressar causa e efeito;
2. Faça comparações quantitativas tendo sempre a frase "comparado com o quê?" em mente;
3. Aumente a credibilidade de um relatório é pela descrição cuidadosa de todas as evidências relevantes: considere as explicações alternativas e os casos contrários;
4. Relate os erros possíveis nos números presentes no gráfico.

Abordando técnicas de prestidigitação, Tufte (2002) ressalta os detalhes usados pelos mágicos para iludir as pessoas e os relaciona com o enganar por meio de maus gráficos:

- Suprima o contexto: nunca diga de antemão o que você quer mostrar, o que é importante;

- Evite que ocorra qualquer análise reflexiva: chame a atenção justamente para onde você deseja que ela não vá – especialmente referindo-se aos casos de "*lixográfico*" (*chart-junk*).

2.1.1 Visualização Científica

É impossível para a comunidade científica atual examinar quantitativamente mais que uma minúscula parcela dos valores que são gerados por simulações ou obtidos por instrumentos de medição em grandes fluxos. Para resolver tal problema emprega-se a visualização de dados, de maneira que valores numéricos particulares não sejam importantes e sim a estrutura global das variáveis que constituem a solução assim, como as inter-relações entre elas (DeFanti *et al.*, 1989). Tal visualização é chamada Científica pois os dados expressos visualmente possuem invariavelmente uma representação no mundo físico: moléculas, imagens médicas, estrutura cerebral, meteorologia, astrofísica, mecânica de elementos finitos e até mesmo matemática (Card *et al.* 1999). Este é o fator que a difere da Visualização de Informação, que também é empregada para representar uma massa extensa de dados, porém abstratos.

2.1.2 Visualização de Informação

"A visualização oferece uma interface entre dois poderosos sistemas de processamento de informação: a mente humana e o computador. Visualização de informação é o nome que se dá ao processo de transformar dados, informação e conhecimento em forma visual para utilizar as capacidades visuais naturais humanas. Com interfaces visuais efetivas podemos interagir rapidamente e eficientemente com grandes quantidades de dados para descobrir características, padrões e tendências ocultas" (Gershon, 1998, p. 9).

Visualização de Informação é a utilização de representações visuais interativas de dados abstratos, apoiadas por computador, com a finalidade de amplificar a cognição. Seus objetivos não são as figuras em si, mas a compreensão, a assimilação rápida de informação ou o monitoramento de vasta quantidade de dados (Card *et al.*, 1999).

Card *et al.* (1999) resumem, no primeiro capítulo de seu trabalho, seis motivos pelos quais a Visualização de Informação pode ampliar a cognição:

1. Ampliação dos Recursos

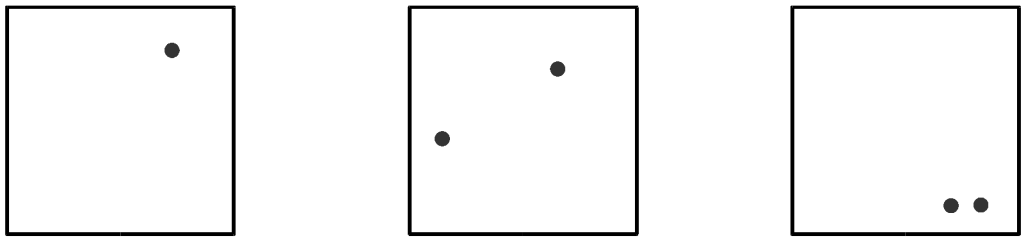
- Liberação de trabalho do sistema cognitivo para o sistema perceptivo, uma vez que algumas inferências cognitivas que são feitas simbolicamente podem ser transcritas em operações perceptivas;
- Expansão da memória de trabalho e de armazenagem, pois informações necessárias já estão visualmente disponíveis para utilização ou para recuperação imediata;
- Processamento perceptivo paralelo: alguns atributos de visualizações podem ser processadas em paralelo, em contraste com texto, que é serial;
- Amplitude de banda para interação hierárquica: sistema visual humano para movimentos combina grande resolução espacial com uma ampla abertura na percepção visual de ambientes;

2. **Redução em Buscas**
   - Localidade de processamento: visualizações agrupam as informações ao dispô-las juntas, reduzindo assim as buscas;
   - Alta densidade de dados: visualizações com freqüência podem representar grandes quantidades de dados em pequenos espaços;
   - Indexação espacial: visualizações podem evitar rótulos simbólicos ao agrupar dados de um objeto;
3. **Ampliação do Reconhecimento de Padrões**
   - Reconhecimento ao invés de recordação: reconhecer a informação gerada por uma visualização é mais fácil que recordar tal informação[2];
   - Abstração e agregação: visualização simplifica e organiza a informação por meio de abstrações e omissões seletivas;
   - Esquematização visual para organização: organizar visualmente dados por relacionamentos estruturais (tempo, dependência etc) evidenciam padrões;
   - Valores, relacionamentos, tendências: visualizações podem ser construídas de modo a evidenciar padrões nestes três níveis;
4. **Inferência Perceptiva**
   - Representações visuais tornam alguns problemas óbvios, uma vez que visualizações têm a capacidade de utilizar inferências perceptivas que são extremamente fáceis de serem realizadas pelos seres humanos;
5. **Monitoramento Perceptivo**
   - Visualização pode permitir o monitoramento de um grande número de eventos em potencial se a representação for organizada de tal modo que tais eventos sejam evidenciados por aparência ou movimento;
6. **Mídia Manipulável**
   - Diferentemente de diagramas estáticos, visualizações podem permitir a exploração de todo o domínio de valores para os parâmetros e amplificar operações dos usuários

O estudo da sintaxe da linguagem visual não é privilégio da ciência. A comunidade das artes já estuda os relacionamentos entre os elementos visuais, notadamente o posicionamento e o seu relacionamento de proximidade dentro de uma área. A importância da proximidade pode ser observada no relacionamento de pontos com o campo em que estão situados e entre si. Um ponto isolado em um campo relaciona-se com o todo, como mostra a Figura 2.2-a, mas ele permanece só e a relação é um estado moderado de intermodificação entre ele e o campo quadrado. A Figura 2.2-b mostra como os dois pontos disputam a atenção em sua interação, criando manifestações comparativamente individuais devido à distância que os separa e, em decorrência disso, dando a impressão de se repelirem mutuamente. Já a Figura 2.2-c mostra uma interação imediata e mais intensa; os pontos se harmonizam e, portanto, se atraem. Quanto maior for a proximidade, maior será sua atração (Dondis, 1997).

[2] Esta propriedade também é relatada por Nielsen (1993) como uma heurística de usabilidade.

Na linguagem visual, os opostos se repelem e os semelhantes se atraem. Assim, o olho completa as conexões que faltam e relaciona automaticamente, com maior força, as unidades semelhantes. O processo perceptivo é demonstrado pelas pistas visuais da Figura 2.3-a, que formam um quadrado (Figura 2.3-b). As pistas foram, então, modificadas na Figura 2.3-c e agora as formas influenciam os elementos que se ligam e a ordem em que se dá a ligação. A Figura 2.3-d mostra possíveis ligações (Dondis, 1997).



(a) (b) (c)

Figura 2.2: Relacionamento de Proximidade

(a) (b) (c) (d)

Figura 2.3: Relacionamento por Semelhança

Propriedades Retinianas

As representações possuem certas propriedades chamadas de retinianas por ser a retina do olho naturalmente sensível a elas. A Tabela 2.1 traz uma relação de cada propriedade, sua classificação, isto é, se é de Extensão ou Diferencial, assim como um exemplo (Card *et al.*, 1999). A classificação refere-se ao fato de a propriedade ser boa para indicar extensão de uma escala em que há uma posição natural em zero – *Extensão* – ou se seu uso principal é para a diferenciação entre marcações / elementos – *Diferencial*.

Algumas propriedades retinianas são mais efetivas que outras para codificar informação. A posição é indubitavelmente a mais efetiva dentre todas. Muitas propriedades são mais efetivas para alguns tipos de dados que para outros: tons de cinza, por exemplo, são efetivos na comparação entre valores ordinais, mas pouco efetivos na representação de variáveis quantitativas. A Tabela 2.1 mostra uma avaliação da efetividade relativa de cada propriedade para representar cada tipo de dado: nominal, ordinal e quantitativo. A avaliação diz se a propriedade é boa para representar determinado dado (+), é pouco efetiva (±) ou se não é capaz de representar o dado satisfatoriamente(-) (Card *et al.*, 1999).

		Nominal	Ordinal	Quantitativo
Extensão	Posição	+	+	+
	Tamanho	+	+	+
	Tons de cinza	-	+	±
Diferencial	Orientação	+	±	±
	Cor	+	±	±
	Textura	+	±	±
	Forma	+	-	-

Tabela 2.1: Propriedades Retinianas
(adaptado de Card *et al.*, 1999)

Um estudo feito por Cleveland e McGill (1984) propõe e avalia codificações gráficas que seriam elementares. Estas codificações seriam: *posição* (ao longo de uma escala comum), *posição* (distribuída ao longo de escalas não-alinhadas), *tamanho*, *direção*, *ângulo*, *área*, *volume*, *curvatura*, *sombreamento* e *cor*. Baseados em informações de fontes variadas – experimentação própria, resultados de psicofísica[3], teoria da psicofísica – os autores supuseram uma ordem nas codificações gráficas, desde aquela que se supõe produzir a maior exatidão de julgamento até aquela com a menor exatidão, como pode ser visto na Figura 2.4. Vale ressaltar que uma combinação de teoria psicofísica e resultados experimentais já havia criado as hipóteses de que julgamentos baseados em tamanho são mais precisos que julgamentos baseados em área, que por sua vez, são mais precisos que

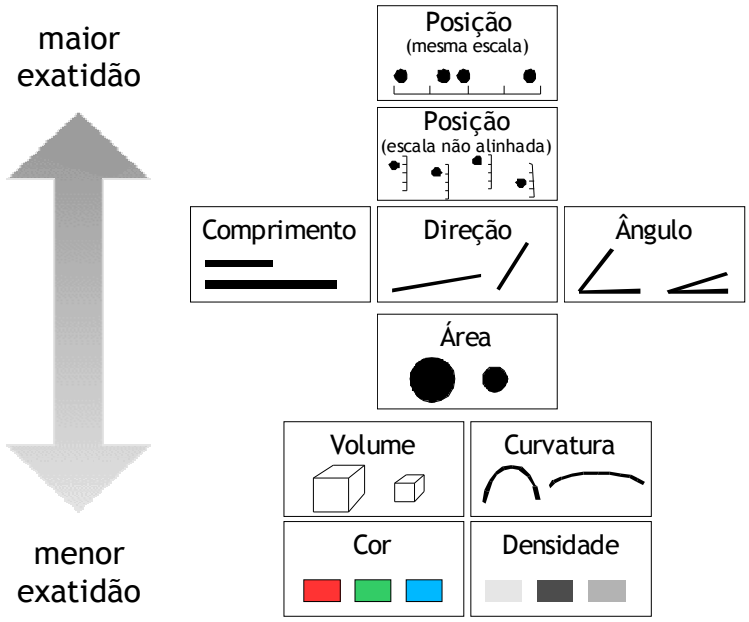


Figura 2.4: Exatidão de Julgamento em Função das Codificações Gráficas Elementares

[3] Estudo das relações entre os fenômenos físicos e os psíquicos (definição do dicionário Michaelis).

julgamentos baseados em volume.

Eles advogam que as representações gráficas deveriam utilizar-se de elementos gráficos com a propriedade "posição" na mais alta posição possível da hierarquia pois assim evocaria julgamentos com maior possibilidade de exatidão e portanto o gráfico maximizaria a habilidade do observador de detectar padrões e organizar a informação quantitativa (Cleveland e McGill, 1986).

Os experimentos de Cleveland e McGill (1986) não validaram todos os postos, mas mostraram que avaliações baseadas em elementos de *posição* foram mais precisos que julgamentos baseados em *tamanho* e *ângulo*. Porém eles ressaltam que o elemento *posição* deve ser subdividido em parcelas e reavaliado pois, para a *posição*, os resultados parecem indicar que a exatidão do julgamento decresce com o aumento da distância entre os elementos.

Um resultado colateral do trabalho de Cleveland e McGill (1986) é que não houve diferença estatística de exatidão de julgamento entre os dois grupos, um composto de mulheres – basicamente donas-de-casa, sem treinamento técnico – e um grupo composto de homens e mulheres com treinamento técnico significativo e com empregos em áreas técnicas. Isto não foi inesperado para os autores uma vez que o objetivo era justamente identificar elementos gráficos que fossem elementares e, portanto, básicos para qualquer observador.

2.2 Recuperação e Visualização de Informação

São descritos a seguir alguns métodos de recuperação de informação que representam, de algum modo, categorias existentes de visualizações.

2.2.1 VR-Vibe

O projeto VR-Vibe (Benford *et al.*, 1999) é um sistema de representação que utiliza métodos estatísticos para analisar bases de documentos. Para representar uma consulta, o VR-Vibe casa cada documento descrito em seu banco de dados com os itens da consulta e calcula uma pontuação a qual representa a atração relativa entre o documento e a consulta. Como sua filosofia envolve a avaliação referente a diversas consultas, esta pontuação é feita também para todas as demais consultas já realizadas anteriormente e ativas no ambiente (Celebourne *et al.*, 1994). Os documentos que atingem um determinado nível para a resposta são representados e sua localização é determinada pela sua pontuação, que é um indicador de atração entre o documento e as consultas. Assim um documento é "atraído" para cada uma das consultas nas quais atinge alta pontuação e localiza-se em algum ponto entre elas.

Cada consulta é composta por uma ou mais palavras-chave e é representada no espaço como um octaedro verde. Os documentos são representados por caixas cujos tamanhos e cores significam maior ou menor relevância do documento frente a todas as

consultas vigentes no momento, conforme a seguinte convenção: cores claras e caixas grandes são documentos muito significativos; cores escuras e caixas pequenas, documentos pouco significativos. O exemplo de uma sessão pode ser observado na Figura 2.5.

Múltiplos usuários podem ser apresentados observando o ambiente segundo a perspectiva que visualizam o ambiente. Cada um dos usuários é representado por uma figura azul semelhante a um martelo.

Não há nenhuma referência de avaliações do sistema feitas por usuários. Portanto não há informação sobre as dificuldades – possivelmente de navegação em ambiente tridimensional – assim como de utilizações alternativas, sugestões etc.

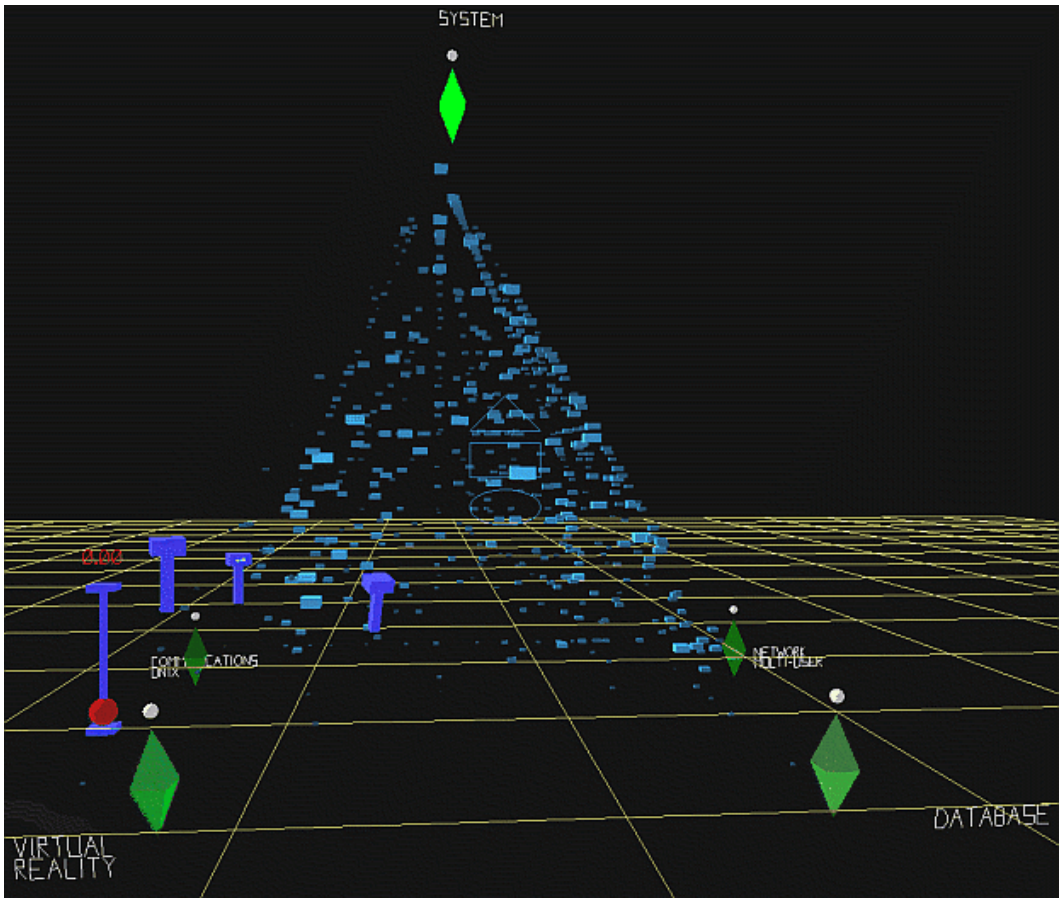


Figura 2.5: Ambiente de Consulta do VR-Vibe
(extraído de Benford *et al.*, 1999)

2.2.2 Scatter/Gather

O sistema *Scatter/Gather* proposto por Hearst e Pedersen (1996), que pode ser observado na Figura 2.6, reúne e apresenta em grupos, aglomerados (*clusters*), os documentos retornados pela consulta que possuam temas similares. Cada grupo é representado por um resumo textual descritivo composto de termos que o caracterizam especificamente, com o intuito de esclarecer qual o seu tópico para o usuário. O próprio sistema cria os grupos em função da similaridade observada entre os documentos recuperados na consulta. A quantidade de aglomerados é arbitrada e seus temas são definidos dinamicamente, em função dos documentos presentes nos resultados da consulta, contrastando com outras abordagens que tentam alocar os documentos em grupos pré-definidos (Chen e Dumais, 2000). Um inconveniente presente nesta abordagem é que nem sempre fica claro para o usuário qual o assunto que causou a reunião de determinados documentos em um grupo. Assim, os aglomerados são rotulados com frases comuns aos documentos e termos freqüentes, no intuito de esclarecer o usuário qual o tópico do grupo.

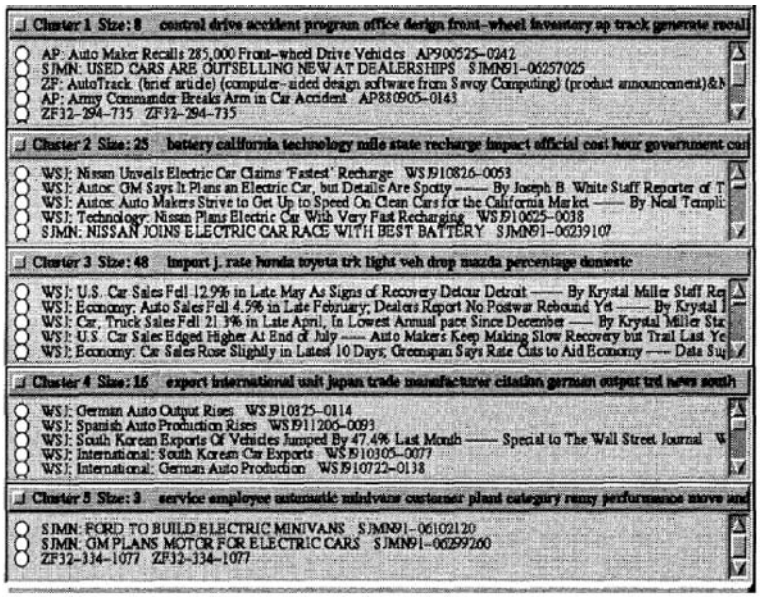


Figura 2.6: Resultados do Sistema *Scatter/Gatter*
(extraído de Hearst e Pedersen, 1996)

A técnica utilizada para avaliar a similaridade entre documentos é conhecido como "cosseno com pesos *tf.idf*". Tal indicador calcula a similaridade em função da direção entre vetores n-dimensionais, nos quais cada coordenada é a freqüência de um termo existente nos documentos.

O sistema foi utilizado para exibir resultados de consultas a uma base de dados padrão composta de documentos governamentais e artigos de revistas estadunidenses.

Os autores do *Scatter/Gather* concluem que sua abordagem produz aglomerados que são talhados especialmente pelas características da consulta, em vez de assumir que os grupos devem possuir um papel classificatório estático. Também consideram que suas conclusões sustentam a Hipótese dos Aglomerados (*Cluster Hypotesis*), que diz que documentos relevantes em uma pesquisa tendem a ser mais similares entre si que entre documentos não relevantes.

2.2.3 Category Interface

A interface desenvolvida por Chen e Dumais (Chen e Dumais, 2000), e por eles referenciada apenas por *category interface* (interface de categoria), estrutura os resultados de mecanismos de busca em categorias. Um exemplo de utilização pode ser observado na Figura 2.7.

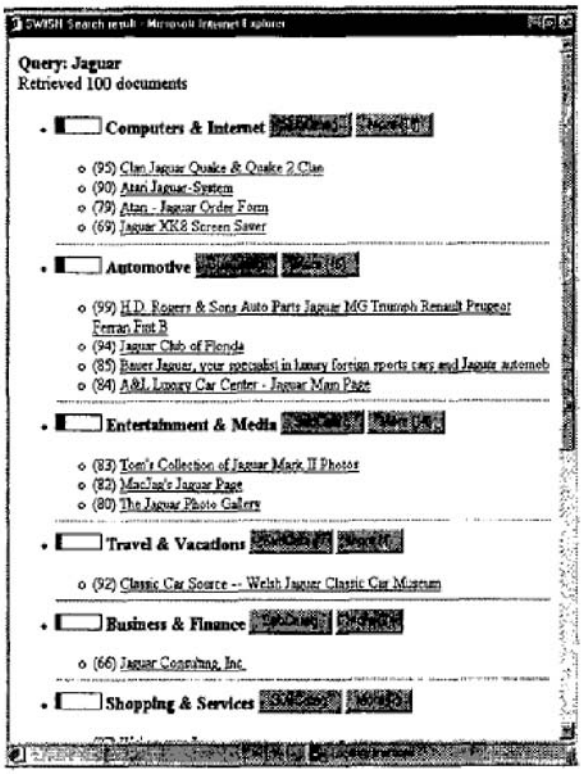


Figura 2.7: Interface do *Category Interface*
(extraído de Chen e Dumais, 2000)

Há distinção entre os métodos que organizam em categorias daqueles que utilizam aglomerados. Ao organizar em aglomerados, os documentos são reunidos por similaridades intrínsecas aos documentos, formando grupos de documentos cujas similaridades sejam as mais altas intra-grupos e as mais baixas inter-grupos. Isto, entretanto, causa o efeito de que nem sempre o motivo que ocasionou a reunião de determinados documentos em um grupo seja aparente. Torna-se geralmente necessário, então exibir algum tipo de rótulo que identifique o aglomerado. A categorização procura ajustar os documentos a uma estrutura de categorias já conhecida e, portanto, fácil de ser reconhecida pelo usuário. A interface apresenta ainda: uma barra de porcentagem antes de cada categoria representando quantos dos documentos da consulta ali se encontram; informações sobre as categorias de nível superior ou inferior e os resumos oferecidos pelos mecanismos de busca, que são exibidos na forma de textos transientes ao se localizar o ponteiro do *mouse* sobre os endereços dos documentos (*links*).

O *category interface* contém um módulo de classificação estatística de texto, que é treinado em um conjunto de documentos representativos de cada categoria. As categorias utilizadas foram as definidas e mantidas por um grupo de 180 editores[4] e são compostas por

[4] www.looksmart.com

13 categorias de primeiro nível e 150 de segundo nível. Assim o sistema fica preparado para classificar imediatamente os documentos nas suas categorias. Os autores testaram esta interface com uma de lista tradicional. Para tal utilizaram tanto critérios subjetivos – questões abertas – quanto critérios objetivos – tempo de busca – e concluíram que o *category interface* foi superior à interface na forma de lista em ambos critérios.

O sistema recebe as palavras-chave da consulta, as remete ao mecanismo de busca e processa a página de resultado. Cada documento indicado na página de resultado é então categorizado, segundo seu resumo, em uma das categorias anteriormente identificadas pelo sistema.

Entretanto uma situação pode reduzir a eficiência do *category interface* àquela da interface de lista. Se a consulta for restrita o suficiente, todos os documentos são classificados como pertencentes à mesma categoria e apresentados como uma lista dentro daquela classificação.

2.2.4 Tilebars

O paradigma de visualização proposto por Hearst (1995) e exibido na Figura 2.8 é bastante particular no sentido de que é o único que se preocupa em representar o tamanho relativo dos documentos, além da distribuição, freqüência e ocorrência dos termos da consulta ao longo dos textos.

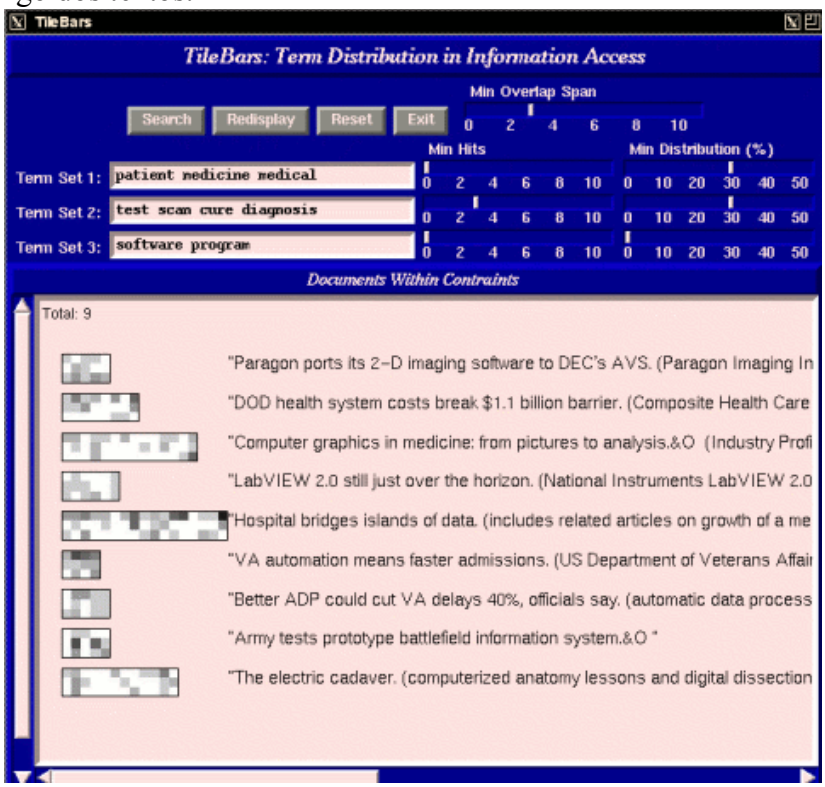


Figura 2.8: Interface do *TileBars*
(extraído de Hearst, 1995)

Utilizando o *TileBars*, o usuário especifica quais são as palavras-chave por conjunto e a consulta retorna uma lista não ordenada de barras gráficas representando documentos presentes no banco de dados – no caso foi exemplificado com um banco de notícias comerciais sobre computadores. As barras apresentam-se superpostas e representam cada um dos conjuntos de palavras-chave. Cada barra está fragmentada em trechos representando partes do documento e os trechos são denotados com tons de cinza, indo do branco até o negro, conforme aumenta a freqüência das palavras-chaves daquele conjunto naquele trecho.

Esta é uma forma interessante de representação, porém requer do usuário uma especificação detalhada dos documentos em palavras-chave. Isso pode ser um problema uma vez que, segundo uma análise feita por Silverstein *et al.* (1999) sobre o comportamento de usuários de mecanismos de busca, 56,4% dos usuários utilizaram um *ou menos* termos na pesquisa e 91,4% utilizaram 2 ou menos termos.

2.2.5 Lighthouse

Leuski e Allan (2000) apresentaram uma interface que exibe simultaneamente tanto a conhecida lista de resultados da consulta quanto uma visualização das similaridades entre os documentos. Esta interface, que pode ser observada na Figura 2.9, representa os documentos como esferas no espaço e a distância entre cada duas é uma função da similaridade entre os documentos que elas representam.

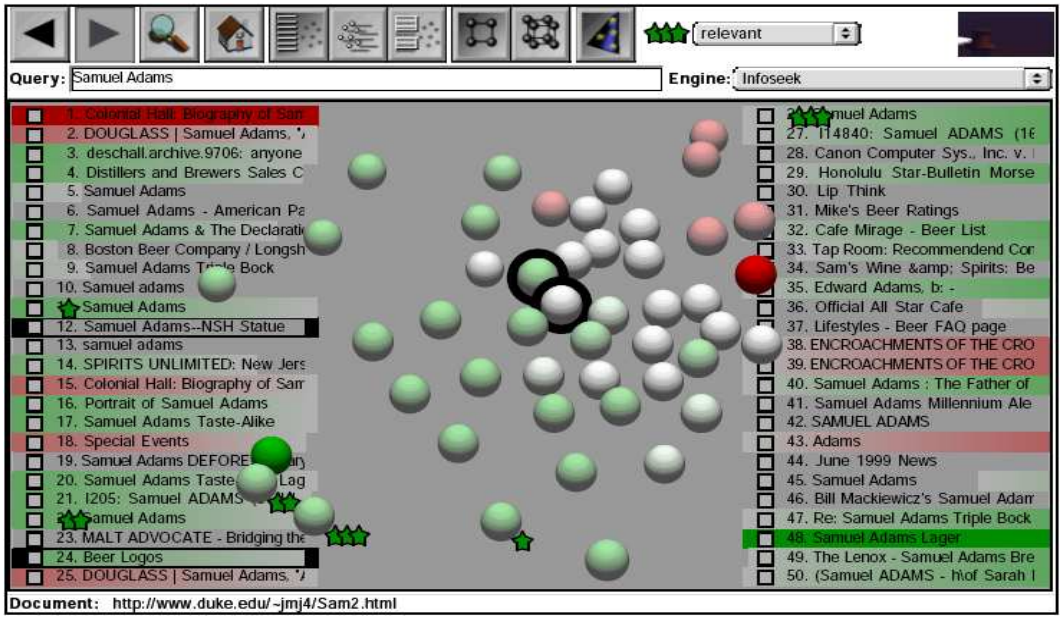


Figura 2.9: Interface de *Lighthouse*
(extraído de Leuski e Allan, 2000)

As esferas que representam os documentos estão vinculadas aos elementos da lista e vice-versa, de modo que ao clicar em um deles o outro também é evidenciado. A exibição das esferas pode ser feita em duas dimensões ou em três, ficando a escolha a cargo do usuário. Os autores relatam, entretanto, que os usuários preferem representações bidimensionais às tridimensionais. Isto confirma o fato, já descrito, de que é necessário um esforço

cognitivo significativo por parte do usuário para recriar um espaço tridimensional a partir de uma imagem bidimensional, sendo que este efeito é ainda mais crítico para usuários inexperientes com computadores (Sebrechts *et al.*,1999).

Os usuários podem identificar documentos como relevantes clicando na caixa de checagem – *checkbox* – ao lado do número do documento. Um clique de *mouse* e tanto a esfera quanto o título são marcados em vermelho, significando "irrelevante". Um segundo clique e o estado passa para "relevante" e a cor da seleção passa para verde. Um terceiro clique e ambos são desmarcados.

Infelizmente os autores não esclarecem quantos usuários avaliaram a interface ou de que modo, fazendo referência apenas a "usuários", "pessoas", "estudo de caso" ou "experimentos com usuários". Deste modo, não é possível saber quantos foram os usuários e como eles fizeram utilização da interface. Detalhes importantes ficam, portanto, sem esclarecimento apropriado:

- As esferas são boas metáforas para documentos? É imediata a compreensão por parte do usuário de que as esferas representam os documentos?
- O fato de ser utilizada a mesma área para exibir a lista e desenhar as esferas, causando sobreposição de elementos, não causa desconforto aos usuários?
- Houve algum tipo de comportamento inesperado? Usuários tentaram, por exemplo, mover alguma esfera de posição ou selecionar um grupo delas?
- É intuitivo para o usuário a utilização das cores feita pelos autores. Por exemplo, a cor vermelha foi facilmente percebida como "irrelevante"?

2.2.6 Kartoo

O sistema Kartoo (Kartoo) é um meta-mecanismo de busca. Ele coleta as palavras-chave do usuário, as submete a diversos mecanismos de busca em paralelo, compila os resultados e os exibe em uma elegante visualização que emprega a metáfora de mapa. Um exemplo de utilização pode ser observado na Figura 2.10, executando uma pesquisa por "medusa". As palavras com maior freqüência aparecem dentro das bolhas relacionando os documentos em que elas aparecem. Assim, ao se passar o *mouse* sobre as palavras, são ligados os documentos nos quais elas aparecem; ao se passar o *mouse* sobre um documento, inversamente, ele é ligado às palavras nele identificadas. Supõe-se que as palavras sejam identificadas pelos resumos retornados pelos diversos mecanismos de busca, pois o Kartoo, segundo esclarece sua página de dúvidas[5], não acessa os documentos diretamente.

[5] www.kartoo.net/a/en/faq.html

Apesar de possuir uma interface com elementos bem distribuídos, a visualização consome muito espaço de tela, sendo que em uma resolução comum de 800×640 *pixels* grande parte da interface fica oculta, notadamente partes das colunas da esquerda e da direita, tentando assim privilegiar a exibição do mapa. Neste caso as barras de rolagem são fundamentais. Problemas acarretados por essa limitação foram detectados no seu teste de usabilidade, discutido na seção 5.2.2. Uma outra limitação é que o Kartoo não é capaz de exibir uma grande quantidade de documentos por vez. Para verificar os outros relacionamentos é necessário utilizar-se do botão "próximo mapa", com freqüência ignorado pelos usuários. Enquanto outros, como o *Lighthouse*, por exemplo, são capazes de exibir cinqüenta documentos por consulta, a visualização do Kartoo já começa a ficar confusa com um número maior que quinze por mapa.

Veremos no próximo capítulo uma caracterização sócio-comportamental dos usuários de mecanismos de busca realizada por meio de um levantamento.

Figura 2.10: Interface do Sistema Kartoo

Capítulo 3

Observação Exploratória da Utilização de Mecanismos de Busca

Foi realizado um levantamento sobre o comportamento dos usuários quanto à utilização de mecanismos de busca. Neste levantamento também foram pesquisadas e avaliadas algumas variáveis sócio-culturais dos usuários envolvidos. Os dados obtidos foram analisados e as conclusões são apresentadas tanto para as variáveis pesquisadas quanto para algumas interações entre elas.

Os mecanismos de busca têm o mesmo padrão de apresentação de respostas: uma lista distribuída em diversas páginas de resultados, conforme se faz necessário devido a quantidade de respostas, onde cada uma delas elenca, em ordem decrescente, as páginas que supostamente contém a informação desejada. A ordem em que aparece cada um dos documentos tidos como resposta à consulta na lista de resultados é dada por métricas particulares internas inerentes a cada mecanismo de busca (Silverstein *et al.* 1999) e para alguns mecanismos de busca estas métricas são alegadamente subjetivas, não havendo um consenso entre os administradores (Kirsch, 1998). Ainda que o usuário saiba exatamente o que procura, é necessário conhecimento do mecanismo de busca subjacente para utilização efetiva dos mecanismos de busca baseados em palavras-chave, uma vez que os mesmos efetuam transformações nas palavras escolhidas, de modo transparente e a revelia do usuário.

Procuramos levantar informações que esclarecessem o posicionamento do usuário, em contexto geográfico e cultural próximos ao nosso, quanto à utilização dos mecanismos de busca. Assim diversas questões ganharam relevo no esforço de caracterizar o relacionamento usuário–mecanismos de busca:

- Com que freqüência os usuários se valem de mecanismos de busca para localizar informações?
- Qual ferramenta ou sistema de mecanismo de busca o usuário utiliza?
- Os usuários de mecanismos de busca encontram a resposta na primeira página de resultados?
- Como os usuários se comportam se a informação desejada não se encontra na primeira página de resultados?

Também procuramos caracterizar o usuário não só com relação ao seu comportamento mas também quanto a características sócio-culturais, idade, sexo e nível de instrução.

3.1 Método da Pesquisa

Uma vez que o universo de usuários de mecanismos de busca é desconhecido e não pode ser estimado sem o apoio das empresas que provém este serviço, tornou-se inviável definir um esquema de referência que possibilitasse uma amostragem probabilística. Uma amostragem probabilística é aquela em que cada elemento passível de ser amostrado possui uma mesma probabilidade *p* diferente de zero de ser selecionado na amostra. Utilizando o esquema de referência, que pretende mapear todos os elementos possíveis, um número determinado destes elementos é aleatoriamente selecionado e avaliado. Este rigor possibilita a utilização de técnicas que permitem a inferência de estatísticas dos elementos amostrados calculadas por meio dos valores obtidos.

Uma vez que não é possível determinar o conjunto de usuários de mecanismos de busca e convocar um sub-conjunto dele a responder o questionário, optou-se por realizar uma amostragem não probabilística baseada na participação espontânea do entrevistado. Embora este método não nos permita extrapolação de resultados para um conjunto além do âmbito dos elementos participantes na amostra, ele é um modo simples e barato de obter algum conhecimento sobre uma população, tendo sempre em mente os limites alcançados pelas conclusões.

Como o objetivo é levantar informações sobre o perfil do usuário de mecanismos de busca, uma amostragem destes elementos é fundamental. Foi realizada então uma pesquisa por participação espontânea que partiu do contexto universitário. Uma página com um formulário eletrônico foi disponibilizada no *site* do grupo ComunIHC – Comunidade de Interface Humano-Computador – (ComunIHC) e passou-se a coletar os dados então fornecidos pelos visitantes. Estes dados foram utilizados para estimar os indicadores sócio-culturais destes usuários assim como seus comportamentos.

3.2 Coleta de Dados

A pesquisa coletou os dados de respondentes voluntários que concordaram em preencher um formulário hospedado no servidor do ComunIHC na Unicamp. Estes respondentes foram informados da pesquisa ou espontaneamente decidiram colaborar ao tomarem conhecimento por meio da notícia (Apêndice II) divulgada na página principal do portal da Unicamp.

A divulgação teve início em 18 de novembro de 2003. Uma chamada no portal da Unicamp remetia à notícia que informava sobre a pesquisa e indicava o endereço do formulário para aqueles que desejassem participar. Mesmo sendo o portal de caráter acadêmico – por tratar-se de uma universidade – sua exposição é de âmbito geral, uma vez que é voltado para a comunidade .

O período de 10 a 12 de dezembro de 2003, em que a notícia ficou em exposição no portal da Unicamp, foi responsável por 47% dos questionários respondidos. Houve também divulgação na forma de uma entrevista para a Rádio CBN, veiculada dia 16/12/04, abordando o assunto e a pesquisa. Em 12/01/04, o formulário foi retirado do ar e substituído por uma nota de agradecimento. Ao total, foram apurados 546 questionários válidos respondidos.

Estima-se que os respondentes sejam integrantes de um ou mais dos conjuntos abaixo:

- Estudantes de pós-graduação ;
- Profissionais de empresas de pesquisa e desenvolvimento da região de Campinas;
- Visitantes genéricos do portal ComunIHC;
- Visitantes genéricos do portal da Unicamp.

3.3 Formulário

O formulário é composto de uma página HTML gerada por um programa em PHP que coleta as informações fornecidas pelo usuário e as envia por correio eletrônico ao coordenador da pesquisa. Uma imagem de sua apresentação encontra-se no Apêndice I.

Propriedades muito desejadas para o formulário foram a sua facilidade e rapidez de preenchimento. O intuito foi evitar tornar o formulário maçante para o respondente, problema que poderia acarretar respostas apressadas, mal consideradas e até mesmo desistência. Assim, vários pontos foram levados em consideração em sua formulação:

- foi feito um esforço para que todo o formulário aparecesse em uma única página do navegador do usuário, evitando que fosse necessário utilizar barras de rolagem;
- procurou-se utilizar a linguagem o mais simples, concisa e objetiva possível;
- nenhuma resposta exigia que o usuário escrevesse algo, todo o formulário poderia ser preenchido e enviado apenas com cliques do *mouse*.

Um experimento piloto foi realizado com um grupo de alunos de pós-graduação e, tendo em vista suas considerações, textos, botões e diagramação foram alterados, visando evitar problemas de compreensão e melhorar a pragmática do preenchimento do formulário.

Foi obtido por fim um formulário que exigia de 20 a 40 segundos para ser preenchido e enviado, com a maior quantidade de perguntas possível, dentro de nossas limitações.

Junto com as respostas foram remetidos também o IP do respondente, a data e a hora em que o formulário foi enviado. A intenção de coletar estes dados suplementares foi a de possibilitar a verificação de integridade e evitar, mesmo que de um modo mínimo, alguns tipos de manipulação.

3.4 Dados

Os formulários preenchidos foram enviados via correio eletrônico para o coordenador da pesquisa. Cada mensagem recebida composta de um formulário respondido continha apenas uma linha com as informações. Um exemplo segue:

DADOS:10.10.10.10[6],12/Dec/2003,08:01:23,1.m,2.s, 3.25,4.7,5.g,6.85,7.n,#

Esta linha foi extraída, tratada e inserida em um SGBD relacional. A partir deste momento esta informação estava disponível para análise.

3.5 Validação

Alguns cuidados foram tomados para evitar que erros, falhas ou má fé permitissem a inserção de dados que não fossem confiáveis. As regras gerais aplicadas foram as seguintes:

- Questionários provenientes do mesmo IP, com diferenças de tempo menor que 1 minuto entre si foram excluídos, à exceção do último;
- Questionários idênticos, provenientes do mesmo IP, com diferenças de tempo menor que 5 minutos entre si foram excluídos, à exceção do primeiro;

Os motivos para estabelecimento destas regras são os seguintes:

- Falha ou sobrecarga no servidor retornam mensagens de erro para o usuário mas, por vezes, acabam por enviar os dados; o respondente imaginando que não conseguiu enviar suas respostas torna a clicar em "enviar" e submete uma outra cópia. Isso pode repetir-se por diversas vezes gerando vários registros iguais com tempos de submissão próximos;
- O usuário fica em dúvida entre uma resposta ou outra, uma vez que o formulário é de alternativas e, para ser "justo", ele envia um formulário com cada resposta;
- Querendo ajudar a pesquisa a recolher uma maior quantidade de resposta, o usuário "compõe" o comportamento de outras pessoas e os submete;
- O respondente simplesmente brinca com o formulário, respondendo a esmo e enviando.

Cada caso em particular foi examinado antes de aplicar as regras gerais de eliminação acima. Há muitas variantes que têm que ser consideradas. Por exemplo, redes internas em geral são servidas por máquinas "proxy", que mascaram o IP do usuário respondendo com um único número IP – o seu – por todas as máquinas da rede interna; isso aparece na pesquisa como diversas respostas provenientes do mesmo IP, como se fossem a mesma pessoa. Estes IPs foram identificados e os formulários validados e aceitos.

3.6 Variáveis

As variáveis necessárias para identificar o perfil sócio-cultural do respondente estão relacionadas na Tabela 3.1 e as variáveis sobre o comportamento do usuário com relação aos mecanismos de busca estão relacionadas na Tabela 3.2.

[6] O número do IP foi aqui alterado para 10.10.10.10 para garantir a anonimidade do respondente.

Variável	Conteúdo
Sexo	Sexo do respondente: Masculino / Feminino
Escolaridade	Nível de instrução do respondente, segundo as categorias: Fundamental, Ensino Médio, Superior e Pós-Graduação
Idade	Idade do respondente segundo as categorias: até 15 anos, de 16 a 25 anos, de 26 a 40 anos e acima de 40 anos

Tabela 3.1: Variáveis Sócio-Culturais

Variável	Conteúdo
Utilização	Freqüência de utilização de mecanismos de busca, assim categorizados : • 1 vez por semana; • 3 vezes por semana e • Todo dia (7 vezes por semana).
Mecanismo de Busca	Qual o mecanismo de busca preferencialmente utilizado, segundo ordem alfabética: • AlltheWeb • Altavista • Google • Outro
Sucesso	Qual o sucesso estimado pelo usuário em obter a informação desejada já na primeira página de resultados: • Raramente (até 15% das vezes) • Algumas Vezes (de 16% a 50% das vezes) • Freqüentemente (de 51% a 85% das vezes) • Quase Sempre (de 86% a 100% das vezes)
Estratégia	Como se comporta quando a informação desejada não se encontra na primeira página de resultados oferecida pelo mecanismo de busca: • Refaz a pesquisa com novas palavras-chave • Requisita página seguinte da consulta corrente • Troca de mecanismo de busca • Desiste de utilizar mecanismos de busca

Tabela 3.2: Variáveis Comportamentais no Contexto de Mecanismos de Busca

3.7 Análise

As variáveis foram examinadas em uma primeira instância uma a uma e depois aos pares, com o intuito de verificar que informações elas poderiam nos fornecer sozinhas ou relacionadas entre si. Cada variável foi tabelada e um gráfico foi traçado para analisar a distribuição dos valores.

Sexo

Como mostra a Figura 3.1, a quantidade de pessoas do sexo masculino que respondeu ao questionário é aproximadamente duas vezes maior que aquela do sexo feminino. A porcentagem duas vezes maior de homens que mulheres pode indicar que:

- Homens são mais propensos a responder questionários na Internet que mulheres;
- Há uma proporção maior de homens utilizando a Internet;
- Por algum motivo, mais homens foram expostos à pesquisa que as mulheres, ou
- Uma composição dos motivos anteriores.

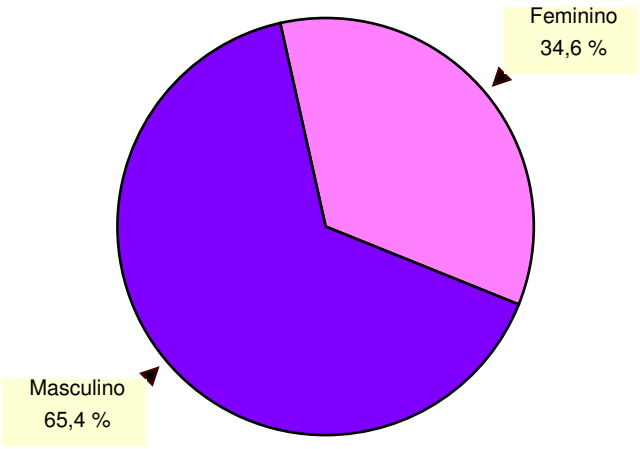


Figura 3.1: Distribuição do *Sexo*

Escolaridade

Por meio da Figura 3.2 constatamos que a quase totalidade dos respondentes têm nível de instrução igual ou maior que o nível superior. Em outras palavras, 92,2% tem nível superior ou pós-graduação, completa ou em andamento. Uma vez que isso é o inverso do retrato social do país, há um forte indício de que a abrangência da pesquisa resume-se a meios de pesquisa e acadêmicos. Os poucos participantes do ensino médio – cerca de 7% – ocorreram durante o período de exposição da chamada da pesquisa no portal da Unicamp.

A Figura indica que, excetuando-se os 6,8% representantes do ensino médio e os 0,9% que não responderam ou têm ensino fundamental, os respondentes dividem-se quase igualmente entre possuir ensino superior e possuir pós-graduação.

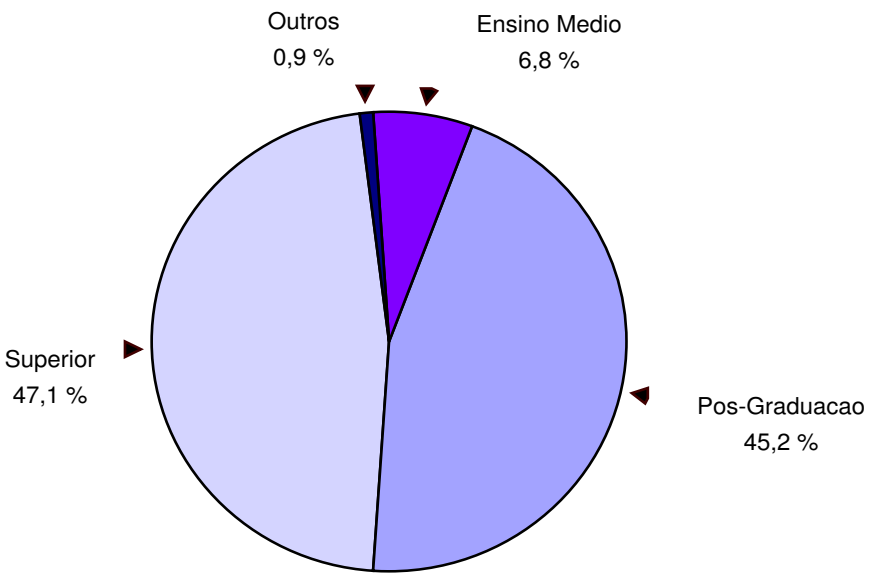


Figura 3.2: Distribuição da *Escolaridade*

Idade

Podemos ver pela Figura 3.3 que a idade dos usuários se distribui ao redor da faixa etária de maior freqüência – que vai de 26 a 40 anos – representando 44,9% da população. A segunda maior classe, de pessoas com idade entre 16 e 25 anos, representa 33,7% da população. Usuários que têm mais que 40 anos representam 21,1% de todos os respondentes.

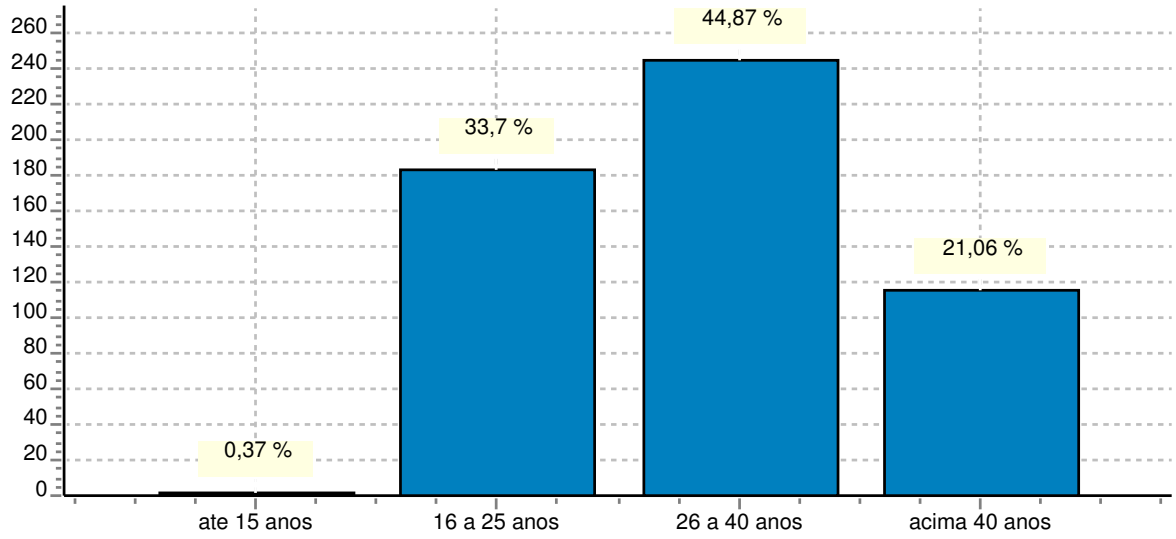


Figura 3.3: Distribuição da *Idade*

Utilização

A taxa de utilização parece indicar que a tendência é de que a maior quantidade de pessoas utiliza mecanismos de busca com freqüência, conforme podemos ver pela Figura 3.4. Assim, cerca de 2/3 dos respondentes utilizam mecanismos de busca todos os dias e essa proporção diminui conforme cai o número de dias por semana em que os mecanismos são utilizados.

Há uma associação indicando que aqueles que utilizam mecanismos de busca três vezes por semana são três vezes mais numerosos que aqueles que utilizam apenas um dia por semana e o mesmo para todo dia – sete vezes por semana – : o número de usuários diários é cerca de sete vezes maior que o usuário com utilização semanal.

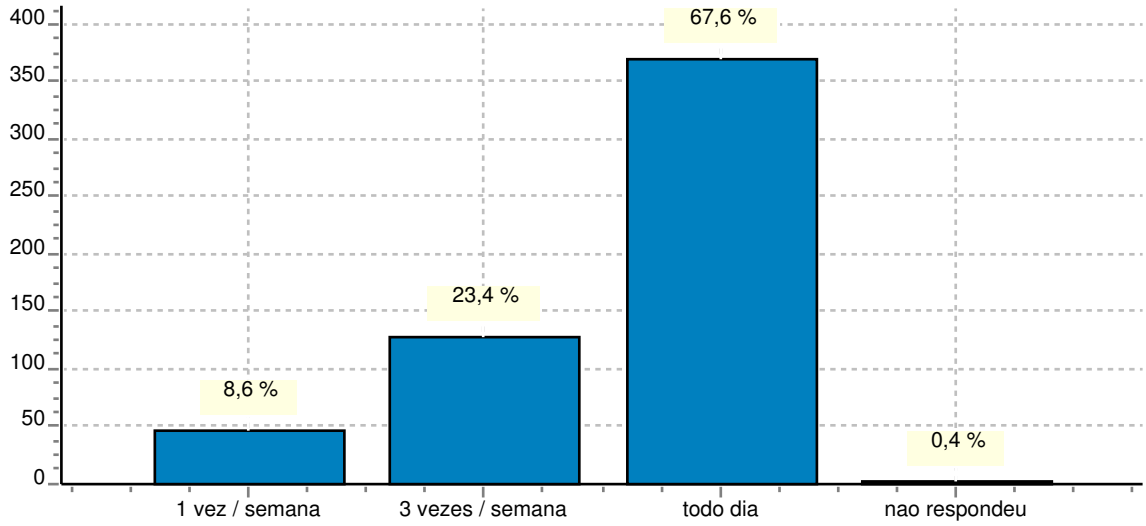


Figura 3.4: Distribuição da *Utilização* de Mecanismos de Busca

Mecanismo de Busca

Os mecanismos de busca escolhidos para serem relacionados foram aqueles que apresentam a maior base de dados indexada segundo o grupo "Search Engine Showdown" (Search Engine Showdown). O ideal seria não o tamanho da base de dados e sim a taxa de utilização diária. Esta informação entretanto não se encontra disponível. A eleição dos três mecanismos em questão parece não ter preterido um outro, desconhecido, que seja de preferência popular uma vez que, somando-se todos os demais mecanismos não nominados na pesquisa, chega-se ao total de apenas 7,1%.

O revelado é que o sistema Google (Google) detém quase 90% da preferência destes usuários. O Altavista (Altavista), antigo líder de preferência, conta com menos que 3% e os demais juntos – Alltheweb (Alltheweb), casos não respondidos e outros mecanismos – compõem cerca de 8% do total. Os valores estão representados na Figura 3.5.

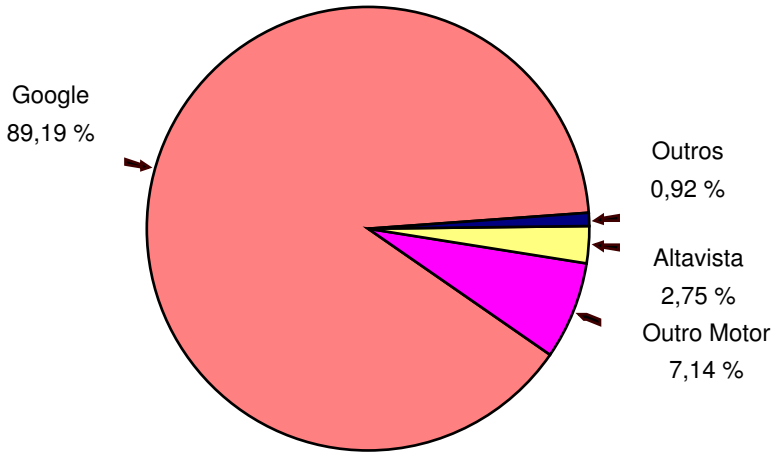
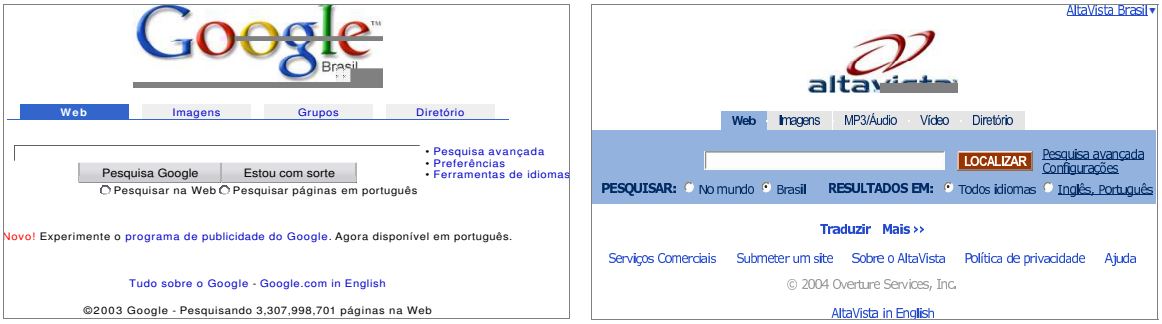


Figura 3.5: Distribuição dos *Mecanismos de Busca*

A Figura 3.6 mostra exemplos da interface dos dois mecanismos de busca identificados pelos usuários como os mais freqüentemente utilizados: Google (Figura 3.6a) e Altavista (Figura 3.6b). Pode-se identificar semelhanças no projeto das interfaces e as suas pragmáticas intencionadas:

- As páginas têm aspecto limpo, com o logotipo evidente. A função é a fixação da marca e um esforço para prover a maior simplicidade de utilização possível;
- A linha para entrada de texto é ampla e central, buscando visibilidade e com isso sugerir a digitação das palavras-chave;
- As abas com outras opções de busca (imagens, grupos etc) são bem destacadas, além do fato de que já há uma escolha clara pré-definida por busca na Internet;
- Opções de menor relevância aparecem, mas com menor ênfase, de modo a não competir com a função primordial desta página do mecanismo de busca, que é "coletar palavras-chave e iniciar pesquisa" e
- Apesar das diferenças de cores e estilos, pode-se verificar que diagramaticamente ambas são extremamente semelhantes, sugerindo que esta disposição dos itens é tida como a mais direta para a entrada de dados da pesquisa.



a) Google b) Altavista

Figura 3.6: Exemplos de Interface de Usuários dos Mecanismos de Busca (visitados em maio de 2004)

Sucesso

Por meio desta variável, procuramos estimar a taxa de sucesso reportada pelo usuário em obter a informação desejada já na primeira página de resultados retornada pelo mecanismo de busca. A Figura 3.7, que mostra a distribuição dos usuários conforme suas taxas de sucesso, sugere que cerca de 40% dos usuários têm a expectativa de que "freqüentemente" – entre 51% e 85% das vezes – obtém a informação desejada na primeira página de resultados.

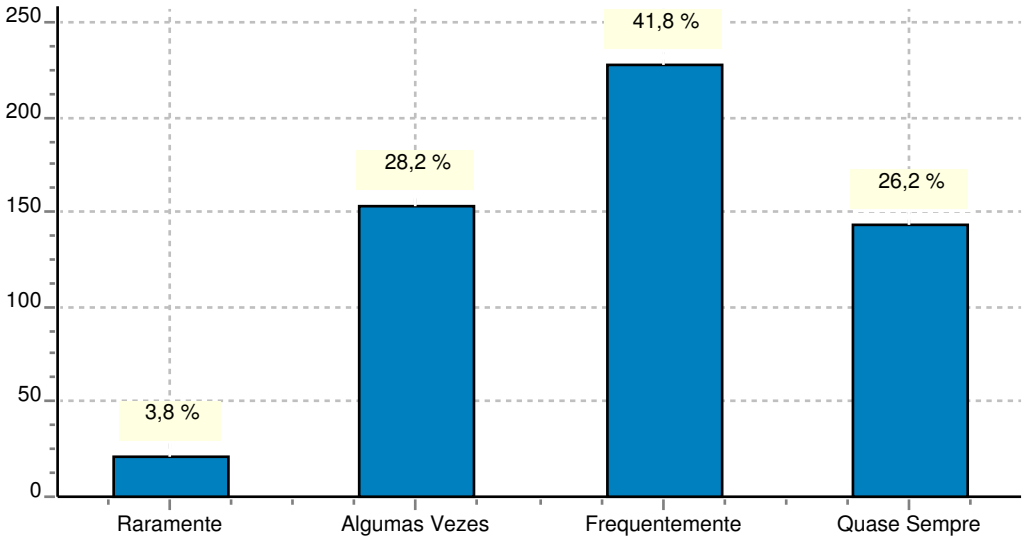


Figura 3.7: Distribuição da Expectativa de *Sucesso* na Primeira Página de Resultados

Cerca de 1/4 dos respondentes alega conseguir já na primeira página a informação procurada; apenas cerca de 4% reportam raramente serem satisfeitos na primeira página.

Podemos notar que 2/3 dos usuários já reportam uma expectativa de mais de 50% de chance de obter a informação na primeira página de resultados. Isso pode ser explicado levando-se em conta o desenvolvimento da habilidade dos usuários em otimizar sua utilização dos motores, ou seja, que aprenderam a entrar com as palavras nos termos do sistema.

Uma análise feita por Silverstein *et al.* (Silverstein *et al.*, 1999) observa que, dentre as sessões de consultas submetidas ao Altavista durante os 43 dias avaliados, 63,7% eram constituídas de apenas uma consulta, na qual apenas uma página de resultados foi examinada: a primeira. Os autores desconhecem a razão deste comportamento e suspeitam de alguns motivos: a necessidade foi satisfeita na primeira página, ocorreu desistência ou havia um desconhecimento dos recursos de refazer a pesquisa ou requisitar a próxima página.

Estratégia

Já vimos que cerca de 2/3 dos usuários têm uma expectativa maior que 50% de encontrar a informação desejada já na primeira página. Mas como se comportam quando isso não ocorre? Como agem frente ao insucesso? Segundo suas respostas, representadas na Figura 3.8, observamos que os usuários dividem-se precisamente entre refazer a pesquisa alterando as palavras-chave e entre requisitar a página de resultados seguinte, utilizando a mesma pesquisa. Supõe-se que isso seja devido à expectativa, por parte do usuário, do fato de as palavras-chave utilizadas serem ou não adequadas. Assim, se ele está satisfeito com as respostas obtidas segundo suas palavras-chave, porém não encontrou a informação na primeira página de resultados, ele se vê motivado a continuar com a mesma pesquisa, requisitando a página seguinte. Por outro lado, se a primeira página não parece promissora, ele refaz a consulta com outras palavras-chave.

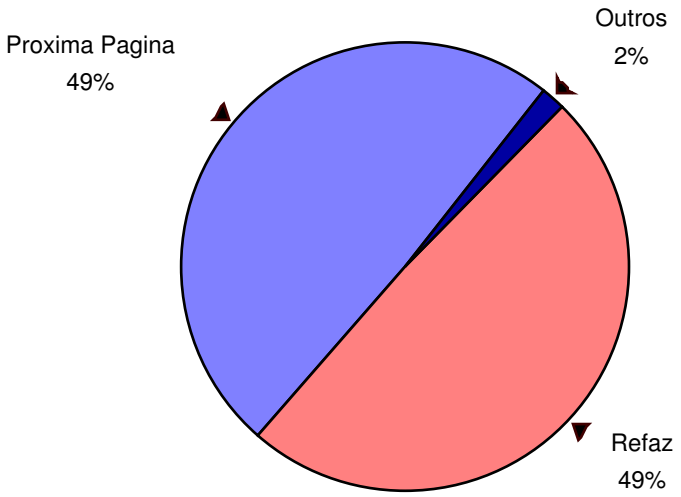


Figura 3.8: *Estratégia* frente Insucesso na Primeira Página de Resultados

Uma parcela residual alega que, frente ao insucesso, ou trocam de mecanismos de busca ou desistem completamente de utilizar este método para obter a informação que desejavam. Estes comportamentos não chegam a somar 1,5% das respostas.

Discussão

Este levantamento foi bastante produtivo na medida em que permitiu identificar algumas características do perfil e comportamento de um conjunto de usuários de mecanismos de busca.

A Figura 3.9 mostra a presença de algum tipo de associação entre freqüência com que mecanismos de busca são utilizados e a taxa de sucesso em obter os resultados logo na primeira página de resposta. De fato, utilizando técnicas de regressão aplicadas a dados categóricos (Everitt, 1977), observamos que há significância estatística para rejeitar a hipótese de que a variável *Sucesso* é independente da variável *Utilização*. Isso quer dizer que a quantidade de pessoas por nível de *Sucesso* não é a mesma, proporcionalmente, por nível de *Utilização* (e vice-versa). A chance de que a rejeição da hipótese de independência entre essas variáveis esteja incorreta devido a uma amostra infeliz é menor que uma em mil ($p < 0{,}001$).

A associação entre *Sucesso* e *Utilização* poderia ser explicada por um possível efeito do aprendizado sobre a utilização desses mecanismos, onde uma maior utilização acarretaria maior conhecimento das particularidades dos mecanismos de busca e isto, por sua vez, uma maior expectativa de sucesso em obter a informação já na primeira página, ou seja, o usuário que possui o modelo mental do sistema tem maior expectativa de sucesso uma vez que sabe formular sua busca nos termos do sistema, a exemplo do que mostra a Teoria da Ação de Norman (Norman, 1983; Norman e Draper, 1986).

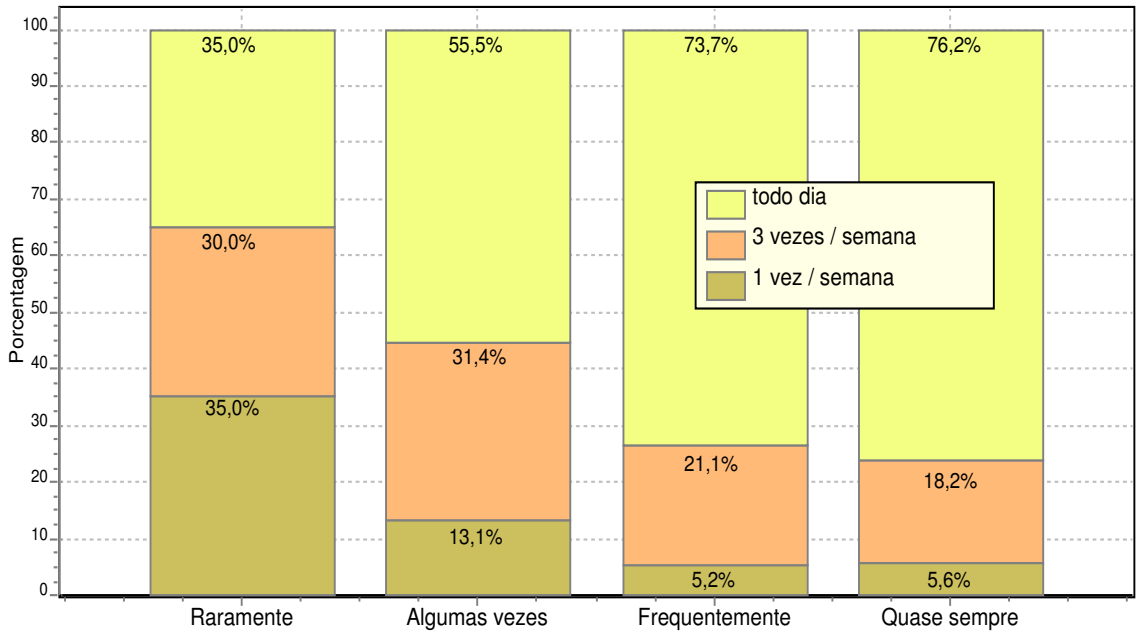


Figura 3.9: Gráfico da Freqüência de *Utilização* pelo *Sucesso* na Primeira Página

O gráfico representado na Figura 3.9 mostra que as categorias que exprimem cada vez mais expectativa de sucesso (raramente → algumas vezes → freqüentemente → quase sempre) são compostas cada vez mais por pessoas que utilizam mecanismos de busca todos os dias. A dependência entre as variáveis permite também que os resultados sejam lidos no sentido contrário: quanto mais dias da semana o usuário alega utilizar mecanismos de busca, mais cresce, proporcionalmente, a expectativa de sucesso em ser satisfeito já na primeira página de resultados.

Pela natureza do plano pelo qual as amostras foram obtidas, o efeito de causalidade – maior taxa de uso implica em maior sucesso ou vice-versa – pode apenas ser suspeitado mas não afirmado. Porém pode-se afirmar que estas variáveis estão associadas, ou seja, o

crescimento em uma delas é acompanhado pelo crescimento – ou decrescimento, caso a associação fosse negativa – da outra variável. Um novo planejamento de experimentos seria necessário para verificar de fato a implicação.

Outras relações de interesse foram identificadas entre as variáveis. Observou-se que a variável *Sexo* não se distribui de modo uniforme entre as categorias de *Idade* ($p < 0{,}001$). A proporção de mulheres, por exemplo, que utilizam mecanismos de busca cresce significativamente de acordo com os níveis de idade, conforme pode ser visto na Tabela 3.3.

Idade	População Feminina
16 a 25 anos	21 %
26 a 40 anos	36 %
Acima de 40 anos	52 %

Tabela 3.3: Distribuição do *Sexo* Feminino pelos Níveis de *Idade*

A expectativa de sucesso (*Sucesso*) não está associada à *Idade* ($p = 0{,}48$). Assim, não há indícios de que a expectativa de sucesso distribui-se de modo que não seja com a mesma proporção dentro das faixas etárias. Não foi verificado, por exemplo, que os mais jovens têm menor (ou maior) expectativa de sucesso que aqueles com mais idade. Tampouco há associação entre *Idade* e *Estratégia* ($p = 0{,}73$): pessoas mais jovens não podem ser distinguidas das mais velhas quanto a procurar pela página seguinte ou refazer completamente a pesquisa.

Não há associação estatisticamente significativa ($p = 0{,}52$) entre *Estratégia* e *Utilização*, ou seja, as pessoas que responderam o questionário adotam, com a mesma proporção, a estratégia de procurar por próxima página ou refazer a consulta independente do fato de utilizarem mecanismos de busca um, três ou todos os dias da semana.

A alta taxa de respostas para usuários que utilizam mecanismos de busca todos os dias e possuem nível superior ou pós-graduação – 340 pessoas, 62,3% da amostra – permite supor que os resultados sejam válidos para este estrato específico da população, tolerando-se um possível viés nos resultados. Este viés seria devido à ausência, por exemplo, de representantes de nível superior de outras partes do país. Supondo, porém, que a classe com nível superior ou pós-graduação seja relativamente homogênea quanto ao seu comportamento com mecanismos de busca, é factível admitir que o viés sofrido pela não representatividade é pequeno e não invalida os resultados.

Uma das análises possíveis de serem realizadas sobre este estrato ocorre na Figura 3.10, onde é examinada a expectativa de *Sucesso* para usuários com *Escolaridade* superior ou pós-graduação que têm *Utilização* de mecanismos de busca diária.

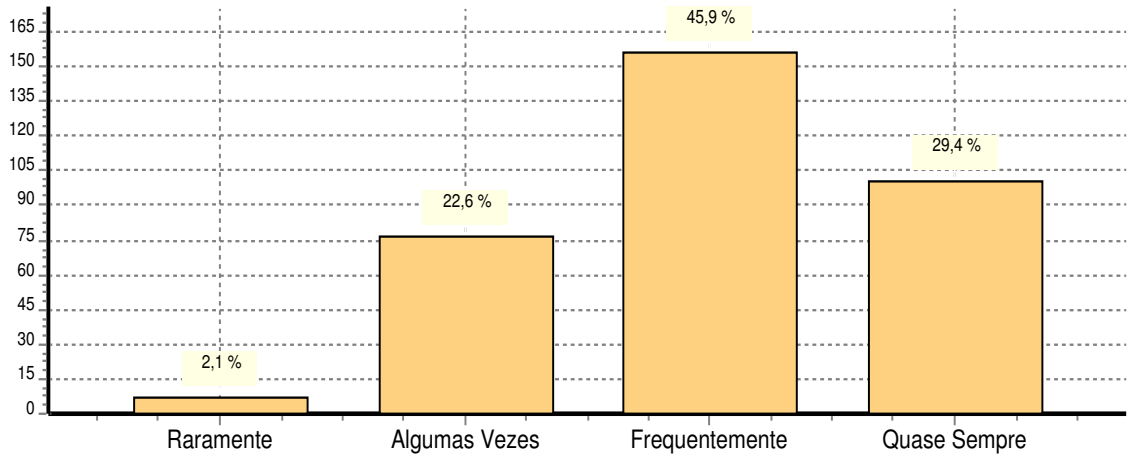


Figura 3.10: Gráfico do *Sucesso* por *Escolaridade* com *Utilização* Diária

Sintetizando os resultados obtidos para o grupo de respondentes, verificamos que:

- Cerca de 2/3 dos usuários é do sexo masculino;
- Há uma associação estatisticamente significativa entre *Sucesso* e *Utilização*;
- A proporção de usuários do sexo feminino cresce conforme crescem os níveis de *Idade*;
- A estratégia de procurar por uma próxima página ou refazer a pesquisa independe da freqüência com que se utiliza mecanismo de busca;
- A expectativa de *Sucesso* não está associada à *Idade*;
- A *idade* não está associada à *Estratégia* de procurar pela página seguinte ou refazer a pesquisa.

O usuário:

- Possui ou está cursando nível superior ou pós-graduação;
- Possui, na maioria, de 26 a 40 anos de idade;
- Utiliza mecanismos de busca todos os dias;
- Prefere o mecanismo de busca Google;
- Estima que freqüentemente – de 51% a 85% das vezes – obtém as informações desejadas já na primeira página de resultados;
- Divide-se entre procurar a próxima página de resultados e refazer a pesquisa com outras palavras-chave quando não consegue as informações na primeira página;

3.8 Considerações

Avaliamos aqui os resultados desta pesquisa que visou caracterizar, mesmo que em apenas um subconjunto, o usuário de mecanismos de busca, utilizando como base dados de seu comportamento fornecidos espontaneamente. Os resultados revelaram que os respondentes – cerca de 90% do total – têm preferência por um sistema de busca específico. Os resultados indicam ainda uma associação estatisticamente significativa entre sucesso na

busca e freqüência de utilização do mecanismo de busca, sugerindo que os respondentes são bem sucedidos na medida em que aprendem a formular sua procura utilizando palavras-chave em termos do modelo mental que criam do funcionamento do sistema.

A ferramenta "mecanismo de busca" é um recurso cotidiano, com desempenho bastante bom – 68% dos participantes da pesquisa têm expectativa maior que 50% de sucesso – mas com espaço disponível para ser melhorado. Um modo de ampliar a expectativa de sucesso poderia ser fornecendo ao usuário informações extras que o auxiliassem. Um exemplo seria prover a próxima página – com o cuidado que o acréscimo de informação não sobrecarregasse o usuário – pois verificamos que metade dos respondentes a solicitam quando o resultado obtido não é satisfatório. Visualização e representação gráfica de resultados de busca seriam soluções recomendáveis de serem avaliadas uma vez que permitem a apresentação de uma grande quantidade de informação que é processada com baixo esforço cognitivo.

Esta pesquisa revelou padrões de comportamento de usuários de mecanismos de busca que os caracterizam como usuários sofisticados, que valorizam – e praticam – o acesso à informação na Internet. Isso pode significar que tais sistemas de busca têm se ajustado a essa categoria de usuários (ou vice-versa).

O próximo capítulo traz uma descrição pormenorizada do sistema ReVEL, com suas propriedades e suas partes componentes.

Capítulo 4

ReVEL - Representação Visual de Elementos de Lista

Este capítulo descreve o ReVEL – Representação Visual de Elementos de Lista – um sistema de software desenvolvido com o intuito de prover uma camada de representação gráfica sobre a lista de respostas obtidas pelo usuário à consulta feita a mecanismos de busca (Zaina e Baranauskas, 2005).

4.1 Modelo Conceitual

A proposta deste trabalho é obter um modo de exibir visualmente, na forma de um diagrama, tanto a relação quanto a intensidade da relação que existem entre os documentos que são tidos como solução para um questionamento feito pelo usuário a um mecanismo de busca. Esta representação, uma vez que é mediada por um sistema computacional, deve obedecer aos preceitos e princípios investigados na disciplina de Interface Humano-Computador.

A representação adotada nesta proposta exibe os documentos como ícones. Cada documento pode ter sua relação com um outro indicada na forma de ligações – arestas – conectando ambos os ícones dos documentos. Se não houver relação entre os documentos ou se esta for menor que um limite determinado dinamicamente durante a análise, nenhuma ligação é exibida entre os ícones destes documentos.

A posição dos ícones é definida e redefinida em tempo real, conforme mais informações vão sendo obtidas sobre eles e seus relacionamentos com os demais. Um algoritmo tenta ajustar a posição de modo que os ícones dos documentos mais relacionados entre si estejam mais próximos e os ícones dos documentos menos assemelhados estejam mais distantes ou com tendência de se afastar uns dos outros. Para evitar que os ícones se movimentem o usuário pode desligar o sistema de auto-posicionamento ou, tratando o ícone de modo individual, "pregá-lo" no lugar. Para fixá-lo no lugar o usuário pode clicar com o botão direito do *mouse* sobre o seu ícone ou selecionar a caixa de checagem – *checkbox* – correspondente na tabela de informações sobre cada documento. O usuário também pode arranjar a disposição dos ícones a seu gosto clicando e arrastando os ícones, a qualquer momento.

Conforme descrito no capítulo anterior, é uma das práticas correntes ao se fazer uma busca usar o recurso de refazer a pesquisa com outras palavras-chave. Uma possibilidade disponibilizada pelo ReVEL é a de manter determinados documentos para a pesquisa seguinte. Caso determinados documentos de uma pesquisa sejam relacionados com o assunto desejado, porém não tragam especificamente a informação desejada, é interessante manter, reter, estes documentos para a pesquisa seguinte e observar quais documentos aparecerão relacionados àqueles mantidos. É uma boa suposição a de que os documentos da nova pesquisa que estiverem ligados aos documentos da pesquisa anterior – retidos por causa de seu assunto – tenham não só o mesmo assunto como espera-se que, agora, tragam a informação procurada.

Um exemplo de início de utilização do sistema pode ser visto na Figura 4.1.

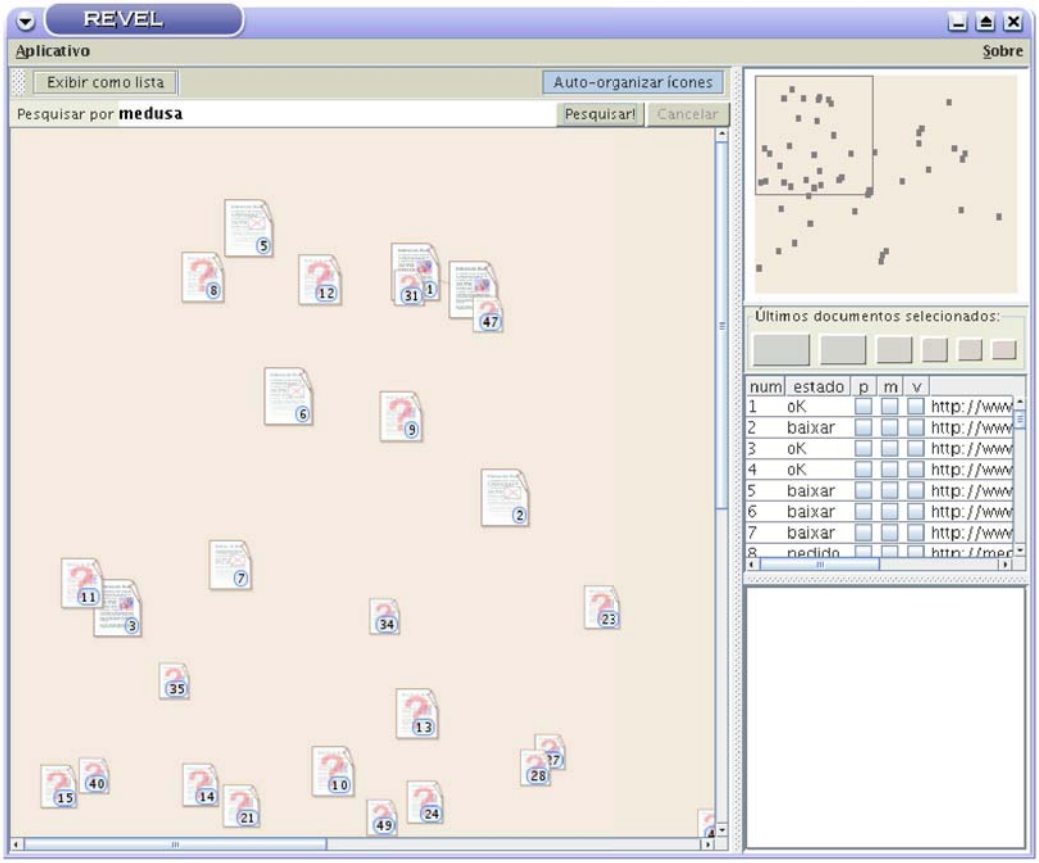


Figura 4.1: Disposição da Aplicação após a Consulta

Os passos realizados pelo usuário na utilização do ReVEL para identificar a informação são variados, porém um modelo genérico para descrever a interação do usuário com o ReVEL pode ser visto na Figura 4.2 abaixo.

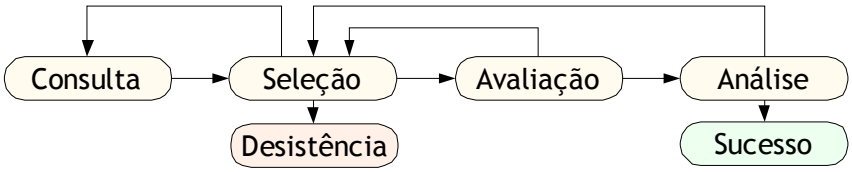


Figura 4.2: Modelo de Interação para Obtenção de Informação no ReVEL

4.2 Operação

Uma consulta em um mecanismo de busca é realizada quando o usuário preenche um campo na página com palavras-chave e ativa a pesquisa, seja pressionando *enter*, seja clicando em um botão que inicie o processo – em geral **pesquisar**, **procurar** ou **encontrar**. O botão costuma estar disposto ao redor da caixa de entrada de palavras-chave, normalmente à direita. Como tal procedimento é praxe nas páginas dos mecanismos de busca, é fundamental que ReVEL ofereça a mesma disposição e o mesmo comportamento, a fim de facilitar a compreensão e a sua utilização pelo usuário.

4.2.1 Execução do Sistema

Uma vez que o ReVEL é um sistema em Java disponível na forma de um pacote `.jar` (*Java Archive*) o modo natural de ser executado em uma janela de terminal é por meio do comando:

```
java -jar revel.jar
```

ou clicando no ícone do sistema em um ambiente gráfico de um computador que possua a máquina virtual Java instalada.

4.2.2 Apresentação da Interface

Iniciada a execução do sistema ReVEL, fica disponível para o usuário uma interface que tem a estrutura representada na Figura 4.3. Seus elementos são, por áreas:

- Controle da Consulta, onde está o campo a ser preenchido para alimentar a consulta;
- Área de Representação, local que exibe a representação gráfica da lista de respostas;
- Visão geral da Área de Representação, onde é exibida uma visão completa, com menos detalhes, da Área de Representação;
- Seleções mais Recentes, botões com os índices dos documentos mais recentemente selecionados;
- Tabela Índice de Documentos, área onde são listados os dados dos documentos e disponibilizados alguns controles sobre a exibição de cada um deles, e
- Resumo do Documento, onde é exibido um resumo do conteúdo do documento.

A seguir, uma explanação mais detalhada de cada área.

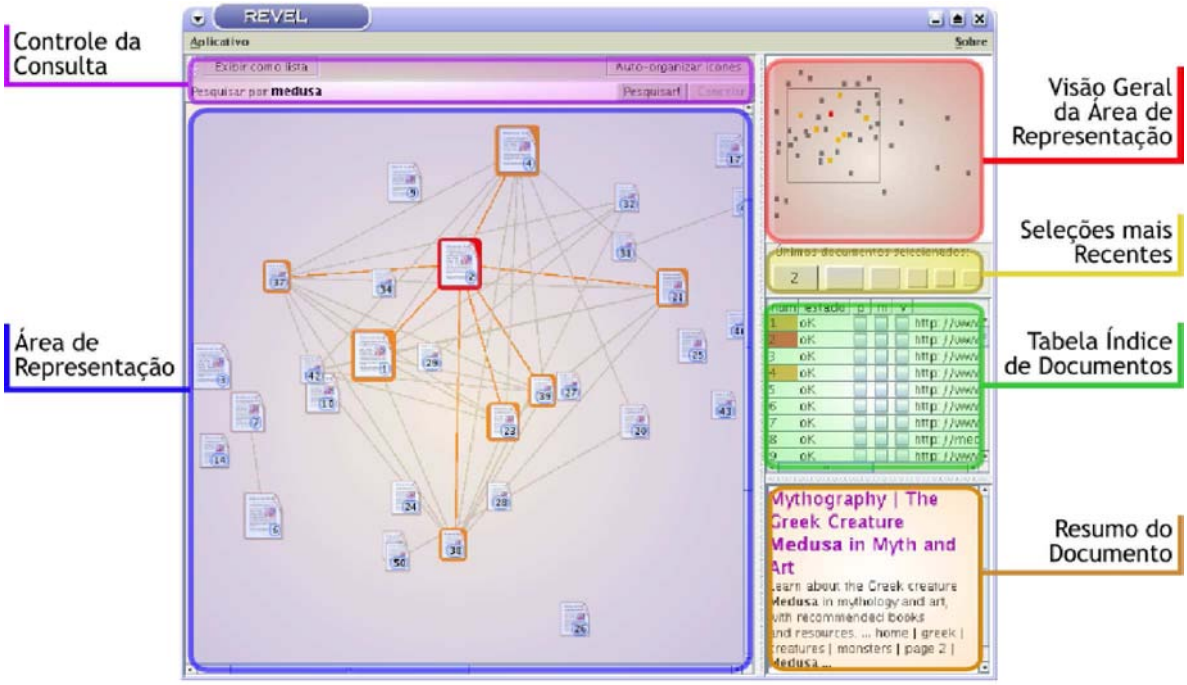


Figura 4.3: Estrutura da Interface

Controle da Consulta

Botões e campo para controle de operação da consulta – Figura 4.4:

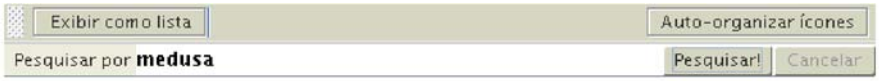


Figura 4.4: Controle da Consulta

- Campo de texto: campo para entrada das palavras-chave a serem pesquisadas. O sistema ReVEL possui mais elementos de interação que uma página de consulta de mecanismos de busca. Para evitar que este acréscimo de complexidade confunda o usuário e impeça que ele comece a utilização, a caixa de entrada de texto mostra-se evidenciada por uma borda vermelha e há um texto em negrito vermelho indicando que é ali o local para se inserir as palavras-chave, como ilustra a Figura 4.5. Vale ressaltar que tal ênfase só existe ao iniciar o sistema e desaparece ao começar a digitar algo ou ao clicar nessa caixa de texto.

Figura 4.5: Caixa de Entrada de Texto Destacada em Vermelho

- Botão **Pesquisar!**: inicia a consulta ao mecanismo de busca utilizando as palavras-chave presentes no campo de texto;
- Botão **Cancelar**: interrompe uma consulta ao mecanismo de busca que esteja em andamento;

- Botão **Exibir como Lista**: abre uma janela e exibe os documentos – obtidos como resposta à consulta – na forma de lista. Um exemplo pode ser visto na Figura 4.6;
- Botão de dois estados **Auto-organizar ícones**: este botão de dois estados – pressionado/liberado – controla a ocorrência ou não da animação dos ícones dos documentos na Área de Representação. A animação é o resultado do esforço de um algoritmo que tenta exibir o grafo composto pelo ícone de cada documento (vértice) e as ligações representando as relações de similaridade deste com os demais documentos (arestas) segundo restrições de peso nas arestas (similaridade entre os documentos). Quando no estado de pressionado, os ícones movem-se segundo disposição ditada pelo algoritmo; quando liberado, o desenho permanece estático.

Figura 4.6: Resultado da Pesquisa Exibido como Lista

Área de Representação

A Área de Representação é onde os documentos obtidos como resposta à consulta ao mecanismo de busca são exibidos graficamente. Um exemplo pode ser visto na Figura 4.7.

Os documentos são representados na forma icônica e aqueles avaliados como similares são representados unidos por arestas de diferentes tamanhos e espessuras. O tamanho de cada aresta é inversamente proporcional à similaridade entre os documentos a que ela é incidente e a espessura diretamente proporcional. Assim, uma aresta curta aproxima dois documentos, sugerindo similaridade, enquanto arestas longas mantêm documentos pouco similares distantes um do outro. Reforçando esta impressão, arestas mais espessas reforçam visualmente a existência de uma relação. Se a similaridade entre

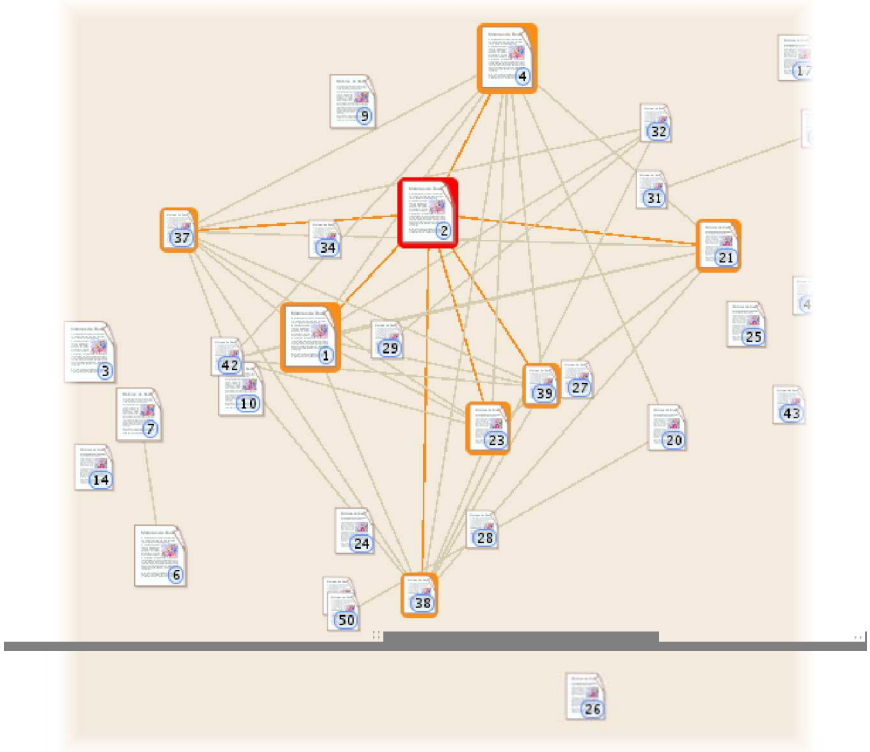


Figura 4.7: Área de Representação

dois documentos for zero ou inferior a um limite calculado, não há aresta entre estes documentos. O valor limite calculado pode alterar-se ao longo do tempo a fim de controlar a densidade de arestas exibidas e, portanto, a clareza na leitura geral da Área de Representação.

Os documentos são representados por ícones que indicam seus estados correntes. Os tipos de ícones possíveis, assim como suas descrições, estão relacionados na Tabela 4.1.

Ícone	Estado	Descrição
	pedido	O documento foi requisitado para ser baixado da Internet.
	baixar	O processo de obtenção está em andamento; a página está sendo obtida (*download*) da Internet neste momento.
	análise	O documento foi obtido, processado e, no momento, estão sendo feitos os cálculos de similaridade entre ele e os demais.
	*ok*	O documento já foi obtido, processado e seus resultados estão disponíveis e em uso pelo ReVEL.
	erro!	O documento não pôde ser obtido por alguma razão externa ao sistema ReVEL.

Tabela 4.1: Tipos de Ícones de Estados dos Documentos

Uma vez que a posição ordinal do documento na lista obtida do mecanismo de busca é importante, pois reflete a expectativa de melhor resposta à consulta dadas as palavras-chave, o tamanho do ícone foi utilizado para representar a posição na lista, segundo a seguinte codificação: quanto maior o posto do documento, mais visível é seu ícone da Área de Representação. Assim, foram criadas quatro categorias relacionando tamanho e posição na lista de resposta. As definições dos limites das categorias assim como os tamanhos a elas relacionados estão descritos na Tabela 4.2.

Ícone	Tamanho (*pixels*)	Posição
	48 x 48	1º ao 5º resultado na lista do mecanismo de busca.
	42 x 42	6º ao 12º resultado.
	36 x 36	13º ao 25º resultado
	30 x 30	acima do 25º resultado.

Tabela 4.2: Tamanhos dos Ícones em Função da Posição na Lista

Além do tamanho do ícone, cada documento é identificado por um número em um retângulo azul em seu canto inferior direito, que é exatamente a sua posição ordinal na lista de resposta. Deste modo, se uma pesquisa retornou dezoito documentos há, na Área de Representação, dezoito ícones numerados de um a dezoito, onde os primeiros cinco ícones de documentos têm o tamanho da primeira categoria – 48 x 48 *pixels* – , o sete seguintes o tamanho da segunda categoria e os seis finais o tamanho da terceira. Além disso, a imagem presente no ícone informa o estado do processamento do documento no momento. Exemplos das possíveis combinações de estado podem ser vistos na Tabela 4.3.

A Área de Representação é inicializada com uma dimensão grande o suficiente para haver espaço para que ocorra o mínimo de oclusão dos ícones de documentos, para que haja espaço para uma manipulação e diagramação que seja confortável para o usuário e também para que não sobrecarregue o algoritmo que tenta organizar os ícones. Uma área com essa dimensão – 1200 por 1000 *pixels* – , entretanto, é grande demais para ser exibida na íntegra na maioria dos monitores. Mesmo oferecendo o recurso das barras de rolagem, faz-se necessário disponibilizar para o usuário um modo de ter uma visão geral de todos os documentos obtidos e representados. Deste modo, foi implementada na interface uma visão geral da Área de Representação.

Ícone	Significado
	O segundo documento da lista de resultados do mecanismo de busca está sendo analisado e está "pregado", isto é, não foi permitido que o algoritmo de melhoria de disposição de ícones o mova de lugar.
	O décimo documento está pronto, ou seja, foi obtido e analisado. O visto verde indica que este documento já foi requisitado pelo usuário para ser visitado no navegador.
	O 46º documento não pôde ser obtido da Internet pois ocorreu um erro. Este ícone está "pregado", isto é, não será movimentado automaticamente pelo algoritmo de melhoria de disposição de ícones.
	Este é o quinto documento selecionado para ser retido, isto é, para ser mantido para a próxima pesquisa, está pronto, foi visitado no navegador (*browser*) e está "pregado" em sua posição na Área de Representação.

Tabela 4.3: Exemplos de Ícones de Documentos da Área de Representação

Visão Geral da Área de Representação

A Área de Representação, por sua dimensão, geralmente é exibida com barras de rolagem. Porém, segundo Plaisant *et al.* (1996), barras de rolagem, mesmo sendo uma solução para falta de espaço, são inadequadas, quando não prejudiciais, para representar visões amplas, uma vez que usuários com freqüência esquecem de examinar a imagem completa ou até ignoram que não é toda a imagem que está sendo exibida. Assim sendo, uma solução seria um resumo, uma imagem que coloque o que está sendo exibido dentro de um contexto geral, mantendo elementos referenciais em cada uma das imagens, para situar a posição de cada elemento – visto em detalhe, na Área de Representação – com relação a um todo, visto no conjunto da Visão Geral.

A Visão Geral da Área de Representação exibe um resumo da área completa disponível para exame e manipulação pelo usuário, que é a Área de Representação. Um exemplo pode ser visto na Figura 4.8.

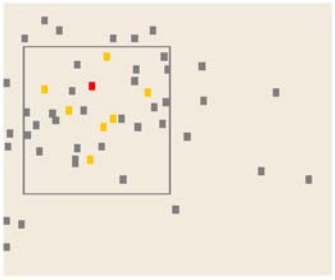


Figura 4.8: Visão Geral da Área de Representação

A Visão Geral é uma imagem em escala menor – pré-definida em 1:6 – que a da Área de Representação e exibe informações suficientes para que o usuário situe o que está sendo exibido na Área de Representação dentro do contexto geral. Ela é provida para tratar a limitação da exibição de imagem quando barras de rolagem são necessárias.

Na Visão Geral são exibidos os documentos e a fração dentro da imagem total referente à Área de Representação. Os documentos são representados como quadrados cinzas. O documento selecionado pelo usuário na Área de Representação é aqui denotado com a mesma cor que lá: vermelho. Os documentos similares ao documento selecionado também são representados em ambas as áreas com a mesma cor: laranja. Um retângulo traçado em cinza-claro indica a parte da área total que aparece naquele momento na Área de Representação. Como exemplo, basta comparar a Figura 4.7 com a região interna do retângulo cinza da Figura 4.8.

Um clique de *mouse* em qualquer parte da Visão Geral centra a exibição da Área de Representação naquela região da imagem completa.

Seleções Mais Recentes

Os botões de Seleções mais Recentes existem para auxiliar o usuário a recordar o número de um documento selecionado em operações passadas. Um exemplo pode ser visto na Figura 4.9.

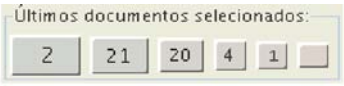


Figura 4.9: Seleções mais Recentes

O usuário pode clicar nos ícones dos documentos na Área de Representação. O respectivo documento é então selecionado. Como retorno ao usuário aparece uma margem vermelha no ícone clicado e o seu resumo aparece na área Resumo do Documento. Também são destacados, porém com a cor laranja, os documentos até o momento calculados como similares – ligados por arestas – ao documento selecionado. A Figura 4.7 exemplifica a seleção de um documento, no caso, o de número dois. Tal procedimento é um recurso para a avaliação do contexto de determinado documento com relação aos demais a ele visualmente vinculados. O usuário pode, porém, desejar rever um documento anteriormente selecionado e não mais recordar qual foi. Os seis botões presentes na área Seleções Mais Recentes são atalhos para os seis últimos documentos selecionados. Ao clicar-se em um deles, o documento cujo número aparece no botão é novamente re-selecionado. O documento mais recentemente selecionado tem seu número no maior botão, mais à esquerda e os anteriores em botões cada vez menores.

Tabela Índice de Documentos

Esta Tabela Índice apresenta o maior conjunto de informações sobre o estado dos documentos manipulados pela aplicação, além de prover um local centralizado para relacionar-se com eles. A Figura 4.10 exibe um trecho da tabela.

Figura 4.10: Tabela Índice de Documentos

A Tabela é composta de seis colunas:

1. Número do documento (**num**): informa a posição ordinal em que tal documento aparece na lista de resultados do mecanismo de busca. Pode indicar que o dado documento está selecionado ou é similar ao selecionado ao aparecer colorido em vermelho ou laranja, respectivamente;
2. Estado atual do documento (**estado**): uma vez que, para verificar a similaridade entre um documento e os outros, é necessário comparar seu conteúdo com o conteúdo dos demais, é necessário obter o documento da Internet. A coluna estado informa a situação atual de cada um dos documentos. Os estados possíveis estão descritos na Tabela 4.1.

As três colunas seguintes são caixas de checagem – *checkboxes* – para examinar ou alterar determinadas características de cada um dos documentos.

3. Documento "pregado" (**p**): Os ícones dos documentos estão sujeitos ao algoritmo que tenta melhorar as suas disposições. Para isso, o algoritmo movimenta os ícones pela Área de Representação. Ao marcar esta caixa de checagem, o ícone fica fixo em sua posição e não mais é movimentado automaticamente pelo algoritmo. Porém o ícone ainda pode ser arrastado com o *mouse* pelo usuário e reposicionado, conforme for de seu interesse. Um documento "pregado" é também indicado por um círculo amarelo com um ponto no centro – – posicionado no canto superior esquerdo do ícone.
4. Documento "mantido" (**m**): Pode ocorrer de uma pesquisa não retornar um documento com a informação procurada porém prover alguns documentos cujos conteúdos seriam relacionados ao que se deseja. Quando uma consulta é realizada, todos os documentos da consulta anterior são descartados em favor dos novos. Ao marcar esta caixa para um documento, ele será mantido para a próxima consulta e será, portanto, possível observar quais documentos da nova consulta são similares aos documentos mantidos da consulta anterior. Um documento mantido para a próxima pesquisa tem seu número, no ícone, alterado e a cor do retângulo que contorna o número passa de azul para vermelho. O novo número será sua posição na seqüência dos documentos selecionados para serem mantidos. Assim, se um determinado documento for o terceiro a ser mantido, seu novo número será três. Se o usuário mudar de idéia e decidir não mais manter um dado documento, ao desmarcar a caixa o documento volta a ter sua cor e número originais.
5. Documento "visitado" (**v**): Esta caixa indica se o documento já foi visitado por um navegador invocado pelo programa. O navegador pode ser invocado ao se clicar duas vezes no ícone do documento na Área de Representação ou ao se marcar esta caixa com

um clique na Tabela. O ícone indica que o documento foi visitado por uma marca de um "visto" verde – ✔ – na sua parte superior, ao centro. Alguns exemplos de ícones possíveis estão representados na Tabela 4.3.

6. Endereço do documento na Internet (**endereço**): Esta linha contém o endereço do documento na Internet, sua URL – *Universal Resource Location.*

Resumo do Documento

Esta área exibe o texto resumido do documento. Este é o texto fornecido pelo mecanismo de busca para contextualizar o documento obtido como resposta da pesquisa feita segundo as palavras-chave fornecidas. Toda vez que um documento é selecionado na Área de Representação ou na Tabela, seu resumo é exibido nesta área. Um exemplo pode ser visto na Figura 4.11, que representa o resumo do documento número dois, quando foi selecionado.

Figura 4.11: Resumo do Documento

4.2.3 Consulta ao Mecanismo de Busca

Informar o usuário continuamente sobre o estado corrente da aplicação é de grande importância e foi listado por Nielsen (1993) como um item de seu conjunto básico de heurísticas para a verificação de usabilidade de interfaces. Assim sendo, diálogos são exibidos durante diversos momentos da consulta informando o estado do sistema, além de mudanças no estado de botões, campos de texto etc.

Uma vez digitadas as palavras-chave e solicitado o início da pesquisa – seja por pressionar *enter*, seja por clicar no botão **Pesquisar!** – alguns elementos da interface se alteram, dando um retorno ao usuário de que sua ação foi reconhecida e passou a ser executada. As alterações são: a caixa de entrada de texto, assim como o botão Pesquisar!, são desabilitados – tornam-se opacos – e o botão **Cancelar** é habilitado. Janelas de diálogos sucedem-se com informações sobre o estado corrente do programa, segundo a seguinte ordem:

1. O ReVEL procura conectar-se com o mecanismo de busca, conforme ilustra a Figura 4.12, a fim de fazer a consulta e obter a página de resultados utilizando as palavras-chave fornecidas;
2. Dependendo do sucesso da consulta, podemos obter três resultados:
   a) O mecanismo de busca foi atingido com sucesso e uma página de resultados não vazia foi obtida. Uma barra de progressão passa a informar o usuário quantos *kilobytes* já foram transferidos, conforme ilustra a Figura 4.13; terminada a transferência, inicia-se a fase de Análise;

b) Houve um problema – provavelmente de conexão – e não foi possível acessar a página do mecanismo de busca. Uma janela é exibida com explicação do problema e uma sugestão para sua correção, conforme também heurísticas de usabilidade de Nielsen (1993). Uma ilustração representando esta situação está na Figura 4.14. O sistema volta então ao seu estado inicial, esperando o usuário executar a consulta novamente ou alterar as palavras-chave;

c) A consulta foi realizada com sucesso, porém o mecanismo de busca não recuperou nenhum documento que satisfizesse a consulta pela(s) palavra(s)-chave utilizada(s). O ReVEL informa com um diálogo de aviso, como ilustra a Figura 4.15, e retorna ao estado de aguardar palavras-chave.

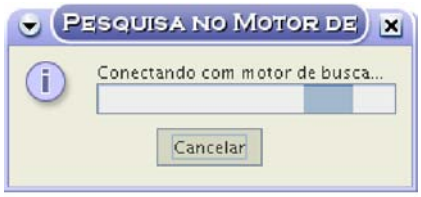


Figura 4.12: Conectando com o Mecanismo de Busca

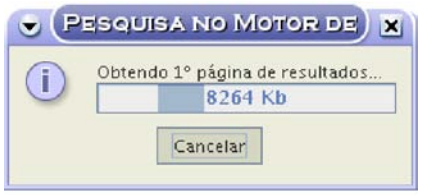


Figura 4.13: Obtendo Página de Resultados da Pesquisa

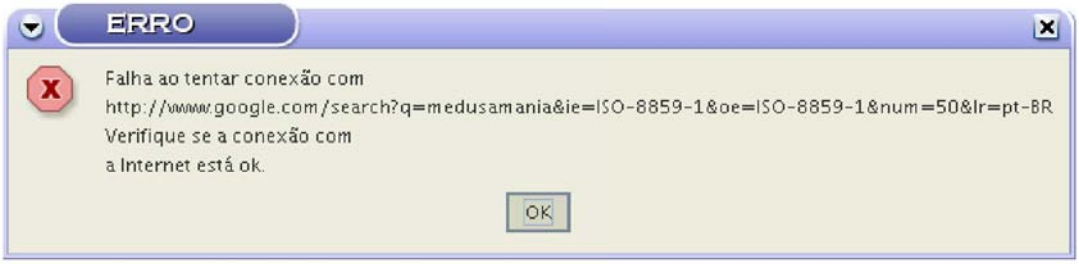


Figura 4.14: Falha ao Tentar Conexão com Mecanismo de Busca

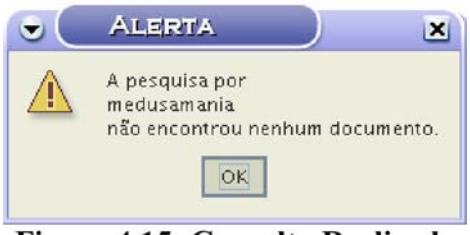


Figura 4.15: Consulta Realizada não Retornou Documentos

4.2.4 Análise dos Resultados

Uma vez que o início do processo de visualização consome tempo para obter os primeiros documentos da Internet e realizar o cálculo de similaridade entre eles para, então, ter algo para exibir, fez-se necessário acrescentar uma janela para informar o usuário que algo estava se passando. Esta janela de espera, conforme vemos na Figura 4.16, mantém o usuário informado do processamento atual da aplicação. Depois de alguns segundos esta janela se fecha e o ReVEL encontra-se pronto para a interação com a visualização. Um exemplo típico deste início de interação com a representação está exibido na Figura 4.1, na página 36.

Figura 4.16: Janela de Espera

Consulta

Insere as palavras-chave e executa a consulta;

Seleção

O usuário seleciona um documento por algum critério – endereço, vínculos a outros documentos, posição etc. – para analisar. Ao clicar em um ícone de documento para exibir seu resumo na área Resumo do Documento, automaticamente ele é destacado como selecionado e todos os outros documentos a ele relacionados também. O ícone clicado aparece destacado com uma margem vermelha e será aqui identificado como seleção primária. Os documentos identificados pelo ReVEL como similares à seleção primária terão seus ícones destacados com uma margem laranja, como pode ser visto na Figura 4.7.
Se após diversas tentativas nenhum documento parecer promissor, o usuário pode decidir refazer a consulta com outras palavras-chave ou então considerar infrutífera a busca e desistir de localizar a informação;

Avaliação

Examina o resumo e, caso seja promissor, invoca o navegador para avaliar o documento. O navegador pode ser invocado para examinar o documento por dois modos: com um duplo-clique no ícone do documento na Área de Representação e ao marcar a caixa de visitado (**v**) na Tabela Índice de Documentos. Uma vez que a página com o documento lhe é exibida pelo navegador, o usuário passa então à fase de Análise, tentando extrair a informação desejada. Caso o resumo não seja promissor, ocorre a volta para o passo de Seleção;

Análise

O usuário determina se as informações contidas no documento exibido satisfazem sua busca. Se ele identifica a informação desejada, considera a busca encerrada com sucesso; caso contrário volta ao passo de seleção;

Desistência

Insatisfeito com a pesquisa, seja por não conseguir localizar a informação desejada em parte ou no todo, o usuário encerra o processo de procura.

Sucesso

O usuário localiza a informação desejada e encerra o processo de procura.

4.2.5 Encerramento de Operação

O programa ReVEL pode ser encerrado a qualquer momento, utilizando-se qualquer um dos três modos disponíveis:

- O botão de encerrar sistema na moldura da janela, provido pelo gerenciador de janelas nos ambientes gráficos – em geral um botão com um "x";
- Por meio do menu, clicando em **Aplicativo** e depois em **Sair**;
- Teclando uma das seqüências conhecidas para encerrar aplicativos em ambientes gráficos: **CTRL-Q**.

Uma janela de diálogo, como ilustrado na Figura 4.17, é exibida. O usuário pode ainda decidir se deseja realmente sair ou não. Para evitar o problema de uma saída indesejada, o botão **Não** está pré-selecionado. Isso por ser um inconveniente menor desejar sair e, por engano, voltar ao sistema do que não desejar encerrar e, por engano, terminar o sistema, quando então os dados e informações são descartados.

Figura 4.17: Diálogo de Encerramento

4.3 Aspectos de Implementação

O sistema ReVEL foi modelado e desenvolvido como sendo o conjunto inter-operante de alguns módulos independentes. Cada módulo, além da interface, é denominado de "gerenciador", tem uma tarefa específica e opera como um processo leve. Assim sendo, a maior parte dos gerenciadores é executada simultaneamente, tentando com isso provocar a menor quantidade possível de interrupções na continuidade do fluxo de operação realizada pelo usuário. A continuidade da interação do usuário com a interface é de grande importância, pois as pessoas têm pouca disposição para esperar um sistema que está ocupado executando alguma tarefa. O ideal é que tarefas demoradas sejam executadas simultaneamente e que o usuário sempre tenha a interface disponível para interagir.

4.3.1 Estrutura Geral

A estrutura do funcionamento de ReVEL está representada na Figura 4.18. Uma descrição geral do funcionamento seria a seguinte:

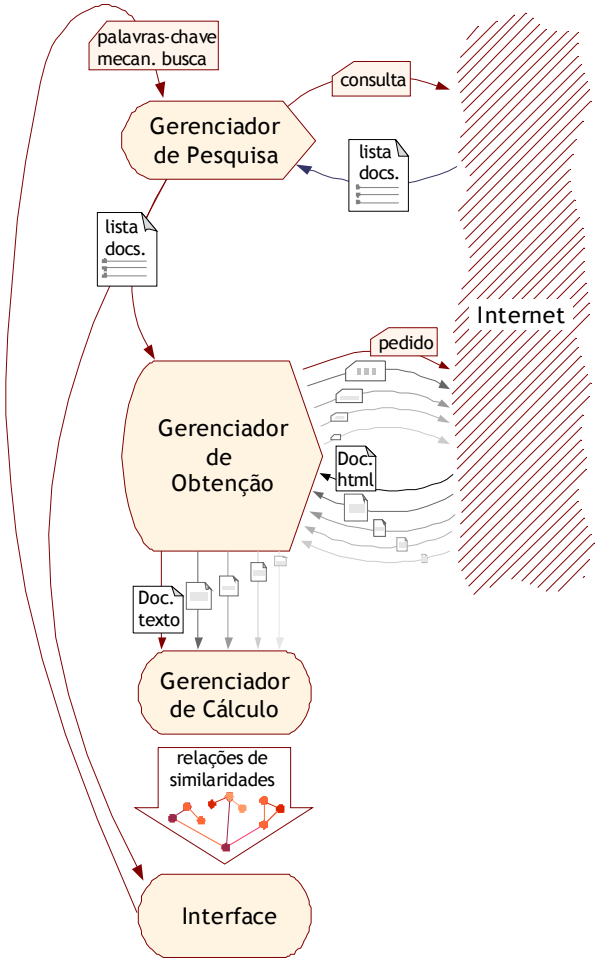


Figura 4.18: Estrutura do Aplicativo ReVEL

1. O usuário, através da Interface, insere as palavras-chave, que junto com o mecanismo de busca, formam a consulta. A consulta é entregue pela Interface ao Gerenciador de Pesquisa;
2. O Gerenciador de Pesquisa consulta o mecanismo de busca na Internet e obtém dele a lista de resposta, uma lista de documentos que supostamente contém um ou mais documentos capazes de suprir a necessidade de informação do usuário. De posse da Lista de Documentos, o Gerenciador de Pesquisa a remete para dois lugares. O primeiro é a Interface, para que o ReVEL realimente, em um momento inicial, o usuário; o segundo é o Gerenciador de Obtenção;
3. O Gerenciador de Obtenção inicia então o processo de obter da Internet cada um dos documentos enumerados na Lista de Documentos recebida. Para que os documentos comecem o mais rápido possível a ser disponibilizados para o Gerenciador de Cálculo – e conseqüentemente para a Interface – há uma restrição de um número máximo de documentos requisitados simultaneamente da Internet. Assim, espera-se que alguns

documentos já estejam sendo processados pelo Gerenciador de Cálculo em segundos após a lista ter sido recebida pelo Gerenciador de Obtenção. O Gerenciador de Obtenção vai, gradativamente, obtendo cada um dos documentos da lista, transformando-o em texto simples – extraindo o texto de dentro do documento HTML, quando necessário – e enviando-o para o Gerenciador de Cálculo;

4. O Gerenciador de Cálculo é responsável por avaliar a similaridade de cada documento recém-obtido contra todos os demais documentos da lista já obtidos e montar uma matriz de similaridade. Seu processamento é simultâneo ao do Gerenciador de Obtenção e ao da Interface. Conforme os resultados de similaridade entre alguns documentos já vão sendo obtidos, estes são repassados à Interface, que os informa para o usuário.

4.3.2 Interface

É pela Interface que o usuário alimenta o sistema das palavras-chave a serem utilizadas na consulta, assim como interage com as representações visuais e a tabela dos resultados.

Uma vez obtidas as palavras-chave, a Interface agrega a elas a definição de um mecanismo de busca e envia este conjunto, empacotado como uma consulta para o Gerenciador de Pesquisa. A presente implementação utiliza o mecanismo de busca Google, porém basicamente qualquer mecanismo de busca que se comunique utilizando documentos HTML pode ser implementado de modo fácil, por meio dos recursos de especialização / generalização do paradigma de Orientação a Objetos sobre o qual o ReVEL foi desenvolvido.

Grafos

A Área de Representação exibe os documentos como grafos, nos quais os documentos são os vértices e as relações de semelhança são arestas. As arestas possuem pesos, que são representados como o inverso de seu comprimento. Assim, uma aresta é tanto menor quanto maior for o seu valor de similaridade, tornando os vértices aos quais ela é incidente mais próximos. Apesar de existir uma gama bastante variada de técnicas para se desenhar grafos, como é possível verificar na bibliografia reunida por Di Battista *et al.* (Di Battista *et al.*, 1994), restrições inerentes ao nosso caso reduzem as alternativas. As restrições possuem graus diferentes de importância e são relacionadas a seguir.

O algoritmo para dispor o grafo deve:

- Possibilitar uma representação gradual, uma vez que os documentos serão obtidos da Internet e também por ser necessário já exibir alguns resultados para o usuário. A maioria dos algoritmos esforça-se por ser rápido ao fornecer o resultado final, mas só é capaz de processar um grafo que não será mais alterado;
- Ser rápido para ser calculado, utilizando o mínimo possível de recursos do computador a fim de que não sobrecarregue a interface e atrapalhe a interação do usuário;
- Ter a possibilidade de lidar com grafos desconexos, pois espera-se que a similaridade reúna os documentos semelhantes em componentes distintos do grafo;
- Lidar com peso nas arestas;

- Supor que sejam grafos simples, com arestas em linha reta, que são mais apropriados para serem desenhados e controlados por ambientes gráficos;

Segundo um estudo sobre facilidade de leitura de grafos (Purchase, 2000), vários parâmetros como simetria, ortogonalidade, cruzamentos de arestas entre outros foram avaliados quanto ao tempo que demandam para serem analisados e quanto ao número de erros cometidos quando perguntas foram feitas sobre propriedades dos grafos. As conclusões indicam que o cruzamento de arestas é o parâmetro que mais atrapalha a leitura de um grafo, embora apenas quando ocorre em grande quantidade. Vários algoritmos foram avaliados e nenhum mostrou exigir estatisticamente mais tempo de resposta que os demais, indicando não haver nenhum que seja mais difícil de ler. Um deles, entretanto, apresentou evidência estatística de possuir uma maior taxa de erros de leitura que os demais. É um algoritmo baseado no princípio de desenho planar de grafos com vértices em grades e arestas retas.

Apesar de não serem estatisticamente melhores que os demais, os grafos desenhados utilizando técnicas baseadas no modelo de "posicionamento por forças" (*force directed)* apresentaram as menores médias para número de erros e tempo de resposta e, portanto, foram escolhidos para dispor os ícones de desenho na Área de representação. Dentre os algoritmos que utilizam posicionamento por forças citados no artigo de Purchase (2000), elegemos o de Fruchterman e Reingold (1991) por preocupar-se com a minimização do cruzamento das arestas e a simplicidade nos cálculos.

O princípio que norteia o método de posicionamento por forças é a metáfora de que os vértices sejam anéis de aço e as arestas molas, transformando o grafo em um sistema mecânico. Os vértices são dispostos em uma posição e liberados, de modo que as forças das molas – arestas – movam o sistema para um estado de mínima energia.

Fruchterman e Reingold (1991) definem seu método como um modelo no qual os vértices são dispostos de modo que, se conectados por arestas, sejam desenhados próximos uns dos outros porém não próximos *demais*. A proximidade dos vértices depende de quantos há e do tamanho da área disponível para desenhá-los. O princípio do algoritmo inspira-se na física de partículas atômicas ou corpos celestiais, onde forças de atração e repulsão se equilibram mantendo a coesão e evitando o colapso.

O único problema com este algoritmo é que ele não supõe peso nas arestas, esforçando-se para distribuir os vértices dos componentes do grafo ao largo da área disponível para desenhá-los, com arestas tendendo a um tamanho uniforme. Alteramos então o algoritmo de modo que o equilíbrio entre as forças de atração e repulsão dos vértices conectados por uma aresta ocorra no exato tamanho da similaridade que há entre os documentos representados por estes vértices. Assim, se eles se encontram mais próximos que a similaridade indicaria a força de repulsão vai tentar afastá-los; se mais distantes, a força de atração será maior e a tendência é de que os vértices se aproximem.

As forças de atração e repulsão do artigo de Fruchterman e Reingold (1991) ficaram então assim definidas:

- **Força de atração**: $f_a(x) = x^2 / k_S$ onde k_s é um coeficiente proporcional à similaridade: $k_S = 400 - 370\, similaridade$, gerando valores que vão de 400 até 30 (*pixels*) conforme a similaridade varia de 0,0 até 1,0;

- **Força de repulsão**: $f_r(x) = -k_a^2 / x$, onde k_a é um coeficiente proporcional à área disponível para desenho e ao número de vértices total: $k_a = \sqrt{(\text{área de desenho} / \text{número de vértices})}$. Está também implementada uma otimização que desconsidera a força de repulsão se a distância x entre os vértices for maior que $2.k_a$.

O equilíbrio entre as forças de atração e repulsão que se estabelece em função da similaridade pode ser visto na Figura 4.19.

O algoritmo alterado satisfaz todas as restrições: dispõe com clareza uma quantidade de documentos – vértices – necessária simultaneamente à obtenção, permite manipulação e representação gradual, é capaz de lidar com grafos simples, desconexos e com pesos nas arestas.

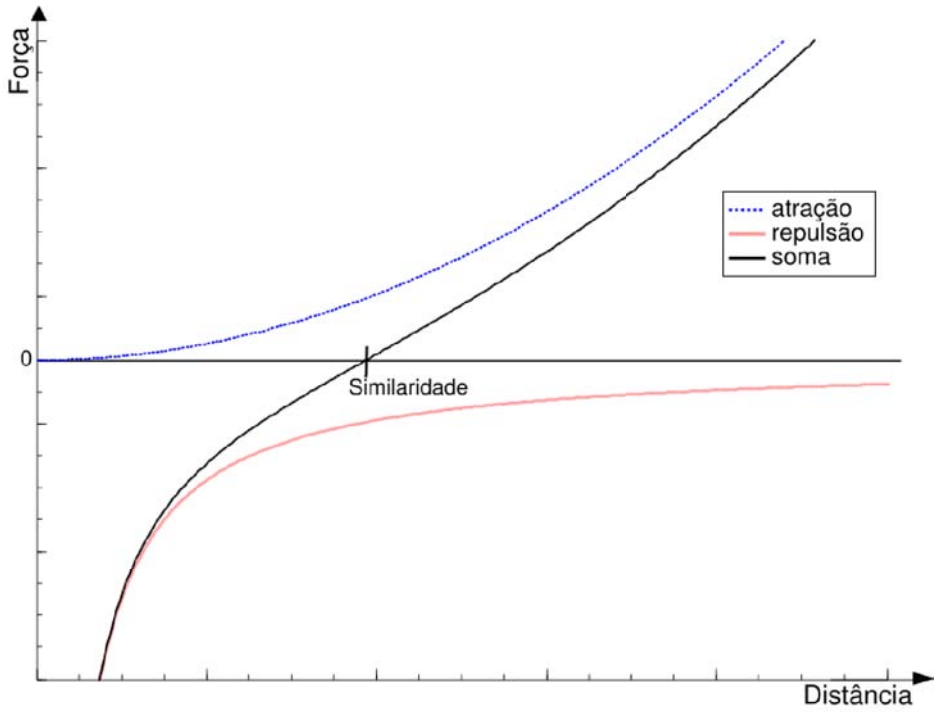


Figura 4.19: Comportamento das Forças em Função da Distância

4.3.3 Gerenciador de Pesquisa

O Gerenciador de Pesquisa está pré-definido para requisitar uma lista de resposta de no máximo 50 documentos. Ele porém foi preparado para requisitar o número máximo de páginas de resultados de cada mecanismo de busca, com o número máximo de documentos por página de resultado. Isto equivale, no caso do Google, a 10 páginas de resultados com 100 itens cada uma, perfazendo um total de uma lista com 1000 documentos. Tal possibilidade não foi sequer testada em virtude do esgotamento de recursos tais como a velocidade de transferência da Internet, espaço de representação na tela, memória de armazenamento além, é claro, da paciência do usuário.

Em posse da página de resultados retornada pelo mecanismo de busca, o Gerenciador de Pesquisa extrai dela uma lista de documentos, onde cada item é composto por um endereço Internet – uma URL – e o resumo do conteúdo do documento. Caso a consulta não tenha retornado nenhum documento, um diálogo é exibido esclarecendo o usuário.

A lista de documentos é enviada ao Gerenciador de Obtenção para que cada elemento da lista seja obtido.

Determinados mecanismos de busca bloqueiam consultas realizadas por terceiros. Para circunscrever tal limitação, o sistema ReVEL apresenta-se como um navegador Mozilla-Firefox sendo executado sobre Linux em uma máquina de arquitetura i386, independente da plataforma ou sistema operacional em que esteja de fato sendo executado. Isso também ajuda a homogeneizar a página de resposta obtida, uma vez que o documento HTML – *HyperText Markup Language* – enviado por alguns mecanismos de busca – Google, por exemplo – é dependente do navegador utilizado na consulta.

4.3.4 Gerenciador de Obtenção

Este módulo é o responsável por controlar a seqüência e o fluxo de obtenção dos itens da lista de documentos da Internet.

Alguns gerenciadores foram construídos para operar com prioridade abaixo do que seria a normal para o escalonamento de processos leves (*threads*) em Java. O Gerenciador de Obtenção é um deles. O motivo geral para reduzir a prioridade é o de evitar que muito recurso de processamento fosse utilizado no gerenciador em detrimento da interface, prejudicando a possibilidade de interação do usuário ou a resposta, para o usuário, de alterações ocorridas no sistema. O Gerenciador de Obtenção possui ainda outro motivo para operar com baixa prioridade: a obtenção de páginas da Internet, em geral, requer menos processamento, uma vez que grande parte da ocupação do computador devida a obtenção é consumida em espera gerada por motivos intrínsecos da comunicação.

Uma das tarefas deste gerenciador é manter um fluxo contínuo de documentos sendo adquiridos em uma boa relação de número de documentos obtidos por tempo. Os extremos possíveis caso não houvesse um balanceamento por um gerenciador seriam: requisitar e controlar a obtenção de todos os documentos ao mesmo tempo ou então requisitar e obter cada um individualmente. O primeiro caso, conseguir todos os documentos aproximadamente ao mesmo tempo, seria ótimo para efetuar os cálculos de similaridade entre eles, porém seria péssimo para o usuário, que estaria aguardando durante todo esse tempo. O segundo caso, obter de modo serial documento após documento, traz o inconveniente de os tempos de atraso de cada um dos documentos somarem-se e acarretarem, ao final, também uma espera tão desconfortável quanto desnecessária para o usuário.

O Gerenciador balanceia a obtenção de documentos em uma taxa configurável de cinco documentos simultaneamente. Este número pode oscilar, independente do valor determinado pela configuração, ao longo da execução do sistema. Isso se dá por uma característica de construção de determinadas páginas da Internet: os controversos quadros internos (*frames*). Os mecanismos de busca não fazem distinção entre a localização de uma palavra-chave estar em um determinado quadro interno ou na página-raíz dos quadros internos e retorna sempre o endereço da página-raiz, o que é bastante razoável, uma vez que ela é a responsável por determinar como os quadros internos serão dispostos e, portanto, como cada quadro compõe o contexto geral. Logo, é uma necessidade que o Gerenciador também esteja preparado para conseguir estes quadros internos. Para acelerar a obtenção, o

limite de conexões simultâneas é expandido em um valor igual ao número de quadros internos e, quando estes são obtidos, o limite é retornado ao seu valor de operação, aquele definido em configuração.

Por fim, este Gerenciador extrai do documento obtido, que na grande maioria das vezes é um arquivo HTML – *HyperText Markup Language* –, o texto necessário para a análise de similaridade e o envia para o Gerenciador de Cálculo.

Algumas considerações foram tomadas quanto aos elementos da lista e em como obtê-los. A primeira consideração é quanto ao tamanho dos documentos. Não seria viável, embora fosse desejável por questão de precisão, obter todo o documento para efetuar os cálculos de similaridade. Alguns documentos têm tamanhos que são inviáveis de serem obtidos por questão de tempo de espera. Então o tamanho máximo determinado para um documento foi arbitrado em 50 Kb. Este valor é grande o suficiente para fornecer uma quantidade de texto que seja razoável para os cálculos de similaridade e de um tamanho que não seja muito demorado para ser obtido da Internet. A segunda consideração é que só seriam utilizados documentos HTML. Há uma gama muito grande de tipos de documentos disponíveis na Internet: documentos de editores de texto, apresentações, arquivos *postscript* de impressão, arquivos PDF etc. Seria inviável desenvolver um interpretador que extraísse de cada um deles seu texto. A solução foi solicitar e utilizar a versão HTML destes documentos, disponibilizada pelo próprio Google.

O Gerenciador de Obtenção conta também com um Banco de documentos. O Banco está configurado para armazenar uma quantidade de documentos de uma consulta para outra. Esta quantidade é proporcional à memória disponível para a máquina virtual Java e limitada a um mínimo de quinze documentos. O motivo da existência deste Banco é para otimizar a obtenção de documentos da Internet. Supõe-se que, de uma pesquisa para outra, haja a mudança de algumas palavras-chave em torno de um mesmo assunto. É, portanto, bastante provável que alguns documentos listados na primeira consulta sejam também elencados na seguinte. Se os documentos estiverem ainda disponíveis, não será necessário requisitá-los da Internet. O Banco substitui os documentos que estão a mais tempo sem serem utilizados por aqueles da consulta recém-descartada.

4.3.5 Gerenciador de Cálculo

O Gerenciador de Cálculo é o responsável por calcular a similaridade entre um dado documento e todos os demais documentos. Sua função é obter um valor que represente o quão similar um dado documento é dos demais. Assim como o Gerenciador de Obtenção, este gerenciador também opera com prioridade abaixo do normal para não onerar o processamento da Interface.

Similaridade

O conceito de similaridade aqui utilizado faz parte do conjunto de conceitos utilizados em Recuperação de Informação. Para a Recuperação de Informação, é importante decidir quais documentos, de um universo de documentos, são relevantes para responder a uma consulta e quais não são. Tal decisão é geralmente dependente de algoritmos de pontuação, que tentam estabelecer uma ordem nos documentos, nos quais os mais relevantes supostamente vêm primeiro. Cada algoritmo é baseado em algumas premissas e, dependendo das premissas, são criados diversos tipos de modelos de recuperação de informação distintos: booleano, vetorial, probabilístico etc.

O modelo adotado no ReVEL foi o vetorial por que, apesar de sua simplicidade e rapidez, é um modelo robusto para coleções de documentos genéricos. Ele gera pontuações que são difíceis de serem melhoradas. Uma grande quantidade de métodos de pontuação alternativos já foi comparada ao modelo vetorial mas o consenso é que, geralmente, o modelo vetorial é superior ou pelo menos tão bom quanto as alternativas conhecidas (Baeza-Yates e Ribeiro-Neto, 1999).

Modelo Vetorial de Similaridade

Sejam $K = \{k_1, k_2, \ldots, k_t\}$ um conjunto de termos e $D = \{d_1, d_2, \ldots, d_t\}$ um conjunto de documentos. A cada termo k_i, $1 \leqslant i \leqslant t$, e cada documento d_j, $1 \leqslant i \leqslant n$, associamos o peso $w_{ij} > 0$ se k_i aparece em d_j e $w_{ij} = 0$ caso contrário.

Definimos o vetor $\vec{d}_j$ como sendo a coleção de pesos associados a d_j, ou seja, $\vec{d}_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$ e a similaridade entre dois documentos d_i e d_j pela correlação entre $\vec{d}_i$ e $\vec{d}_j$ quantificada pelo cosseno do ângulo entre eles, isto é

$$sim(d_i, d_j) = \frac{\langle \vec{d}_i, \vec{d}_j \rangle}{\|\vec{d}_i\| . \|\vec{d}_j\|} = \frac{\sum_{m=1}^{t} (w_{m,i} . w_{m,j})}{\sqrt{\sum_{m=1}^{t} w_{m,i}^2} \cdot \sqrt{\sum_{m=1}^{t} w_{m,j}^2}}$$

Uma consulta q é um vetor de pesos associados aos termos. Podemos considerá-la como o vetor $\vec{d}_q = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$ associado a um documento d_q. Podemos portanto avaliar a similaridade entre um documento e uma consulta q. Observe que $0 \leqslant sim(d_j, q) \leqslant 1$, pois $w_{ij} \geqslant 0 \; \forall \; i,j$.

É necessário agora definir como calcular os pesos. Uma proposta é calcular o peso por meio do produto de dois fatores. O primeiro, procura determinar quais as características que definem, que têm em comum, os documentos que pertencem ao conjunto de resposta, ao conjunto que satisfaz a consulta. O segundo, procura determinar quais as características que mais distinguem os documentos que pertencem ao conjunto de resposta dos demais. Assim, o primeiro fator quantifica a semelhança dentre os documentos do conjunto resposta e é chamado de *fator tf* (*term frequency*). O segundo fator quantifica a dissemelhança dentre os documentos da coleção e é chamado de *idf* (*inverse document frequency*). A motivação para o *idf* é que termos muito freqüentes na coleção são pouco úteis para distinguir documentos uns dos outros e devem, portanto, acarretar um peso total menor.

Daremos agora as definições para os fatores *tf* e *idf*. Seja n o número total de documentos na coleção e n_i o número de documentos nos quais o termo k_i aparece. Seja $f_{i,j}$, $1 \leqslant i \leqslant t, \; 1 \leqslant j \leqslant n$ a freqüência bruta, isto é, o número de vezes que o termo k_i aparece no documento d_j. Seja $fmax_i = \max_{1 \leqslant m \leqslant n} \{f_{i,m}\}$. Definimos a freqüência normalizada $\overline{f}_{i,j}$ como sendo:

$$\overline{f}_{i,j} = \frac{f_{i,j}}{fmax_i}, \qquad 1 \leqslant i \leqslant t, \;\; 1 \leqslant j \leqslant n$$

O *idf* para o termo k_i é dado pelo logaritmo da relação entre o número total de documentos na coleção e o número de documentos nos quais o termo aparece:

$$idf_i = \log \frac{n}{n_i}, \qquad 1 \leqslant i \leqslant t$$

Podemos agora calcular os pesos $w_{i,j}$ da seguinte maneira:

$$w_{i,j} = \overline{f}_{i,j} . idf_i, \qquad 1 \leqslant i \leqslant t, \quad 1 \leqslant j \leqslant n$$

Tais estratégias são chamadas de esquemas *tf·idf* (*term-frequency – inverse document frequency*). Segundo Baeza-Yates e Ribeiro-Neto (1999), os melhores esquemas conhecidos para ponderação de termos utilizam esta estratégia ou variações desta fórmula.

Adaptações no Método *tf·idf*

Algumas adaptações tiveram que ser feitas no método de cálculo de similaridades. O método de cálculo *tf·idf* supõe que todos os documentos que compõem a coleção estejam disponíveis no momento de calcular as similaridades de d_j. Isto não ocorre no ReVEL, uma vez que os documentos vão sendo gradativamente obtidos da Internet. O modo de contornar esta limitação foi a de tirar um instantâneo da similaridade de d_j com a coleção disponível no momento.

Uma estrutura de dados no ReVEL contém os valores das similaridade $sim(d_j, d_p)$. A cada novo documento d_{p+1} obtido, esta estrutura é atualizada e são calculadas e acrescidas as similaridades $sim(d_1, d_{p+1})$, $sim(d_2, d_{p+1})$, ... , $sim(d_p, d_{p+1})$. Assim, a cada novo documento d_{p+1} obtido, as suas similaridades com os demais vão sendo cada vez mais próximas da similaridade real, calculada se toda a coleção estivesse disponível. O único documento cujos valores calculados serão os corretos será o último, d_N, infelizmente aquele considerado o menos relevante pelo mecanismo de busca. Resultados empíricos, dos quais a Figura 4.20 é um exemplo, mostram que o valor da similaridade entre os documentos vai se estabilizando perto do valor final, calculado com a coleção completa, conforme o valor de p se aproxima de N. Para corrigir o problema do desvio na similaridade dos primeiro documentos tidos como mais relevantes, o Gerenciador de Cálculo recoloca na fila de cálculo os primeiros $min\ \{x, n\}$ documentos e os reprocessa. O valor x está pré-definido em 15 (documentos), que corresponde a uma página e meia de resposta.

O reprocessamento dos primeiros x documentos ocorre em duas situações: caso o número de documentos na coleção ultrapasse $2x$ ou caso o tempo decorrido desde o início do primeiro cálculo até um dado momento ultrapasse 40 segundos. O critério de tempo é importante quando há demora para obter os documentos e quando o número de documentos na coleção é inferior a $2x$, uma vez que o Gerenciador de Cálculo não sabe, *a priori*, de quantos documentos a coleção é composta.

Resultados empíricos revelaram que, apesar de o resultado do cálculo inicial ser uma aproximação bastante ruim para o valor do *tf·idf* calculado para a coleção inteira, já é um valor útil como resultado para a Interface exibir para o usuário. Além disso, o recálculo aproxima bastante o novo valor do que seria o valor correto do *tf·idf* . A Figura 4.20 exibe os valores de similaridade entre o primeiro documento de uma consulta por "medusa" contra os demais 49 documentos. Pode-se perceber que após o valor de 25 documentos na

coleção o valor calculado com a coleção incompleta é bastante próximo do valor calculado com a coleção completa. Apesar de ser um caso particular, é bastante representativo do comportamento geral dos valores de de *tf·idf* calculados para outras pesquisas.

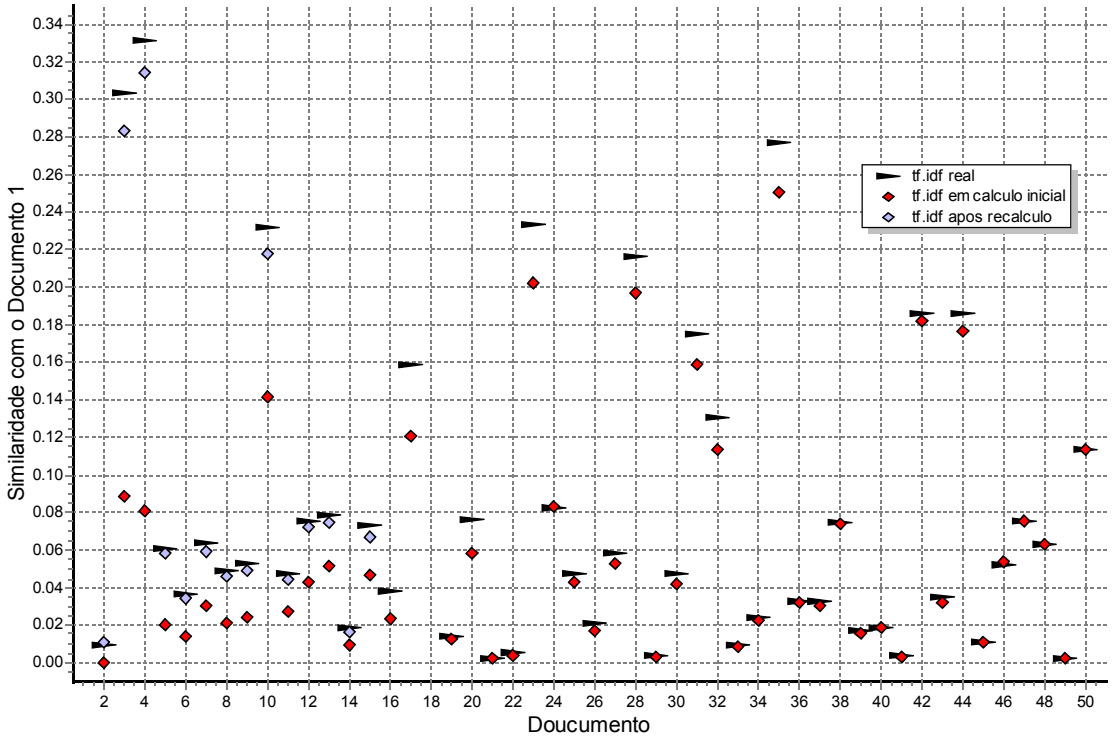


Figura 4.20: Comportamento de *tf·idf* em Função do Número de Documentos

A estrutura de dados onde o Gerenciador de Cálculo mantém os valores de similaridades é compartilhada com a Interface que tem, portanto, sempre os valores mais recentes ao seu dispor.

Estatísticas do desenvolvimento do sistema foram coletadas e apresentam-se no Anexo VI.

Vamos então verificar, no próximo capítulo, por meio de testes de usabilidade, como o sistema é visto e avaliado pelos usuários.

Capítulo 5

Avaliações com Usuários

É fundamental verificar qual a compreensão que os usuários têm da interface e, portanto, qual a utilização que eles fazem dela. Para isso é empregada a técnica de testes de usabilidade, que reportam a perspectiva que os usuários têm da interface.

5.1 Testes de Usabilidade

Testes de Usabilidade são testes feitos com usuários visando atingir um de dois objetivos: avaliar questões ligadas ao desempenho ou saber quais aspectos da interface estão bons ou ruins e como podem ser melhorados (Rocha e Baranauskas, 2003).

Os objetivos podem ser atingidos avaliando os fatores que caracterizam a usabilidade de um sistema, ou seja: facilidade de aprendizado, facilidade de uso, eficiência de uso e produtividade, satisfação do usuário, flexibilidade, utilidade e segurança no uso (Nielsen, 1993). Em geral um teste visa priorizar alguns poucos fatores destes descritos acima – facilidade de aprendizado e de uso, por exemplo – uma vez que alguns deles com freqüência podem ser contraditórios como, por exemplo, facilidade de aprendizado / produtividade.

Um teste em geral é composto de quatro etapas, como pode ser visto na Tabela 5.1.

1. **Preparação:** quando garante-se que o ambiente e os equipamentos estejam prontos antes do usuário chegar. Atenção especial deve ser dada para garantir que nenhum efeito residual de um teste anterior – pré-seleções, avisos, resultados anteriores etc. – estejam nos equipamentos a serem utilizados;

2. **Introdução:** aqui os usuários são apresentados à situação de teste e colocados a vontade. Também é importante esclarecer o usuário quanto a diversos detalhes do teste, principalmente que:

- O propósito do teste é avaliar o sistema e não o usuário;
- Os usuários não devem se preocupar em ferir sentimentos dos desenvolvedores com suas observações;
- Os resultados do teste servirão para melhorar a interface do usuário;
- O sistema é confidencial e não deve ser comentado com outros (possíveis usuários em futuros testes);
- A participação é voluntária e o usuário pode desistir a qualquer momento;
- O anonimato do participante estará sempre garantido;
- Há a necessidade da gravação de áudio ou vídeo e porquê;
- O usuário pode fazer qualquer pergunta durante o teste, mas nem sempre o avaliador poderá respondê-la;
- O usuário receberá instruções específicas sobre a forma do teste – falar em voz alta, executar as tarefas o mais rápido que puder etc.

3. **Teste:** é a etapa em que a coleta de informação sobre a utilização ocorre:

- O avaliador pede ao usuário que leia, preencha e assine o termo de sigilo do teste – uma réplica do termo utilizado encontra-se no Apêndice III;
- Algumas vezes o usuário responde um questionário, que objetiva conhecer especificidades do usuário;
- Executa-se o teste: o usuário recebe as tarefas, as lê e confirma sua compreensão; inicia cada tarefa, já sob gravação; o(s) avaliador(es) tomam notas e questionam o usuário;

4. **Sessão Final:** depois de esgotadas as tarefas ou o tempo definido para tanto, os participantes são convidados a fazer comentários, emitir opiniões ou preencher um questionário final.

Tabela 5.1: Etapas do Teste de Usabilidade
(itens adaptados de Rocha e Baranauskas, 2003 e de Prates e Barbosa, 2003)

5.2 Testes com Mecanismos de Busca

Testes de usabilidade foram executados com dez alunos de pós-graduação de diferentes áreas: dois de Geografia, um de Engenharia Civil e os demais de Ciência da Computação. Foram entrevistados homens e mulheres em igual quantidade. Antes de iniciar o teste, foi pedido ao usuário que preenchesse um termo de sigilo, que garante a anonimidade do respondente e cujo texto encontra-se no Apêndice III. Logo após foram lidas as Instruções que informam os motivos, o conteúdo e as fases do teste. As instruções encontram-se no Apêndice IV.

O teste com cada usuário foi composto de duas fases, pode ser observada na Tabela 5.2, cada uma com as mesmas tarefas porém utilizando um modo de busca diferente: um era o método de lista comum e para ele foi padronizada a utilização do Google, com o navegador Internet Explorer; a outra era com um tipo de visualização – ReVEL ou Kartoo. Em função de qual visualização era utilizada nós temos duas categorias: Google/Kartoo e Google/ReVEL. A ordem em que o usuário executava cada fase foi aleatorizada – sorteava-se qual o modo de busca que iria iniciar o teste – até o momento em que foi necessário

			1ª Fase			2ª Fase		
			Questões			Questões		
	Usuário	Mecanismo	1	2	3	1	2	3
Kartoo	$U_{K,1}$	Google	✓	✓	✓			
		Kartoo				—	—	✓
	$U_{K,2}$	Google	✓	✓	✓			
		Kartoo				✓	✓	✓
	$U_{K,3}$	Google	✓	—	✓			
		Kartoo				✓	✓	—
	$U_{K,4}$	Google				✓	✗	✗
		Kartoo	✓	—	—			
ReVEL	$U_{R,1}$	Google	✓	✓	✓			
		ReVEL				✓	✓	✓
	$U_{R,2}$	Google				✓	—	✓
		ReVEL	✓	—	✓			
	$U_{R,3}$	Google				✓	✓	✓
		ReVEL	✓	✓	✓			
	$U_{R,4}$	Google	✓	✓	✓			
		ReVEL				✓	✓	✓
	$U_{R,5}$	Google				✓	—	✓
		ReVEL	✓	—	✓			
	$U_{R,6}$	Google				✓	—	—
		ReVEL	✓	—	✓			

Tabela 5.2: Respostas do Teste de Usabilidade
(✓ : resposta correta; ✗ : resposta incorreta; — : desistência)

atribuir a ordem, a fim de garantir que houvesse um mesmo número de usuários iniciando com o modo comum – Google – e com visualização – ReVEL/Kartoo. Como o número de usuários foi dez, não foi possível garantir que houvesse um mesmo número de homens e mulheres iniciando com cada modo; assim, três homens e duas mulheres iniciaram com a fase do Google e dois homens e três mulheres iniciaram o teste com a fase da visualização. Uma síntese de quais usuários iniciaram em qual fase pode ser observada na Tabela 5.2. Os usuários $U_{K,2}$, $U_{K,4}$, $U_{R,2}$, $U_{R,4}$, $U_{R,6}$ são do sexo feminino; os demais, do sexo masculino. Para evitar uma identificação dos respondentes e assim garantir seu sigilo, podemos apenas informar que os usuários $U_{K,1}$, $U_{K,2}$, $U_{K,3}$ e $U_{K,4}$ são da pós-graduação em Computação.

Embora a finalidade da avaliação não seja de cunho estatístico, foi utilizada a mesma estrutura a que se recorre para realizar experimentos estatísticos, denominada Planejamento de Experimentos. Tal estrutura – aleatorização, distribuição igual de elementos por categorias etc. – é um recurso para minimizar ou eliminar a possibilidade de que uma modelagem falha do experimento – usuários de um mesmo sexo, todos usuários iniciando com Google etc. – introduza erros sistemáticos nos resultados, acarretando um viés não identificável nas conclusões.

Cada fase foi composta de três tarefas, supostamente de nível crescente de dificuldade. A primeira tarefa era encontrar um endereço Internet – URL – no qual tivesse o nome do herói grego que matou o monstro Medusa. Como segunda tarefa, o usuário deveria descobrir qual o autor de determinados versos e a que obra eles pertenciam. Os versos em questão são de Pedro Calderón de la Barca, da obra "A Vida é Sonho":

Que é a vida? Um frenesi.
Que é a vida? Uma ilusão,
uma sombra, uma ficção;
o maior bem é tristonho,
porque toda a vida é sonho
e os sonhos, sonhos são.

A terceira tarefa pedia que o usuário localizasse um valor numérico em especial: o Produto Interno Bruto brasileiro para o ano de 2003. As respostas podem ser vistas na Tabela 5.2, sendo que " ✓ " denota resposta correta, " ✗ " denota resposta incorreta e " — " significa desistência.

Os testes com os alunos de Geografia ocorreram na sala de Informática dos professores do Departamento de Geografia da Universidade do Estado de São Paulo (Unesp) campus de Presidente Prudente no dia quatro de janeiro de 2005. O computador utilizado possuía 1,7 Ghz de velocidade de CPU, 256 Mbytes de memória e utilizava o sistema operacional Windows Me. Os alunos da Ciência da Computação e da Engenharia Civil realizaram o teste na sala de alunos de doutorado do Instituto de Computação da Universidade Estadual de Campinas (Unicamp), utilizando um computador com 1,5 Ghz, 512 Mbytes de memória e utilizando o sistema operacional Windows XP, ocorrendo entre 28 de janeiro de 2005 e 23 de fevereiro do mesmo ano.

Os testes Google/Kartoo foram realizados nas dependências do NIED – Núcleo de Informática Aplicada à Educação – NIED – da UNICAMP. O computador possuía um processador de 800 Mhz de velocidade da CPU, 256 Mbytes de memória e o sistema operacional era um Windows 98.

Os testes foram filmados e, à exceção dos alunos de Geografia, os relatórios internos de utilização do ReVEL (*logs*) foram coletados. Estes relatórios, são gerados e gravados em disco pelo ReVEL com objetivo de prover dados para depuração do sistema, porém os eventos registrados ajudaram bastante a coletar informações sobre o comportamentos dos usuários e o desempenho do sistema. Infelizmente, os alunos de Geografia não tiveram seus relatórios coletados pois, aparentemente, o ReVEL não obteve permissão de escrita em disco no laboratório da Unesp onde o teste foi efetuado.

Terminadas as tarefas, seja por terem sido completadas ou por desistência, foi requisitado ao usuário que preenchesse um questionário de perguntas abertas sobre suas impressões. O questionário encontra-se reproduzido no Apêndice V.

Ficou evidente, principalmente pela utilização do usuário $U_{R,3}$, que a capacidade de expressar a consulta nos termos do mecanismo de busca é um diferencial que possibilita melhorar as chances de localizar a informação. Aqueles usuários com conhecimento da sintaxe de consulta, usaram recursos como `"* matou górgona"`, significando "páginas na quais haja uma palavra qualquer seguida do texto 'matou górgona' ", ou mesmo a expressão `"que é a vida" "um frenesi"` significando "páginas nas quais haja o texto explícito 'que é a vida' e o texto explícito 'um frenesi' e não apenas estas palavras isoladas – "que", "vida" e "frenesi" – que, em qualquer ordem na página, podem estar se referindo a um outro assunto completamente diverso. Isto pode indicar que uma ferramenta útil a ser considerada como acréscimo ao sistema poderia ser o recurso de manipular visualmente o texto de entrada de modo a prover para o usuário um meio interativo de inserir recursos de sintaxe à frase de consulta.

5.2.1 Google

A qualidade mais reportada do modo de lista representado pelo Google é justamente sua simplicidade, como ilustrado na Figura 3.5, página 27. Soma-se a isso o fato de a interface ser aquela utilizada para navegação Internet e, portanto, já conhecida, como pode ser observado pelos comentários dos usuários: "não tem necessidade de aprendizado prévio", "claro como pesquisar", "pouca dificuldade de compreensão/utilização". O modo texto não foi considerado problema por três usuários: "sinto mais segurança pois estou mais acostumado com lista", "gosto do visual simples e disposição ao longo da página" e "falta de recursos gráficos não incomoda", embora um usuário relate que a visualização – no caso, ReVEL – tem "mais funcionalidade".

Quanto a lista de resposta, um usuário relatou que, independente dos itens que a componham, ele *sempre* seleciona o primeiro para examinar. Provavelmente explorando este comportamento é que o Google criou o botão "estou com sorte" que não retorna a lista de resultados e sim já o primeiro documento.

Dois recursos do Google foram valorizados. Um deles explicitamente, quando o usuário comentou que "gosto muito do *cache* exibindo as palavras buscadas", referindo-se à exibição de uma página especial, cópia do documento original disponível na Internet. Nesta cópia, as palavras-chave utilizadas aparecem destacadas, cada uma colorida com uma cor diferente, facilitando sua localização dentro do documento. Outro recurso que mostrou-se valioso, embora o usuário não a tenha relatado explicitamente mas a tenha utilizado, é a sugestão de alternativa ortográfica. Ocorreu de este usuário digitar "herói grego matou górgona meduza" e o Google, em sua página de resposta, sugeriu: "não desejaria procurar por 'herói grego matou górgona medu**s**a' ".

Alguns problemas variados também foram relatados:

- "muito texto e sem relacionamento entre os elementos selecionados";
- "perda de credibilidade pela questão da propaganda paga";
- "não sei como ele organiza os *sites* encontrados pois, às vezes informações mais atuais são listadas por último; não sei os critérios utilizado para listar os *sites*";
- " 'estou com sorte' não é intuitiva (não sei a diferença entre ela e a opção *default*)."

5.2.2 Kartoo

A interface do Karto agradou esteticamente - "achei muito engraçadinho", "interface é limpa e leve", "gostei do contraste e das cores" – embora não fosse unanimidade: "fiquei desorientado, aparece tudo jogado". O mapa foi compreendido – "achei que o 'mapa' é melhor e mais fácil que a lista que o Google apresenta", porém, também não por todos: "não entendi o desenho de fundo". A sensação mais comum foi de desorientação "me causou desconforto, meio perdido".

A grande quantidade de conceitos do mapa – tamanho do desenho do documento, desenho no fundo, palavras na visualização, desenhos de documentos sobrepostos etc. – talvez tenha dificultado a compreensão de cada um dos elementos isoladamente:

- "resultado de busca é complexo";
- "ainda não compreendi o que os *links* significam, nem as diferentes representações (gráficas) para o resultado da busca";
- "dificuldade de entender a lógica da pesquisa e o significado dos elementos";
- "algumas linhas, papéis e 'nomes' ou 'termos' aparecem entre os *sites* e não sei para que servem";

Usuários também expressaram o desejo de utilizar uma busca avançada mas não identificaram como: "não achei opção para busca avançada".

Um conflito surgiu quanto à movimentação pela interface. Como o ambiente de pesquisa é o navegador mas o Kartoo é uma aplicação em *Shockwave/Flash*[7], houve confusão quanto à função do botão "retornar" do navegador, o qual os usuários tentavam utilizar para voltar à pesquisa anterior, que não estava disponível por ser uma instância anterior de utilização da aplicação.

A barra de rolagem, necessária para exibição da interface na maior parte dos monitores – resolução de tela de 800×640 *pixels* – não foi percebida e, portanto, a coluna da direita ficou oculta para todos os usuários.

5.2.3 ReVEL

O teste revelou alguns detalhes interessantes sobre o entendimento que os usuários tiveram dos componentes da Interface.

[7] www.macromedia.com

Os ícones foram prontamente identificados como os documentos recuperados na consulta mas a numeração de cada um deles, não. Nem todos os usuários identificaram os números presentes nos ícones como a posição daquele documento na lista de resposta. Um estudante de Geografia até mesmo supôs que os números representassem a ordem cronológica dos documentos. Também houve equivocos na interpretação das arestas. um dos estudantes supôs que as arestas significavam "a continuação do documento", como se representassem um tipo de "próxima página", mesmo não sendo as arestas representadas como vetores orientados.

Metade dos alunos – aqueles de Computação – relataram que o posicionamento automático dos ícones "perturba". Um deles acrescentou que a movimentação, que ocorre de tempo em tempo, dá a impressão "que o programa não está pronto, ainda está trabalhando".

Dentre os componentes da Interface, o mais imediatamente reconhecível foi a Visão Geral, provavelmente por ser um recurso comum em interfaces que representam na tela grandes áreas de visualização. Alguns alunos interagiram com a Visão Geral como o esperado, clicando no novo local de visualização dentro da Visão Geral para tê-la representada na Área de Representação. Outros, porém, tentaram clicar e arrastar o retângulo que representa, na Visão Geral, a região exibida naquele momento na Área de Representação.

As Dicas (*tooltips* ou *hover texts*), textos explicativos que aparecem por um determinado tempo quando o ponteiro do *mouse* fica parado sobre um componente, foram tidas como úteis por três dos usuários. Um deles, ao identificar que o ReVEL tinha o recurso, passou a explorar a interface para ver que dica havia para cada um dos componentes. Um usuário compreendeu o conceito por trás dos botões de Seleções Mais Recentes por meio das dicas. Entretanto um dos usuários, perito na utilização de computadores, simplesmente ignorou as dicas. Perguntado sobre seu comportamento, ele disse que "geralmente elas não dizem nada de novo".

Ao acaso, aconteceu de um dos usuários ser daltônico. Ele reclamou que há tons de vermelho em demasia na interface e isso causa – em daltônicos – um cansaço muito grande. Sugeriu que se optasse por tons de azul e deu como exemplo a interface do Orkut[8].

Dentre os usuários, um relatou que achava mais fácil localizar informação usando o sistema e um deles relatou que, apesar de a Tabela (Índice de Documentos) ser uma grande ajuda, "a visualização gráfica não ajuda".

Surpreendentemente, a Tabela foi amplamente utilizada por todos os usuários. Os três alunos de Computação a expandiram até o seu máximo e reclamaram que ela não podia ser ampliada ainda mais. Um deles acrescentou que "é legal poder ver todos os endereços" (URLs). Em geral, a Tabela foi o elemento mais utilizado, como se a visualização fosse um recurso auxiliar da Tabela e não o contrário, ao menos por construção.

Um usuário gostou da possibilidade de ver a lista de resultados original e outro expressou sua satisfação pela facilidade com que se pode requisitar que o navegador abra um documento. Isso se dá por que basta clicar na caixa de checagem de documento Visitado (**v**) da Tabela que uma nova janela é aberta, com o navegador exibindo o dado documento. Isto requer apenas um clique.

[8] www.orkut.com

Os documentos que apresentaram erros na obtenção são prevenidos de serem automaticamente movimentados – "pregados" – e alinhados ao longo do topo esquerdo superior da Área de Representação. Um usuário interpretou tais ícones como "os documentos clicados". Mais tarde este usuário corretamente os identificou por perceber que eram os mesmos marcados com "erro!" na Tabela.

Um comportamento curioso apareceu separando os usuários de Geografia dos demais: quando consultando uma página no navegador, os usuários da área de Geografia a liam até o fim, procurando pela informação, contrastando com o comportamento dos usuários de Computação e Engenharia Civil, que apenas examinavam a página superficialmente.

Como a obtenção dos documentos ocorre em um processo leve e a lista de resultados é quase que imediatamente disponível para avaliação do usuário, nenhum dos usuários reclamou de demora no tempo de obtenção dos documentos. Talvez tenha ficado oculto o fato de que o processo ainda não esteja completo: mesmo que já tenham sido apresentados os ícones dos documentos, nem todos os seus relacionamentos foram calculados e, portanto, representados. A disposição automática dos ícones ocorre durante todo o processo: a partir do início da consulta até o encerramento do ReVEL, quando ocorre o início de uma nova consulta ou até ser desabilitado. Todos usuários concluíram e acertaram as tarefas 1 e 3 mas metade desistiu da tarefa 2, que não era suposta ser a mais difícil. Isto provavelmente ocorreu por serem os versos muito citados e, com freqüência, errôneamente atribuídos, confundindo assim os usuários.

Alguns Dados Quantitativos

A possibilidade de análise do relatório de depuração – *log* – do ReVEL colhido durante o teste com os alunos da Computação permitiu uma avaliação mais quantitativa de alguns parâmetros. A Tabela 5.3 traz, compilando valores da Tabela 5.4, o número médio de consultas por tarefa. É possível perceber que, para estes quatro usuários, realmente a segunda tarefa apresentou mais dificuldade, exigindo um número maior, na média, de consultas.

Tarefa	Número Médio de Consultas	Desvio Padrão
1	1,25	0,50
2	3,50	1,29
3	1,75	0,96
Total	2,17	1,34

Tabela 5.3: Número Médio de Consulta por Tarefa

Todos os valores coletados apresentam-se sintetizados na Tabela 5.4. Por meio destes relatórios quantificamos:

- hora em que o consulta foi iniciada – coluna *Início da Consulta*;
- momento em que o último documento foi obtido – coluna *Fim da Obtenção*;
- o tempo total consumido para obter os documentos da Internet – coluna *Tempo Total de Obtenção*;
- tempo em que a consulta se encerrou (foi iniciada uma outra consulta ou foi terminado o teste) – coluna *Fim da Consulta*;

- o tempo em que o usuário esteve avaliando os documentos (diferença entre a 4ª e a 1ª colunas) – coluna *Tempo da Consulta*;
- quantos documentos a consulta retornou – coluna *Documentos obtidos*;
- para quantos documentos o sistema teve tempo de calcular a similaridade antes que o usuário decidisse por alterar, refazer ou reiniciar a consulta – coluna *Documentos Calculados*;
- se a consulta foi truncada, isto é, se o ReVEL teve tempo de realizar todas suas operações antes que o usuário decidisse por uma nova consulta – coluna *Consulta foi truncada?* ;
- sobre qual tarefa foi a consulta – coluna *Tarefa* e
- quais as palavras-chave foram utilizadas na consulta – coluna *Palavras-chave*.

Avaliar o tempo da consulta não era um dos objetivos do teste de usabilidade. Porém analisando a Tabela 5.4 foi possível calcularmos que o tempo médio de cada consulta é de três minutos e onze segundos, que é um tempo bastante longo. Se levarmos em consideração, porém, que diversos usuários ficaram explorando a interface e que ela lhes foi apresentada naquele momento, parece ser um tempo razoável.

No próximo capítulo apresentam-se sintetizados pontos fundamentais deste trabalho, na forma de discussões, conclusões. Algumas propostas de extensões também são apresentadas.

	Início da Consulta	Fim da Obtenção	Tempo Total de Obtenção	Fim da Consulta	Tempo da Consulta	Docs. Obtidos	Docs. Calculados	Consulta foi truncada?	Tarefa	Palavras-chave
$U_{R,1}$	16:12:53	16:15:00	02:07	16:18:24	05:31	50	43	s	1	herói medusa
	16:18:24	16:18:32	00:08	16:21:29	03:05	2	2	n	2	poesia "que é a vida? um frenesi."
	16:21:29	16:22:44	01:15	16:23:40	02:11	45	18	s	2	"Pedro Calderon de la Barca" frenesi
	16:27:39	16:28:26	00:47	16:29:32	01:53	32	32	n	3	PIB Brasil 2003
$U_{R,2}$	15:32:23	15:33:33	01:10	15:33:33	01:10	50	50	n	1	górgona medusa
	15:52:21	15:53:59	01:38	15:54:02	01:41	50	50	n	2	que é a vida? um frenesi
	16:01:53	16:02:55	01:02	16:03:42	01:49	50	50	n	2	calderón de la barca
	16:03:42	16:06:06	02:24	16:06:06	02:24	28	23	s	2	calderón de la barca obras
	16:06:06	16:09:34	03:28	16:09:34	03:28	20	13	s	2	poesias de calderón de la barca
	16:11:32	16:11:46	00:14	16:12:59	01:27	4	4	n	3	produto interno bruto do brassil 2003
	16:15:42	16:16:58	01:16	16:17:05	01:23	38	38	n	3	pib 2003
	16:17:05	16:18:10	01:05	16:18:10	01:05	32	32	s	3	valor pib 2003
$U_{R,3}$	16:01:53	16:02:06	00:13	16:05:44	03:51	8	8	n	1	matou Medusa
	16:15:52	16:16:21	00:29	16:19:07	03:15	3	3	n	2	"que é a vida? um frenesi" autor
	16:19:07	16:20:42	01:35	16:22:54	03:47	3	3	n	2	"que é a vida" "um frenesi" Pedro Calderón
	16:23:32	16:23:50	00:18	16:28:57	05:25	50	50	n	2	"que é a vida" "um frenesi"
	16:29:41	16:30:45	01:04	16:31:54	02:13	40	18	s	3	"o valor do produto interno bruto no ano de 2003 é"
$U_{R,4}$	17:37:39	17:38:26	00:47	17:58:00	03:14	50	50	n	1	górgona medusa
	17:57:59	17:58:11	00:12	18:00:15	02:16	50	50	n	1	matou Medusa
	18:00:15	18:02:45	02:30	18:03:30	03:15	50	50	n	2	Que é a vida? Um frenesi
	18:03:30	18:05:08	01:38	18:07:48	04:18	50	50	n	2	Caldéron de la Barca teatro
	18:07:48	18:08:03	00:15	18:12:13	04:25	45	45	n	2	obras Caldéron de la Barca
	18:12:18	18:12:53	00:35	18:16:11	03:53	48	48	n	2	Calderon de la Barca sonho
	18:16:11	18:16:47	00:36	18:20:29	04:18	44	44	n	2	Calderon de la Barca A vida é sonho
	18:20:29	18:20:53	00:24	18:24:55	04:26	30	30	n	3	PIB Brasil 2003
	18:24:55	18:25:16	00:21	18:32:00	07:05	32	32	n	3	PIB Brasil 2003 foi

Tabela 5.4: Alguns Valores dos Testes de Usabilidade

Capítulo 6

Discussão, Conclusões e Trabalhos Futuros

A grande quantidade de informação disponível atualmente na Internet torna, com freqüência, localizar uma determinada informação uma tarefa árdua, quando logra ser produtiva. Diversos recursos têm sido produzidos na literatura acadêmica com a finalidade de sanar ou, ao menos, mitigar tal problema.

A técnica de Visualização de Informação possibilita explorar as capacidades visuais humanas para, com mais facilidade, absorver e compreender grandes quantidades de dados abstratos. Este trabalho apresentou o *design* e desenvolvimento do ReVEL, um sistema de *software* que provê uma camada de representação gráfica sobre a lista de respostas obtidas de um mecanismo de busca.

Destacamos alguns dentre os motivos alegados de que a Visualização pode ampliar a cognição, que foram utilizados no ReVEL e como:

- **Ampliação dos Recursos:**
  A exibição de todos documentos relacionados por similaridade em uma rede – grafo conexo – explicita quais documentos estão relacionados entre si e com qual intensidade, poupando esforço cognitivo de, por exemplo, avaliar barras que indicam quão significativa é a categoria para o conjunto de palavras-chave, como visto em Chen e Dumais (2000). Evita também o problema de documentos que localizam-se no limiar da definição de categorias e são incluídos em ambas (Chen e Dumais, 2000) pois, no grafo, se um documento A está relacionado a um B e B está relacionado a outro C, fica evidente que B é similar a A e C porém A e C não são similares entre si por não estarem ligados. Outra ampliação de recurso refere-se à expansão da memória de trabalho, pois agora estão à disposição cinqüenta documentos em lugar dos dez normalmente oferecidos pelos mecanismos de busca. Ainda há o recurso de recobrar os seis últimos documentos selecionados, na ferramenta Seleções mais Recentes descrita na seção 4.2.2;
- **Redução em Buscas**
  O agrupamento dos documentos feito pelo grafo, dispõe juntos os documentos, evitando o movimento de vai-e-vem em uma lista, quando os documentos estão

dispersos. Também consegue-se alta densidade de dados, exibindo cinqüenta – ou mais – documentos, com as mesmas informações disponibilizadas pelo mecanismo de busca, só que simultaneamente;

- **Mídia Manipulável**
  Tendo um paralelo imediato dentro de IHC através da Manipulação Direta (Shneiderman, 1983), a capacidade descrita como "mídia manipulável" é expressa no ReVEL pela possibilidade de verificar, com resposta imediata na interface, as interações realizadas pelo usuário, possibilitando-lhe alterar parâmetros – seleção e posicionamento de documentos, reter documento para próxima consulta etc – e vê-los imediatamente serem efetivados;

As propriedades retinianas e os elementos gráficos elementares, descritos na seção 2.1.2, foram observadas e agregadas ao ReVEL:

- **Posição:**
  Segundo o comparativo compilado por Card *et al.* (1999) das propriedades retinianas, a posição entre os elementos é, dentre todas, a mais efetiva para codificar informação. Esta também é a conclusão de um estudo feito por Cleveland e McGill (1984). Fica claro, portanto, que a razão de ser empregada, no sistema ReVEL, a melhor propriedade para codificar a similaridade entre documentos – o fator mais importante a ser representado – seja o mais efetivo. Não bastasse o próprio bom senso para justificar esta decisão, ainda haveria o Princípio de Ordenação por Importância de Mackinley (1986): "codifique a informação mais importante do modo mais efetivo".
- **Tamanho**:
  A segunda melhor propriedade para representar todo tipo de informação foi destinada a representar o quanto determinado documento é significativo para as palavras-chave escolhidas. Assim, utilizando a posição do documento na lista de resposta como um estimador de quão significativo ele é, grafamos os documentos tão grandes quanto mais próximos do topo da lista, em categorias de tamanhos, conforme mostra a Tabela 4.2 na página 41.

O levantamento inicial realizado com usuários de ferramentas de busca foi importante para identificar características e hábitos destes usuários, além de embasar algumas decisões de *design* e de desenvolvimento do ReVEL:

- Ficou evidente ser o Google o mecanismo de busca preferencial, logo ele foi o escolhido para representar resultados em forma de lista neste trabalho;
- A avaliação da pragmática da interface de entrada de dados Google/Altavista vista na Figura 3.6, na página 28, sugeriu que a mesma abordagem fosse utilizada no ReVEL, conforme exibe a Figura 4.5 à página 38;
- A estratégia do usuário frente ao insucesso de obter a informação na primeira página de resultados era de que quase metade dos usuários – 49% – buscassem a página seguinte. Logo seria importante o ReVEL já apresentar a "página seguinte". Uma vez que o sistema recupera cinqüenta documentos, isso equivale à página inicial e mais quatro "páginas seguintes";

- Somando as porcentagens das expectativa de "quase sempre" e "freqüentemente" obter sucesso já na primeira página de resposta temos um total de 67%, que é bastante próximo do encontrado por Silverstein *et al.* (1999), que verificaram que 63,7% das sessões pesquisadas eram compostas apenas de uma consulta.

Um sistema foi então desenvolvido e avaliado. Uma comparação entre o sistema ReVEL (Zaina e Baranauskas, 2005) e os similares foi compilada e pode ser analisada na Tabela 6.1, na qual estão representadas as principais características dos sistemas disponíveis para visualização de recuperação de informação ora abordados.

		VR-Vibe	***Scatter/ Gather***	***Category Interface***	***TileBars***	***Light-house***	***Kartoo***	***ReVEL***
Consulta à Internet (e não a bancos de dados ou de documentos)		-	-	✓	-	✓	✓	✓
Apresentação dos resultados em forma gráfica		✓	-	-	✓	✓	✓	✓
Possibilidade de reutilizar resultados parciais de uma sessão de consulta para a seguinte		✓	-	-	-	-	-	✓
Usuários podem manipular dinamicamente a visualização		✓	-	-	-	-	✓	✓
Avaliação de todo o texto e não apenas de uma parcela (sumário, resumo) do documento para determinar a similaridade		-	-	-	✓	✓	-	✓
Agrupa Documentos por:	categorias (pré-definidas)	-	✓	✓	-	-	-	-
	similaridade	-	-	-	-	✓	-	✓
	presença em diversos mecanismos de busca	-	-	-	-	-	✓	-

Tabela 6.1: Comparativo entre Sistemas de Visualizações para Resultados de Consultas

A avaliação com os usuários revelou algumas peculiaridades que não foram imaginadas no projeto ou desenvolvimento:

- O mecanismo de posicionamento automático de ícones de documentos foi opinado como incômodo por alguns usuários;
- A Tabela Índice de Documentos foi amplamente utilizada, revelando que é importante para estes usuários analisar os endereços nos quais os documentos se encontram;
- Contrariamente às expectativas, nenhum usuário reclamou do tempo de espera de obtenção dos documentos em função, provavelmente, do fato de a interface não ficar indisponível para o usuário;

Um gráfico comparativo com as porcentagens das respostas dos usuários dos Testes de Usabilidade pode ser examinado na Figura 6.1. Podemos ver que o sistema ReVEL apresentou proporcionalmente o maior índice de respostas corretas e uma taxa de desistência similar àquela apresentada pelo Google. Estes resultados não têm respaldo estatístico porém, indicam que se for realizado um experimento com rigor estatístico – tamanho amostral suficiente, aleatorização de participantes etc – estes resultados podem vir a ser confirmados estatisticamente.

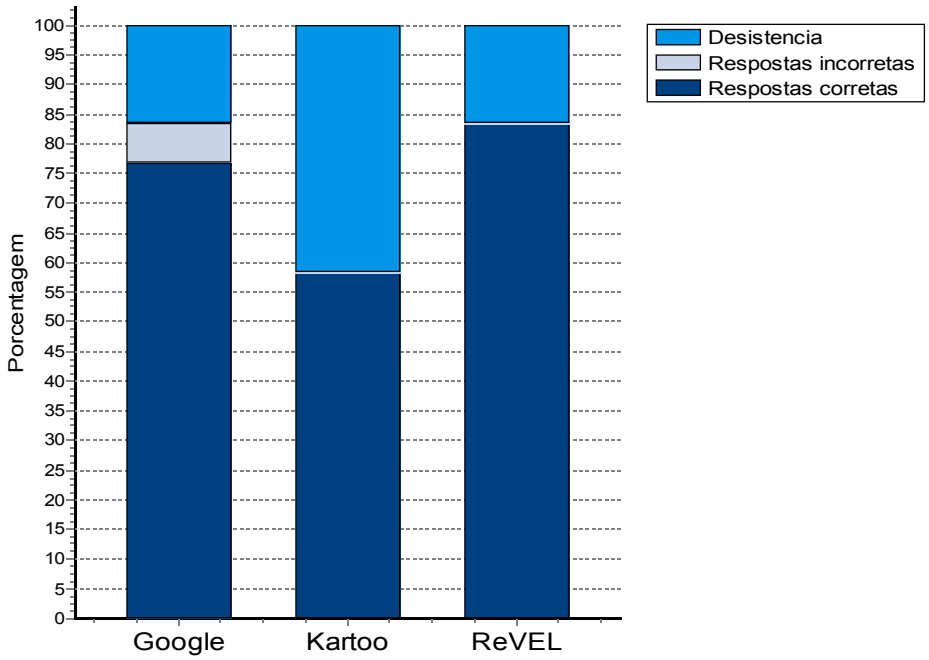


Figura 6.1: Comparativo entre Google, Kartoo e ReVEL

Como trabalhos futuros, podemos estudar o modelo mental dos usuários para mecanismos de busca. Também pode-se explorar os benefícios e requisitos sugeridos pelos usuários – alteração de cores, tamanho e posição de elementos da interface etc. Seria também interessante adaptar o sistema ReVEL, originalmente em Java, para módulos *applet*[9] Java e assim integrá-lo com o navegador. Um comentário feito por um dos usuários no Teste de Usabilidade – "eu sempre clico no primeiro documento" – sugeriu que, talvez por uma questão de pragmática, o primeiro documento resultante da consulta já devesse vir selecionado. Assim a rede de similares já iniciaria a se formar, tendo este documento em evidência.

A quantidade de informação disponível hoje em dia cresce rapidamente. O meio de acessá-la é mais e mais através de sistemas computacionais. É, portanto, uma necessidade que sejam estudados modos de tornar possível a todo ser humano um acesso simples, eficaz, prazeroso a estas informações.

[9] pequeno aplicativo que pode ser integrado à páginas HTML.

Referências

Alltheweb	www.alltheweb.com
Altavista	www.altavista.com
Baeza-Yates e Ribeiro-Neto (1999)	Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier (1999). *Modern Information Retrieval*. ACM Press / Addison Wesley.
Benford *et al.* (1999)	Benford, Steven D.; Taylor, Ian; Brailsford, David; Koleva, Boreana; Craven, Mike; Fraser, Mike; Reynard, Gail; Greenhalgh, Chris (1999). Three Dimensional Visualization of th World Wide Web. *ACM Computing Surveys*, dezembro, volume 31, número 4(es), pp. 1-16.
Card *et al.* (1999)	Card, Stuart K.; Mackinlay, Jock D.; Shneiderman, Ben (Eds.) (1999). *Readings in Information Visualization: Using Vision to Think*. Editora Morgan Kaufman, São Francisco.
Celebourne *et al.* (1994)	Colebourne, Andy; Mariani, John; Rodden, Tom; Twidale, Michael; Benford, Steve; Ingram, Rob; Snowdon, Dave (1994). Populated Information Terrains: Supporting the Cooperative Browsing of On-line Information. *University of Lancaster Research Report CSCW/13/1994*, obtido de citeseer.ist.psu.edu/213585.html.
Chen e Dumais (2000)	Chen, Hao; Dumais, Susan (2000). Bringing Order to the Web: Automatically Categorizing Search Results. *CHI Letters*, vol. 2, issue 1, pp. 145-150.
Cleveland e McGill (1984)	Cleveland, William S.; McGill, Robert (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, setembro, volume 79, número 387, pp. 531-553.

Cohen (1997) — Cohen, Jonathan D. (1997). Drawing Graphs to Convey Proximity: An Incremental Arrangement Method. *ACM Transactions on Computer-Human Interaction*, vol. 4, nº 3, setembro, pp. 197-229.

ComunIHC — Comunidade de IHC, www.comunihc.unicamp.br.

DeFanti *et al.* (1989) — DeFanti, T.A.; Brown, M. D.; McCormick, B. H. (1989). Visualization – Expanding Scientific and Engineering Research Opportunities. *IEEE Computer*, 22(8), pp.12-25.

Di Battista *et al.* (1994) — Di Battista, Giuseppe; Eades, Peter; Tamassia, Roberto; Tollis, Ioannis G. (1994). Algorithms for Drawing Graphs: an Annotated Bibliography. *Computational Geometry: Theory and Applications 4,* pp. 235-282.

Dondis (1997) — Dondis, Donis A. (1997). *Sintaxe da Linguagem Visual*. 2ª edição. Editora Martins Fontes, pp. 236.

Everitt (1977) — Everitt, B. S. (1977). *The Analysis of Contingency Tables*, Editora Chapman and Hall Ltd.

Fruchterman e Reingold (1991) — Fruchterman, Thomas M. J.; Reingold, Edward M. (1991). Graph Drawing by Force-directed Placement. *Software – Practice and Experience,* vol 21 (1 1), pp. 1129-1164.

Gershon (1998) — Gershon, Nahum; Eick, S. G.; Card, Stuart (1998). Information Visualization. *Interactions*, março/abril de 1998, pp. 9-15.

Google — www.google.com

Hearst (1995) — Hearst, Marti A. (1995). *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 59 - 66.

Hearst e Pedersen (1996) — Hearst, M.A.; Pedersen, J.O. (1996). Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '96), pp. 76-84.

Kartoo — www.kartoo.com

Keim (2001) — Keim, Daniel A. (2001). Visual Explorations of Large Data Sets. In *Communications of the ACM*, agosto de 2001, vol. 44, nº 8, pp. 39-44.

Kirsch (1998) — Kirsch, S.(1998). *Infoseek's Experiences Searching the Internet*, ACM SIGIR, Volume 32, número 2, pp. 3-7.

Larkin e Simon (1987)	Larkin, Jill H.; Simon, Herbert A. (1987). Why a Diagram is (Sometimes) Worth Ten Thousand Words. In *Cognitive Science*, 11 (1), pp. 65-100.
Leuski e Allan (2000)	Leuski, Anton; Allan, James (2000). Lighthouse: Showing the Way to Relevant Information. *Proceedings of IEEE Symposium on Information Visualization (InfoVis'00)*, pp. 125-130.
Lommel (1978)	Lommel, Andreas (1976). *A Arte Pré-Histórica e Primitiva*, pág. 14.
Mackinley (1986)	Mackinley, Jock D. (1986). Automating the Design of Graphical Presentations of Relational Information. *ACM Transactions on Graphics*, 5(2), pp. 110-141.
Muramatsu e Pratt (2001)	Muramatsu, J.; Pratt, W. (2001) Transparent Queries: Investigating Users' Mental Models of Search Engines, *Proceedings of SIGIR'01*, pp. 217-224.
Myers *et al.* (1996)	Myers, Brad; Hollan, Jim; Cruz, Isabel (1996). Strategic Directions id Human-Computer Interaction. *ACM Computing Surveys*, volume 28, nº 4, dezembro, pp. 794-809.
Nielsen (1993)	Nielsen, J. (1993). *Usability Engineering*. Academic Press.
Norman e Draper (1986)	Norman, D. A.; Draper, S. W. (eds) (1986), *User Centered System Design: New Perspectives on Human-Computer Interaction.* Editora Lawrence Erlbaum Associate Publishers.
Plaisant *et al.* (1996)	Plaisant, Catherine; Milash, Brett; Rose, Anne; Widoff, Seth; Shneiderman, Ben (1996). Lifelines:Visualizing Personal Histories. *Proceedings of CHI'96*, ACM Conference on Human Factors in Computing Systems, Nova York, pp. 221-227.
Prates e Barbosa (2003)	Prates, Raquel Costa; Barbosa, Simone Diniz Junqueira (2003). Avaliação de Interfaces de Usuário – Conceitos e Métodos. *Anais do XXIII Congresso da Sociedade Brasileira de Computação / XXII JAI – Livro Texto*, pp. 246-293.
Purchase (2000)	Purchase, H. C. (2000). Effective Information Visualization: a Study of Graph Drawing Aesthetics and Algorithms. *Interacting with Computers*, 13, pp. 147-162.
Robertson *et al.* (1993)	Robertson, George G.; Card, Stuart K.; Mackinlay, Jock D. (1993). Information Visualization Using 3D Interactive Animation. *Communications of the ACM*, abril de 1993, volume 36, nº 4, pp. 57-71.

Rocha e Baranauskas (2003)	Baranauskas, Maria Cecília Calani; Rocha, Heloisa Vieira (2003). Avaliação de Interfaces. *Design e Avaliação de Interfaces Humano-Computador*. Nucleo de Informática Aplicada à Educação – NIED, pp. 163-213.
Salton e Buckley (1991)	Salton, Gerard; Buckley, Chris (1991). Automatic Text Structuring and Retrieval – Experiments in Automatic Encyclopedia Searching. *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 21-30.
Shneiderman (1983)	Shneiderman, B. (1983). Direct Manipulation: a Step Beyond Programming Languages. *IEEE Computer*, 16(8), pp. 57-69.
Search Engine Showdown	www.searchengineshowdown.com/stats/size.shtml, visitado em 23/02/2004.
Sebrechts *et al.* (1999)	Sebrechts, Marc M.; Cugini, John V.; Vasilaskis, Joanna; Miller, Michael S.; Laskowski, Sharon J. (1999).Visualization of Search Results: a Comparative Evaluation of Text, 2D and 3D Interfaces. *Proceedings of ACM SIGIR*, pp. 3-10.
Silverstein *et al.* (1999)	Silverstein, C.; Henzinger, M.; Marais, H.; Moricz, M. (1999). Analysis of a Very Large Web Search Engine Query Log, *ACM SIGIR*, Volume 33, número 1, pp. 6-12.
Tukey (1977)	Tuckey, J. W. (1977). *Exploratory Data Analysis*. Editora Addison-Wesley.
Tufte (1983)	Tufte, Edward R. (1983). *The Visual Display of Quantitative Information.* Editora Graphics Press.
Tufte (2002)	Tufte, Edward R. (2002). *Visual Explanations – Images and Quantities, Evidence and Narrative.* Editora Graphics Press, 5ª edição.
Zaina e Baranauskas (2005)	Zaina, Cláudio M.; Baranauskas, M. Cecília Calani (2005). Revealing Relationships in Search Engine Results. *Proceedings of the CLIHC 2005*, em publicação.
Wise *et al.* (1995)	Wise, James A.; Thomas, James J.; Pennock, Kelly; Lantrip, David; Pottier, Marc; Schur, Anne; Crow, Vern (1995). Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents. *Proceedings of InfoVis'95, IEEE Symposium on Information Visualization*, pp. 51-58.

Apêndices

Apêndice I: Formulário Eletrônico

Pesquisa sobre Comportamento de Usuários de Motores de Busca

1 - Sexo

○ masculino ○ feminino

2 - Escolaridade (cursando no momento)

○ fundamental ○ médio ○ superior ○ pós-graduação

3 - Idade

○ até 15 anos ○ 16 a 25 ○ 26 a 40 ○ mais de 40 anos

4 - Uso de *sites* de busca:

○ não uso (e já pode clicar em "enviar")
○ cerca de 1 vez por semana
○ 3 vezes por semana
○ todo dia

5 - O *site* de busca que costumo usar é:

○ AlltheWeb ○ Altavista ○ Google ○ Outro

6 - O *site* de busca já traz na primeira página de resultados o que procuro:

○ raramente (0 a 15% das vezes) ○ freqüentemente (51% a 85%)
○ algumas vezes (16% a 50%) ○ quase sempre (86 a 100% das vezes)

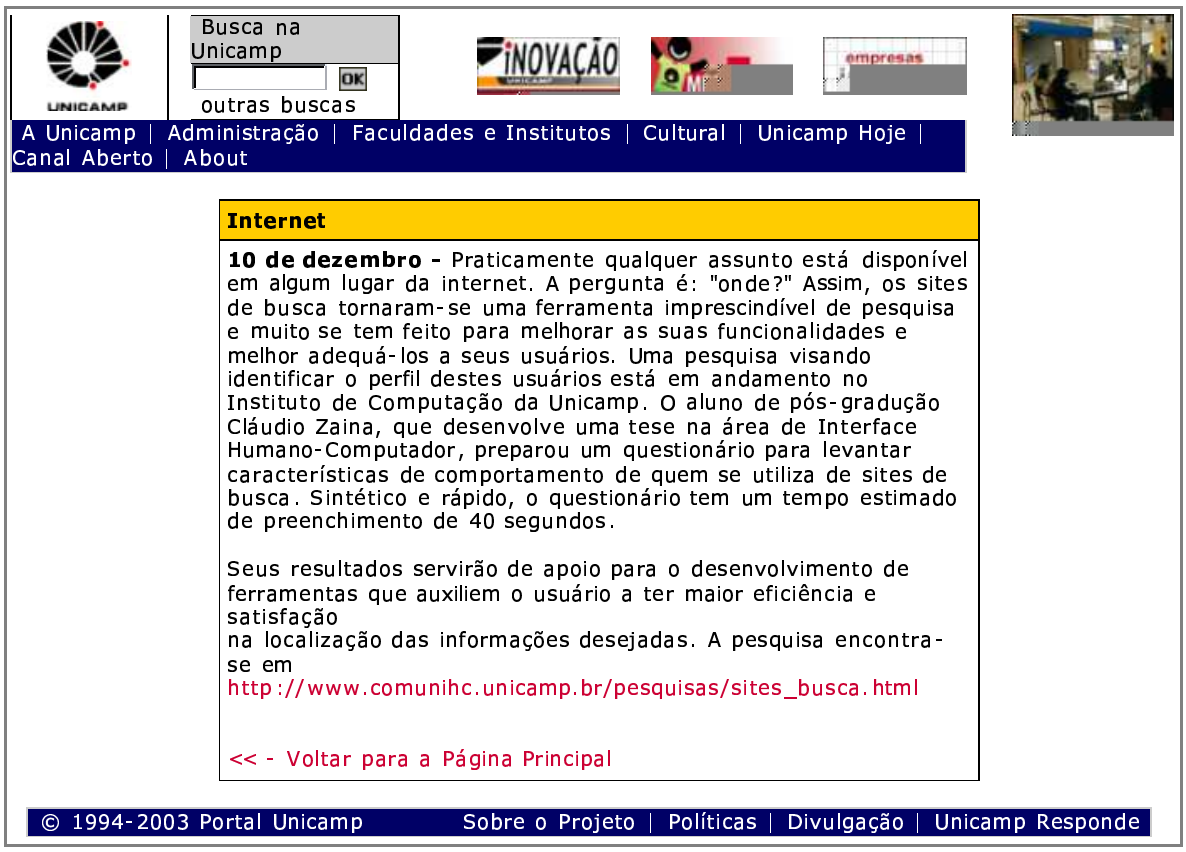
7 - Se não acho a resposta na primeira página de resultados:

○ escolho outras palavras e refaço a pesquisa
○ vou para a próxima página de resultados
○ desisto deste site de busca e vou para outro
○ desisto de achar em sites de busca o que estava procurando

E N V I A R
limpar

Obrigado!

Apêndice II: Chamada da Pesquisa no Portal da Unicamp



Apêndice III: Termo de Sigilo

Termo de Sigilo

Eu, _____, declaro para todos fins que concordo em participar do Teste de Usabilidade a ocorrer _____. Estou consciente que os resultados serão utilizados para fins de pesquisa e que minha anonimidade pessoal será sempre e de todo modo resguardada.

Data: _____ de _____ de _____

Assinatura: _____

Apêndice IV: Instruções para o Teste de Usabilidade

Teste de Usabilidade para Mecanismos de Busca

Definições

O teste propõe-se a avaliar características da interface que estejam de acordo com o entendimento, pelo usuário, das funcionalidades providas pelo sistema, assim como levantar possíveis divergências entre entre esse entendimento e as funcionalidades.

Objetivo

O objetivos do teste é identificar quais aspectos são positivos e quais são negativos nas interfaces dos dois motores de busca examinados: Google, com resultados em forma de lista e um Aplicativo, com resultados representados de modo gráfico.

Preparação

O teste ocorrerá em laboratórios do Instituto de Computação da Universidade de Campinas e será composto de três tarefas que serão executadas para cada um dos dois métodos, para cada usuário.

O computador disponível opera sistema operacional Windows XP, possui 512 Kb de memória RAM e 1200 MHz como velocidade da CPU. Estará disponível o Internet Explorer para ser utilizado como navegador.

Recursos Humanos

Os usuários serão alunos voluntários do Instituto de Computação.

O experimentador será um aluno de mestrado em Interface Humano-Computador pela Universidade de Campinas.

Tarefas

As tarefas requisitadas dos usuários são as seguintes:

Tarefa 1: Achar via motor de busca ou aplicativo, conforme requisitado, uma página que informe qual o herói grego matou a górgona Medusa.

Tarefa 2: Obter via pesquisa no motor de busca ou aplicativo, conforme solicitado, quem compôs e qual o nome da obra na qual estão os versos:

"Que é a vida? Um frenesi.
Que é a vida? Uma ilusão,
uma sombra, uma ficção;
o maior bem é tristonho,
porque toda a vida é sonho
e os sonhos, sonhos são."

Tarefa 3: Obter, através de pesquisa em motor de busca ou aplicativo, o valor do Produto Interno Bruto do Brasil em 2003.

O experimentador tentará avaliar dificuldades e concepções do usuário durante a execução das tarefas porém lhe será vedado auxiliá-lo a realizá-las.

Teste

Inicia com a leitura, concordância e preenchimento do termo de sigilo.

O teste será composto de três tarefas, repetidas para o motores de busca e para o aplicativo para cada usuário. Após a aplicação do teste será requisitado o preenchimento de um questionário abordando detalhes da interação do usuário com cada motor de busca.

Estado Inicial

O computador terá disponível e em execução um navegador, o Internet Explorer. Quando for utilizado o aplicativo, este estará em operação, aguardando a utilização pelo usuário.

O método a ser utilizado em primeiro lugar terá sido sorteado e será indicado na folha de Tarefas, que será entregue aos usuários. As instruções pertinentes ao teste serão dadas.

Esclarecimentos aos Usuários

Os esclarecimentos dados aos usuários antes do início do teste são os seguintes:

- o objetivo é avaliar o sistema e não o usuário;
- é permitido ao usuário encerrar o teste a qualquer momento, informando o motivo da desistência;
- é altamente desejado que os usuários sempre relatem seus contentamentos, descontentamentos e surpresas quanto ao funcionamento e operação do sistema;
- pede-se sigilo quanto às questões das tarefas para com os outros usuários, a fim de não influenciá-los;
- os usuários podem fazer perguntas durante o andamento do teste porém é possível que, por motivos de não invalidar o experimento, nem sempre estas poderão ser respondidas pelo examinador e
- será requisitado que os usuários relatem suas dúvidas, opiniões e considerações.

Questionário

Após o término das tarefas, será requisitado de cada usuário que preencha um questionário, respondendo a perguntas que auxiliem os experimentadores a avaliar a opinião do usuário.

As perguntas são as seguintes e referem-se sempre aos dois métodos em questão:

1 Impressões Gerais;
2 Descreva o procedimento utilizado para realizar a pesquisa no (Google / Aplicativo);
3 Como você supõe que seja construída a resposta fornecida para sua consulta (Google / Aplicativo)?
4 Quais as facilidades e dificuldades para realizar a pesquisa você reportaria para cada um dos motores de busca?

Apêndice V: Questionário do Teste de Usabilidade

Questões do Teste de Usabilidade para Mecanismos de Busca

1- Impressões Gerais

Google:

Aplicativo:

2- Descreva o procedimento utilizado para realizar a pesquisa no

Google:

Aplicativo:

3- Como você supõe que seja construída a resposta fornecida para sua consulta?

Google:

Aplicativo:

4- Quais as facilidades e dificuldades para realizar a pesquisa você reportaria para cada um dos métodos?

Google:

Aplicativo:

Apêndice VI: Estatísticas de Desenvolvimento do Sistema

A versão que foi utilizada nos testes com usuários, considerada bastante estável, foi codificada em Java[10], utilizando o paradigma de Orientação a Objetos, sobre o ambiente de desenvolvimento Eclipse[11]. O sistema foi implementado por um único desenvolvedor e é constituído por 41 classes, 415 funções e 4.194 linhas de código efetivas, ou seja, não considerando comentários. Foi feito o acompanhamento de diversas estatísticas do desenvolvimento, medidas a partir do momento em que os módulos foram reunidos para compor o sistema. O período vai de 24 de maio de 2004 até 18 de dezembro de 2004.

A Figura 6.2 mostra a evolução do desenvolvimento do programa estimada pelo seu tamanho, em número de linhas de código efetivas, ou seja, sem considerar comentários. São visíveis diversas reengenharias nas quais, por um curto período de tempo, o código foi alterado sofrendo contrações.

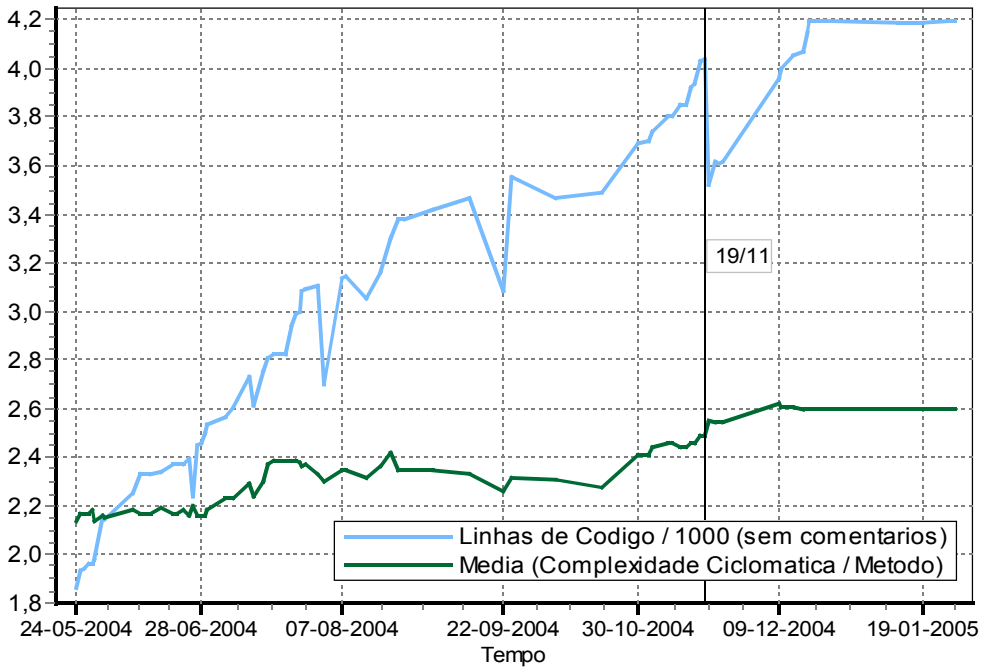


Figura 6.2: Crescimento do Sistema ao Longo do Tempo

Pode-se perceber, examinando a Figura 6.3, a ocorrência de uma reengenharia no dia 19/11/2004, quando o sistema perdeu um quarto de suas classes em cerca de 500 linhas de código, provocados por uma reestruturação dos métodos das classes. Esta reestruturação elevou a média de métodos por classe de algo em torno de 8 para cerca de 10 métodos por classe.

[10] java.sun.com

[11] www.eclipse.org

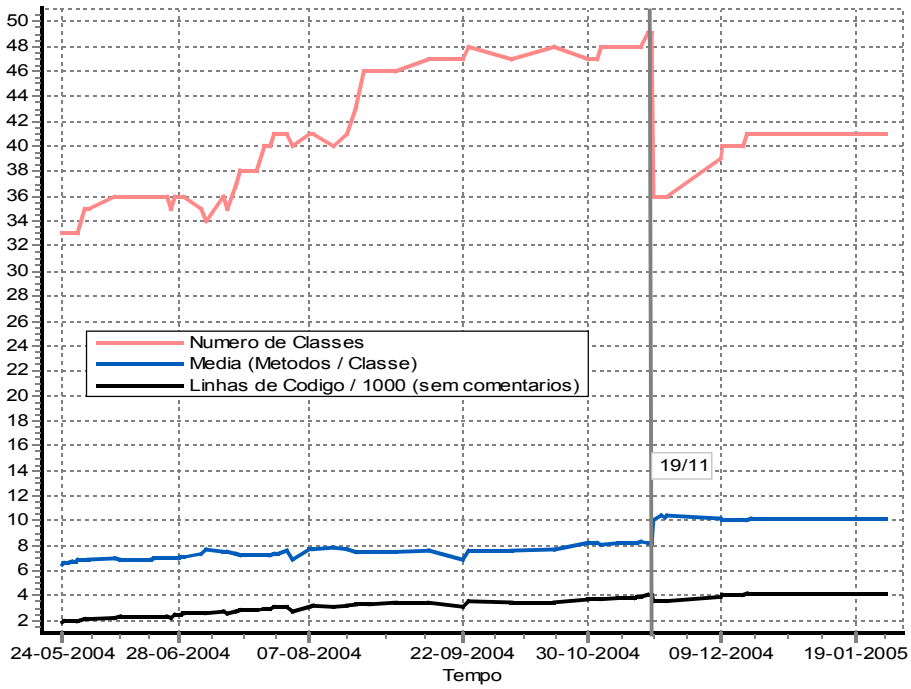


Figura 6.3: Acompanhamento de Estatísticas de Desenvolvimento