Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciências
Escola Superior de Desenho Industrial

Márcia Severo Lunardi

Visualização em nuvens de texto como apoio à busca exploratória na web

Rio de Janeiro
2008

Márcia Severo Lunardi

Visualização em nuvens de texto como apoio à busca exploratória na web

Dissertação apresentada como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Design da Universidade do Estado do Rio de Janeiro. Área de concentração: Design.

Orientador Prof. Dr. André Soares Monat

Rio de Janeiro
2008

CATALOGAÇÃO NA FONTE
UERJ / REDE SIRIUS / CTC/G

L961 Lunardi, Márcia Severo.
Visualização em nuvens de texto como apoio à busca exploratória na web / Márcia Severo Lunardi. – 2008.
112 f.

Orientador: Prof. Dr. André Soares Monat.
Dissertação (Mestrado) – Universidade do Estado do Rio de Janeiro, Escola Superior de Desenho Industrial.

Bibliografia.
1. Design da informação – Teses. 2. Interfaces gráficas de usuários – Teses. 3. Sistemas de busca – Teses. I. Monat, André Soares. II. Universidade do Estado do Rio de Janeiro. Escola Superior de Desenho Industrial. III. Título.

CDU 004.514

_____ _____

Assinatura Data

Márcia Severo Lunardi

Visualização em nuvens de texto como apoio à busca exploratória na web

Dissertação apresentada como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Design da Universidade do Estado do Rio de Janeiro. Área de concentração: Design.

Aprovada em 27 de março de 2008
Banca examinadora:

_____

Prof. Dr. André Soares Monat (Orientador)
Escola Superior de Desenho Industrial – UERJ

_____

Prof. Dr. Washington Dias Lessa
Escola Superior de Desenho Industrial – UERJ

_____

Profª. Drª. Priscila Lena Farias
Centro Universitário SENAC-SP

Rio de Janeiro
2009

AGRADECIMENTOS

A minha família maravilhosa, meus pais Telmo e Zoila, e irmãos, Marcela e Duda, pela torcida e pelo carinho em todos os momentos da minha vida.

Um agradecimento especial a minha mãe que provou que “mãe é mãe”, a única pessoa a quem eu poderia confiar um bebê recém nascido para que eu pudesse assistir às aulas com tranqüilidade. Seu apoio foi fundamental.

Ao meu sogro e mestre José, obrigada pelas dicas e pela revisão atenta desse texto, e aos meus cunhados, Victor e Dolfo, que tanto admiro pela dedicação ao mundo acadêmico, três grandes exemplos. Obrigada também por terem sido ótimas babás.

A minha filha Vitória que me acompanhou durante todo o mestrado, primeiro às aulas, ainda na da minha barriga e depois servindo de incentivo para eu prosseguir. Espero do fundo do coração que isso um dia seja motivo de orgulho para ela.

Ao meu marido Zeca pelo entusiasmo com que compartilhou todos os momentos dessa pesquisa, pela força nos momentos de angústia e pela paciência aos meus questionamentos intermináveis, sempre com muito amor e carinho. Tanto a sua admiração por mim como a que sinto por você é sempre um grande incentivo.

Ao Zeca, de novo, que permitiu que eu pudesse conceber qualquer sistema livremente garantindo a todo o momento que implementaria o que quer que eu imaginasse. E foi o que fez de forma brilhante ao desenvolver o sistema idealizado nessa dissertação.

Ao meu orientador André Monat, por sempre me incentivar a buscar o novo, o estado da arte, sem medo de arriscar.

Um agradecimento especial ao professor Washington Dias Lessa pela forma instigante com que apresenta as questões do design e suas fronteiras. As discussões de sua aula foram enriquecedoras.

Aos meus colegas do mestrado, todos brilhantes, cada um ao seu modo, que tornaram nossos encontros sempre tão agradáveis. Foram momentos inesquecíveis.

RESUMO

LUNARDI, Márcia Severo. Visualização em nuvens de texto com apoio à busca exploratória na web. 2008. 112 f. Dissertação (Mestrado em Design) - Escola Superior de Desenho Industrial, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2008.

A presente dissertação é o resultado de uma pesquisa que avalia as vantagens da utilização de nuvens de texto para apresentar os resultados de um sistema de busca na web. Uma nuvem de texto é uma técnica de visualização de informações textuais e tem como principal objetivo proporcionar um resumo de um ou mais conteúdos em uma única tela. Em uma consulta na web, os resultados aparecem listados em diversas páginas. Através de uma nuvem de texto integrada a um sistema de busca é possível a visualização de uma síntese, de um resumo automático, do conteúdo dos resultados listados em várias páginas sem que elas tenham que ser percorridas e os sites acessados individualmente. A nuvem de texto nesse contexto funciona como uma ferramenta auxiliar para que o usuário possa gerenciar a grande carga de informação que é disponibilizada nos resultados das consultas. Dessa forma os resultados podem ser vistos em contexto e, ainda, as palavras que compõem a nuvem, podem ser utilizadas como palavras-chave adicionais para complementar uma consulta inicial. Essa pesquisa foi desenvolvida em duas fases. A primeira consistiu no desenvolvimento de uma aplicação integrada a um sistema de buscas para mostrar seus resultados em nuvens de texto. A segunda fase foi a avaliação dessa aplicação, focada principalmente em buscas exploratórias, que são aquelas em que os objetivos dos usuários não são bem definidos ou o conhecimento sobre o assunto pesquisado é vago.

Palavras-chave: Nuvens de Texto. Visualização de Informações. Design da Informação. Design de Interface. Sistemas de Busca. Busca Exploratória. Recuperação de Informações.

ABSTRACT

This dissertation presents the results of a research that evaluates the advantages of text clouds to the visualization of web search results. A text cloud is a visualization technique for texts and textual data in general. Its main purpose is to enhance comprehension of a large body of text by summarizing it automatically and is generally applied for managing information overload. While continual improvements in search technology have made it possible to quickly find relevant information on the web, few search engines do anything to organize or to summarize the contents of such responses beyond ranking the items in a list. In exploratory searches, users may be forced to scroll through many pages to identify the information they seek and are generally not provided with any way to visualize the totality of the results returned. This research is divided in two parts. Part one describes the development of an application that generates text clouds for the summarization of search results from the standard result list provided by the Yahoo search engine. The second part describes the evaluation of this application. Adapted to this specific context, a text cloud is generated from the text of the first sites returned by the search engine according to its relevance algorithms. The benefit of this application is that it enables users to obtain a visual overview of the main results at once. From this overview the users can obtain keywords to navigate to potential relevant subjects that otherwise would be hidden deep down in the response list. Also, users can realize by visualizing the results in context that his initial query term was not the best choice.

Keywords: Text Clouds. Information Visualization. Information Design. Interface Design. Web Search. Exploratory Search. Information Retrieval.

Lista de Figuras

Lista de Tabelas

Sumário

1. Introdução

A presente dissertação é o resultado de uma pesquisa que avalia as vantagens da utilização de uma forma de visualização de informações para apresentar os resultados de um sistema de busca na *web*.

A pesquisa foi desenvolvida em duas fases. A primeira fase consistiu no desenvolvimento de uma aplicação integrada a um sistema de buscas para mostrar seus resultados. A segunda fase foi a avaliação dessa aplicação com a finalidade de testar a hipótese dessa pesquisa.

A forma de visualização escolhida para essa avaliação foi uma nuvem de texto, mais conhecida por seu nome original em inglês, *text cloud* ou ainda, *tag cloud*. Apesar de ter se tornado bastante popular recentemente na *web*, essa forma de visualização ainda não apresenta estudos sobre a sua utilização na visualização de resultados de sistemas de busca.

Com a internet presente, cada vez mais, na vida das pessoas, atividades cotidianas são planejadas e decisões são tomadas, muitas vezes baseadas em consultas à *web*. Para realizar essas consultas as pessoas utilizam os sistemas de busca que estão entre as páginas mais acessadas hoje em dia. São eles que permitem que os usuários encontrem as informações que procuram em meio a uma infinidade de documentos. No entanto, com o crescimento exponencial de informações disponibilizadas na rede, as consultas apresentam cada vez mais resultados. Os resultados aparecem listados de forma textual e sequencialmente em inúmeras páginas, exigindo dos usuários um esforço extra no sentido de refinar e localizar o que procuram.

A área de buscas na *web* tem sido objeto de pesquisa sob vários enfoques, com contribuições de diversos campos do saber. As pesquisas têm-se dado principalmente no âmbito da Interação Humano-Computador, Ciência da Computação, Recuperação de Informações, Mineração de Dados e Design de Interfaces.

Ainda quanto à questão da abundância de informações na *web*, Meirelles e Moura (2007) ressaltam a posição de Bonsiepe quanto à importância do design como mediador entre usuário e informação, no sentido de viabilizar a aquisição de conhecimento e o seu compartilhamento:

> ...“conhecimento como experiência acumulada precisa ser comunicado e compartilhado entre indivíduos. O processo de comunicar e compartilhar está ligado a apresentação do conhecimento e essa é, ou pode vir a ser, um assunto do Design.” (Bonsiepe, 2000 apud Meirelles e Moura, 2007)

1.1. O problema da pesquisa

Apesar de os sistemas de busca oferecerem recursos de pesquisa avançada que possibilitam um refinamento das consultas, pesquisas demonstram que os usuários além de não utilizarem esses recursos usam em média apenas duas ou três palavras-chave por consulta, e na maioria das vezes não olham além da primeira página de resultados.

Os principais estudos sobre o comportamento dos usuários utilizando sistemas de busca foram conduzidos por Silverstein et al. (1999), Hoscher e Strube (2000), Jansen et al. (2000, 2005), Wolfram et al. (2001) e Spink et al. (2001, 2002).

Os principais sistemas de busca adotam algoritmos próprios para localizar e listar os resultados por ordem de relevância. Esses algoritmos vêm sendo constantemente aprimorados e atendem bem ao seu propósito quando os usuários têm objetivos bem definidos (Bates, M. J., 1989).

No entanto, quando os usuários têm conhecimentos reduzidos sobre o assunto da consulta, objetivos pouco definidos ou muito complexos, eles não sabem que palavra-chave utilizar. Geralmente, para contornar esse problema adotam estratégias que envolvem utilizar palavras-chave genéricas como uma tentativa de se obter nos resultados palavras-chave mais específicas para uma nova consulta (Baldonado, M. Q. W. e Winograd, T., 1997; Bates, M. J., 1989; e, Pirolli, P. e Card, S., 1995).

Esse tipo de estratégia utilizada pelos usuários demanda tempo e esforço cognitivo para se chegar ao objetivo pretendido e tem recebido especial atenção por parte da comunidade científica. Foco de diversos estudos, esse comportamento é classificado como busca exploratória (Marchioni, 2006).

Análises recentes sobre o objetivo das consultas sugerem que cerca de 20 a 30% de todas as consultas realizadas na *web* são exploratórias por natureza (Rose & Levinson, 2004).

Diante da demanda por sistemas de busca que atendam melhor os usuários nessas situações, novas propostas vem sendo apresentadas. Uma das principais linhas de estudos tem se concentrado na categorização automática dos resultados.

A partir do agrupamento de temas a categorização de documentos tem-se mostrado vantajosa pelo fato de proporcionar a visualização de um conjunto maior de resultados por

página e também por mostrar esses resultados em contexto. (Dumais, Cutrell & Chen, 2001).

A partir da constatação da importância do contexto na localização de resultados de busca surgiram recentemente propostas de sistemas de busca, ainda em suas versões beta, que utilizam técnicas de visualização de informações para mostrar graficamente os seus resultados.

Técnicas de visualização de informações atendem ao objetivo de aumentar a compreensão dos resultados comunicando informações contextuais através de variações na forma como os dados são mostrados e também apresentam soluções para a disposição de grandes quantidades de resultados de busca em uma única tela.

Um exemplo dessa nova geração de sistemas de busca que apresenta uma visualização gráfica de seus resultados é o Grokker[1]. Em uma pesquisa conduzida por Rivadeneira e Bederson (2003) não foram constatadas diferenças na comparação entre as duas interfaces, a gráfica e a textual desse sistema, em termos de eficiência. Entretanto, o nível de satisfação e aceitação foi mais alto com relação à interface gráfica.

Apesar dos experimentos nessa área se encontrarem ainda em estágios iniciais, essas novas propostas apontam para um novo caminho a ser explorado, o da visualização de resultados de sistemas de busca utilizando outras técnicas e sob diferentes abordagens.

1.2. Motivação

De forma resumida, a motivação para a realização da pesquisa aqui relatada surgiu a partir da soma das seguintes constatações:

Os sistemas de busca têm importância crescente na localização de informações na *web*.

Os usuários de sistemas de busca utilizam apenas duas ou três palavras-chave por consulta, e na maioria das vezes não olham além da primeira página de resultados.

Os sistemas de busca não atendem bem nos casos de busca exploratória, sendo que estas correspondem a até um terço de todas as consultas realizadas na *web*.

Existe uma nova geração de sistemas de busca que utilizam técnicas de visualização para mostrar seus resultados, porém de forma ainda bastante embrionária.

[1] http://www. Grokker.com. Esse sistema de buscas conta com a colaboração de Ben Shneiderman. Ben Shneiderman é diretor e fundador do Laboratório de Interação Humano-Computador e Membro do Instituto de Estudos Avançados de Computação da Universidade de Maryland sendo uma das principais referências da área de visualização de informações.

Ainda existe espaço para a experimentação de novas formas de visualização de resultados de sistema de busca.

Se, por um lado, os resultados das pesquisas já empreendidas nessa área fornecem subsídios para a investigação proposta nessa dissertação, por outro, evidenciam a ausência de estudos cujo enfoque principal está no design gráfico e da informação.

1.3. O objetivo

O objetivo geral proposto por essa pesquisa foi testar uma nova forma de visualização de informações para mostrar os resultados de sistemas de busca. A forma de visualização escolhida foi a nuvem de texto, mais conhecida como *text cloud*, seu nome original em inglês, ou ainda *tag cloud*.

A área de visualização de informações estuda formas e técnicas de representação visual para revelar tendências e relações entre dados estatísticos levando a percepção de uma nova dimensão da informação. Essa é uma área de estudos multidisciplinar influenciada pelas ciências cognitivas, ciência da computação, matemática, estatística, design, entre outras (Ware, 2000).

A nuvem de texto é uma forma de visualização de informações que se tornou bastante popular na *web* nos últimos anos. No entanto, ainda não existem estudos sobre a sua utilização na visualização de resultados de sistemas de busca abertos.

Como seu próprio nome indica, a nuvem de texto é uma forma de visualização de dados lingüísticos. Basicamente uma nuvem de texto mostra a freqüência com que as palavras aparecem em um determinado texto e tem como objetivo principal proporcionar uma compreensão rápida, uma espécie de resumo desse conteúdo.

Além de mostrar as palavras mais freqüentes de um texto, uma nuvem mostra essas palavras em tamanhos variados. As palavras com fontes maiores são as mais freqüentes e as com fontes menores as menos freqüentes segundo uma escala. Uma nuvem de texto é uma lista hierarquizada visualmente e sua abordagem é puramente estatística.

O benefício que uma nuvem de texto oferece ao ser integrada a um sistema de busca é que ela proporciona uma espécie de resumo dos resultados em uma única tela. Graças a sua capacidade de mostrar uma grande densidade de palavras, uma nuvem é bastante adequada para representar grandes quantidades de texto, que é o caso dos resultados de uma busca na *web*.

Os dados extraídos dos textos dos resultados e comunicados por uma nuvem nada mais são do que uma listagem de palavras e o número total de vezes que ela aparece. No entanto, uma nuvem de textos tem características que permitem que seja visualizada uma dimensão da informação que não poderia ser percebida nos resultados paginados. A nuvem permite que o contexto dos resultados seja visualizado.

Encurtar o caminho entre o buscar e o achar é o grande desafio dos sistemas de busca e diversas áreas de pesquisa vêm convergindo esforços no sentido de pensar novas formas para que a tarefa de localizar uma informação na *web* seja mais bem sucedida.

Esse estudo está inserido na linha de pesquisa denominada **"***Design e Tecnologia"* que abrange as áreas de design e informática; hipertexto e navegação inteligente; e estrutura da informação. Faz parte do Programa de Pós-Graduação em Design da Escola Superior de Desenho Industrial, ESDI, da Universidade do Estado do Rio de Janeiro, UERJ.

Foi adotada nesse trabalho uma postura que reconhece a importância do design também na concepção de projetos tidos como multidisciplinares, principalmente nas áreas de tecnologias emergentes.

Essa postura está de acordo com as considerações de Meirelles e Moura (2007) a respeito da *web* e de seus novos paradigmas projetuais e informacionais:

> "Atuar no campo do Design frente às mídias digitais e interativas exigirá do profissional agir muito mais como um mediador dos processos de comunicação entre usuários-sistema e usuários-usuários. Esse fato requer pensar amplamente nas possibilidades de uso de equipamentos digitais e suas interfaces, na forma como a informação é distribuída e também na possibilidade de aplicação da linguagem hipermidiática e dos elementos projetuais na oferta e facilitação de interação para o usuário, nas conseqüências sociais decorrentes da mudança de atitude dele e – por que não? – no desenvolvimento de projetos que contemplem novos modelos de negócios emergentes desse contexto.
>
> Faz-se necessário, então, pensar os projetos considerando a imprevisibilidade do uso e a fluidez da informação, a favor da conectividade de idéias e da construção de conhecimento. O design da informação será, portanto, cada vez mais fundamental e indispensável no processo de construção e novos paradigmas de interação."

1.4. A hipótese e avaliações

De acordo com Santos (2002), a hipótese é "uma verdade provisória fundamental para qualquer processo de investigação científica, pois consiste no lançamento de uma afirmação a respeito de algo ainda desconhecido ou, pelo menos, não satisfatoriamente conhecido". Marconi e Lakatos (2000) também apresentam que a hipótese constitui-se em uma "suposta, provável e provisória resposta a um problema, cuja adequação (comprovação = sustentabilidade ou validez) será verificada através da pesquisa".

A hipótese dessa investigação é que:

A visualização dos resultados de um sistema de busca em uma nuvem de texto pode auxiliar os usuários a encontrar o que procuram facilitando a construção de consultas em buscas exploratórias.

Para testar essa hipótese foi construída uma aplicação integrada ao sistema de busca Yahoo[2] que funciona da seguinte forma:

1. Quando um usuário digita uma ou mais palavras no campo de busca do sistema Yahoo, e pressiona a barra de espaço, a aplicação cria automaticamente uma nuvem de texto a partir do conteúdo dos 40 primeiros resultados que seriam listados para aquela consulta.

2. Em seguida, apresenta essa nuvem de texto dos resultados na interface do Yahoo, e

3. Permite que essas palavras sejam acrescentadas à consulta inicial do usuário para que este então, possa submeter a consulta reformulada ao sistema.

A partir da construção dessa aplicação foi possível realizar duas formas de avaliação a partir de métodos empíricos. Primeiro foi feita uma avaliação cooperativa da aplicação. Por se tratar de um conceito novo, essa avaliação inicial foi usada para registrar a compreensão do aplicativo da nuvem de texto pelos participantes. Também atendeu ao objetivo de familiarizar esse participante com o aplicativo para o experimento controlado.

Depois foi conduzido um experimento controlado que comparou as mesmas consultas sendo realizadas no Yahoo padrão, e no Yahoo com o auxílio da nuvem de texto.

O experimento controlado avaliou especificamente:

Se os usuários utilizaram palavras da nuvem de texto para refinar suas consultas, diminuindo assim o esforço cognitivo empregado.

E nos casos positivos, se as buscas foram concluídas de forma mais rápida e satisfatória com o auxílio da nuvem.

1.5. A estrutura da dissertação

A presente dissertação está estruturada em 8 capítulos, os quais apresentam o referencial teórico pertinente à pesquisa, a metodologia utilizada e, finalmente, as conclusões. Para que se possa ter uma compreensão geral desse trabalho, a seguir, é apresentado, através de um breve resumo, o conteúdo de cada capítulo.

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

Nesse capítulo foram apresentados os conteúdos introdutórios e os principais elementos da pesquisa: o problema, a motivação, o objetivo a hipótese e o experimento realizado. Nesse capítulo, também será apresentada, a seguir, a estrutura da dissertação.

Capítulo 2 - Nuvens de texto: uma nova forma de visualização de informações na *web*

Nesse capítulo as nuvens de texto são introduzidas. Como essa é uma forma de visualização de informações que surgiu apenas recentemente na *web* é explicado em detalhes como ela é construída, sua forma de apresentação, principais características e aplicações correntes. Também é feita uma diferenciação entre as nuvens de texto e as nuvens de *tags*, similares em alguns aspectos.

São abordados ainda, aspectos gerais sobre o tema visualização de informações e sobre a visualização de dados lingüísticos especificamente.

Capítulo 3 - Análise das nuvens de texto sob a perspectiva do design

Nesse capítulo são abordados aspectos relacionados ao design das nuvens de texto que como qualquer outra forma de representação visual pode ser refinada para comunicar com clareza, precisão e eficiência idéias complexas.

É feita uma análise das nuvens de texto sob a perspectiva do design segundo Tufte (2001), um dos mais importantes especialistas em infografia da atualidade, e Bertin (1986), o primeiro estudioso a propor uma base teórica consistente para o que hoje chamamos de visualização de informações. Além da análise sob a ótica desses dois autores, rebatendo as Leis da Gestalt sobre uma nuvem de texto, foi feita uma leitura visual da sua forma.

Capítulo 4 - Sobre os sistemas de busca

Essa pesquisa propõe uma nova aplicação para uma nuvem de texto com a sua integração a um sistema de busca na *web*. Logo, nesse capítulo é feita uma apresentação e uma contextualização desse objeto de estudo. É explicado o que é um sistema de busca e a sua importância. Em seguida é apresentado um panorama atual do mercado mostrando os principais sistemas comerciais e o número de páginas indexadas por eles. E, finalmente, é explicado como os sistemas de busca na *web* funcionam: como é feita a análise e a indexação das páginas, o armazenamento das páginas indexadas e a recuperação das páginas por ocasião da consulta pelos usuários.

Capítulo 5 - Comportamento de interação dos usuários com sistemas de busca na *web*

Nesse capítulo são apresentadas as principais pesquisas que foram realizadas sobre o comportamento de interação dos usuários com sistemas de busca cujos resultados reforçam e justificam a necessidade de propostas como a apresentada por essa dissertação.

Primeiro são apresentados os principais estudos quantitativos realizados nessa área. Esses estudos analisaram os registros das consultas realizadas pelos usuários de quatro grandes sistemas de busca ao longo de um determinado período.

Em seguida são apresentados estudos que identificaram as estratégias de busca adotadas pelos usuários quando eles não têm objetivos definidos ou não sabem que palavras-chave utilizar. Esse comportamento é conhecido como comportamento de busca exploratória.

Além desse comportamento, esse capítulo aborda a importância do contexto nos resultados de busca na *web.* Foi identificado que quando os usuários iniciam uma busca exploratória a falta de uma visão geral dos resultados é particularmente problemática.

Finalmente, são apresentados alguns sistemas de busca na *web* que de forma ainda experimental integram algum tipo de visualização gráfica para mostrar seus resultados.

Capítulo 6 - A aplicação proposta

Uma aplicação foi desenvolvida para testar a hipótese defendida nessa dissertação. A nuvem de texto nessa aplicação tem a finalidade de apresentar um resumo dos principais resultados retornados por um sistema de busca para uma dada consulta. Esse capítulo descreve em detalhes essa aplicação.

Primeiro é explicada a teoria por traz da construção da nuvem de texto nessa aplicação e, também, é explicado o funcionamento e a forma de interação com a mesma. Em seguida são abordados os aspectos relacionados à construção gráfica da nuvem de texto.

Finalmente, como o foco principal dessa dissertação é o design da informação e não o desenvolvimento do sistema, as etapas do desenvolvimento da aplicação são descritas de forma breve, apenas para que se tenha um entendimento geral do que foi feito e das tecnologias e *softwares* utilizados.

Capítulo 7 - Delineamento da pesquisa

Esse capítulo apresenta o delineamento da pesquisa. Seu tema, problema, hipótese e metodologia.

A pesquisa relatada nessa dissertação foi desenvolvida em duas fases. A primeira fase consistiu na construção da aplicação, que gera uma nuvem de texto a partir dos resultados do sistema de busca Yahoo, relatada no capítulo 6.

A segunda fase da pesquisa foi a avaliação da aplicação com a finalidade de testar a hipótese dessa dissertação.

Para testar a aplicação desenvolvida foram adotadas duas formas de avaliação. Primeiro foi feita uma avaliação cooperativa, e posteriormente foi realizado um experimento controlado que comparou o desempenho do sistema Yahoo padrão, com a aplicação proposta integrada a esse mesmo sistema.

Capítulo 8 - Conclusões e futuros trabalhos

O oitavo capítulo dessa dissertação apresenta os resultados da pesquisa, suas conclusões gerais e contribuições para o meio acadêmico, especialmente para o campo do Design. Também são sugeridos desdobramentos futuros para o aprimoramento da utilização de nuvens de texto em sistemas de busca, e possíveis variações dessa aplicação.

Bibliografia: Listagem com as referências bibliográficas utilizadas neste trabalho.

Anexos: Formulários e material utilizados na fase de aplicação da pesquisa.

2. Nuvens de texto: uma nova forma de visualização de informações na *web*

2.1. Diferença entre uma nuvem de texto e uma nuvem de *tags*

A nuvem de texto (*text cloud*) é uma forma de visualização de informações que mostra a freqüência com que as palavras aparecem em um determinado texto. Recentemente ela se tornou bastante popular na *web* e o seu sucesso se deve principalmente ao grande uso em redes sociais[3].

Geralmente, nas redes sociais, as nuvens são compostas por *tags*. *Tags* são palavras-chave atribuídas pelos usuários para classificar determinadas páginas na *web*. Logo, as nuvens de *tags* (*tag clouds*), como são conhecidas, envolvem taxonomias criadas e compartilhadas pelos usuários.

Tradicionalmente, a classificação de conteúdos é baseada em um vocabulário controlado e em categorias estruturadas hierarquicamente em uma taxonomia[4]. Taxonomias grandes e consistentes têm que ser mantidas e controladas por especialistas.

O uso de *tags* segue uma abordagem completamente diferente com palavras escolhidas livremente pelos usuários. Quando esse processo de categorização é feito de forma colaborativa por diferentes usuários, ele é conhecido como "*folsksonomy*", termo originalmente cunhado pelo arquiteto de informação Thomas Vander Wal, que é um neologismo resultado da combinação dos termos em inglês *folks* + *taxonomy*, algo como taxonomia do povo, das pessoas (Quintarelli, 2005). Um exemplo de serviço baseado em redes sociais que utiliza *tag clouds* de forma colaborativa é o *site* Del.icio.us[5]. O Del.icio.us foi lançado em 2003 pelo desenvolvedor Joshua Schachter e foi o primeiro a utilizar *tags*.

[3] Redes sociais são grupos de pessoas que tem interesses e objetivos em comuns a serem compartilhados. Na *web* elas utilizam como ponto de encontro e canal de comunicação *sites*, fóruns de discussão ou listas de e-mails.

[4] Uma taxonomia, assim como um vocabulário controlado, é um sistema de classificação hierárquica que define os relacionamentos entre termos. Esses relacionamentos podem ser semânticos e também conceituais. Taxonomias e vocabulários controlados têm como objetivo tornar a linguagem menos ambígua conectando conceitos e capturando a relação entre objetos do mundo real (Smith, 2007).

[5] http://www.delicious.com. O Delicious é um gerenciador de *sites* favoritos que é mantido e atualizado via *web* pelos seus usuários. Ele também permite o compartilhamento e a classificação dos favoritos através de *tags* que permitem a interação com outros usuários do serviço criando uma comunidade. O Delicious se auto define como um sistema de "favoritos sociais" (*social bookmarks*).

Lançou também outras tendências que se tornaria mais tarde a base para a web 2.0, como o acesso fácil a sua base de dados e serviços, e a extensão de tecnologias existentes com componentes sociais (Smith, 2007).

O Pew Internet & American Life Project[6] publicou uma pesquisa em janeiro de 2007 sobre o hábito de se atribuir *tags*, mais conhecido como *tagging*, que indica que 28% dos usuários da internet nos E.U.A. já categorizaram conteúdos *online* como fotos, textos ou postagens em *blogs* (Lee, 2007).

As nuvens de texto, diferentemente das nuvens de *tags*, são compostas, basicamente, como seu próprio nome indica, por textos e tem como objetivo principal proporcionar uma compreensão rápida do conteúdo de um determinado texto ou conjunto de textos, a partir de um resumo dado pelas palavras que aparecem com mais freqüência. Seja a primeira página de um jornal *online* com seus títulos, chamadas e matérias, um discurso político, uma música, um livro ou um poema.

Apesar de existir essa diferença entre as nuvens de texto e as nuvens de *tags* é importante registrar que muitas vezes ambas são chamadas genericamente de *tag clouds*.

Uma nuvem de texto também pode ser usada como um sistema de acesso e navegação na *web* da mesma forma que as *tag clouds*.

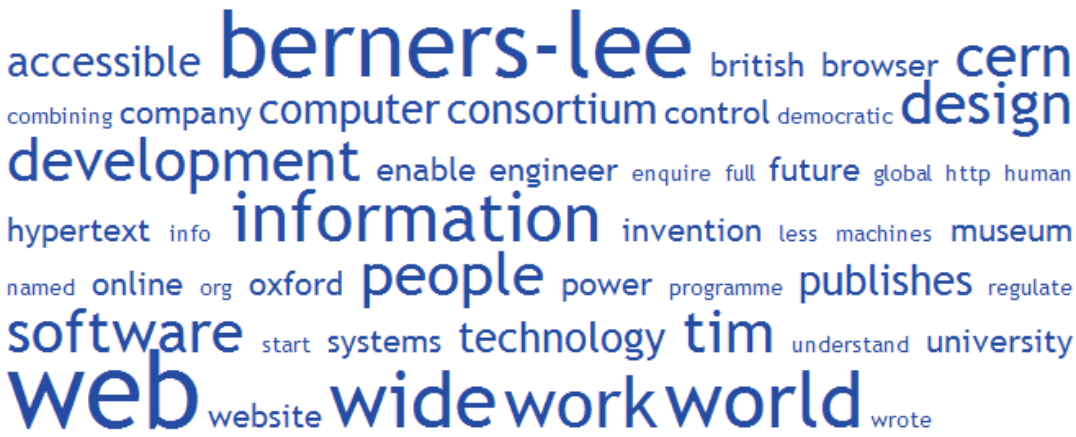


Figura 1 - Nuvem de texto gerada a partir do texto sobre Tim Berners-Lee extraído do *site Design Museum London*. (www.designmuseum.org)

[6] http://www.pewinternet.org/

Figura 2 – *Tag cloud* com termos relacionados a Web 2.0. (www.wikipedia.org/wiki/Tag_cloud)

Na visualização clássica de uma nuvem as palavras são dispostas seqüencialmente em ordem alfabética, e apresentam tamanhos variados de uma mesma fonte tipográfica, diretamente proporcionais ao número de vezes que aparecem no texto. Palavras de fontes maiores indicam que elas aparecem mais vezes no texto, e de fontes menores menos vezes. Outras formas de apresentação de uma nuvem são: por ordem de freqüência das palavras nas quais as palavras de fontes maiores aparecem primeiro e a ordem espacial onde as palavras são arrumadas de modo a preencher todos os espaços a partir do algoritmo de Feinberg (Rivadeneira et al. 2007).

Figura 3 - (a) ordem alfabética, (b) ordem por freqüência e (c) ordem espacial. (Rivadeneira et al. 2007)

Outros exemplos de utilização das nuvens de texto foram apresentados pelo jornal *The New York Times*[7] *online*, pelo *site* Globo. com[8] e pelo *site* especializado em pesquisas políticas Pollster.com[9].

O jornal *The New York Times* mostra as nuvens das palavras mais procuradas pelos seus leitores nas últimas vinte e quatro horas, na última semana e no último mês. Nesse caso, o tamanho da fonte das palavras é proporcional ao número de vezes em que essa palavra foi consultada.

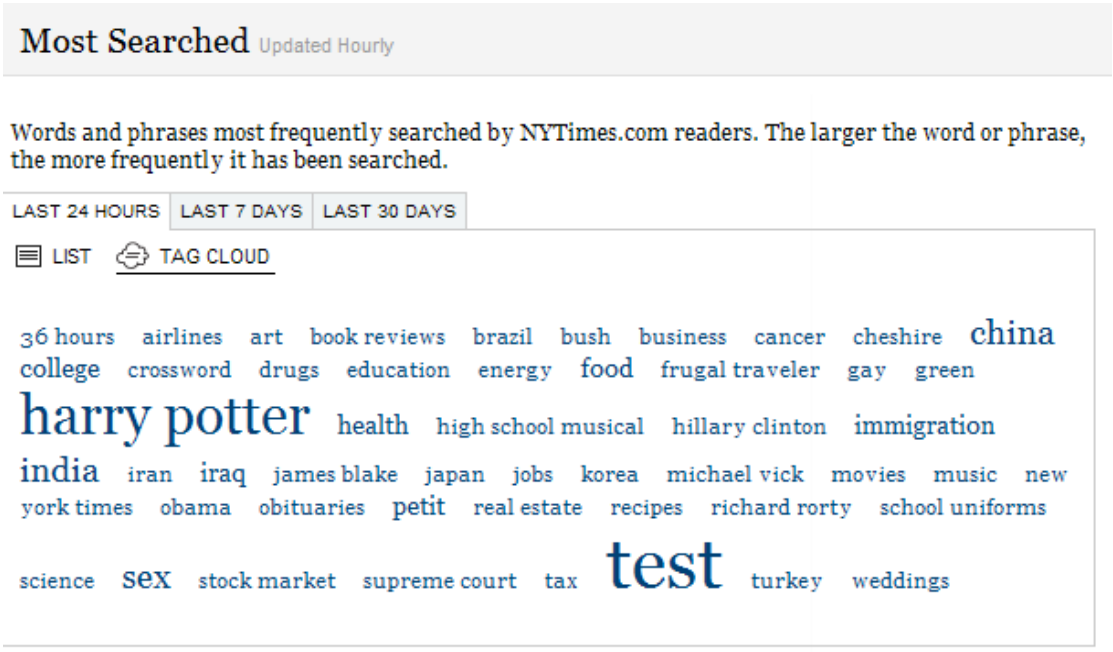
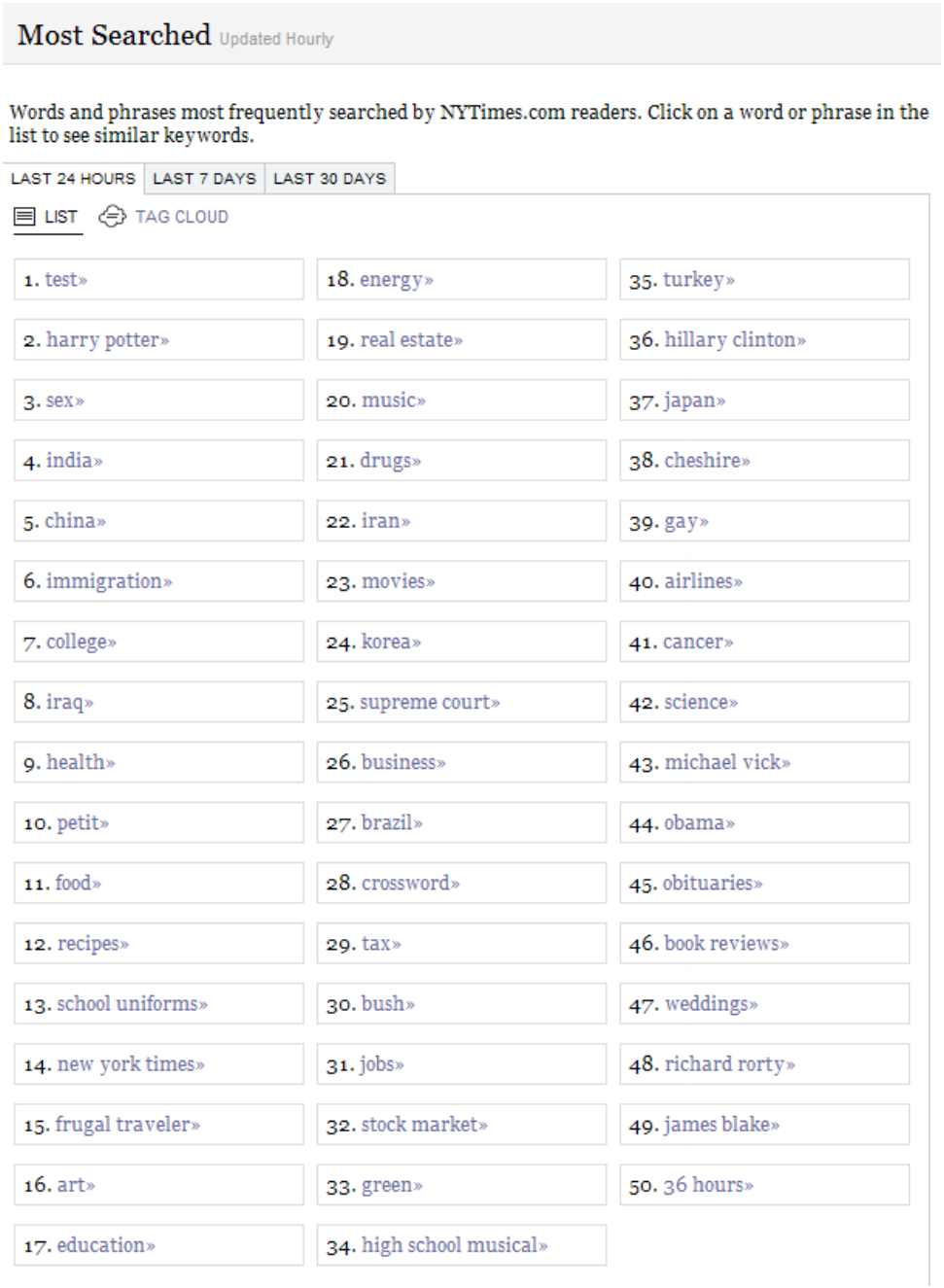


Figura 4 - Nuvem das palavras mais freqüentes que foram pesquisadas na versão *online* do jornal *The New York Times* no dia 28/07/2007. (www.nytimes.com)

[7] http://www.nytimes.com/

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

[9] http://pollster.com/blogs/tag_clouds_for_the_democratic.php



Most Searched Updated Hourly

Words and phrases most frequently searched by NYTimes.com readers. Click on a word or phrase in the list to see similar keywords.

LAST 24 HOURS | LAST 7 DAYS | LAST 30 DAYS

LIST | TAG CLOUD

1. test»	18. energy»	35. turkey»
2. harry potter»	19. real estate»	36. hillary clinton»
3. sex»	20. music»	37. japan»
4. india»	21. drugs»	38. cheshire»
5. china»	22. iran»	39. gay»
6. immigration»	23. movies»	40. airlines»
7. college»	24. korea»	41. cancer»
8. iraq»	25. supreme court»	42. science»
9. health»	26. business»	43. michael vick»
10. petit»	27. brazil»	44. obama»
11. food»	28. crossword»	45. obituaries»
12. recipes»	29. tax»	46. book reviews»
13. school uniforms»	30. bush»	47. weddings»
14. new york times»	31. jobs»	48. richard rorty»
15. frugal traveler»	32. stock market»	49. james blake»
16. art»	33. green»	50. 36 hours»
17. education»	34. high school musical»	

Figura 5 - Lista das palavras mais freqüentes que foram pesquisadas na versão *online* do jornal *The New York Times* no dia 28/07/2007. (www.nytimes.com)

Seguindo a idéia do *The New York Times*, o portal Globo.com no dia 30 de agosto de 2007 estreou o seu novo *layout*, e nesta nova versão também passou a utilizar uma nuvem de texto na página principal mostrando as palavras mais buscadas pelos leitores.

Figura 6 - Primeira página do portal Globo.com no dia do lançamento do seu novo *layout*, 30/08/2007. Destaque para a nuvem de texto dos assuntos mais buscados, utilizada pela primeira vez.

(www.globo.com)

Recentemente, o *site* Pollster.com publicou um conjunto de nuvens, cada uma delas representando o discurso de um dos candidatos à eleição presidencial do partido democrata americano. As nuvens foram geradas a partir dos discursos transcritos de um debate ocorrido entre os candidatos em 27/04/2007.

Nesse caso elas foram utilizadas para permitir uma visualização comparativa e interpretativa dos discursos mostrando uma visão geral do conteúdo abordado por cada candidato.

Figura 7 - Nuvens de textos dos discursos dos candidatos à eleição presidencial do partido democrata americano. (www.pollster.com)

As nuvens de texto se tornaram tão populares que recentemente elas serviram de inspiração para a primeira página da edição impressa do jornal *The Vancouver Sun*.


THE VANCOUVER SUN

FINAL EDITION www.vancouversun.com SERIOUSLY WESTCOAST SINCE 1912 WEDNESDAY, JANUARY 3, 2007

B.C. PROPERTY VALUES UP 23%

The bigger the letters, the larger the growth in property values for 2006. Anmore was the largest, Whistler was the smallest. Full chart on A2.

Anmore Belcarra Bowen Island Burnaby
Cariboo Central Saanich Central Vancouver Island
Coquitlam Courtenay East Kootenay Esquimalt
Fort St. John Fraser Valley Gibsons Golden
Gulf Islands Kamloops Kelowna Nelson/Trail
New Westminster North Fraser North Vancouver (city)
North Vancouver (district) Oak Bay Parksville Peace River
Penticton Port Coquitlam Port Moody
Powell River Prince George Revelstoke
Richmond/Delta Saanich Sidney Sooke Squamish
Surrey/White Rock Tofino Vancouver Vernon
Victoria West Vancouver Whistler

BY DERRICK PENNER
VANCOUVER SUN

A record amount of new construction and the sixth year of a hot real estate market pushed the value of British Columbia real estate to $808 billion, according to BC Assessment's 2007 property roll.

That is a 23-per-cent increase from 2006.

New construction added 106,734 new properties to the roll and pumped $19 billion into assessments, the authority reported Tuesday.

However, $131 billion of the $150-billion gain was driven by hot property markets in communities around the province.

"It's crazy in B.C. all over the place," said Bill Levis, BC Assessment's manager of audit and research.

"[Assessment increases] are not just isolated to the Lower Mainland or Victoria. We've had increases in Dawson Creek, Prince George, Terrace ... places that historically are more level.

"B.C. seems to be the place people want to live."

The Vernon assessment region, which includes Vernon, Armstrong, Salmon Arm and surrounding communities, had the biggest assessment gain at just over 30 per cent to $13.3 billion.

The Penticton assessment area was another big gainer, with its roll increasing just more than 2 per cent to almost $18 billion.

The Vancouver assessment area's property roll increased 25.8 per cent to almost $156 billion. Most of that was within the city of Vancouver itself.

Whistler is the only municipality to buck the trend. Although the resort community's assessment is up 0.9 per cent to just

Figura 8 - Primeira página do jornal *The Vancouver Sun* do dia 03/01/2007.

2.2. Visualização de informações

O campo de atuação da área de visualização de informações estuda formas e técnicas de representação visual para revelar tendências e relações entre dados estatísticos levando a percepção de uma nova dimensão da informação. Essa é uma área de estudos multidisciplinar influenciada pelas ciências cognitivas, ciência da computação, matemática, estatística, design, entre outras (Ware, 2000).

A utilização de gráficos para representar dados estatísticos é antiga e a série de gráficos de Playfair de 1786 é reconhecida como um marco inicial da área por diversos pesquisadores. Esses foram os primeiros gráficos que utilizaram linhas e áreas para representar dados abstratos (Tufte, 2001).

No entanto, o termo visualização de informações é relativamente novo, foi utilizado pela primeira vez em 1989 em uma publicação sobre interação humano-computador (Robertson et al., 1989) e, desde então, é associado à representação visual de dados auxiliados pelo

uso de computadores. Segundo Card et al. (1999) visualização de informações se define como sendo "o uso mediado por computadores de representações visuais de dados abstratos e interativos para ampliar a cognição".

2.3. Visualização de dados lingüísticos

Como já foi dito, a nuvem de texto é uma forma de visualização de informações textuais. Diferentemente dos dados puramente estatísticos, dados provenientes de linguagem natural apresentam alguns desafios próprios para a visualização. Os textos são conceitos abstratos representados de diversas maneiras diferentes, os dados são nominais (desordenados), com ambigüidade de significados, e a interpretação semântica depende do contexto e da compreensão cultural compartilhada (Hearst, 2002).

Apesar de apresentar grandes desafios, técnicas de visualização de informações têm sido aplicadas à linguagem natural para auxiliar na recuperação de informações, na mineração de dados textuais, na análise de discursos e na geração de resumos de conteúdo.

Um texto é uma seqüência de dados codificados que chegam ao receptor de maneira mais lenta do que uma imagem. No entanto, autores (Triesman e Gormican, 1988; Duncan e Humphreys, 1989; e, Chaud e Yeh, 1995, apud Ware, 2000) defendem que técnicas de visualização de informações podem ser muito efetivas, principalmente para codificar padrões e relações nos dados lingüísticos, usando as propriedades gráficas pré-conscientes[10]. Dessa forma as conexões podem ser exploradas de forma mais rápida sem que seja necessário o tempo e a atenção requeridos na leitura.

A motivação para estudos na área de visualização de dados lingüísticos vem da constatação de que os leitores da internet são mais impacientes que os leitores de materiais impressos:

> "A impaciência do leitor digital vem da cultura, não da natureza da tela. Os usuários de sites têm expectativas diferentes dos usuários de impressos. Eles querem sentir-se "produtivos", não contemplativos; não querem processar, querem buscar; esperam ser desapontados, distraídos e atrasados por pistas falsas." (Lupton, 2006)

Segundo Jacob Nielsen (2002) os usuários da internet não gostam de ler (...) Querem continuar movendo-se e clicando.

[10] Estudos das propriedades do olho humano e das capacidades visuais indicam que existe um processamento pré-consciente de informações relacionadas à forma que é automático. Propriedades como cor, tamanho, contraste, linhas, orientação são visualizadas pelo processo pré-consciente. O autor também resalta que não existe uma propriedade que seja mais eficiente para ser percebida por esse processo porque elas sempre dependerão da sua intensidade e do contexto em que estão inseridas (Ware, 2000).

Técnicas de visualização de informações têm sido aplicadas aos dados lingüísticos com diferentes graus de sofisticação, seja na análise computacional (na transformação dos dados), seja na estrutura ou no método de visualização propriamente dito. A visualização de informações de dados lingüísticos é uma área nova, com cerca de apenas dez anos de existência e apenas recentemente a aplicação desse tipo de visualização migrou para áreas de uso mais populares e comerciais (Card et al., 1999). Uma dessas áreas é a de recuperação de informações na *web*, ou seja, nos sistemas de busca. Alguns exemplos de sistemas de busca que utilizam algum tipo de visualização de dados textuais são o KartOOVISU[11], o Grokker[12] e o Quintura[13]. A abordagem adotada por esses sistemas será tratada no capítulo 5 dessa dissertação.

2.4. Nuvem de texto como uma forma eficiente de visualização de informações

Ware (2002) define cinco vantagens que a visualização de informações propicia quando utilizada de forma eficiente:

- **Compreensão:** a visualização permite a compreensão de grande quantidade de informação.
- **Percepção:** a visualização revela propriedades do dado que não podem ser antecipadas.
- **Controle de qualidade:** a visualização permite o controle de qualidade dos dados porque os problemas se tornam imediatamente aparentes.
- **Foco + contexto:** a visualização facilita a compreensão de um aspecto em escala pequena no contexto geral de maior escala dos dados.
- **Interpretação:** a visualização apóia a formação de hipóteses que propiciam futuras investigações.

A nuvem de texto como forma de visualização de informações propicia todas essas vantagens. Além disso, a boa legibilidade das nuvens e a capacidade de mostrar uma grande densidade de palavras são bastante adequadas para grandes quantidades de texto.

As nuvens atendem ao objetivo de aumentar a compreensão comunicando informações contextuais através de variações na forma como os dados são visualizados, proporcionando a visualização de dimensões adicionais da informação de forma a tornar o contexto explícito.

[11] http://beta.kvisu.com

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

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

Isso vai de encontro à definição de Bürdek (2006), que diz que o Design é uma disciplina que não produz apenas realidades materiais, mas especialmente preenche funções comunicativas.

Os dados extraídos dos textos e comunicados por uma nuvem nada mais são do que uma listagem de palavras e o número de vezes que ela aparece em um texto com um diferencial qualitativo proporcionado pela forma de visualização. Uma nuvem de texto é uma lista hierarquizada visualmente.

Quando esses dados são apresentados em forma de nuvem, é possível se perceber a importância de uma determinada palavra em comparação ao todo, no caso o número total de palavras. Essa informação adicional comunica a importância semântica, ou o contexto das palavras mostradas. Uma nuvem de texto é uma proposta visual para se comunicar relações importantes e dimensões adicionais de significados dentro das limitações de um espaço plano.

Uma nuvem é uma visualização filtrada de um universo multidimensional e pode mostrar relações mais complexas do que aquelas apresentadas por uma listagem que possui apenas uma dimensão.

Essa capacidade de mostrar mais de uma dimensão permite que uma nuvem seja o reflexo da estrutura do texto que ela representa. Dessa forma, elas não se restringem simplesmente a representar uma enumeração das palavras de um determinado conteúdo, que é a principal conquista de uma lista associada a valores.

As variações visuais como a proximidade, profundidade, brilho, intensidade e cor nas nuvens funcionam como pistas para um possível mapa de relacionamentos semânticos.

2.5. Diferenças na forma de navegação e leitura entre nuvens e listas

Nuvens de texto e de *tags* também costumam ser usadas como sistema de navegação e acesso à *web*. Nesse caso as palavras que compõem a nuvem são também *hiperlinks*. A diferença entre a navegação baseada em nuvens e a navegação tradicional em lista, como os menus verticais a esquerda das páginas, é que as nuvens não são lineares e não são desenhadas para serem consumidas de forma linear.

A forma como a nuvem é apresentada permite múltiplos pontos de entrada usando diferenciadores visuais como cor e tamanho da fonte tipográfica para atrair a atenção do usuário. Dessa maneira as nuvens diminuem o esforço visual e cognitivo por parte do usuário já que este não tem que percorrer uma série de itens até encontrar o que deseja. Uma nuvem permite que o usuário vá direto a qualquer ponto de interesse e destaca as palavras que são possivelmente mais relevantes devido a sua alta freqüência.

Os sistemas de navegação baseados em listas são ordenados de diversas formas. Esses sistemas ainda incorporam hierarquias que são mostradas como submenus. Muitos oferecem opções variadas para que o próprio usuário ordene a lista da sua forma preferida e apresentam soluções “customizadas”. No entanto, esses sistemas de navegação continuam fundamentalmente sendo baseados em listas e, seja qual for o padrão adotado, eles são desenvolvidos para serem lidos linearmente.

Os resultados apresentados por sistemas de busca também são estruturados em forma de listas, que na maioria das vezes são distribuídas em várias páginas. Da mesma forma que a estrutura das listas encoraja interações lineares, interfaces baseadas no uso de nuvens encorajam diferentes tipos de comportamentos de busca e localização de informações. É importante ressaltar que nada impede que as nuvens também sejam lidas de forma linear, da esquerda para a direita, no padrão ocidental de leitura.

Dois estudos de avaliação de *tag clouds* apresentados por Halvey e Keane (2007) e Rivadeneira et al. (2007) confirmaram a leitura não linear das nuvens. O primeiro estudo indicou que as nuvens não são lidas linearmente e sim varridas visualmente. O segundo, observou o fato de a leitura ser realizada por quadrantes reafirmando a constatação da pesquisa anterior.

3. Análise das nuvens de texto sob a perspectiva do design

> …"Eu acredito que esses mesmos princípios devam ser aplicados de forma muito mais abrangentes, por todo o MIT e também por todo o nosso sistema universitário de maneira geral. Pelo menos no MIT, já existe há muitos anos uma consciência da necessidade de se combinar as áreas humanas e científicas no nível curricular. No entanto, apesar das nossas melhores intenções, o modelo de treinamento nessa área permanece sendo algo como a área de humanas abraçando a de tecnologia, ou vice-versa ... Para nós não é suficiente simplesmente produzir um tecnologista que é consciente do contexto cultural da tecnologia ou um especialista em humanidades que pode falar fluentemente sobre tecnologia ... O que é necessário é uma verdadeira fusão da sensibilidade artística com a do engenheiro em uma única pessoa." [14]
>
> John Maeda, 1998 (Fry, 2004)

No capítulo anterior apresentou-se a forma de visualização de informações conhecida como nuvem de texto, assim como suas principais características e possibilidades de aplicação.

Essa dissertação apresenta uma nova aplicação para as nuvens de texto. Integrá-la a um sistema de busca na *web* para, dessa forma, mostrar um resumo dos principais resultados do sistema para uma determinada consulta. A principal função de uma nuvem de texto nesse contexto é prover um auxílio cognitivo ao usuário do sistema de busca. Para atingir esse objetivo, é fundamental que a apresentação da nuvem esteja tecnicamente bem estruturada.

O sucesso dessa aplicação está diretamente relacionado à construção de nuvens que sejam de entendimento rápido e claro, além de esteticamente agradáveis. A escolha da fonte tipográfica, proporção ideal entre as palavras, espacejamento e alinhamento, entre outros aspectos, devem ser criteriosos.

[14] Considerações a respeito da combinação da computação com o design. John Maeda é um designer gráfico, artista visual e cientista da computação reconhecido mundialmente. É diretor associado de pesquisas do MIT Media Lab do Massachusetts Institute of Technology. Foi apontado pela revista Esquire como umas das 21 pessoas mais importantes do século XXI. Biografia disponível em: http://www.media.mit.edu/people/bio_maeda.html

Outros fatores também têm impacto na eficácia dessa aplicação. Entretanto a análise de uma nuvem de texto sob a ótica de algumas das principais teorias do design é fundamental para assegurar que o desempenho da aplicação proposta não seja afetado de forma negativa por uma nuvem graficamente mal estruturada.

Logo, nesse capítulo, é feita uma análise das nuvens de texto sob a perspectiva do design segundo Tufte (2001), um dos mais importantes especialistas em infografia da atualidade, e Bertin (1986), o primeiro estudioso a propor uma base teórica consistente para o que hoje chamamos de visualização de informações. Além da análise sob a ótica desses dois autores, rebatendo as Leis da Gestalt sobre uma nuvem de texto, foi feita uma leitura visual da sua forma.

Além da análise teórica apresentada nesse capítulo, no capítulo 6, que trata especificamente da aplicação desenvolvida, outros aspectos relacionados à construção gráfica da nuvem de texto, gerada pela aplicação, também serão abordados de forma mais detalhada.

A apresentação da nuvem na interface da aplicação proposta é apenas uma das etapas no seu processo de desenvolvimento. No entanto, merece um destaque maior por se tratar da área específica do programa de pós-graduação da qual essa dissertação faz parte.

O processo de desenvolvimento de uma aplicação dessa natureza, de informação computacional, segue uma série de etapas (Fry, 2004). O processo como um todo é multidisciplinar, mas, cada etapa tem seus alicerces em áreas bem definidas.

Fry (2004), ao descrever esse processo, destaca quatro áreas distintas como mostra a figura abaixo.

Figura 9 – O processo de desenvolvimento de um sistema de informação computacional segundo Fry (2004).

No entanto, as etapas de representação e refinamento não podem estar separadas da etapa de interação. Todas elas são etapas do design da interface do sistema e devem ser trabalhadas em conjunto. Todas essas etapas são especializações da área do design. Uma divisão mais adequada para o processo descrito por Fry (2004) é proposta abaixo.

Figura 10 - Nova leitura do processo de desenvolvimento de um sistema de informação computacional proposto por Fry (2004).

1. AQUISIÇÃO – Essa etapa trata da questão da obtenção dos dados, seja de um arquivo em disco ou de uma fonte disponível em uma rede.

2. ANÁLISE – Nessa etapa é feita a análise estrutural dos dados e a separação e ordenação em categorias.

3. FILTRAGEM – É a etapa aonde é feita a remoção dos dados que não são relevantes.

4. MINERAÇÃO – É feita a aplicação de métodos estatísticos ou de mineração de dados para discernir padrões ou alocar os dados em um contexto matemático.

5. REPRESENTAÇÃO – Nessa etapa é definida a forma como os dados serão exibidos visualmente.

6. REFINAMENTO – Nessa etapa são feitos melhoramentos na forma básica de apresentação dos dados para que estes se tornem mais claros e esteticamente equilibrados.

7. INTERAÇÃO – Nessa etapa são adicionados métodos de manipulação dos dados ou de controles de visibilidade.

Nesse capítulo, trataremos de alguns aspectos teóricos relacionados às etapas cinco, seis e sete do processo, cujos alicerces se encontram na área do design gráfico, da visualização de informações e interação homem-computador.

No capítulo 6, todas essas etapas serão descritas de forma detalhada à medida que o processo de desenvolvimento da aplicação proposta é apresentado.

Uma nuvem de texto é uma forma de visualização de informações que apenas recentemente se tornou popular na *web,* e por essa razão foi alvo de poucas avaliações até o momento.

Dois estudos recentes da área de design de interfaces foram apresentados por Halvey e Keane (2007) e por Rivadeneira et al. (2007) avaliando as nuvens de texto.

O primeiro comparou a eficiência de listas e nuvens na localização de determinadas palavras e avaliou a influência do tamanho das fontes e da ordenação das palavras na execução da tarefa. A pesquisa indicou que houve uma variação expressiva no tempo levado para a realização das tarefas propostas em testes com fontes de tamanhos diferentes, indicando a necessidade de uma seleção cautelosa das mesmas. Tanto nas nuvens quanto nas listas a ordem de apresentação que se mostrou mais eficiente foi a alfabética. Os itens apresentados no canto superior esquerdo foram localizados mais facilmente, o que é esperado em culturas ocidentais, que lêem da esquerda para a direita, no entanto, os itens localizados no meio ou no canto inferior direito ficaram em segundo lugar na ordem de localização indicando que as nuvens não são lidas linearmente e sim varridas visualmente.

O segundo estudo avaliou a eficiência de diferentes formatações de nuvens na execução de variadas tarefas. Em um primeiro experimento observou o fato de a leitura ser realizada por quadrantes reafirmando a constatação da pesquisa anterior que indica que as nuvens não são lidas linearmente. No experimento seguinte, o estudo avaliou o impacto do tamanho da fonte e de diferentes formas de apresentação das nuvens na memória dos usuários. O efeito do tamanho das fontes na memória foi confirmado sem surpresas, fontes maiores foram mais lembradas e quanto à apresentação da nuvem, a apresentação espacial foi a mais lembrada.

As abordagens desses dois estudos, apesar de avaliar aspectos relacionados aos elementos visuais e à forma das nuvens, são mais focadas na execução de tarefas e na usabilidade das interfaces.

As principais teorias conhecidas e aplicadas há tempos na área do design gráfico podem ser aplicadas às nuvens de texto desde que alguns aspectos e particularidades das nuvens em si, e do meio (computador) sejam observados. Design é escolha, e princípios guiam escolhas dentre as várias opções possíveis.

Uma análise dessa natureza confirma que uma nuvem é uma representação visual que como qualquer outra, pode ser refinada para comunicar da melhor forma idéias complexas com clareza, precisão e eficiência.

Esse é um dos aspectos chave da excelência gráfica, um conjunto de princípios apresentados por Tufte (2001) em *The Visual Display of Quantitative Information*.

Esses princípios foram apontados a partir de uma análise da prática gráfica dos últimos dois séculos, desde Playfair. William Playfair foi o primeiro a conceber e publicar vários tipos diferentes de gráficos estatísticos. É considerado o inventor das formas gráficas mais conhecidas como o gráfico de barras, o gráfico de pizza, o de série temporal entre outras. O seu trabalho *The Commercial and Political Atlas*, publicado em 1786, é considerado um marco histórico.

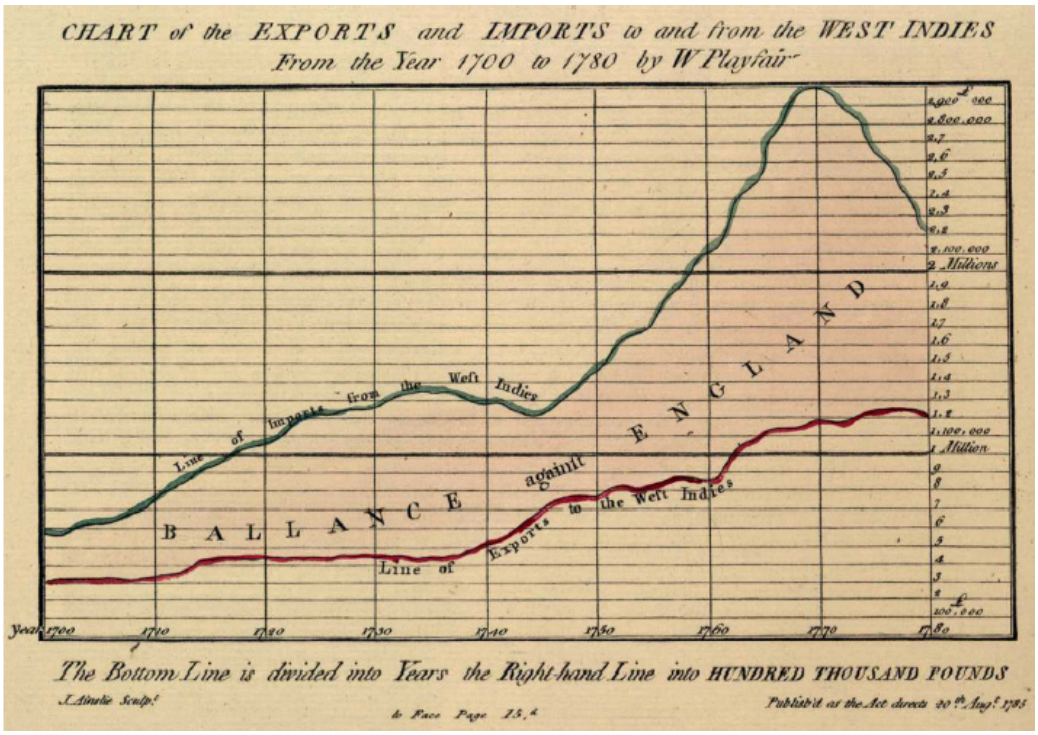
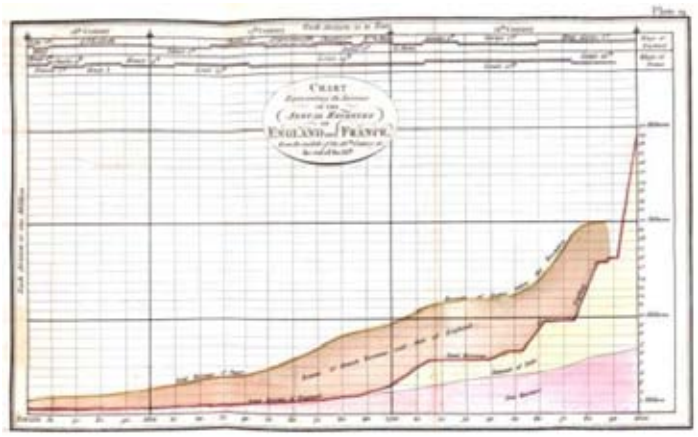
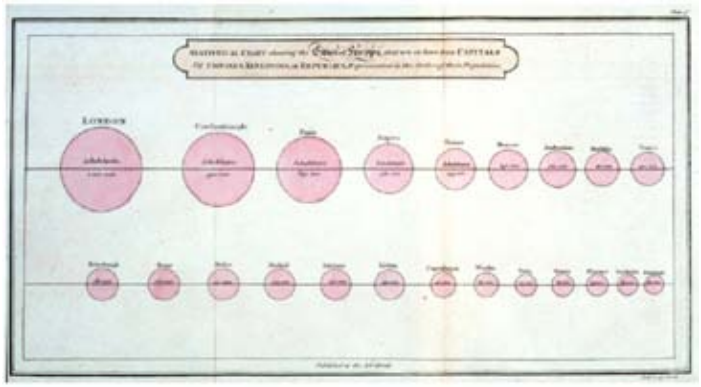


Figura 11 - Prancha 4, *The Commercial and Political Atlas* de Playfair, 1786. (retirado de Spence, 2006)



O uso de cores para diferenciar categorias.



O uso de círculos para codificar quantidade.

Figura 12 - Outros gráficos de Playfair. (retirado de Spence, 2006)

A aplicação correta de diretrizes de design proporciona formas de visualização que são mais prazerosas de serem utilizadas e mais naturais de serem interpretadas, fazendo com que o máximo de informação a partir dos dados seja revelado com o mínimo de distrações e interferências.

Apesar de esses princípios terem sido desenvolvidos para o design de diagramas impressos (por exemplo, em jornais e revistas), eles podem ser aplicados ao design de diversas formas de visualização de informações, entre elas, as de nuvens de texto, principalmente como uma forma de avaliação qualitativa.

Logo, os princípios da excelência gráfica que devem ser observados indicam que uma boa representação gráfica:

- Pode ser refinada para comunicar idéias complexas com mais clareza, precisão e eficiência.
- Induzir o observador a pensar principalmente sobre a substância e não sobre a metodologia, o design ou a tecnologia envolvida na execução do gráfico.
- Encorajar o olhar a comparar diferentes pedaços de dados.
- Revelar a informação em diferentes níveis de detalhe.

Na sua obra Tufte (2001) também fala da importância da integridade gráfica dos dados, ou seja, uma representação gráfica deve evitar distorções na informação que o dado tem a transmitir, mostrar a verdade sobre os dados. Um dos princípios da integridade que deve ser especialmente observado no caso das nuvens de texto diz respeito à representação de números como sendo uma medida física na superfície do gráfico que deve ser diretamente proporcional à medida representada.

Nas nuvens, no entanto, a integridade gráfica dos dados pode ficar comprometida uma vez que os comprimentos das palavras variam aleatoriamente, assim como a sua posição dentro da nuvem. Esse aspecto será visto em detalhes mais a frente.

Tufte (2001) também apresenta uma discussão teórica a respeito dos dados apresentados graficamente.

Um ponto levantado é que os elementos do gráfico devem ser multifuncionais, servindo a mais de um propósito. As palavras em uma nuvem de texto são elementos que além de representar seu próprio significado, representam uma quantidade associada ao seu tamanho de fonte.

Ele apresenta o poema *Easter Wings* de George Herbert (1593-1633) como um exemplo de representação visual que utiliza palavras de forma multifuncional. Esse poema foi escrito 150 anos antes de Playfair e já utiliza o espaço, no caso o comprimento das linhas, para

representar quantidade. Além de ordenar e transmitir a mensagem do texto, as linhas longas descrevem prosperidade, abundância, generosidade, levantar vôo. As curtas falam de pobreza e as intermediárias indicam transição e mudança.

Easter-wings.

Lord, who createdst man in wealth and store,
Though foolishly he lost the same,
Decaying more an more,
Till he became
Most poore:
With thee
O let me rise
As larks, harmoniously,
And sing this day thy victories:
Then shall the fall further the flight in me.

My tender age in sorrow did beginne:
And still with sicknesses and shame
Thou didst so punish sinne,
That I became
Most thinne.
With thee
Let me combine
And feel this day thy victorie:
For, if I imp my wing on thine,
Affliction shall advance the flight in me.

Figura 13 - Poema *Easter Wings* de George Herbert (1593-1633). (retirado de Tufte, 2001)

3.1. Níveis de informação de uma nuvem de texto

Segundo Bertin (1986), os dados por si só não fornecem a informação. É necessário ver as relações que o conjunto de dados estabelece. A informação útil para a decisão é dada pelas relações de conjunto.

O objetivo de uma representação gráfica é apresentar um nível superior da informação. A informação útil não é um aumento da quantidade de uma informação, mas, ao contrário, uma redução dessa quantidade por reagrupamentos pertinentes. Estes reagrupamentos correspondem aos novos grupos definidos pelo conjunto de relações que os dados estabelecem entre si.

Uma nuvem de texto é uma representação gráfica a partir de uma tabela de dois atributos. Um deles é uma lista de palavras e o outro é o número de vezes que ela aparece em um

texto. Essa tabela fornece apenas uma seqüência linear de dados e pode ser bastante extensa. As relações que esses dados mantêm entre si, ou seja, as relações úteis à reflexão e a uma eventual tomada de decisão não são explícitas.

No entanto, uma nuvem de texto permite que sejam visualizados outros conjuntos e relações entre os dados que não podem ser visualizados na tabela.

Bertin (1986) afirma que: "A percepção visual é espacial, e permite qualquer um utilizar um novo princípio de reclassificação: a tomada em consideração, simultânea, de muitos elementos".

Os olhos percebem semelhanças e conjuntos. E o reagrupamento do que se assemelha simplifica a informação. Em uma nuvem de texto os conjuntos formados por palavras com fontes tipográficas de tamanhos semelhantes, mesmo não estando justapostos, são facilmente identificáveis.

Também é possível comparar a relação entre os diversos conjuntos formados por palavras de fontes de tamanhos diferentes, e ainda, a relação de tamanho entre os diferentes conjuntos e a proporção deles em relação ao conjunto geral, o conjunto de todas as palavras mais freqüentes. Segundo Bertin (1986):

> "É preciso substituir a noção de quantidade de informações pela noção de níveis de informação que se exprimem verbalmente pelo nível das questões e graficamente pelos níveis de leitura. A informação é uma relação. Mas esta relação pode ser estabelecida entre elementos, subconjuntos, ou conjuntos."

O nível elementar da informação é a relação que existe entre um elemento X e um elemento Y. No caso, cada palavra e o número de vezes que ela aparece na tabela de dados.

O nível do conjunto de informações é o nível mais elevado de conhecimento. A partir da tabela de dados composta por palavras e freqüências não se pode chegar a esse nível de leitura. A nuvem de texto permite esse nível de leitura mais elevado, e mantém a legibilidade do nível elementar.

3.2. Variáveis visuais de uma nuvem de texto

Para transcrever as relações de semelhança, de ordem e de proporcionalidade Bertin (1986) sistematizou as variáveis visuais que o olho pode perceber: localização, tamanho, valor, textura, cor, orientação e forma. Ele separa essas variáveis em dois grupos, as variáveis da imagem e as variáveis de separação da imagem.

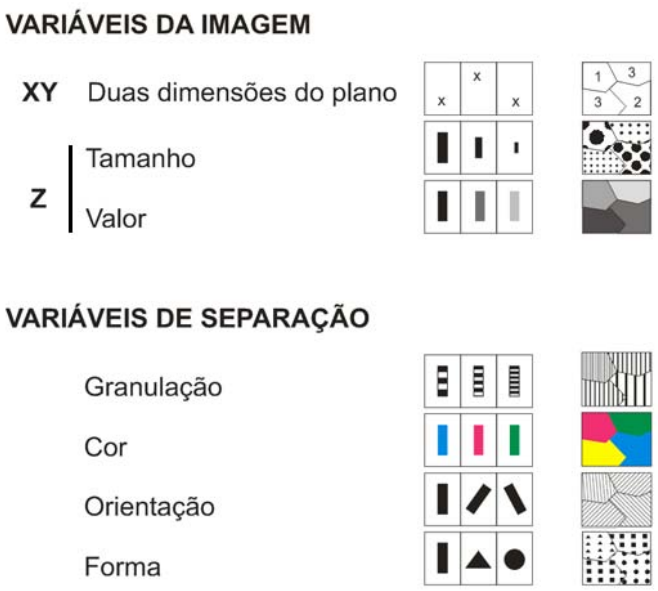


Figura 14 - Variáveis visuais de Bertin. (adaptado de Bertin, 1986)

As variáveis da imagem são as duas dimensões do plano XY e as variações de Z. O componente expresso por Z se eleva acima do plano como um relevo. Esse relevo constrói uma imagem visível somente quando os números são transcritos pelo tamanho ou valor. É a terceira dimensão da imagem.

As propriedades do plano permitem que se visualizem grupos diferentes e semelhantes e a ordem e a proporção entre duas grandezas pela diferença de tamanho entre elas. O plano da imagem também é contínuo e homogêneo. É possível se construir uma imagem na imagem, mas não se pode abstrair uma parte da imagem.

Transcrevendo essa teoria para uma nuvem de texto, o plano se refere à nuvem como um todo, um objeto único cuja origem do nome se dá exatamente pela ilusão de relevo formado pelo tamanho e valor associado às palavras, que faz com que essa representação se assemelhe a uma nuvem no sentido literal da palavra.

A localização de uma variável visual, no caso, as palavras em uma nuvem, é dada através das dimensões X e Y do plano. Nas nuvens de texto a localização dos elementos de maior peso, é aleatória uma vez que a ordenação das palavras é alfabética e não se sabe quais as palavras que vão apresentar a maior freqüência.

A variável visual de maior importância em uma nuvem de texto é o tamanho da fonte tipográfica. O tamanho é quantitativo. É o tamanho da fonte que reflete a quantidade numérica associada a uma determinada palavra. Palavras representadas por fontes maiores indicam que elas aparecem mais vezes no texto que originou uma determinada nuvem.

No entanto, o tamanho é uma variável que quando é utilizada em palavras apresenta um problema que gera distorções na percepção de uma nuvem de texto. O tamanho da fonte tipográfica pode ser controlado e associado a uma escala de ordem numérica, mas, além do tamanho das fontes, as palavras têm comprimentos diferentes, e essa variação de comprimento afeta o seu destaque em uma nuvem.

Em nuvens de texto palavras longas sempre terão mais destaque que palavras curtas mesmo nos casos em que estejam representando a mesma importância numérica porque nas nuvens de texto as palavras mais longas ocupam uma área maior.

O valor, quando utilizado nas nuvens de texto também deve estar associado à freqüência das palavras e serve para reforçar esse aspecto. O valor é dado pela variação na tonalidade de uma única cor. Essa variação é obtida através da adição gradativa de branco ou preto a essa cor. Em nuvens de texto de fundo branco, cores mais escuras são utilizadas para representar ordens de grandeza maiores e cores mais claras, grandezas menores.

O valor não é quantitativo, mas tamanho e valor são ordenados. Não há necessidade de legenda para colocar em ordem os tamanhos ou valores. Eles criam uma ordem espontânea que deve evidentemente corresponder à ordem do componente. Em uma transcrição ordenada, a legenda serve apenas para definir verbalmente os limites dos patamares.

A área de um gráfico pode ser dividida em duas áreas complementares: a área do desenho onde está contida a representação gráfica propriamente dita e a área exterior onde normalmente estão posicionadas as componentes de auxílio à leitura (título, legenda e identificações) (Silva, 2003).

Normalmente as nuvens de texto apresentadas na *web* mostram apenas a área do desenho. Quando existe um componente de auxílio à leitura este se resume muitas vezes ao título. As nuvens de texto se propõem a ser auto-explicativas.

No entanto, a ausência de informações de auxílio à leitura é um problema grave em nuvens de texto, exatamente porque não define os limites dos patamares representados. E se torna mais grave ainda por se tratar de uma forma de representação nova sobre a qual existe muito pouca informação sobre sua construção e sobre o que exatamente representa.

Tamanho e valor também são valores dissociativos, ou seja, eles têm uma visibilidade variável. Quanto maior o tamanho do elemento maior é a sua visibilidade. Elementos menores por sua vez são menos visíveis. Da mesma forma o valor torna os elementos mais ou menos visíveis. O fato de esses valores serem dissociativos também é fundamental para o “aspecto de nuvem” da representação, da formação do relevo e da sensação de profundidade.

Logo, as propriedades relevantes das variáveis visuais de tamanho e valor, para uma nuvem de texto são: quantitativas, ordenáveis e dissociativas.

As variáveis de separação fazem com que os relevos desapareçam nos gráficos. A terceira dimensão em uma nuvem de texto faz parte da sua natureza, logo, de maneira geral não se faz necessária à aplicação de variáveis de separação. As variáveis de separação são: granulação, cor, orientação e forma. Elas separam de fato apenas os elementos entre si eliminando toda ordem.

Em nuvens, apenas a cor em alguns casos é usada como variável de separação. Um exemplo dessa aplicação pode ser visto quando as palavras em uma nuvem de texto também são elementos de navegação na *web* e funcionam como *links* para outras páginas. Nesse caso, os *links* visitados passam a ter outra cor. Dessa forma eles são separados das outras palavras indicando que já foram clicados.

3.3. Leitura visual da forma

> "A 'fórmula' fundamental da teoria da Gestalt poderia ser expressa da seguinte maneira: existem totalidades, cujo comportamento não é determinado pelos seus elementos individuais, mas nos quais os processos parciais são eles mesmos determinados pela natureza intrínseca do todo".
>
> Max Wertheimer [15]

Muitos dos conceitos e fatores de organização formal estudados pelos psicólogos da Gestalt[16] coincidem com a preocupação prática projetual relativa à concepção da aplicação para o experimento proposto por essa dissertação.

Rebatendo as Leis da Gestalt sobre uma nuvem de texto, é possível se fazer uma leitura visual da forma desse tipo de representação e identificar o que pode ser adotado na construção de uma nuvem de modo a garantir uma melhor pregnância da forma.

A pregnância da forma é a lei básica da percepção visual da Gestalt. Uma boa pregnância pressupõe que a organização formal do objeto, no sentido psicológico, tenderá a ser sempre a melhor possível do ponto de vista estrutural. Naturalmente, quanto pior ou mais confusa

[15] Max Wertheimer (1880/1943), psicólogo, foi um dos fundadores da Gestalt juntamente com Kurt Koffka (1886/1941) e Wolfgang Köhler (1887/1967).

[16] A Gestalt é uma Escola de Psicologia Experimental que teve seu início por volta de 1910. O movimento gestaltista atuou principalmente no campo da teoria da forma, com contribuição relevante aos estudos da percepção, linguagem, inteligência, aprendizagem, memória, motivação, conduta exploratória e dinâmica de grupos sociais. Por meio de numerosos estudos e pesquisas experimentais, os gestaltistas formularam suas teorias acerca dos campos mencionados (Gomes Filho, 2004).

for a organização visual da forma do objeto menor será o seu grau de pregnância (Gomes Filho, 2004).

Nem todas as Leis da Gestalt são aplicáveis as nuvens de texto. A proposta dessa análise não é enquadrar uma nuvem dentro de um molde, mas sim, identificar o que pode ser aplicado dessas leis para que esta tenha uma melhor organização formal, e identificar possíveis relações entre as partes da nuvem que ficam prejudicadas pelos princípios intrínsecos da sua construção.

A Teoria da Gestalt propõe que o cérebro humano tende automaticamente a desmembrar a imagem em diferentes partes, organizá-las de acordo com semelhanças de forma, tamanho, cor, textura etc., que por sua vez serão reagrupadas de novo num conjunto gráfico que possibilita a compreensão do significado exposto.

As Leis da Gestalt estabelecem relações através das quais as partes da imagem são agrupadas na percepção visual, além da pregnância, são elas: unidade, segregação, unificação, fechamento, continuidade, proximidade e semelhança.

Antes de tudo, as nuvens de texto podem apresentar tanto formas geométricas, como podem ser irregulares. Os formatos mais freqüentes são os retângulos, verticais ou horizontais formados pelos textos justificados ou o formato irregular, de um texto com alinhamento centralizado.

Outros formatos geométricos não são adequados uma vez que a densidade de palavras ideal para uma visualização adequada de uma nuvem não permite que se construa um formato com contornos bem definidos a partir de palavras. Alguns formatos como triângulos, por exemplo, também poderiam sugerir uma espécie de hierarquia não existente nesse tipo de representação.

Uma nuvem de texto bem desenhada deve ser percebida antes de tudo como uma unidade, um elemento que se encerra em si mesmo e se assemelha a uma nuvem real. O uso de apenas uma cor para as fontes tipográficas e a variação de valor dessa cor, caso se deseje reforçar o tamanho das fontes contribui para essa unidade. Outro elemento fundamental para a garantia da unidade de uma nuvem de texto é o uso de uma mesma fonte tipográfica em toda a composição.

As subunidades que compõem uma nuvem, ou seja, as palavras e conjuntos de palavras com mesmo tamanho de fonte devem ser percebidas como volumes dentro do todo.

Em uma nuvem de texto se tem o controle do número de subunidades que vão aparecer. Elas são definidas pelo número de patamares escolhidos para serem associados aos tamanhos das fontes das palavras. Todavia, não se tem o controle da posição nem do tamanho dessas subunidades dentro da unidade principal. Por essa razão, a segregação

das subunidades fica prejudicada. Essa segregação só pode ser feita pelo agrupamento proporcionado pela semelhança dos tamanhos das palavras.

Um outro fator que geralmente age em comum com a semelhança criando um reforço mútuo é a proximidade. Elementos óticos próximos uns dos outros tendem a ser vistos juntos.

A unidade principal da nuvem é garantida pela proximidade dos elementos, contudo, as subunidades que aparecem em uma nuvem são agrupadas de forma aleatória.

Uma palavra com fonte tipográfica de um determinado tamanho próxima de palavras com fontes menores vai ter mais destaque na nuvem. Da mesma forma que se essa mesma palavra aparecer perto de outras palavras com fonte de tamanho semelhante ela receberá menos destaque isoladamente. No entanto, o grupo de palavras receberá mais destaque como um todo.

Uma nuvem de texto é uma composição cujo centro de interesse, varia a cada nova apresentação. O centro de interesse é a área que atrai primeiro a atenção, a mais importante comparada aos outros elementos da composição. O centro de interesse de uma nuvem de texto será determinado pela localização das palavras com fontes maiores. Na eventualidade dessas palavras aparecerem próximas de palavras de tamanho semelhante mais força terá esse centro de interesse. Geralmente, uma nuvem de texto, ordenada alfabeticamente, irá apresentar mais de um centro de interesse, distribuídos de forma desigual.

Além disso, a localização de uma palavra irá afetar de forma diferente a sua percepção. Uma palavra posicionada no centro de uma nuvem vai chamar mais atenção do que essa mesma palavra posicionada em um canto dessa mesma nuvem.

Concluindo, é possível se dizer que a pregnância de uma nuvem de texto, como um todo, é média. A apreensão da unidade principal é fácil e rápida. No entanto, alguns atributos da “boa Gestalt” não são contemplados na percepção das subunidades. A segregação das subunidades é prejudicada pela impossibilidade de se reforçar a semelhança dos elementos pela proximidade, uma vez que a formação dessas subunidades é aleatória na composição de uma nuvem.

A pregnância, no entanto, de uma nuvem pode ser melhorada se a densidade das palavras for bem calculada, se a escolha do tamanho máximo e mínimo das fontes forem proporcionais ao tamanho da nuvem, e se a entrelinha, distância da linha de base de uma linha tipográfica para outra, for bem estudada assegurando dessa forma uma melhor harmonia, legibilidade e equilíbrio na composição.

4. Sobre os sistemas de busca

> "Os gregos faziam suas perguntas ao oráculo de Delfos no famoso templo de Apolo. Esse era um local de troca de idéias e informação. Os oráculos da era da informação são os sistemas de busca na *web* onde milhões de usuários submetem consultas diariamente as suas bases de dados."
>
> (Paltrige, 1999)

O objetivo desse capítulo é descrever o funcionamento básico de um sistema de busca na *web* passando por suas principais etapas: a análise e a indexação das páginas, o armazenamento das páginas indexadas e a recuperação das páginas que preenchem os requisitos indicados pelo usuário por ocasião da consulta.

A compreensão desse funcionamento é de fundamental importância uma vez que a aplicação desenvolvida e testada por essa pesquisa interage diretamente com um sistema de busca como um *plugin*. Um *plugin* é um sistema que executa certas funções, geralmente específicas, sob demanda. Aplicações em geral suportam *plugins* permitindo que desenvolvedores autônomos criem novas funcionalidades para os mesmos.

Também é apresentado um breve panorama sobre o crescimento da informação na *web*, o perfil de utilização dos usuários e a importância dos sistemas de busca nesse contexto.

4.1. A importância dos sistemas de busca na *web*

As tecnologias de informação e de comunicação abrem novas perspectivas à sociedade do futuro. A Internet possibilita hoje uma difusão rápida da informação que, uma vez produzida, circula instantaneamente, e pode ser acessada por toda a sociedade.

Segundo a pesquisa de julho de 2007 do Internet Systems Consortium, Inc[17] já existem 489 milhões de *sites* disponíveis na Internet.

[17] http://www.isc.org

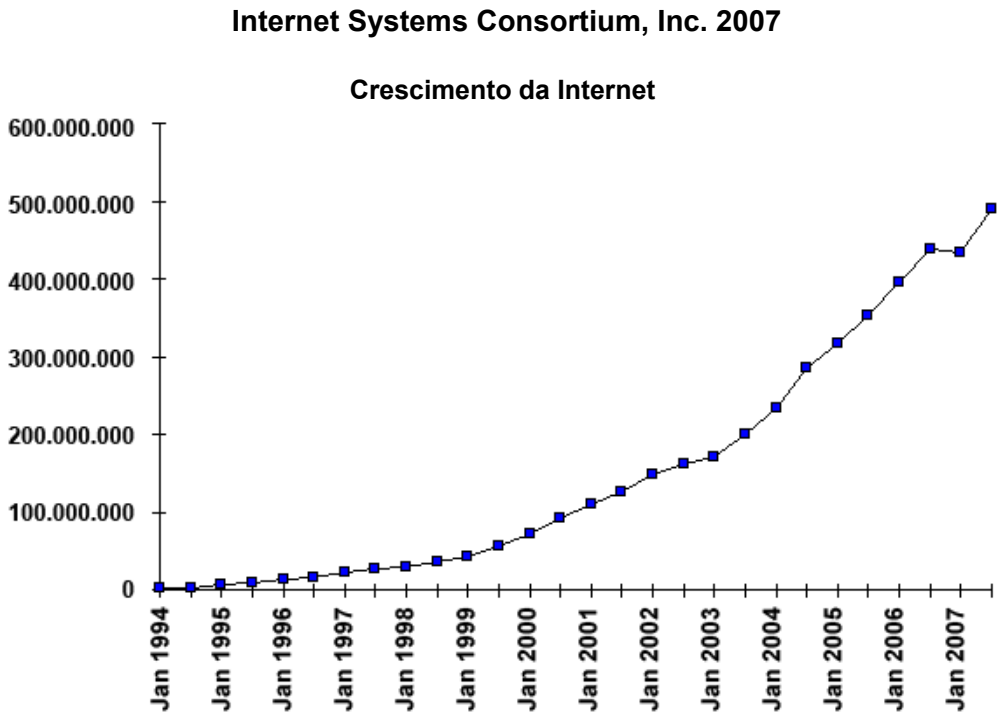


Figura 15 - Crescimento do número total de *sites* disponíveis na Internet. (www.isc.org)

A quantidade de informações disponíveis na Internet é cada vez maior e a sua utilização é cada dia mais presente na vida de todos. Atividades cotidianas são planejadas e decisões são tomadas baseadas em consultas à Internet.

A tabela abaixo mostra a utilização média da Internet pela população mundial no mês de agosto de 2007, segundo o Instituto Nielsen Netratings[18]. Foi verificado que cada pessoa gasta em média quase uma hora por dia na Internet.

Nielsen Netratings

Uso Mundial da Internet (Agosto de 2007)

Número médio de sessões por pessoa	34
Número médio de domínios visitados por pessoa	69
Número médio de páginas visitadas por pessoa	1.518
Número médio de páginas visitadas por sessão	44
Tempo total gasto na Internet por pessoa	31:25:53
Tempo médio gasto por sessão	00:56:06
Tempo médio dedicado a cada página visitada	0:00:45

Tabela 2 - Uso Mundial da Internet em agosto de 2007. (www.nielsen-netratings.com)

[18] http://www.nielsen-netratings.com/

Para se localizar algo em meio a toda essa informação disponível, é imprescindível um mecanismo que permita a pesquisa nesse universo. Esse é o objetivo dos sistemas de busca que surgiram logo após a criação da Internet. Os sistemas de busca se propõem a permitir a pesquisa e a localização de dados específicos em meio a toda essa informação crescente, apresentando os resultados de forma organizada, rápida e eficiente.

4.2. Panorama atual do mercado dos sistemas de busca

Um levantamento feito durante a execução dessa dissertação mostra que o mercado de sistemas de busca na *web*, nos Estados Unidos, se encontra dividido da seguinte forma: Google (53,7%), Yahoo! (22,7%), MSN/Windows Live (8,9%), AOL (5,4%) e Ask (1,8%). Menos de 7,5 % dos usuários de sistemas de busca utilizam outros sistemas diferentes dos citados (Nielsen//NetRatings, 2007).

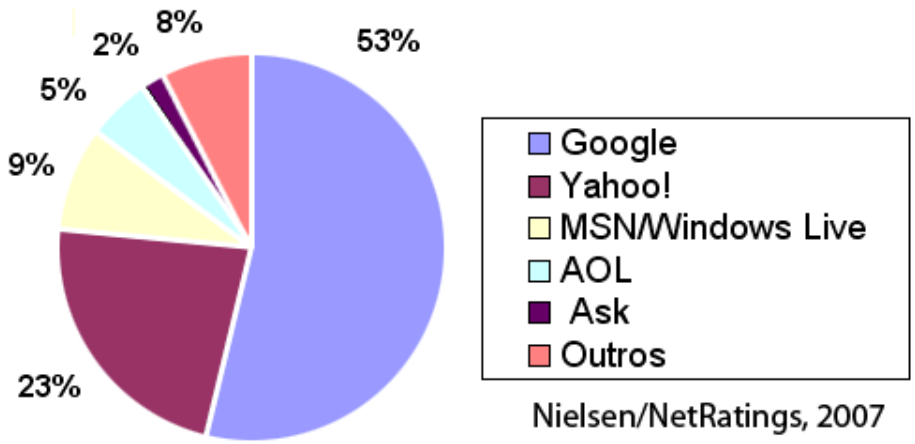


Figura 16 - Distribuição do mercado dos sistemas de busca nos E.U.A. em maio de 2007. (www.nielsen-netratings.com)

A tabela abaixo mostra a quantidade de páginas indexadas pelos maiores sistemas de busca. Esses números foram relatados pelas próprias empresas e embora não reflitam o momento atual, permitem que se tenha uma dimensão do universo que abrangem. Nos últimos anos os principais sistemas mudaram suas estratégias e deixaram de divulgar esses dados (Sullivan, 2004).

Sistemas de busca	Número de páginas indexadas
Google	8.1 bilhões
MSN	5.0 bilhões
Yahoo!	4.2 bilhões (estimado)
Ask	2.5 bilhões

Tabela 2 – Número de páginas indexadas pelos maiores sistemas de busca. (Sullivan, 2004)

4.3. Recuperação de informação

Os sistemas de busca nada mais são do que sistemas de recuperação de informação. Recuperação da Informação (RI) é uma subárea da ciência da computação que trata da representação, armazenamento, organização e do acesso aos itens de informação (Baeza-Yates e Ribeiro Neto, 1999). Um item de informação é geralmente constituído de texto como documentos, páginas na *web*, livros, entre outros, podendo conter outros tipos de dados, como fotografias, gráficos e figuras. Segundo Macedo (2001), o principal objetivo de um sistema de RI, é a seleção, num universo de documentos disponíveis, do conjunto de documentos relevantes para uma necessidade de informação do usuário.

As interfaces homem-máquina existentes atualmente, não permitem que um sistema de RI obtenha as informações diretamente da mente do usuário. Portanto, o usuário precisa traduzir a sua necessidade de informação utilizando uma linguagem formal específica de um determinado sistema, o que representa uma das grandes dificuldades para o usuário.

A necessidade de informação do usuário deve ser traduzida para uma consulta que possa ser processada por um sistema de RI. A consulta deverá ser formulada através de uma palavra-chave ou um conjunto de palavras-chave. O tipo de formalização exigida para a consulta, depende também do tipo de sistema de RI que estiver sendo utilizado pelo usuário.

A dificuldade na formulação da necessidade de informação pelo usuário, ocorre também em grande parte por se tratar de uma “necessidade visceral” (Macedo, 2001), ou seja, o usuário está consciente que precisa da informação, mas não consegue nem sequer a sua definição em linguagem natural. Portanto, transpô-la, para a linguagem suportada pelo sistema de RI é muito mais difícil. Consequentemente, o usuário pode vir a formular uma consulta inadequada e a probabilidade do sistema retornar documentos não relevantes para a necessidade do usuário aumenta.

Através da consulta formulada pelo usuário, o sistema de recuperação de informação é capaz de selecionar as informações (documentos) relevantes para a necessidade do usuário. A forma utilizada pelo sistema para selecionar a informação relevante é identificar a similaridade entre as informações armazenadas no sistema com a necessidade de informação descrita na expressão da consulta. Conforme Wives (2002) esta comparação pode ser problemática, porque um documento pode ser relevante à consulta do usuário, mas não ser relevante para o usuário (que pode ter formulado incorretamente a sua necessidade de informação). Após determinar quais os documentos de uma coleção são relevantes à consulta do usuário, os sistemas de RI retornam o resultado da consulta em uma lista, onde os documentos estão ordenados de acordo com um grau de relevância.

4.4. Sistemas de recuperação de informação na *web*

Um sistema de busca na *web* é um tipo de sistema de recuperação de informação desenvolvido especificamente para o ambiente da Internet. A interface desse tipo de sistema é apresentada como mais uma página na *web*. Basicamente é um *site* desenvolvido para ajudar as pessoas a encontrarem outros *sites* na Internet.

Muitos sistemas de RI são fechados, ou seja, funcionam em ambientes que permitem o controle do que é introduzido no sistema. No caso dos sistemas de busca na *web* esse controle não existe.

Existe uma série de termos utilizados para denominar os sistemas de recuperação de informação na *web*. Em português, o termo mais comum é sistema de busca que provém do inglês *search engine*. Também são utilizados os termos mecanismo de busca, ferramenta de busca, ou ainda, motor de busca como sinônimos (Moura 2000).

Muitas vezes o termo sistema de busca é usado de forma genérica e engloba também os diretórios, os metabuscadores e os sistemas híbridos. Estes sistemas, embora tenham modos de funcionamento diferentes, tratam do mesmo problema, que é a recuperação, em um universo de documentos, do conjunto de documentos relevantes para uma necessidade da informação do usuário (Macedo, 2001). A seguir será feita uma breve descrição de cada um deles.

Mecanismos de busca

Nessa dissertação, o termo sistema de busca é utilizado para definir os sistemas cuja indexação das páginas é feita exclusivamente através de programas de computadores. Esses sistemas também são conhecidos como mecanismos de busca. Os mecanismos de busca utilizam robôs, que percorrem a *web* a fim de encontrar as páginas (descobrimento das informações), uma base de dados onde armazenam referências da informação indexada e uma interface que permite ao usuário efetuar sua consulta e apresentar os

resultados obtidos. Os principais sistemas de busca apontados na tabela 2 funcionam dessa forma. O funcionamento desses sistemas será detalhado adiante.

Diretórios

Os diretórios foram os primeiros sistemas propostos para organizar e localizar as informações na *web* (Céndon, 2001). Nos diretórios a indexação das páginas é realizada manualmente. Os editores, como são conhecidos os profissionais especializados que fazem essa indexação (geralmente documentaristas e bibliotecários), escrevem ou revisam, e aprovam previamente os textos descritivos sobre os sites, que uma vez indexados, são organizados em uma estrutura hierárquica de categorias. Um exemplo de diretório é o *Open Directory Project* [19] cujos editores são usuários voluntários.

Metabuscadores

Os sistemas de metabusca são sistemas que localizam a informação em outros sistemas de busca simultaneamente e combinam os resultados encontrados em uma só lista de resultados (Blattman, 2000). Estes sistemas não possuem uma base de dados própria, uma vez que utilizam os dados de outros sistemas de busca.

Sistemas híbridos

Os sistemas híbridos, como o próprio nome indica, apresentam resultados que são provenientes tanto de diretórios como de mecanismos de busca, indexados pelo próprio sistema ou ainda ou por outros sistemas.

4.5. Funcionamento dos sistemas de busca na *web*

Os sistemas de busca na *web* cuja indexação das páginas é feita exclusivamente através de programas de computadores (também conhecidos como mecanismos de busca) apresentam diferenças entre si no seu funcionamento, mas de maneira geral todos realizam três tarefas básicas:

- Rastreiam os *sites* na *web* em busca de palavras-chave,
- Indexam esses *sites* e,
- Permitem que os usuários dos sistemas pesquisem por palavras ou combinações de palavras em seus índices.

[19] http://www.dmoz.org/

Rastreamento

Um sistema de busca rastreia os *sites* na *web*, principalmente seus textos, em busca de palavras que sejam uma amostra representativa dos seus conteúdos. Outras informações como a posição em que as palavras aparecem, seja no título, subtítulo das páginas, ou nos seus meta-dados, também são consideradas durante esse rastreamento.

Os meta-dados são dados que não são visíveis e podem ser inseridos no código fonte da página. Eles são geralmente conceitos e palavras-chave que melhor definem o conteúdo do *site,* segundo os critérios do próprio autor.

O rastreamento das páginas é feito de forma automática por *softwares* robôs chamados de *spiders.* Os robôs fazem esse mapeamento seguindo *links* de uma página a outra de forma recursiva em um processo contínuo conhecido como *web crawling*.

Os pontos de partida são os servidores de hospedagem mais utilizados e as páginas mais populares, denominados de endereços sementes. O robô começa seu rastreamento em um *site* com alto índice de acesso, seguindo cada *link* encontrado. Dessa forma, o sistema de rastreamento rapidamente se espalha cobrindo as páginas mais utilizadas da Internet.

O Google[20], sistema de busca de maior popularidade na *web*, começou como um sistema de busca acadêmico. No artigo que descreve como o sistema foi construído, Sergey Brin e Laurence Page (1998) dão um exemplo de quão rápido os robôs podem funcionar. Eles construíram o sistema inicial para utilizar *spiders* múltiplos, geralmente eram três de cada vez. Cada *spider* podia manter simultaneamente conexão com 300 páginas na Internet. Nos momentos de pico o sistema podia rastrear cerca de 100 páginas por segundo.

Indexação

O rastreamento das páginas é contínuo, uma vez que o conteúdo disponível na *web* muda constantemente e seu crescimento é exponencial.

Depois que uma página é rastreada o sistema de busca indexa e armazena a informação coletada na sua base de dados para ser acessada posteriormente. Na base de dados podem ser encontrados endereços das páginas, títulos, cabeçalhos, resumos, tamanho, e as palavras contidas nos documentos.

Os robôs ao visitarem uma página, primeiro verificam se a mesma já foi visitada anteriormente ou se é uma página nova para ele. Caso, a página já tenha sido indexada, o robô verifica se ocorreu alguma modificação desde a última visita, e se ocorreu, atualiza a informação armazenada sobre a página na base de dados.

[20] http://www.google.com

A maioria dos robôs possui um período de tempo predeterminado para revisitar os *sites* por eles indexados visando detectar as mudanças ocorridas. Conforme Macedo (2001) destaca, os robôs estão documentados de forma superficial na literatura principalmente por questões de sigilo comercial.

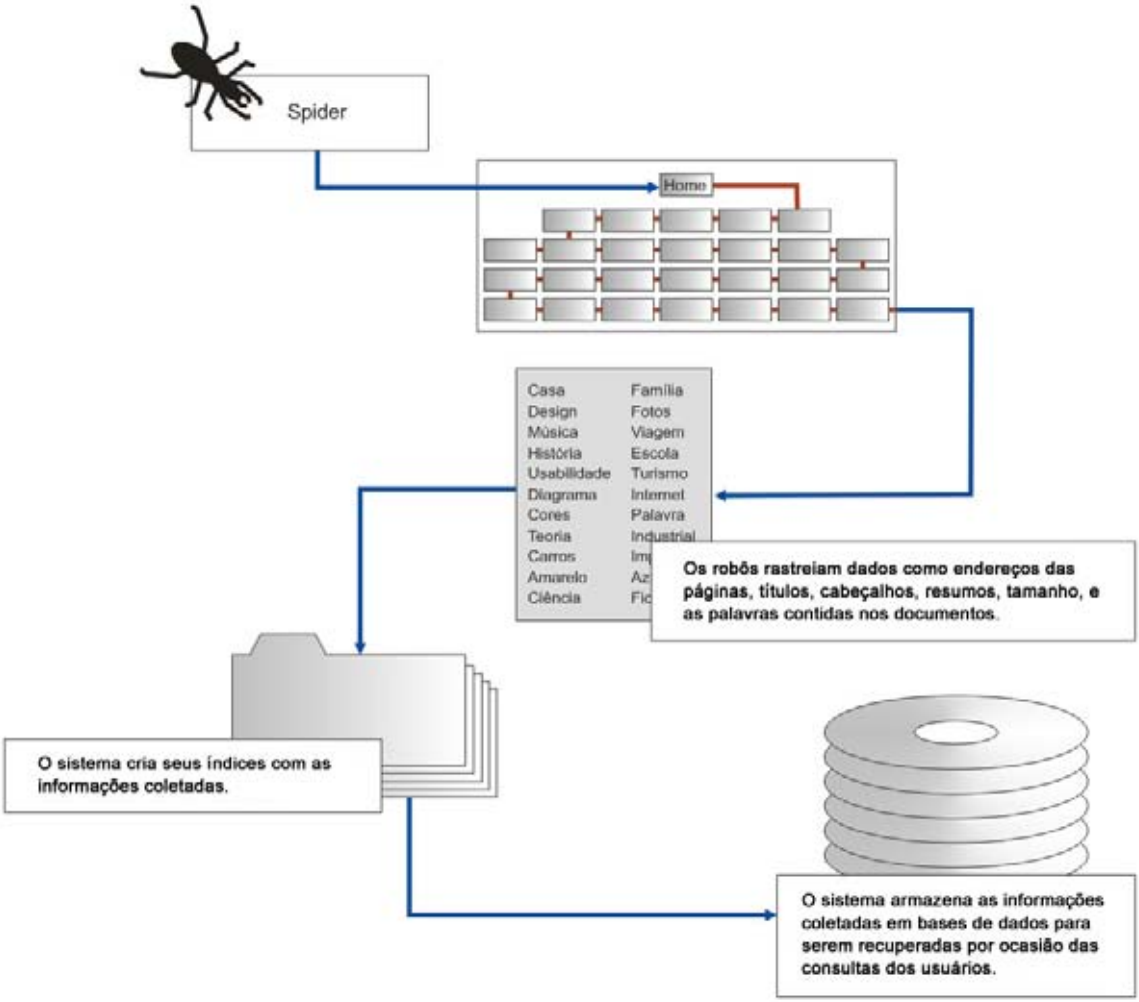


Figura 17 - Processo de indexação de informações em um sistema de busca na *web*.

A pesquisa em um sistema de busca e a construção do resultado

A indexação das informações armazenadas tem o objetivo de permitir que essas informações sejam localizadas da maneira rápida e eficiente.

Para realizar uma pesquisa em um sistema de busca o usuário precisa construir uma consulta e submeter ao sistema. Essa consulta é feita em uma interface que é responsável pela interação do usuário com o sistema. Vale destacar que ao realizar uma pesquisa, o usuário não está pesquisando diretamente a *web* e sim uma base de dados a partir de um *site* da *web*.

Essa consulta pode ser composta por apenas uma palavra ou pode ser mais complexa utilizando várias palavras, expressões e operadores lógicos que permitem que a pesquisa seja refinada ou estendida.

Quando um usuário faz uma consulta em um sistema utilizando palavras-chave, o sistema faz o processamento das expressões e palavras utilizadas por esse usuário, e recupera as páginas indexadas com palavras coincidentes em sua base de dados em tempo real.

Nesse processamento, os sistemas de busca também utilizam algoritmos próprios que calculam a relevância dos *sites* indexados para aquela consulta. A relevância do *site* é que determina o seu posicionamento, o seu *rank*, na listagem dos resultados. Os *sites* considerados mais relevantes são listados primeiro.

Alguns dos critérios adotados pelos sistemas para construir o *rank* dos resultados são listados abaixo:

- Número de páginas que contêm *links* para uma outra página. Quanto mais *links* uma página tiver apontando para ela, maior é a importância atribuída à mesma.
- Número de vezes em que as palavras-chave usadas na pesquisa surgem ao longo da página de resultado.
- Proximidade entre as palavras pesquisadas. Quanto mais próximas às palavras estiverem entre si, maior relevância é atribuída à página.
- Número de vezes em que as palavras-chave aparecem nos títulos, subtítulos e nas primeiras linhas da página, lugares considerados mais nobres.

Resumindo, são esses os três componentes que estão estreitamente associados às três funções básicas de um sistema de busca: a análise e a indexação das páginas, o armazenamento das páginas indexadas e a recuperação das páginas que preenchem os requisitos indicados pelo usuário por ocasião da consulta.

A *web* é heterogênea, e os sistemas de busca conforme foi mencionado anteriormente também, rastreiam, armazenam e indexam arquivos em outros formatos como vídeos e imagens. No entanto, essa descrição do funcionamento de um sistema de busca se detém à indexação, ao armazenamento e à recuperação de textos, objeto de estudo dessa dissertação.

5. Comportamento de interação dos usuários com sistemas de busca na *web*

No capítulo anterior foi explicado o funcionamento básico dos sistemas de busca na *web* passando por suas principais etapas: a análise e a indexação das páginas, o armazenamento das páginas indexadas e a recuperação das páginas que preenchem os requisitos indicados pelo usuário por ocasião da consulta.

Nesse capítulo são apresentados os principais estudos realizados sobre o comportamento dos usuários dos sistemas e busca.

Primeiro é apresentada uma série de pesquisas onde foram analisados os *query logs* de quatro sistemas de busca: Excite[21], Altavista[22], Fireball[23] e AlltheWeb[24]. *Query logs* são os registros de todas as consultas, realizadas em um sistema de busca em um determinado período.

Os estudos do Excite analisaram quatro amostras coletadas entre 1997 e 2001. Os estudos do Altavista analisaram duas amostras, de 1998 e 2002. E os estudos do Fireball e Alltheweb amostras entre 1998 e 2002.

As principais constatações desses estudos sobre o comportamento de busca dos usuários foram que as consultas tendem a ser curtas com em média duas ou três palavras. Os usuários raramente utilizam as ferramentas de busca avançada ou operadores lógicos para aumentar a precisão das consultas, e ainda, os usuários de sistemas de busca geralmente não olham além da primeira página de resultados.

Um quadro contendo um resumo desses três estudos se encontra na página 49.

Estudos de *query logs* analisam basicamente as consultas realizadas pelos usuários na *web* de forma quantitativa. No entanto, existe um outro comportamento de busca que não é medido quantitativamente e tem sido alvo de diversas pesquisas recentes. Esse

[21] http://www.excite.com/

[22] http://www.altavista.com

[23] http://www.fireball.de

[24] http:// www.alltheweb.com

comportamento é denominado comportamento de busca exploratória. Abordado recentemente em duas renomadas conferências internacionais: SIGIR2006[25] e CHI2007[26] foi também o tema principal da publicação *Communications of the ACM*, da *Association for Computing Machinery* de abril de 2006.

Cerca de 20 a 30% de todas as consultas realizadas *na web* são exploratórias por natureza (Rose & Levinson, 2004) e nessa dissertação é apresentada uma proposta que integra uma nuvem de texto a uma interface de sistema de busca cujo objetivo principal é auxiliar os usuários, em buscas exploratórias, a construir suas consultas oferecendo palavras-chave que possam complementar sua consulta inicial.

Além desse comportamento, esse capitulo aborda a importância do contexto nos resultados de busca na *web.* Quando os usuários iniciam uma busca exploratória a falta de uma visão geral dos resultados é particularmente problemática.

Os resultados de todos esses estudos servem de embasamento e justificam a proposta apresentada por essa dissertação.

5.1. Principais pesquisas realizadas

5.1.1. Os estudos do Excite

Os estudos mais conhecidos sobre o comportamento dos usuários utilizando sistemas de busca foram conduzidos por Jansen et al. (2000), Spink et al. (2001, 2002) e Wolfram et al. (2001). Em todos eles foram analisados os *query logs* do sistema de busca Excite.

Query logs são os registros de todas as consultas, realizadas em um sistema de busca em um determinado período. Uma consulta pode ser composta por uma palavra ou um conjunto de palavras. Ela é computada cada vez que o usuário insere uma ou mais palavras no campo de busca e submete a consulta ao sistema apertando o botão destinado a esse fim, com o objetivo de obter uma lista de resultados.

O primeiro desses estudos foi realizado por Jansen et al. (2000) e teve como objeto de análise os *query logs* do período de um dia de março de 1997. Foram analisadas 51.473 consultas de 18.113 usuários. Esse estudo pioneiro foi o responsável por indicar muitas das peculiaridades que diferenciam uma busca na *web* das buscas realizadas nos sistemas tradicionais de recuperação de informação.

Entre as diferenças indicadas podemos destacar que as consultas realizadas nos sistemas de busca na *web* geralmente são vagas e as sessões curtas. O uso de recursos avançados

[25] Conference on Research & Development on Information Retrieval

[26] Computer-Human Interaction Conference

é mínimo assim como o uso dos operadores lógicos. Uma sessão de busca é composta por uma ou várias consultas, realizadas entre o momento em que o sistema é iniciado e fechado.

O tamanho médio das consultas, relatado na pesquisa, foi de apenas 2,21 palavras, com menos de 4% das consultas contendo seis palavras ou mais. Foi constatado que os usuários utilizavam muito pouco os recursos avançados de busca do sistema, assim como os operadores lógicos ou outros modificadores de consultas. O uso desses recursos apareceu em menos de 18% das consultas. Paralelo a isso, as sessões de busca realizadas foram curtas com uma média de 2,8 consultas por sessão. O número médio de páginas de resultados que foram visitadas foi de apenas 2,35 por consulta e mais de 50% dos usuários não acessaram resultados além da primeira página.

Foi constatado também que algumas poucas palavras foram consultadas com muita freqüência e as demais formaram uma longa lista de termos pouco utilizados. Por exemplo, de um total de 21.862 termos utilizados apenas uma vez por usuários diferentes, 9.790 apareceram apenas uma única vez nas consultas em geral.

Um segundo estudo envolvendo os *query logs* do Excite foi realizado por Spink et al. (2001). Foi adotada uma metodologia semelhante à adotada no primeiro estudo, só que dessa vez foram analisadas 1.025.910 consultas de 211.063 usuários realizadas em setembro de 1997.

Os resultados, em alguns aspectos, foram consistentes com os do estudo anterior, como por exemplo, a média de palavras por consulta que ficou em 2,16. No entanto, outros comportamentos se alteraram. Houve um aumento no número de usuários que foram além da primeira página de resultados retornadas pelo sistema (71%) e houve uma redução no uso de recursos avançados, com o uso de operadores lógicos aparecendo em menos de 5% das consultas.

O terceiro estudo dessa série foi realizado por Wolfram et al. (2001). Foi baseado na análise de mais de um milhão de consultas realizadas por 200.000 usuários do Excite no ano de 1999. Esse estudo mostrou poucas variações estatísticas com relação ao estudo inicial feito com os dados de 1997. As consultas continuaram tendo duas palavras em média e o uso de recursos avançados continuou sendo raro, em cerca de 8% das consultas.

No entanto, foi observado um aumento nas consultas por apenas uma palavra. Nos estudos com os dados de 1997 as consultas com apenas um termo correspondiam a 48% do total. Nesse estudo utilizando os dados de 1999 as consultas envolvendo apenas uma palavra atingiram um total de 60%. E ainda, 43% dos usuários não se aventuraram além da primeira página de resultados.

Esse dado vai de encontro aos resultados do estudo de março de 1997, onde aproximadamente 50% dos usuários não foram além da primeira página. Já os resultados do estudo de setembro apontaram uma freqüência menor, de 29%.

Um último estudo com os *query logs* do Excite realizado por Spink et al. (2002) analisou 1.025.910 consultas de 262.015 sessões de busca de mais de 200.000 usuários realizadas em abril de 2001.

Esse estudo também mostrou poucas mudanças com relação aos resultados dos estudos anteriores indicando uma constância no comportamento dos usuários do sistema de busca. Houve um pequeno aumento no número médio de palavras por consulta 2,6 e um pequeno aumento no percentual de usuários vendo apenas uma única página de resultados, 50%.

5.1.2. Os estudos do Altavista

Em 1998, o Altavista era o sistema de busca na *web* mais utilizado, e um estudo conduzido por Silverstein et al. (1999), baseado nos *query logs* de 43 dias de acesso entre agosto e setembro daquele mesmo ano indicou resultados que confirmaram as tendências dos estudos com o sistema de busca Excite. A média de 2,35 palavras por consulta, menos de 15% das buscas levando os usuários a se aventurar além da primeira página de resultados e o uso de recursos avançados de busca, sendo utilizados em apenas 20% das consultas.

Um estudo de acompanhamento, avaliando ainda o Altavista, foi realizado por Jansen et al. (2005). Foi baseado na análise de 1.073.388 consultas de 369.350 sessões de busca coletadas durante um período de 24 horas em setembro de 2002. Esse estudo indicou um pequeno aumento no tamanho das consultas, para 2,92 palavras em média e um aumento significante na freqüência com que os usuários olharam além da primeira página de resultados. De 15% em 1998 para 27% em 2002.

No estudo de 1998, 77% das sessões de busca realizadas eram constituídas de apenas uma consulta. Em 2002, mais de 50% das sessões continham duas ou mais consultas com os usuários modificando a consulta inicial em 52% das vezes. Em 1998 a modificação das consultas originais acontecia em 20% dos casos.

5.1.3. Os estudos do Fireball e AlltheWeb

Outros estudos realizados com dois grandes sistemas de busca europeus, Fireball e AlltheWeb puderam avaliar os padrões de busca entre os usuários de uma comunidade restrita.

O estudo do Fireball usou dados de 31 dias do mês de janeiro de 1998. Cobriu mais de 16 milhões de consultas e três milhões de páginas na *web* (Hoscher e Strube 2000). Os resultados indicaram uma média de palavras por consulta de 1,66. Essa média é menor do que as obtidas nos estudos do Excite e do Altavista. Quanto à utilização de recursos avançados de busca e o número de páginas de resultados visitadas a pesquisa apresentou resultados compatíveis com os obtidos nos outros estudos. Mais de 97% das consultas realizadas no Fireball não utilizaram nenhum tipo de operador e 59% dos usuários focaram suas atenções apenas na primeira página de resultados.

O estudo do sistema de busca AlltheWeb, realizado por Jansen e Spink (2005) foi baseado em dois conjuntos de dados, um de fevereiro de 2001 e o outro de maio de 2002. Cada conjunto de dados foi baseado em períodos de 24 horas de uso e cobriram aproximadamente 200.000 usuários e um milhão de consultas.

Esse estudo destacou um declínio na média de palavras por consulta do ano de 2001 para o ano de 2002, com um aumento nas consultas compostas por apenas uma palavra de 25% em 2001 para 33% em 2002. A duração das sessões também diminuiu.

5.1.4. Resumo dos estudos

Em resumo, as principais constatações desses estudos sobre o comportamento de busca na *web* foram:

1. As consultas tendem a ser curtas e logo são potencialmente vagas e ambíguas; a maioria dos estudos reporta consultas com em média duas ou três palavras.

2. Os usuários raramente utilizam as ferramentas de busca avançadas, como por exemplo, os operadores booleanos para aumentar a precisão das consultas. Essas facilidades são utilizadas em menos de 20% das consultas, e em estudos mais recentes, em menos de 10%.

3. Os usuários de sistemas de busca geralmente não olham além da primeira página de resultados de uma consulta.

4. Apesar de esses estudos não serem muito recentes, eles avaliaram um longo período, e foi constatado que não houve muita mudança no comportamento dos usuários.

Tabela 3 – Quadro resumo com os dados dos estudos de *query logs* apresentados no capítulo 5.

QUADRO RESUMO DOS ESTUDOS									
Sistema de busca	Excite	Excite	Excite	Excite	Altavista	Altavista	Fireball	AlltheWeb	
Ano de publicação do estudo	2000	2001	2001	2002	1999	2005	2000	2005	
Data da coleta dos dados	Abr 1997	Set 1997	1999	2001	1998	2002	1998	2001	2002
N°de consultas analisadas	51.473	1.025.910	1.025.910	1.025.910	993.208.159	1.073.388	16 milhões	451.551	957.303
N°médio de palavras por consulta	2,21	2,16	2,4	2,6	2,35	2,92	1,66	2,4	2,3
Percentagem de uso de recursos avançados e operadores lógicos nas consultas	18%	5%	8%	10%	20%	20%	3%	1%	1%
Percentagem de usuários que visitaram apenas a primeira página de resultados	50%	29%	43%	50%	85%	73%	59%	83%	76%

5.2. Comportamento de busca exploratória

Os estudos de *query logs* do Excite, Altavista, Fireball e AlltheWeb fazem uma análise quantitativa das consultas realizadas na *web* pelos usuários. Eles analisam exclusivamente o número das palavras-chave utilizadas nas buscas nesses sistemas. No entanto, existem outros comportamentos de busca que não estão diretamente relacionados ao uso de palavras-chave nas consultas e não são medidos quantitativamente.

Um desses comportamentos é o denominado comportamento de busca exploratória. O termo busca exploratória foi cunhado por pesquisadores das áreas de Recuperação da Informação, Interação Humano-Computador e Visualização da Informação para denominar um comportamento de busca bastante freqüente que é descrito abaixo (Ryen W. White, Bill Kules e Benjamin B. Bederson, 2005).

De acordo com Bates, M. J. (1989), os sistemas de busca funcionam bem quando as necessidades de informação dos usuários são bem definidas. No entanto, eles não funcionam tão bem nas situações em que os usuários não possuem o conhecimento necessário ou o domínio do contexto para formular as suas consultas. São nessas situações que os usuários iniciam as buscas exploratórias.

Estudos identificaram que os usuários geralmente desenvolvem estratégias para lidar e compensar essas situações onde as necessidades de informação são vagas (Baldonado, M. Q. W. e Winograd, T., 1997; Bates, M. J., 1989; Pirolli, P. e Card, S., 1995). Por exemplo, os usuários submetem uma consulta tentativa e passam a navegar a partir dos resultados retornados contando apenas com sua habilidade em interpretar pistas contextuais e navegar entre documentos. Resumindo, exploram a informação disponível procurando seletivamente e obtendo passivamente pistas sobre qual o próximo passo a ser seguido. O objetivo da exploração de informações é o refinamento de uma necessidade vaga de informação que através da interação com diversas fontes de informação leva a um entendimento maior do problema.

Um estudo realizado por Teevan et al. (2004) também mostrou que ao invés de ir direto ao assunto através das consultas por palavras-chave, os usuários de sistemas de busca na *web* navegam até o assunto de interesse através de pequenos passos. Esses passos são localizados, e são dados apoiados no conhecimento contextualizado desses usuários, mesmo quando eles sabem exatamente o que estão procurando.

Esse comportamento foi bastante comum principalmente entre os participantes da pesquisa que tinham pouco conhecimento a respeito da estrutura da informação consultada.

Algumas vantagens foram observadas nesse comportamento de busca realizado através de pequenos passos. Dessa forma os usuários não precisam especificar muito, o que estão procurando, e também conseguem observar os resultados em um contexto que permite que ele seja compreendido de uma maneira mais clara.

Segundo Marchioni (2006), essa estratégia envolve várias consultas e uma pesquisa exploratória e interativa nos resultados obtidos. Torna necessária uma escolha seletiva de um caminho a ser percorrido de forma a se obter pistas sobre os passos seguintes. Por esse motivo, esse comportamento foi denominado busca exploratória e compreende uma mistura de sorte, aprendizagem e investigação.

Interfaces de sistemas que visam auxiliar esse processo de busca exploratória já vêm sendo desenvolvidas há algum tempo (Hearst, 2000) e continuam sendo alvo de discussões e estudos recentes. O que mostra que ainda existem oportunidades nas áreas de pesquisa e de desenvolvimento para melhorar as interfaces de busca atuais.

5.3. A importância do contexto nos resultados de busca na *web*

Os sistemas de busca são bastante eficientes na geração de listas de resultados relevantes. E quando os usuários realizam consultas por temas conhecidos, geralmente encontram o que procuram na primeira página de resultados.

Mas no caso de buscas exploratórias mais sofisticadas uma lista de resultados pode não ser a forma mais eficiente para auxiliar essa tarefa (White, Kules, Drucker, & schraefel, 2006).

Apesar da dificuldade de se quantificar a ocorrência de buscas de caráter exploratório, análises recentes sobre o objetivo das consultas sugerem que cerca de 20 a 30% de todas as consultas realizadas *na web* são exploratórias por natureza (Rose & Levinson, 2004).

Quando os usuários iniciam uma busca exploratória a falta de uma visão geral dos resultados é particularmente problemática (Baldonado, M. Q. W. e Winograd, T., 1997; Bates, M. J., 1989; Pirolli, P. e Card, S., 1995; Teevan et al., 2004).

A fim de proporcionar uma visão geral dos resultados em contexto, alguns estudos vêm sendo realizados investigando o uso de categorias semânticas geradas de forma automática para organizar as várias páginas de resultados retornadas pelos sistemas de busca.

Dentre esses estudos podemos destacar o realizado por Dumais, Cutrell, & Chen (2001). Nesse estudo foram desenvolvidas e avaliadas sete interfaces que organizaram os resultados apresentados pelos sistemas de busca tradicionais em categorias. Os resultados continuaram sendo listados segundo a ordem de relevância do sistema. No entanto, eles foram agrupados visualmente mostrando os itens de uma mesma categoria juntos.

Essas interfaces foram comparadas com as interfaces tradicionais onde os resultados são apresentados em listas. E em todas as comparações, as interfaces organizadas por categorias, foram mais eficientes. Nelas os usuários encontraram os resultados procurados de forma mais rápida.

O Clusty[27] é um exemplo de sistema de busca que apresenta os resultados organizados em categorias.

Figura 18 - Sistema de busca Clusty. Resultados organizados em categorias para a consulta "cars". No destaque aparecem as categorias apresentadas. (www.clusty.com)

5.4. Visualização gráfica de resultados de busca na *web*

O sucesso dos resultados apresentados por esses estudos abriu caminho para o desenvolvimento de uma nova área de pesquisa. Recentemente surgiram sistemas de

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

busca que apresentam formas de visualização gráfica de seus resultados, que por sua vez são organizados em categorias semânticas.

Essas formas de visualização têm sempre o objetivo de mostrar uma visão geral dos resultados das consultas e o contexto desses resultados.

Um desses sistemas de busca que apresenta a visualização dos seus resultados é o Grokker[28]. O Grokker agrupa os resultados em uma hierarquia e apresenta as categorias em um diagrama de Euler. Os círculos maiores representam as categorias principais e os menores as subcategorias desses grupos. Os usuários podem explorar os resultados através de um *zoom* nas categorias.

Figura 19 - Visualização dos resultados de busca mostrados pelo Grokker. (www.grokker.com)

O KartOOVISU[29], na sua versão beta, é um sistema de meta-busca, que apresenta os resultados obtidos em forma de mapa conceitual e permite visualizar as relações semânticas que interligam as informações.

[28] http://www.grokker.com
[29] http://beta.kvisu.com/

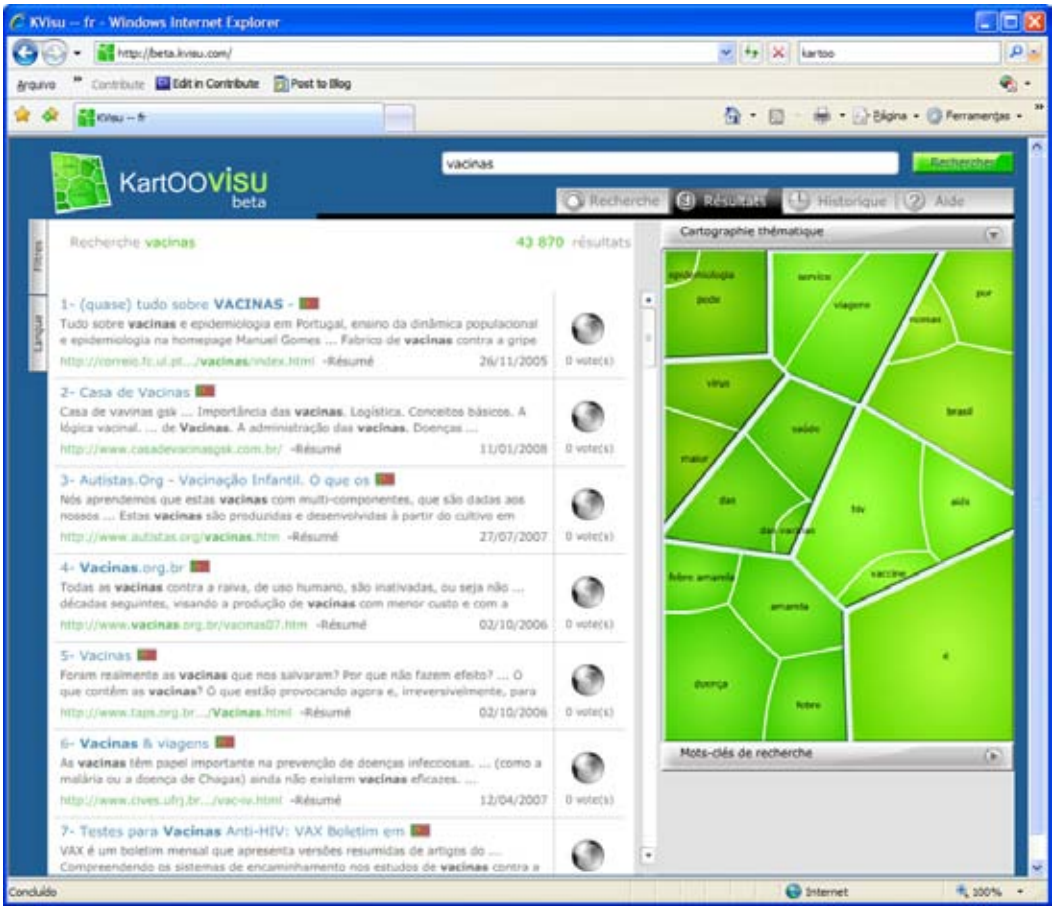


Figura 20 - Visualização dos resultados de busca mostrados pelo KartOOVISU. (beta.kvisu.com)

O sistema de busca Quintura[30], também em sua versão beta, é a proposta visual que mais se assemelha a aplicação desenvolvida nessa pesquisa.

O Quintura apresenta um mapa visual semântico. Esse mapa mostra as palavras que são contextualmente relacionadas com as palavras utilizadas na consulta e é construído a partir de resumos retornados por um sistema de busca externo. As palavras são dispostas segundo a sua proximidade semântica.

Em 2007 o Quintura foi apontado como o Sistema de Busca Alternativo do Ano pelo site especializado AltSearchEngines[31].

[30] http://www,quintura.com

[31] http://www.altsearchengines.com/

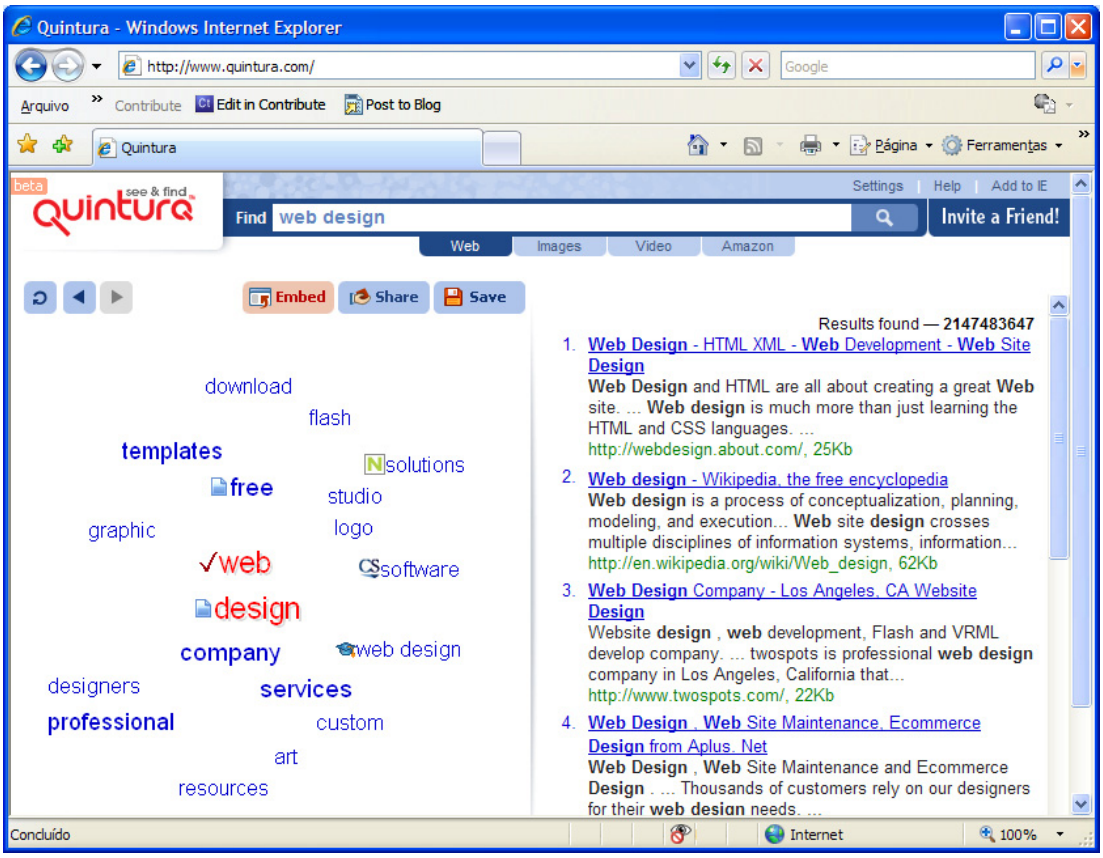


Figura 21 - Visualização dos resultados de busca mostrados pelo Quintura. (www.quintura.com)

Resultados de estudos comparativos entre a apresentação textual e a apresentação gráfica de resultados de busca, variaram em função da tarefa desempenhada pelos usuários. Algumas tarefas favoreceram as interfaces gráficas enquanto outras, as textuais (Sebrechts et al., 1999; e Becks e Minkenbererg, 2002).

Na pesquisa conduzida por Rivadeneira e Bederson (2003) não foram constatadas diferenças de eficiência na comparação entre as duas interfaces, no entanto, o nível de satisfação e aceitação foi mais alto com relação à interface gráfica.

Rivadeneira e Bederson (2003) ressaltam que o design de diferentes representações visuais para mostrar os resultados de busca é desafiador.

Todos esses sistemas de busca visuais ainda se encontram em suas versões beta e estão ainda em fase de aprimoramento. Todavia, essa nova geração de sistemas de busca abre caminho para novas experimentações como a apresentada por essa pesquisa.

A aplicação proposta por essa pesquisa apresenta os resultados de um sistema de busca em contexto, pois esse é um benefício intrínseco das técnicas de visualização de

informações. Porém nesse momento a nossa abordagem é puramente estatística. Apesar de bastante promissora, a categorização automática de conteúdos é bastante complexa e nem tudo está resolvido ainda.

Alguns problemas observados na visualização gráfica dos resultados dos sistemas citados muitas vezes vêm de uma categorização mal estruturada e não da visualização propriamente dita.

Vocabulário controlado e ontologias

É importante ressaltar que além das pesquisas envolvendo a categorização automática de resultados de busca, outras pesquisas vêm sendo realizadas envolvendo o uso de vocabulário controlado associado à recuperação de informações na *web*.

Um vocabulário controlado é um tipo de sistema de classificação que define os termos mais utilizados sobre num determinado assunto e suas relações. Esses relacionamentos podem ser semânticos e também conceituais. É usado também para acabar com a ambigüidade entre termos de múltiplos significados. Sua função é manter a consistência entre o conteúdo e servir como referência para as diferentes entidades que alimentam o banco de dados desse conteúdo. Um vocabulário controlado também pode conter um dicionário de sinônimos e um glossário.

A chamada *web* semântica é baseada na construção de vocabulários controlados chamados de ontologias.

As aplicações mais imediatas para *web* semântica visam categorizar informação e aumentar a qualidade do resultado das ferramentas de busca através de resolução de ambigüidade e contextualização da informação.

Em Ciência da Computação e Ciência da Informação, uma ontologia é um modelo de dados que representa um conjunto de conceitos dentro de um domínio e os relacionamentos entre estes.

A *web* semântica é uma extensão da *web* atual, que permitirá aos computadores e humanos trabalharem em cooperação[32]. A *web* semântica interliga significados de palavras e, neste âmbito, tem como finalidade conseguir atribuir um significado, um sentido aos conteúdos publicados na internet de modo que seja perceptível tanto pelo humano como pelo computador.

A idéia da *web* semântica surgiu em 2001, quando Tim Berners-Lee, James Hendler e Ora Lassila publicaram um artigo na revista Scientific American, intitulado: “Web Semântica: um novo formato de conteúdo para a Web que tem significado para computadores vai iniciar uma revolução de novas possibilidades”.

[32] http://www.w3.org/2001/sw/SW-FAQ#What1

Apesar de promissora essa anunciada revolução tem-se mostrado lenta e bastante complexa para ser implementada de forma abrangente. Os principais desafios estão na construção, verificação, evolução e integração de ontologias que precisam ser automatizadas e interativas e, também, na geração das anotações semânticas para descrever e recuperar os arquivos publicados na web.

6. A aplicação proposta

> "Projete uma interface homem-máquina de acordo com as habilidades e as fraquezas da humanidade, e você terá ajudado o usuário não apenas a realizar uma tarefa, mas a ser alguém mais feliz e produtivo."
>
> Jef Raskin, 2000 (Lupton, 2006)

6.1. A visualização de resultados de sistemas de busca em nuvens de texto

A presente dissertação é o resultado de uma pesquisa que avalia as vantagens da utilização de uma forma de visualização de informações para apresentar os resultados de um sistema de busca na *web*. A forma de visualização escolhida para essa avaliação foi a nuvem de texto.

Apesar de bastante popular na *web*, essa forma de visualização ainda não apresenta estudos sobre a sua utilização na visualização de resultados de sistemas de busca abertos.

A hipótese dessa investigação é que:

A visualização dos resultados de um sistema de busca em uma nuvem de texto pode auxiliar os usuários a encontrar o que procuram facilitando a construção de consultas em buscas exploratórias.

Para testar essa hipótese foi necessária a construção de uma aplicação que permite a visualização de resultados de um sistema de busca em nuvem de texto.

6.2. A teoria por trás da construção da nuvem na aplicação

Em uma consulta na *web*, os resultados aparecem listados em diversas páginas. Através de uma nuvem de texto integrada a um sistema de busca é possível a visualização de uma síntese, de um resumo automático, do conteúdo dos resultados listados em várias páginas sem que elas tenham que ser percorridas e os *sites* acessados individualmente.

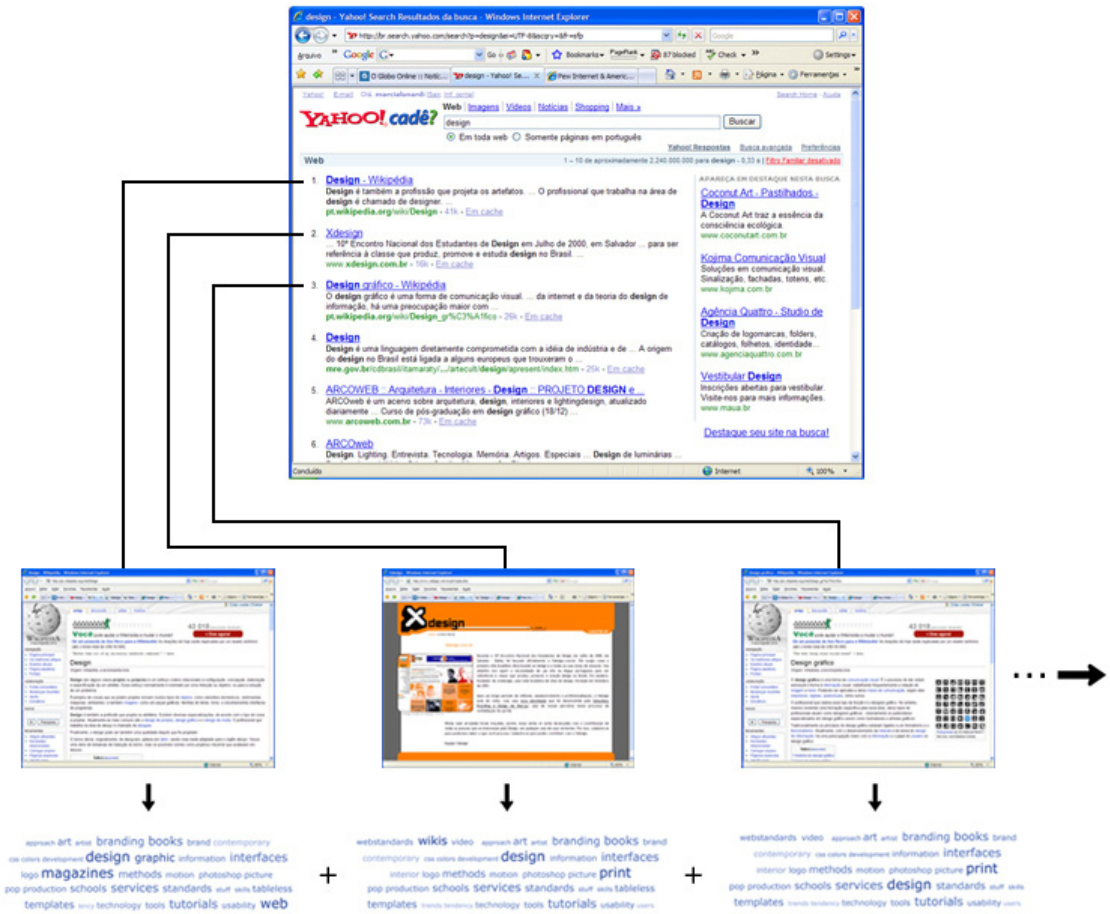


Figura 22 - Construção da nuvem de texto dos resultados.

A nuvem de texto nesse contexto funciona como uma ferramenta auxiliar para que o usuário possa gerenciar a grande carga de informação que é disponibilizada nos resultados das consultas, e ainda, as palavras que compõem a nuvem, podem ser utilizadas como palavras-chave adicionais para complementar uma consulta inicial.

A nuvem de textos de um resultado de busca é formada a partir de uma consulta inicial. Cada resultado corresponde a uma página. A partir do conteúdo de cada página, é construída uma nuvem de texto com suas palavras mais freqüentes.

A nuvem dos resultados corresponde à soma das nuvens das principais páginas listadas pelo sistema de busca segundo seu algoritmo de relevância.

Se uma nuvem de texto de uma página na *web* é uma espécie de resumo do conteúdo dessa página, uma nuvem de texto construída a partir de várias páginas de um resultado de

busca, somadas, permite que se tenha uma idéia geral do resultado daquela consulta sem que seja necessário percorrer as páginas listadas. Dessa forma, também é possível se perceber pelo contexto em que a palavra está inserida, se ela foi a escolha mais adequada para uma determinada consulta.

Todas as páginas dos resultados apresentam em comum a palavra utilizada na consulta inicial. Essa palavra, em cada página de resultados está inserida em um contexto diferente. Dentro de cada contexto diferente, estarão presentes palavras que fazem parte daquele universo semântico específico. Várias páginas podem também estar relacionadas com um mesmo contexto, apresentando palavras repetidas.

Uma vez somada essas nuvens, ou seja, somadas as freqüências das palavras das nuvens formadas a partir dos textos das primeiras páginas de resultados, temos uma nuvem que representa o universo dos principais resultados retornados pelo sistema de busca.

A nuvem de resultados não representa uma intercessão de palavras comuns. Apenas a palavra utilizada na consulta inicial é que certamente estará presente em todas as páginas. A nuvem apresenta as palavras mais freqüentes que resultam da soma de várias nuvens. Uma única página de resultados pode apresentar uma mesma palavra repetida diversas vezes ao longo do texto fazendo com que essa palavra apareça na nuvem de resultados.

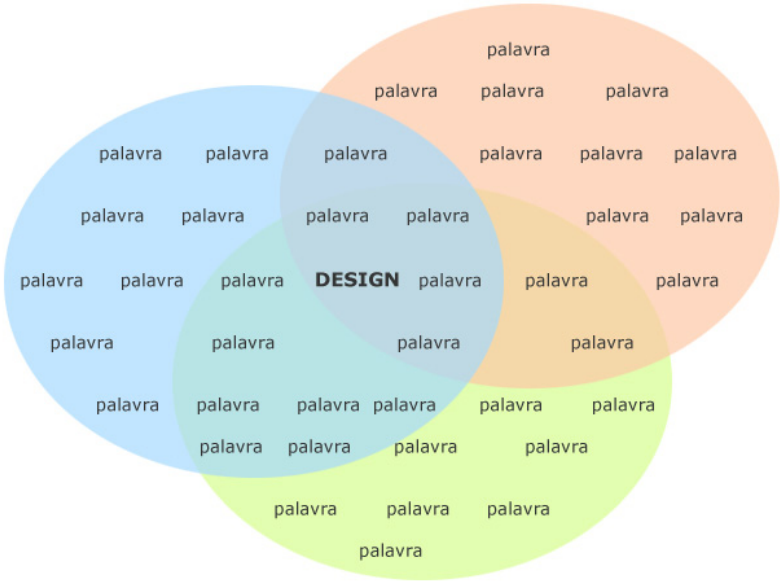


Figura 23 - Construção de uma nuvem a partir de três páginas de resultado para a consulta “DESIGN”.

A palavra design certamente estará presente na nuvem, no entanto, outras palavras com alta freqüência, mesmo que estejam em apenas uma das páginas do resultado, farão parte da composição da nuvem.

Por exemplo, em uma busca pela palavra DESIGN, possivelmente a nuvem de texto conterá palavras como COR, GRÁFICO, LOGOMARCA, WEB, entre outras. Isso, porque essas palavras aparecem com freqüência em textos que contém a palavra design.

Eventualmente, alguns textos vão conter a palavra DESIGN em outro contexto. Por exemplo, “DESIGN de estruturas frigoríficas”. A nuvem de textos dessa página provavelmente vai conter palavras como SUÍNOCULTURA, RAÇÃO, RESFRIAMENTO, e assim por diante.

No entanto, como a aplicação proposta gera uma nuvem que é baseada em um volume grande de textos, a freqüência dessas palavras fora de contexto acaba sendo diluída.

A aplicação proposta por essa pesquisa tem uma abordagem puramente estatística.

No entanto, é importante ressaltar que existem linhas de pesquisa envolvendo a categorização automática de resultados de busca, e outras ainda, envolvendo o uso de vocabulário controlado associado à recuperação de informações na *web* como foi mencionado no capítulo 5.

6.3. O funcionamento da aplicação

Para testar a hipótese dessa pesquisa foi desenvolvida uma aplicação funcional utilizando o sistema de busca Yahoo[33].

Quando o usuário digita uma palavra no campo de busca do sistema e pressiona a barra de espaço, automaticamente, uma nuvem de texto aparece na página principal.

Essa nuvem de texto é construída a partir do conteúdo, das 40 primeiras páginas que seriam listadas nos resultados do Yahoo para aquela consulta.

Para gerar essa nuvem de textos, a aplicação desenvolvida varre o conteúdo das 40 primeiras páginas listadas pelo Yahoo. Extrai o texto da linguagem de marcação e demais códigos de programação, e cria um índice com as palavras dos textos e o número de vezes que elas aparecem.

Dessa listagem inicial são eliminadas palavras freqüentes da língua portuguesa, como artigos e preposições, que não agregam valor na diferenciação entre um tema e outro em uma nuvem.

A nuvem de textos que é gerada quando uma palavra é digitada no campo de busca do sistema é um resumo dos resultados mais relevantes segundo o Yahoo para aquela

[33] http://www.yahoo.com

consulta. Entretanto, esse resumo aparece antes mesmo que a consulta seja submetida em definitivo.

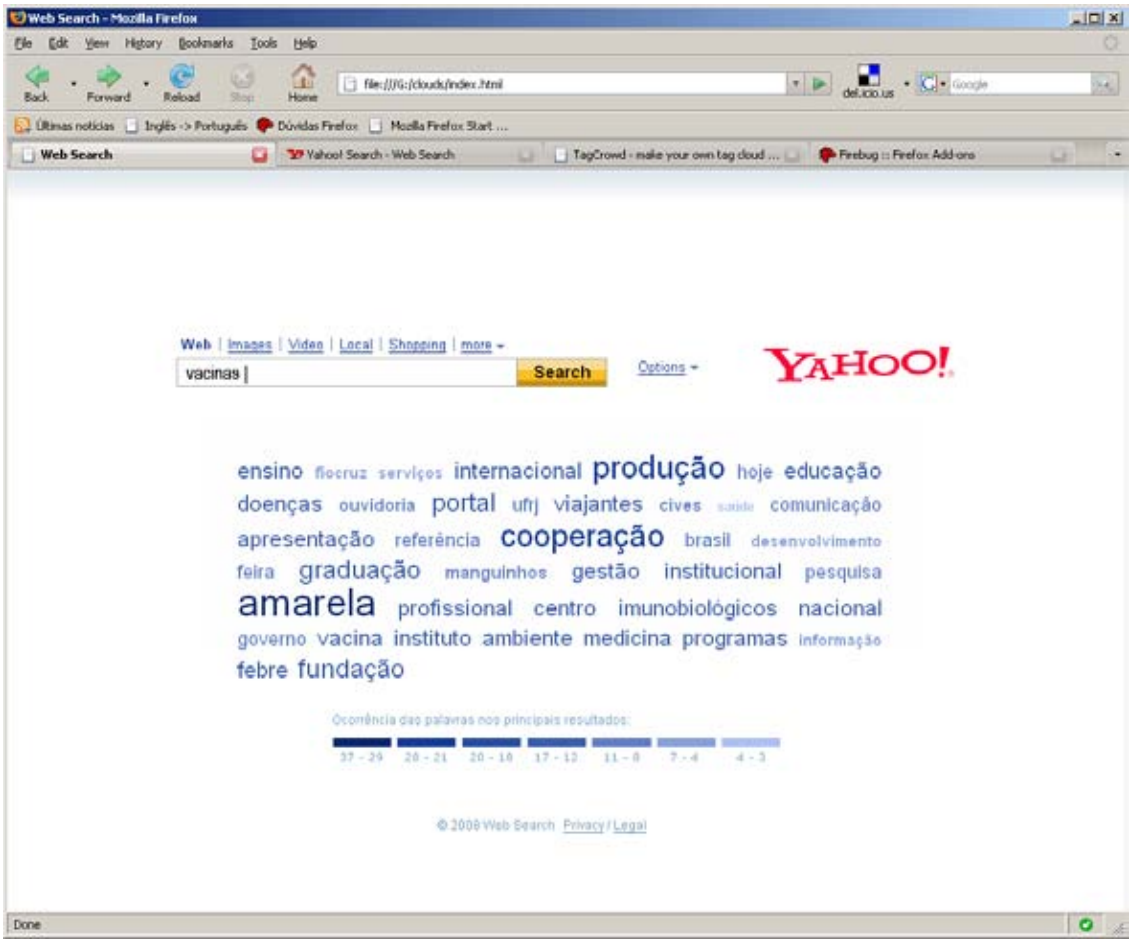


Figura 24 - A nuvem gerada pela aplicação na interface do Yahoo.

A partir desse resumo, o usuário pode obter sugestões de palavras-chave para complementar a consulta inicial e, então, submetê-la ao sistema, ou ainda, apenas pressionar a barra de espaço para obter uma nova nuvem. Para adicionar uma palavra da nuvem à consulta inicial, basta clicar sobre a palavra desejada para que esta automaticamente seja incluída no campo de busca ao lado da palavra inicial. O número de palavras que podem ser adicionadas ao campo de busca é ilimitado.

Cada vez que uma palavra nova é adicionada à consulta, e a barra de espaço é pressionada, a nuvem é reformulada. Nesse caso, o sistema retorna páginas que contêm individualmente todas as palavras inseridas no campo de busca e uma nova nuvem é gerada a partir dos resultados listados para essa nova consulta.

6.4. A construção gráfica da nuvem de texto gerada pela aplicação

> "Integrar fisicamente forma e conteúdo talvez tenha sido o impulso mais persistente da arte e do design do século XX. Os poetas dadaístas e futuristas, por exemplo, usaram a tipografia para criar textos cujo conteúdo era inseparável do leiaute concreto de letras específicas em uma página.
>
> No século XXI, forma e conteúdo voltaram a ser separados. Folhas de estilo, por exemplo, impelem os designers a pensarem global e sistematicamente ao invés de se concentrarem na construção fixa de uma superfície particular. Esse modo de pensar permite que o conteúdo seja reformatado para dispositivos e usuários diversos e o prepara para a sua pós-vida, enquanto os meios de estocagem eletrônica iniciam seus ciclos de decadência e obsolescência." (Lupton, 2006)

Na aplicação desenvolvida para essa pesquisa, seguimos essa tendência. Nela, forma e conteúdo são separados. O conteúdo da nuvem vem de uma base de dados e muda a cada consulta. A forma é toda pré-definida em uma única folha de estilo[34] e tem que ser pensada para todas as possíveis variações que esse conteúdo possa apresentar.

6.4.1. Lei de potência

Um padrão observado em nuvens de texto e *tag clouds* é conhecido como lei de potência. Uma lei de potência implica que pequenas ocorrências são muito comuns ao contrário de grandes instâncias que são muito raras apesar de existirem.

Um exemplo típico é a distribuição de renda em nosso país, onde muitos possuem muito pouco e raros são aqueles indivíduos que detêm a maior parte da riqueza. Outro exemplo é a ocorrência de vários terremotos de pequena intensidade e poucos de grande intensidade.

Este comportamento algumas vezes é chamado de lei de Zipf, outras vezes é chamado de lei de Pareto e outras tantas simplesmente de lei de potência.

Os três termos são utilizados para descrever fenômenos em que grandes eventos são raros, assim como pequenos eventos são muito comuns.

[34] Folha de estilo, de *cascading style sheet*, ou simplesmente CSS, é uma linguagem utilizada para definir a apresentação de documentos escritos em uma linguagem de marcação, como HTML. Seu principal benefício é justamente prover a separação entre o formato e o conteúdo de um documento.

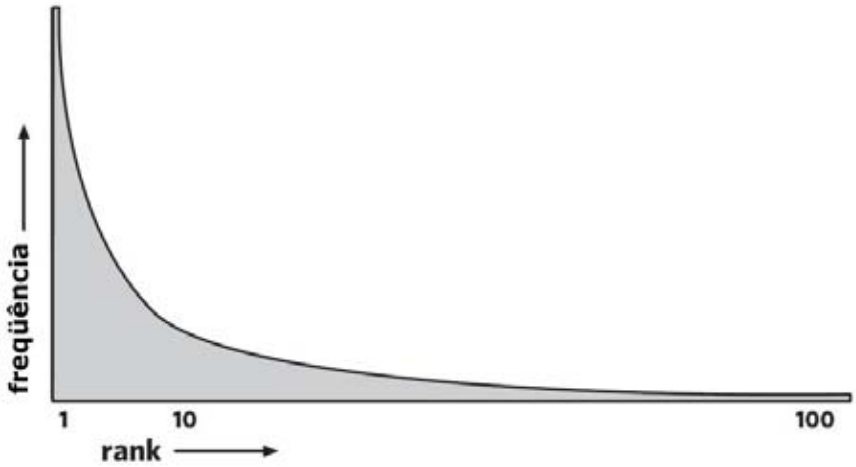


Figura 25 - Típica curva de uma lei de potência.

A lei de Zipf (1932) aborda especificamente freqüência de palavras. Em seu trabalho inicial de 1932, George Kingsley Zipf, professor de lingüística de Harvard, apresentou um estudo em que determinava a freqüência do uso das palavras e, assim, construiu um *rank* onde em primeiro lugar estava a palavra mais utilizada, em segundo aquela imediatamente após, e assim por diante. Sua lei afirmava que a enésima maior ocorrência de uma palavra era inversamente proporcional ao seu *rank*. Por sua vez, Vilfredo Pareto, 1848-1923, em seu trabalho estava interessado na distribuição de renda na Itália em 1906.

O predomínio da lei de potência influencia em muitas das escolhas relacionadas ao design da nuvem. Por exemplo, a alta freqüência de algumas poucas palavras é preciso ser compensada.

Na construção de uma nuvem de texto a primeira decisão a ser tomada é a de quantas palavras irão fazer parte da sua composição. Essa decisão está relacionada com o espaço destinado na interface para a exibição da nuvem. Optamos por mostrar 40 palavras com alinhamento justificado. Destinamos uma área de 650 pixels de largura por 600 pixels de altura, aproximadamente, para a exibição da nuvem. A altura é variável, pois o número de linhas varia em função do tamanho das palavras.

Seguindo a lei de potência, a lista de palavras mais freqüentes de um resultado de busca na *web* tem muitas palavras com freqüências baixas. Logo, optamos por definir uma freqüência mínima para que a palavra fosse incluída na nuvem. A freqüência mínima adotada foi três.

6.4.2. A escala da nuvem

Se a escala de tamanho das fontes das palavras em uma nuvem fosse baseada unicamente nas suas freqüências, algumas palavras iriam aparecer absurdamente grandes e outras extremamente pequenas.

A figura 26, segundo Smith (2007), mostra a escala da nuvem de *tags* mais populares do site Flickr[35].

Nesse caso, é possível ver o tamanho de três *tags* polulares: "*wedding*", "*architecture*", e "*rome*" na nuvem e o número respectivo de fotos (a freqüência) de cada uma delas. A terceira linha, *to-scale size*, mostra o tamanho da *tag* diretamente proporcional ao número de fotos. O tamanho da *tag* "*rome*" foi definido como tamanho mínimo na escala.

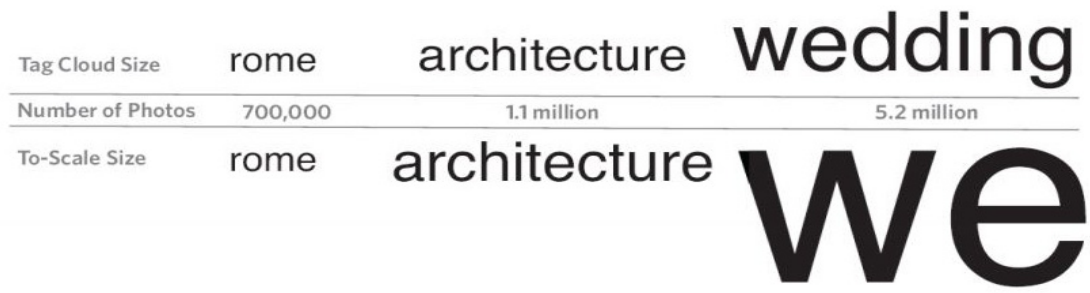


Figura 26 - A escala da nuvem das *tags* mais populares do Flickr. A escala é ajustada a fim de garantir a leiturabilidade da nuvem. (retirado de Smith, 2007)

A escolha da escala é uma questão estética e matemática e é preciso chegar a uma fórmula que crie a melhor nuvem para atender ao seu propósito.

Em uma nuvem de texto típica, que segue a curva da lei de potência, é preciso chegar a um equilíbrio entre a legibilidade das palavras de fonte menor e a precisão dos tamanhos de fonte selecionados para refletir as freqüências. Nesse caso, a leiturabilidade[36] da nuvem deve ser priorizada.

A escala adotada para a nuvem de texto na aplicação foi tamanho de fonte de 12 pixels para o patamar mínimo e tamanho de 36 pixels para o patamar máximo. O equilíbrio entre o tamanho máximo de fonte e mínimo deve ser observado para que as palavras com o tamanho máximo não sobreponham as do tamanho mínimo. A entrelinha, distância da linha de base de uma linha tipográfica para outra, também é de extrema importância na composição desse equilíbrio, e para essa aplicação foi adotada entrelinha de 36 pixels.

[35] http://www.flickr.com/

[36] Legibilidade, do inglês *legibility* em tipografia está associada ao tipo, ao alfabeto; é uma característica intrínseca deste. Diz-se que é a facilidade com que reconhecemos e distinguimos as letras, os sinais etc. Algumas letras podem causar confusão, como o **b** e o **h**, **l** e **I**, **C** e **G** ou **c** e **e**, por exemplo, e assim apresentarem legibilidade ruim em determinados tipos.

Leiturabilidade, do inglês *readability* está associado ao tipo composto, ou seja, ao alfabeto em uso, considerando-se seu tamanho, alinhamento, largura da coluna, espaçamento entre as linhas e entre as letras, cores, layout da página etc. É a capacidade de ler, adequadamente, palavras e textos. Aplica-se tanto a textos curtos como longos, embora seja um termo normalmente associado ao conforto visual, à ergonomia de textos mais longos.

Uma vez definidos os patamares máximo e mínimo da escala, o passo seguinte na construção de uma nuvem é a definição do número de patamares para essa escala. Para a nuvem da aplicação foi definido o número de sete patamares. Para os sete patamares foi associada uma escala de valor de azul conforme detalha a tabela 4.

Tamanho da fonte em pixels	Tamanho da fonte Arial	Escala de valor de azul	Valores hexadecimais
12 13	palavra palavra		# ACC1F3
14 15	palavra palavra		# 86A0DC
16 18	palavra palavra		# 607EC5
21	palavra		# 395CAE
24	palavra		# 264CA2
30	palavra		# 133B97
36	palavra		# 072771

Tabela 4 - Escala adotada na nuvem da aplicação.

A distribuição das freqüências entre esses patamares pode ser feita segundo vários métodos. Para a construção de uma nuvem os métodos de distribuição mais relevantes são o proporcional e o linear.

Escala proporcional

Na escala proporcional, o tamanho das palavras[37] é diretamente proporcional a sua freqüência. Logo, é feita uma distribuição entre a freqüência máxima e a freqüência mínima pelo número de patamares escolhido para a nuvem. Essa distribuição gera uma nuvem que

[37] Nesse capitulo, para fins didáticos a expressão “tamanho da palavra” é utilizada genericamente para se referir ao tamanho da fonte tipográfica da palavra.

reflete a curva típica de lei de potência vista na figura 25, com poucas palavras de tamanho grande, e muitas de tamanho pequeno.

Escala linear

Outro método para a distribuição das freqüências é a escala linear. Nesse método a curva de lei de potência é transformada em uma reta através de uma função logarítmica. Essa mudança faz com que a diferença entre a maior e a menor palavra passe a ser linear em vez de exponencial.

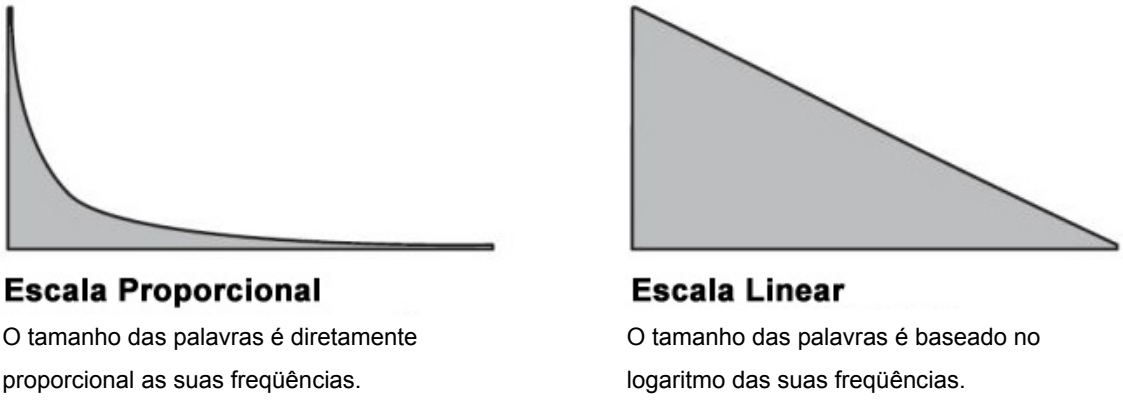


Figura 27 - A escala proporcional pode resultar em poucas palavras grandes e muitas pequenas. A escala linear levanta o meio da distribuição suavizando as diferenças.

A tabela abaixo mostra uma lista de palavras com suas respectivas freqüências. O tamanho direto é o tamanho atribuído segundo uma escala proporcional. O tamanho linear é o tamanho atribuído para essas palavras segundo a função logarítmica da escala linear. Como pode ser visto na última coluna, as palavras com freqüência moderada ficam maiores usando a escala linear.

Palavra	Freqüência	Tamanho (direto)	LOG	Tamanho (linear)
Design	120	48	4.78	48
Web2.0	43	24	3.76	38
Internet	34	22	3.52	36
Cultura	12	15	2.48	26
Colaboração	4	12	1.39	15

Tabela 5 - Comparação do tamanho das palavras segundo as escalas proporcional e linear.

A diferença entre uma nuvem de texto com distribuição proporcional e uma com distribuição linear pode ser visualizada nas figuras 28 e 29. Ambas foram geradas a partir da mesma lista de palavras e freqüências da tabela 6.

Palavra	Freqüência	Palavra	Freqüência	Palavra	Freqüência
design	120	gaming	23	science	5
ux	68	google	14	lists	4
ia	65	tv	14	innovation	3
socialsoftware	54	culture	12	miscellaneous	3
tags	46	comix	7	complexity	3
web2.0	43	statistics	6	facets	3
business	34	art	5	networks	3

Tabela 6 - Lista de palavras e respectivas freqüências que serviu de base para a geração das nuvens com distribuição proporcional e linear nas figuras 28 e 29.

Figura 28 - Nuvem de texto com distribuição proporcional.

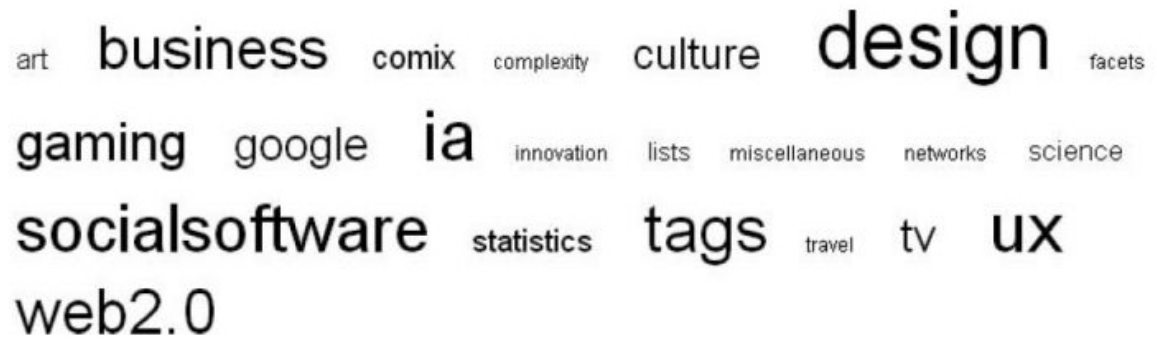


Figura 29 - Nuvem de texto com distribuição linear.

De acordo com Smith (2007), a distribuição linear gera nuvens mais atrativas e mais fáceis de ler. Outro benefício aparece nos casos em que a nuvem oferece algum tipo de interação ou é usada como mecanismo de navegação. Nesse caso, as palavras ficam maiores e são

mais fáceis de serem clicadas. No entanto, se a precisão for fundamental para a nuvem em questão deve-se considerar a distribuição proporcional.

Como a lista de palavras mais freqüentes em um resultado de busca segue a lei de potência, optamos por adotar a distribuição linear na escala da nuvem gerada pelo aplicativo desenvolvido.

A fonte digital escolhida para a nuvem da aplicação foi a Arial. A Arial é uma fonte sem serifa projetada em 1982 por Robin Nicholas e Patricia Saunders. A fonte Arial foi escolhida por sua boa legibilidade e, também, por ela fazer parte do sistema operacional Windows, uma vez que a aplicação desenvolvida utilizar as fontes do sistema operacional do computador do usuário. Na ausência da fonte Arial o aplicativo utiliza automaticamente a fonte Verdana como segunda opção. A Verdana é uma fonte sem serifa projetada por Matthew Carter especialmente para a tela e, também, é distribuída com o sistema operacional Windows.

6.5. O desenvolvimento da aplicação

> "Os designers devem conhecer melhor os serviços *online* e pensar além da aparência do *site*, buscar a implementação de APIs e organização de conteúdo, que não devem ser tratados como blocos desvinculados e autônomos, mas sim pensados como um todo." (Meirelles e Moura, 2007)

Como o foco principal dessa dissertação está no design da informação e na construção da nuvem na interface gráfica da aplicação, e não o seu desenvolvimento propriamente dito, suas etapas serão descritas de forma breve, apenas para que se tenha um entendimento geral do que foi feito e das tecnologias e *softwares* utilizados.

A aplicação para essa pesquisa foi desenvolvida na linguagem de programação Java[38] a partir de um serviço do sistema de busca Yahoo, disponibilizado livremente e de forma gratuita e, também, é considerada um serviço.

Um serviço *web* do inglês *web service* é uma aplicação modular que pode ser utilizada através da Internet. Os consumidores desses serviços são outras aplicações que se comunicam, normalmente através do protocolo HTTP, usando padrões de XML para trocar dados e informações.

O serviço desenvolvido nessa pesquisa faz o *download* do conteúdo das 40 primeiras páginas do *set* de resultados do Yahoo para uma dada consulta, para o disco rígido do

[38]Java é uma linguagem de programação orientada a objeto desenvolvida pela empresa Sun Microsystems.

computador e monta uma nuvem de textos, logo abaixo do campo de busca do sistema assim que uma palavra é digitada no mesmo.

Essa nuvem de textos apresenta as palavras retornadas por outra aplicação, externa ao Yahoo. O serviço de montagem da nuvem de textos funciona como uma busca dentro de outra base de dados indexada, que utiliza o Apache Lucene[39].

O Apache Lucene é um *software* de busca que possui uma *Application Programming Interface* (API) de indexação de documentos. É um *software* de código aberto da Apache Software Foundation também escrito em Java. O Lucene contém apenas o núcleo de um sistema de busca. Para o Lucene não importa a origem dos dados, seu formato, ou mesmo a linguagem em que foi escrito, desde que esses dados possam ser convertido para texto. Isto significa que o Lucene pode ser utilizado para indexar e buscar dados gravados em arquivos, páginas *web* em servidores remotos, documentos gravados no sistema de arquivos local, arquivos textos ou arquivos PDF, ou qualquer outro formato do qual possa ser extraído informação textual.

Para a aplicação desenvolvida nessa pesquisa, os dados indexados no Lucene são os arquivos em HTML e na língua portuguesa do *set* de resultados do Yahoo, que são copiados para o disco rígido do computador por ocasião da consulta.

A aplicação utiliza esse índice gerado pelo Lucene para montar uma lista de palavras e suas respectivas ocorrências. Essa lista é filtrada para que algumas palavras como artigos e preposições sejam retiradas.

Finalmente essa lista é tabulada no formato de uma nuvem de textos. Essa nuvem de textos é construída através de Javascript e formatada por um arquivo CSS.

Javascript é uma linguagem de programação que foi criada para atender principalmente as necessidades de interação com páginas na *web*.

Cascading Style Sheets, ou simplesmente CSS, é uma linguagem de estilo utilizada para definir a apresentação de documentos escritos em uma linguagem de marcação, como HTML. Seu principal benefício é prover a separação entre o formato e o conteúdo de um documento.

Logo, a formatação da nuvem de textos na aplicação não está definida dentro do documento, ela foi definida em um arquivo externo. Basicamente, nesse arquivo foi definido o design da interface da aplicação. Elementos como tamanho e alinhamento da nuvem de textos, entrelinhas, fontes digitais utilizadas, cores, e escala de tamanho das palavras mostradas.

[39] http://lucene.apache.org/

A nuvem gerada pela aplicação dessa pesquisa é exibida em uma cópia da tela do Yahoo que acessa suas funcionalidades, no entanto ela pode vir a ser distribuida como um *plugin* para o Yahoo. Um acessório que complementa as suas funcionalidades.

Um *plugin* é um programa de computador que interage com uma aplicação principal. Um navegador, por exemplo, e permite que certas funções, geralmente específicas, sejam executadas sob demanda.

Aplicações em geral suportam *plugins* permitindo que desenvolvedores autônomos criem novas funcionalidades.

7. Delineamento da pesquisa

7.1. Tema

Cruz e Ribeiro (2004) afirmam que selecionar um tema significa encontrar um objeto de estudo que mereça ser investigado cientificamente e que tenha condições de ser formulado e delimitado em função da pesquisa. De acordo com essa afirmação, essa pesquisa de mestrado apresenta o seguinte tema: a visualização de resultados de sistema de busca na *web* em nuvem de texto.

7.2. Problema

Pesquisas demonstram que os usuários de sistemas de busca utilizam em média apenas duas ou três palavras-chave por consulta e, na maioria das vezes, não olham além da primeira página de resultados. Os recursos de pesquisa avançada que possibilitam um refinamento das consultas também não são utilizados.

Os sistemas de busca atendem bem ao seu propósito quando os usuários têm objetivos bem definidos. No entanto, quando os usuários têm conhecimentos reduzidos sobre o assunto da consulta, objetivos pouco claros ou muito complexos, eles não sabem que palavra-chave utilizar. Geralmente, para contornar esse problema adotam estratégias que envolvem utilizar palavras-chave genéricas como uma tentativa de se obter nos resultados palavras-chave mais específicas para uma nova consulta (Baldonado, M. Q. W. e Winograd, T., 1997; Bates, M. J., 1989; e, Pirolli, P. e Card, S., 1995).

Esse tipo de estratégia utilizada pelos usuários demanda tempo e esforço cognitivo para se chegar ao objetivo pretendido e tem recebido especial atenção por parte da comunidade científica. Foco de diversos estudos, esse comportamento é classificado como busca exploratória (Marchioni, 2006).

Análises recentes sobre o objetivo das consultas sugerem que cerca de 20 a 30% de todas as consultas realizadas *na web* são exploratórias por natureza (Rose & Levinson, 2004).

Diante dessa demanda por sistemas de busca que atendam melhor os usuários nessas situações, novas propostas vêm sendo investigadas. Uma linha de estudos tem-se

concentrado na categorização automática dos resultados. Essa abordagem ainda requer aprimoramentos, mas demonstrou ser vantajosa pelo fato de proporcionar a visualização de um conjunto maior de resultados por página e também por mostrar esses resultados em contexto.

Técnicas de visualização de informações atendem bem esses dois propósitos, e partindo dessa constatação, surgiram novos sistemas de busca, com propostas ainda embrionárias, que utilizam técnicas de visualização de informações para mostrar graficamente os seus resultados.

Apesar de essas tentativas não se mostrarem ainda satisfatórias, elas apontam para um novo caminho a ser explorado. O da visualização de resultados de sistemas de busca.

A presente dissertação é o resultado de uma pesquisa que avalia as vantagens da utilização de uma forma de visualização de informações para apresentar os resultados de um sistema de busca na *web*. A forma de visualização escolhida para essa avaliação foi a nuvem de texto. Apesar de bastante popular na *web*, essa forma de visualização ainda não apresenta estudos sobre a sua utilização na visualização de resultados de sistemas de busca.

7.3. Hipótese

De acordo com Santos (2002), a hipótese é uma verdade provisória "fundamental para qualquer processo de investigação científica, pois consiste no lançamento de uma afirmação a respeito de algo ainda desconhecido ou, pelo menos, não satisfatoriamente conhecido". Marconi e Lakatos (2000) também afirmam que a hipótese constitui-se em uma "suposta, provável e provisória resposta a um problema, cuja adequação (comprovação = sustentabilidade ou validez) será verificada através da pesquisa".

A hipótese dessa investigação é que:

A visualização dos resultados de um sistema de busca em uma nuvem de texto pode auxiliar os usuários a encontrar o que procuram facilitando a construção de consultas em buscas exploratórias.

7.4. Metodologia da pesquisa

A pesquisa relatada nessa dissertação foi desenvolvida em duas fases. A primeira fase consistiu na construção da aplicação, que gera uma nuvem de texto a partir dos resultados do sistema de busca Yahoo, relatada no capítulo 6.

A segunda fase da pesquisa foi a avaliação da aplicação com a finalidade de testar a hipótese dessa dissertação.

Para testar a hipótese dessa dissertação optamos pelos métodos empíricos, que são métodos de avaliação que envolvem a utilização de participantes (Jordan 1998).

Dentre os métodos empíricos conhecidos decidimos realizar dois tipos de avaliação. Primeiro uma avaliação cooperativa, e em seguida um experimento controlado. Ambos os métodos de avaliação foram aplicados com o mesmo grupo de participantes.

A avaliação cooperativa foi realiza com dois propósitos. Primeiro para uma avaliação inicial da compreensão do aplicativo da nuvem de texto pelos participantes, e também, com o objetivo de familiarizar esse participante com o aplicativo para o experimento controlado.

No experimento controlado, foi feita uma comparação entre o Yahoo padrão, que serviu como controle, e o Yahoo com a aplicação da nuvem de texto. As mesmas consultas foram realizadas nos dois sistemas com o propósito de verificar possíveis vantagens agregadas pela nuvem nesse contexto.

7.4.1. Seleção dos participantes

Para as avaliações realizadas por essa pesquisa foram selecionados dez participantes que foram divididos em dois grupos homogêneos. Esses dez participantes foram selecionados a partir de uma entrevista semi-estruturada (ANEXO I), que foi aplicada com 18 voluntários de ambos os sexos e faixa-etária variada.

A entrevista foi dividida em duas partes. A primeira identificou o perfil de uso da internet dos entrevistados. Através dessas perguntas foram eliminados os entrevistados que utilizavam ferramentas de busca avançada.

A segunda parte da entrevista avaliou o grau de *expertise* dos entrevistados com relação aos temas abordados nas tarefas da pesquisa.

Foram eliminados os entrevistados que detinham possíveis conhecimentos avançados a respeito de qualquer um dos três assuntos. Culinária regional brasileira, vacinas infantis e cidades gaúchas, respectivamente.

A partir desse questionário oito entrevistados foram descartados. Os entrevistados que não foram eliminados foram divididos em dois grupos de constituição semelhante para a realização do experimento controlado. O objetivo dessa entrevista foi selecionar uma amostra de usuários ditos “comuns”. O principal aspecto observado na seleção dos participantes foi a possibilidade de dividi-los em grupos homogêneos, uma vez que as mesmas tarefas seriam desempenhadas pelos grupos e comparadas. Dessa forma, também

foram consideradas na elaboração dos grupos a mesma proporção na distribuição da formação acadêmica, idade, e sexo dos participantes.

7.4.2. Avaliação cooperativa

De acordo com Monk et al. (1993), em uma avaliação cooperativa os participantes desempenham tarefas previamente formuladas e, juntos, com o pesquisador, avaliam a usabilidade de um determinado sistema. Os participantes são encorajados a perguntar sobre o processo de interação com o sistema e o pesquisador faz perguntas sobre o entendimento do participante em relação ao mesmo. Isso faz com que o procedimento pareça bastante natural para o participante e exige menos recursos que outros métodos de teste mais formais.

Uma de suas principais vantagens, que foi decisiva para a escolha desse método, é a possibilidade de se trabalhar com protótipos e simulações parciais, caso do aplicativo avaliado nessa pesquisa. Segundo Monk et al. (1993), recomenda-se a avaliação cooperativa para produtos que necessitam de aprimoramento técnico, para protótipos em um estágio intermediário ou para protótipos funcionando em sua plenitude.

Esse método, além da facilidade de utilização, também promove o máximo de *feedback* sobre como o projeto deve ser reformulado.

Como a utilização de nuvem de texto para a visualização de resultados de sistemas de busca é um conceito novo, a avaliação cooperativa foi realizada com todos os dez participantes selecionados para a pesquisa. Como esses mesmos dez participantes fizeram parte do experimento controlado, essa avaliação também serviu para familiarizá-los previamente com o aplicativo.

Principais passos da avaliação cooperativa:

Com os participantes da pesquisa pré-selecionados foram definidas duas tarefas para serem executadas pelos mesmos. Foi pedido aos participantes que respondessem duas questões exploratórias utilizando a aplicação desenvolvida, questão 1, sobre uma fruta nativa brasileira e a questão 2, sobre vacinas infantis (ANEXO III).

Cada tarefa foi explicada pelo pesquisador, oralmente, e a questão foi entregue, impressa, em um cartão.

Todos os participantes realizaram as tarefas no mesmo ambiente físico, utilizando o mesmo computador, com as mesmas configurações, e a mesma velocidade de conexão com a internet.

Todas as sessões foram filmadas.

Durante a execução das tarefas os participantes foram encorajados a narrar o que faziam e a explicar os passos que tomavam. Durante as sessões, os participantes também foram questionados ativamente a respeito das suas intenções e expectativas a fim de se coletar informações qualitativas a respeito de possíveis dúvidas e dificuldades quanto à utilização do aplicativo.

7.4.3. Experimento controlado

Para o experimento controlado, foram utilizados os mesmos dez participantes da avaliação cooperativa. Na avaliação cooperativa eles foram tratados como um grupo só, sem distinção. No entanto, na seleção prévia desses participantes eles foram distribuídos em dois grupos homogêneos, grupo A e grupo B.

Como tarefa, todos os participantes tiveram que responder outras duas questões de caráter exploratório utilizando o sistema de busca Yahoo. A questão 3, sobre cidades gaúchas, e a questão 4, sobre culinária regional brasileira (ANEXO IV).

Para cada questão só existia uma resposta correta que poderia ser obtida em diferentes *sites* e a partir de palavras-chave variadas.

Os 5 participantes do grupo A, responderam a questão 3 utilizando o sistema de busca Yahoo padrão e a questão 4 utilizando a versão do Yahoo com o aplicativo da nuvem de texto.

Os 5 participantes do grupo B, responderam a questão 4 utilizando o sistema de busca Yahoo padrão e a questão 3 utilizando a versão do Yahoo com o aplicativo da nuvem de texto.

Dessa forma, cada questão foi respondida 10 vezes, 5 em cada versão do Yahoo.

A questão 3 (sobre cidades gaúchas) foi respondida 5 vezes no Yahoo padrão e 5 vezes no aplicativo da nuvem de texto.

E a questão 4 (sobre culinária regional brasileira) foi respondida 5 vezes no Yahoo padrão e 5 vezes no aplicativo da nuvem de texto.

O Yahoo padrão foi utilizado no experimento como variável de controle. Com essa metodologia adotada foi possível comparar a mesma questão sendo respondida com e sem o auxílio da nuvem de texto, variável independente.

Nesse experimento, os participantes receberam as instruções por escrito (ANEXO II) e as questões foram entregues em cartões separados (ANEXO IV).

Nesse experimento não houve a interferência do pesquisador. Todos os participantes tiveram a oportunidade de se familiarizar previamente com o funcionamento da aplicação da nuvem de texto durante a avaliação cooperativa.

Como na avaliação cooperativa, todos os participantes realizaram a tarefa no mesmo ambiente físico, utilizando o mesmo computador, com as mesmas configurações, e a mesma velocidade de conexão com a internet.

Todas as sessões foram filmadas.

Esse experimento controlado verificou especificamente:

1. Se os usuários utilizaram palavras da nuvem de texto para refinar suas consultas.

2. E nos casos positivos, se as buscas foram concluídas de forma mais rápida e satisfatória com o auxílio da nuvem comparadas às buscas realizadas no controle.

Foi medido o tempo levado para a conclusão de todas as questões e também o número de páginas acessadas até a obtenção da resposta correta.

Para medir o grau de satisfação dos participantes foi aplicado um questionário com perguntas abertas no final da sessão (ANEXO VI).

8. Resultados

Nesse capítulo são relatados os resultados das avaliações realizadas com a aplicação desenvolvida, as conclusões gerais dessa pesquisa e futuros trabalhos.

Alguns aspectos observados no experimento controlado são relatados junto com os resultados da avaliação cooperativa quando reforçam o que está sendo abordado. Também são relatados junto com os resultados da avaliação cooperativa alguns aspectos qualitativos que foram coletados via formulário após o experimento controlado.

Durante a aplicação da pesquisa, alguns participantes, pré-selecionados pela entrevista, foram eliminados por conhecerem a resposta de uma ou das duas questões propostas como tarefa. Nesse caso, novos participantes foram selecionados para recompor os grupos.

8.1. Resultados da avaliação cooperativa

Quanto à utilização da aplicação e execução das tarefas

Durante a avaliação cooperativa nenhum participante apresentou dificuldades na utilização da aplicação. Entretanto, mesmo após as explicações sobre o seu funcionamento, metade dos usuários apresentou uma tendência a apertar a tecla *enter*, em vez de espaço, com o intuito de gerar uma nova nuvem, após a inserção do último termo da consulta. Essa forma de interação é um do comportamentos padrão para se submeter uma consulta a um sistema de busca e sugere que essa questão específica seja revista com relação à usabilidade da aplicação.

Quanto a execução das tarefas, foi observado que os participantes com menor formação acadêmica apresentaram mais dificuldades na construção das consultas e na localização dos resultados. Muitas dessas dificuldades vêm da má interpretação das questões e da falta de conhecimentos gerais. Mas de maneira geral as tarefas foram concluídas sem problemas.

Quanto à aceitação da aplicação

A aceitação da aplicação é uma medida subjetiva da satisfação dos participantes e foi obtida a partir de comentários expontâneos, feitos por estes, durante o seu uso. A aceitação foi muito boa por todos os participantes. Os participantes utilizaram adjetivos como “muito interessante” e “legal” várias vezes durante as sessões e mesmo sendo informados sobre a possibilidade de utilizar a busca padrão do Yahoo a qualquer momento, todos insistiram em utilizar a nuvem.

Um participante resaltou que a aplicação da nuvem deve ser vista como um valor adicional do sistema de busca: “mesmo não sendo útil para todas as consultas sempre é mais uma ferramenta disponível da qual se pode tirar proveito”.

Aspectos específicos observados

Já a partir das primeiras consultas ficou bastante evidente que a tarefa influência diretamente na utilidade e na eficácia da nuvem.

As nuvens ajudaram nos casos onde as respostas eram objetivas e compostas por apenas um termo, caso da questão 1 (ANEXO III), cuja resposta era CERRADO e também da questão 4 (ANEXO IV), do experimento controlado, cuja resposta era PARÁ. Nesses casos a resposta apareceu na primeira nuvem para a maioria das consultas formuladas.

Os participantes avaliaram que as nuvens foram bastante úteis nesses casos tornando as buscas mais objetivas e a identificação dos resultados mais fácil. Também mencionaram como positivo o fato da nuvem apresentar não só os resutado de forma resumida mas também informações adicionais sobre o assunto, se referindo ao contexto.

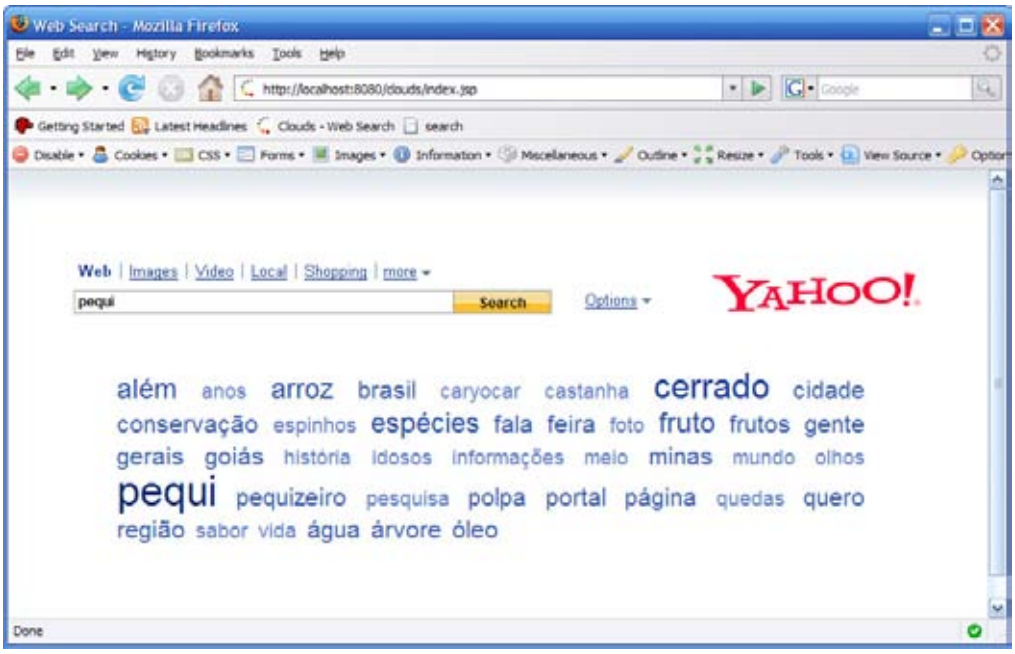


Figura 30 - Nuvem de resultados para a consulta pelo termo PEQUI.

A palavra PARÁ, resposta da questão 4 (ANEXO IV), levou mais tempo para ser localizada na nuvem, confirmando que palavras mais curtas ficam menos evidentes em meio a outras palavras, mesmo quando o tamanho da fonte da palavra em questão é um dos maiores da escala. A questão da posição da palavra na nuvem também pode ter influenciado na sua localização, entretanto esse aspecto requer uma investigação específica.

Figura 31 - Nuvem de resultados para a consulta pelo termo TACACÁ.

No caso dessa questão, mais de um participante, novamente, enfatizou o benefício da nuvem em oferecer o contexto associado ao TACACÁ, desconhecido para eles. Na nuvem, os participantes puderam identificar não só a região e o estado desse prato típico, como também, seus ingredientes.

Foi identificado no discurso dos participantes que as nuvens ajudaram a diminuir o esforço cognitivo nas buscas através de frases como: “é mais fácil”, “é mais direto”, “é mais resumido”, “está tudo em uma página”, “não tenho que ficar navegando para achar o que quero”.

No caso da questão 2 (ANEXO III), sobre vacinas infantis, foi constatado que a resposta por ser complexa jamais apareceria na nuvem. A nuvem no máximo poderia sugerir novas palavras para serem adicionadas a consulta inicial. No entanto, os participantes não perceberam isso de imediato o que causou uma certa frustração após a formulação de várias nuvens.

Mesmo informados sobre a possibilidade de utilizar a busca padrão do Yahoo a qualquer momento, todos os participantes insistiram em utilizar a nuvem. Essa insistência permitiu que outro aspecto fosse observado. Por ser uma busca mais complexa os usuários foram reformulando e adicionando mais palavras-chave a consulta inicial. Apesar disso as nuvens

geradas a partir das consultas com três ou mais palavras-chave apresentaram menos contexto e foram consideradas de pior qualidade.

Contudo, de maneira geral, os participantes foram unânimes em informar que as nuvens no mínimo ajudaram sugerindo palavras-chaves para refinar ou reformular suas consultas.

Conhecimento prévio

O aspecto mais significante que foi observado durante a avaliação cooperativa foi a associação do conhecimento prévio dos participantes com a utilização da aplicação.

Alguns participantes não apresentavam nenhum conhecimento prévio, mínimo requerido, para a obtenção da resposta das questões. Nesses casos, mesmo a resposta aparecendo na nuvem ela não foi identificada. Por exemplo, para responder a questão 1 (ANEXO III), que perguntava qual é a vegetação típica da fruta pequi, o participante deveria saber identificar previamente no mínimo alguns tipos de vegetação e saber por conseqüência que cerrado era um deles.

Já nos casos em que os participantes tinham apenas uma vaga idéia sobre a resposta, a nuvem já ajudou, pois foi possível a partir das palavras sugeridas fazer algumas associações. Um exemplo desse tipo de associação foi visto durante o experimento controlado e relatado posteriormente pelo participante. O participante que estava respondendo a questão 3 (ANEXO IV) viu a palavra GONÇALVES e respondeu “Gonçalves Dias” pois teve uma vaga lembrança de uma série de televisão que se passava naquela cidade do sul do Brasil. Em seguida viu a palavra BENTO e corrigiu sua resposta.

Figura 32 - Nuvem de resultados para a consulta pelos termos VALE DOS VINHEDOS RS.

Essa questão também permitiu que fosse observado mais um aspecto. Como a aplicação trata as palavras e suas respectivas freqüências individualmente, as palavras compostas aparecem separadas na nuvem.

A resposta para essa questão era BENTO GONÇALVES, e o resultado apareceu na primeira nuvem na maior parte das consultas, no entanto, os termos BENTO e GONÇALVES apareceram separados, tratados individualmente. Os participantes localizaram a resposta, mas para isso se apoiaram no seu conhecimento prévio despertado pela associação das duas palavras.

Outro aspecto observado foi o fato de nomes próprios, normalmente iniciados por letra maiúscula, aparecerem nas nuvens com letra minúscula dificultando sua identificação.

Outras consultas livres

Durante a avaliação cooperativa, alguns participantes manifestaram interesse em realizar outras consultas cujos temas eram áreas específicas de seu conhecimento. Essas consultas foram incentivadas e permitiram a coleta de informações adicionais.

Por se tratar de assuntos de interesse dos usuários eles identificaram termos relacionados com a busca inicial nas nuvens que aparentemente não tinham nenhuma relação com o assunto.

Como exemplo podemos citar a consulta realizada por um dos participantes pelo termo BANANA. Nessa nuvem puderam ser identificados dois universos, o da banana fruta, daí as palavras associadas EMBRAPA, BANANEIRA, FRUTA, DOCE, NANICA, RECEITA, BRASIL, etc. E o universo do carnaval associado à banda Chiclete com Banana, daí as palavras CHICLETE, CARNAVAL, MÚSICA, DISCO, etc.

Esse participante reformulou sua consulta adicionando a palavra CHICLETE da nuvem, refinando o contexto da sua consulta para sua área de interesse e obteve uma nova nuvem com palavras como: BANDA, BLOCO, SALVADOR, BAIANA, TRIO entre outras.

Nessa nuvem, um leigo sobre o assunto não associaria as palavras BELL, MARQUES, NANA, CAMALEÃO ao carnaval e nem à banda Chiclete com Banana, no entanto, o participante identificou relações em todas essas palavras. Informou que BELL e MARQUES são cantores, NANA se refere ao bloco NANA BANANA e identificou CAMALEÃO como sendo outro bloco do carnaval baiano.

Esse reconhecimento do contexto de um assunto conhecido se repetiu com outros participantes e gerou uma identificação positiva e satisfação em utilizar a aplicação.

Figura 33 - Nuvem de resultados para a consulta pelos termos BANANA CHICLETE.

Por outro lado, essas consultas livres reforçaram que a eficácia da nuvem depende da tarefa realizada. Um participante testou duas consultas: CAPITAL BRASIL e CAPITAL PAÍS acreditando que a resposta Brasília apareceria em evidência. Todavia, a resposta sequer apareceu na nuvem. Essa consulta permitiu a seguinte análise: nos casos onde as palavras são associadas a diversos assuntos, ou seja a diferentes universos semânticos, as nuvens não apresentam nenhum contexto específico. Nesse caso específico, as palavras CAPITAL, BRASIL e PAÍS são utilizadas em diversos contextos diferentes como por exemplo, mercado de capitais, capital financeiro, etc.

Outras considerações

Alguns participantes de forma isolada fizeram outras considerações relevantes.

Um participante sugeriu como melhoria a possibilidade de se selecionar ou excluir da nuvem um determinado universo semântico.

Outro, mencionou que a repetição de determinadas palavras que aparecem no singular e no plural limita o número de novas palavras diferentes na nuvem. Caso de VINHO e VINHOS e VINÍCOLA e VINÍCOLAS na nuvem da figura 32, por exemplo.

Foi mencionado também que essa ferramenta poderia ser útil para crianças indicando um novo caminho de investigação para essa pesquisa.

8.2. Resultados do experimento controlado

O experimento controlado verificou especificamente dois pontos: primeiro se os usuários utilizaram palavras da nuvem de texto para refinar suas consultas, e nos casos positivos, se as buscas foram concluídas de forma mais rápida e satisfatória com o auxílio da nuvem comparadas às buscas realizadas no Yahoo padrão.

A utilização das palavras da nuvem de texto para refinar as consultas só pode ser mensurada nesse experimento nos casos em que as palavras foram adicionadas à consulta através da interação direta dos participantes com a nuvem.

Nas consultas realizadas foi observado que 3 participantes utilizaram um total de 5 palavras das nuvens de resultados interagindo com a aplicação (ANEXO VII). Além disso, no questionário qualitativo aplicado após o experimento outros participantes mencionaram que também utilizaram palavras-chave da nuvem para complementar suas consultas sem, no entanto, interagir com a aplicação. Nesses casos o participante viu a palavra na nuvem e digitou essa palavra no campo de busca do sistema.

Esse número não é considerado baixo se for observado que as respostas foram obtidas na primeira nuvem de resultados pela maioria dos participantes sem que fosse necessária sua reformulação (ANEXO VII).

O tempo levado para a conclusão de todas as questões e também o número de páginas acessadas até a obtenção da resposta correta foram medidos.

Tempo de execução das tarefas

Comparando o tempo de execução das mesmas tarefas sendo realizadas com o auxílio das nuvens e no Yahoo padrão, foi observado que a questão 4 (ANEXO IV), foi respondida mais rápido por 60% dos participantes no Yahoo padrão contra 40% com o auxílio das nuvens. Já a questão 3 foi respondida em menos tempo no Yahoo padrão por todos os participantes.

Tabela comparativa do tempo de execução da Questão 3 (Bento Gonçalves)		
	Yahoo Nuvem	**Yahoo Padrão**
1	4:10 min	2:00 min
2	19:00 min	4:30 min
3	5:00 min	2:30 min
4	6:10 min	1:50 min
5	4:20 min	3:20 min
Tempo médio	**7:44 min**	**2:50 min**

Tabela 7 - Tempo de execução da questão 3 no Yahoo Nuvem e no Yahoo Padrão.

Gráfico do tempo de execução da Questão 3 (Bento Gonçalves)

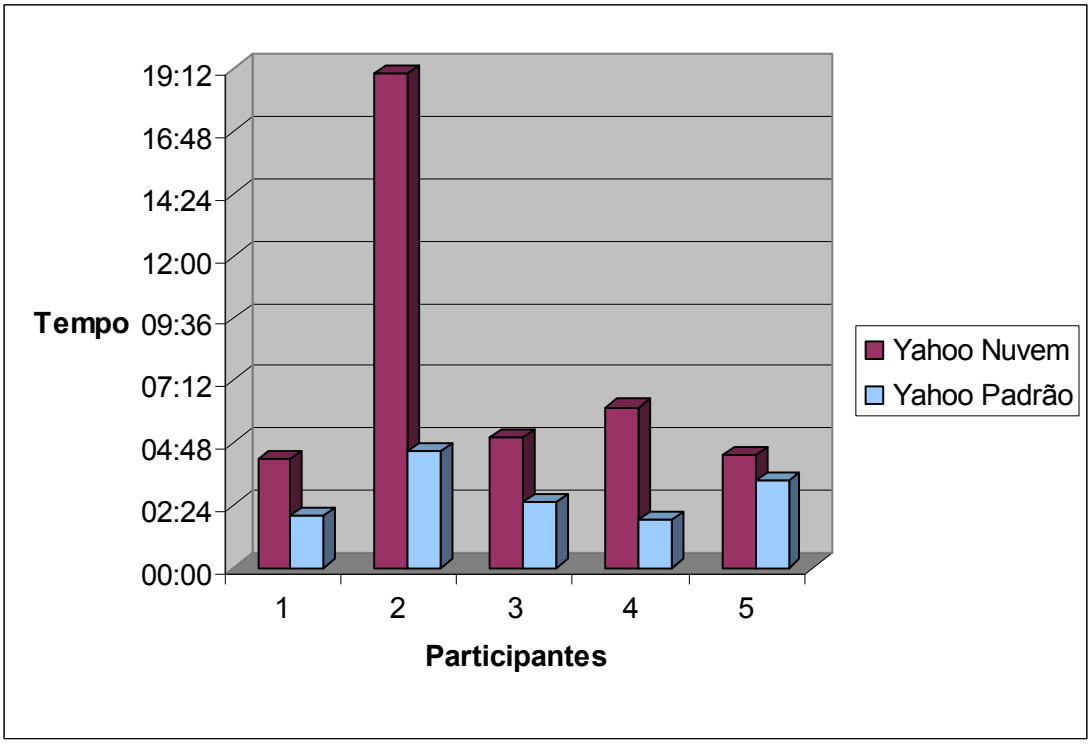


Figura 34 - Gráfico do tempo de execução da questão 3.

Tabela comparativa do tempo de execução da Questão 4 (Pará)		
	Yahoo Nuvem	**Yahoo Padrão**
1	2:40 min	2:20 min
2	4:10 min	1:00 min
3	1:20 min	50 seg
4	1:30 min	2:00 min
5	50 seg	1:30 min
Tempo médio	**2:06 min**	**1:32 min**

Tabela 8 - Tempo de execução da questão 4 no Yahoo Nuvem e no Yahoo Padrão.

Gráfico do tempo de execução da Questão 4 (Pará)

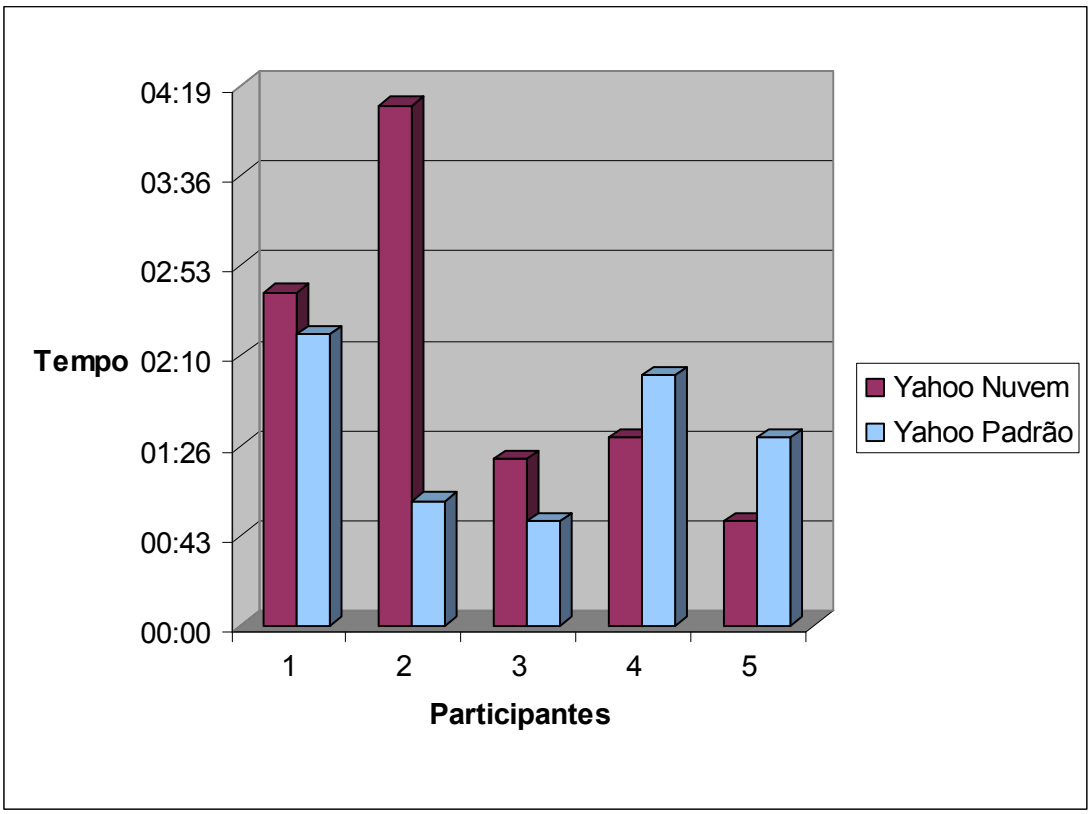


Figura 35 - Gráfico do tempo de execução da questão 4.

Número de páginas acessadas até a obtenção das respostas

Com relação ao número de páginas acessadas até a obtenção das respostas foi observado que esse número foi inferior em todos os casos onde as questões foram respondidas com o auxílio das nuvens. Além disso, em 80% dos casos os participantes obtiveram a resposta da questão diretamente na nuvem, sem sequer acessar alguma página de resultados.

É importante ressaltar que nesses casos onde nenhuma página de resultados foi acessada, as nuvens foram reformuladas por 4 participantes para a questão 3 (ANEXO IV), o que reforça que a eficácia da nuvem depende da tarefa realizada (ANEXO VII).

Tabela comparativa do número de páginas acessadas até a obtenção da resposta da Questão 3 (Bento Gonçalves)		
	Yahoo Nuvem	**Yahoo Padrão**
1	0	1
2	1	4
3	0	2
4	0	2
5	0	4
Número médio	0,2	2,6

Tabela 9 - Número de páginas acessadas até a obtenção da resposta da questão 3 no Yahoo Nuvem e no Yahoo Padrão.

Tabela comparativa do número de páginas acessadas até a obtenção da resposta da Questão 4 (Pará)		
	Yahoo Nuvem	**Yahoo Padrão**
1	0	1
2	1	1
3	0	1
4	0	2
5	0	1
Número médio	0,2	1,2

Tabela 10 - Número de páginas acessadas até a obtenção da resposta da questão 4 no Yahoo Nuvem e no Yahoo Padrão.

8.3. Resumo dos resultados

Resumindo os resultados em pontos positivos e negativos temos o seguinte retrato:

Pontos positivos:

- A aceitação da aplicação foi muito boa por todos os participantes envolvidos na pesquisa e sua utilização foi compreendida com bastante facilidade.
- Foi constatado que a tarefa influência diretamente na utilidade e na eficácia da nuvem. (Esse aspecto é positivo em alguns casos e negativo em outros).

- O aproveitamento das nuvens está ligado diretamente ao conhecimento geral do usuário e requer um conhecimento prévio mínimo em relação ao assunto da consulta. (Esse aspecto é positivo em alguns casos e negativo em outros).

- As nuvens se mostraram bastante eficazes nos casos onde as respostas são objetivas e compostas por apenas um termo.

- Foi identificado como positivo, também, o fato da nuvem apresentar os resultado de forma resumida e também informações adicionais sobre o contexto do assunto.

- As nuvens apresentam mais contexto nas consultas compostas por apenas um ou dois termos, porém isso atende à maioria das consultas.

- Nas consultas cujas respostas são mais complexas as nuvens tem sua eficácia diminuida. Nesses casos a nuvem não apresenta a resposta, apenas apresenta novas palavras, em contexto, para serem adicionadas a consulta inicial. (Esse aspecto é positivo em alguns casos e negativo em outros).

- De maneira geral, os participantes foram unânimes em informar que as nuvens no mínimo ajudaram sugerindo palavras-chaves para refinar ou reformular suas consultas.

- A utilização da nuvem reduziu o esforço cognitivo dos usuários e o número de páginas visitadas até a obtenção dos resultados.

Pontos negativos:

- A aplicação trata as palavras e suas respectivas freqüências individualmente, logo, as palavras compostas aparecem separadas na nuvem.

- As consultas compostas por palavras-chave que são associadas a diversos assuntos, ou seja a diferentes universos semânticos, não apresentam nenhum contexto específico.

- O tempo de conclusão das tarefas não foi menor com o auxílio das nuvens na maioria dos casos.

- A medida que mais termos são adicionados a consulta o contexto vai se perdendo apresentando nuvens de pior qualidade.

8.4. Conclusões

Podemos concluir a partir dessa pesquisa que a visualização dos resultados de um sistema de busca em uma nuvem de texto pode auxiliar os usuários a encontrar o que procuram facilitando a construção de consultas em buscas exploratórias.

Porém, a eficácia da nuvem está ligada diretamente à tarefa em questão e, também, ao conhecimento prévio do usuário em relação ao assunto da consulta.

A conclusão das tarefas não foi mais rápida com o auxílio das nuvens, porém, a sua utilização reduziu o esforço cognitivo dos usuários e o número de páginas visitadas até a obtenção dos resultados.

Foi possível constatar que o desenvolvimento de aplicações para a visualização de resultados de sistemas de busca é algo bastante desafiador, porém diante do entusiasmo com que a aplicação foi recebida pode-se dizer que essa área é de fato um caminho que deve continuar sendo explorado.

A visualização dos resultados em nuvens deve ser vista não como uma alternativa, mas sim como mais uma possibilidade, mais uma ferramenta disponível da qual os usuários podem tirar proveito para realizar suas consultas.

8.5. Futuros trabalhos

Essa pesquisa apresenta um conceito novo. Um novo uso para a técnica de visualização de informações conhecida como nuvem de texto. A aplicação utilizada na pesquisa foi desenvolvida exclusivamente para esse fim e, mesmo funcionando em sua plenitude, trata-se de um protótipo inicial.

Logo, podemos sugerir como futuros trabalhos melhorias na aplicação desenvolvida e a ampliação da pesquisa aumentando o número de usuários e também o número de tarefas.

Podem ser investigadas outras formas de interação com a aplicação visando melhorar a sua usabilidade.

Com relação ao layout da interface, hoje, a nuvem da aplicação aparece apenas na página inicial do sistema de busca. Uma outra possibilidade que poderia ser investigada seria a apresentação da nuvem, na mesma tela, junto com os resultados retornados para uma consulta.

Um desdobramento seguinte dessa perquisa seria no sentido de incorporar algum tipo de vocabulário controlado ou ontologia à aplicação. Essa abordagem abre caminho para novas possibilidades como a de se selecionar ou excluir da nuvem determinados universos semânticos.

O uso de vocabulários controlados também permite alguns tratamentos específicos solucionando por exemplo o problema das palavras que aparecem no singular e no plural e das palavras compostas que têm seus termos tratados de forma isolada.

Um outro desdobramento permitido a partir do uso de vocabulários controlados seria trabalhar com o *feedback* implícito dos usuários que é uma linha de pesquisa bastante em voga na área de recuperação de informações. O *feedback* implícito dos usuários é obtido a partir da sua interação com os resultados, analisando o seu comportamento em relação ao serviço, de forma automática mediante apenas o seu consentimento.

Dessa forma a aplicação poderia retornar nuvens contendo apenas as áreas de interesse do usuário sem ambigüidades.

E finalmente, outra possibilidade a ser investigada seria o uso da aplicação por crianças, adaptando a nuvem a sua realidade utilizando outras cores, tamanhos maiores de fonte e também incorporando vocabulário controlado.

Bibliografia

BAEZA-YATES, R. & RIBEIRO NETO, B. **Modern Information Retrieval**. ACM Press, Addison Wesley, 1999.

BALDONADO, M. Q. W. & WINOGRAD, T. **SenseMaker: An information-exploration interface supporting the contextual evolution of a user's interests**, In Proceedings of the Conference on User Interface and Software Technology, p. 11-18, 1997.

BATES, M. J. **The design of browsing and berrypicking techniques for the online search interface.** Online Review v.13, n.5, p. 407-424, 1989.

BECKS, A., SEELING, C. & MINKENBERG, R. **Benefits of Document Maps for Text Access Knowledge Management: A Comparative Study.** Proceedings of the ACM Symposium on Applied Computing (SAC2002), Madrid, Spain, p. 621-626, 2002.

BERTIN, Jacques. **Neográfica e o Tratamento Gráfico da Informação.** Curitiba: Editora da Universidade do Paraná, 1986.

BREITMAN, Karin. **Web Semântica: A internet do Futuro.** LTC, 2005.

BRIN, Sergey & PAGE, Laurence. **The anatomy of a large-scale hypertextual Web search engine.** In: Proceedings of the 7th International World Wide Web conference on Computer Networks, p.107-117, Brisbane, Australia,1998.

BÜRDEK, Bernhard E. **Design: História, Teoria e Prática do Design de Produtos.** São Paulo: Editora Edgard Blücher, 2006.

CARD, Stuart K, MACKINLAY, Jock D. & SHNEIDERMAN, Ben. **Readings in Information Visualization: Using Vision to Think.** Morgan Kauffman, 1999.

CRUZ, Carla; RIBEIRO, Uirá. **Metodologia científica: teoria e prática.** 2. ed. Rio de Janeiro: Axcel Books, 2004.

DANTAS, Marta. **Dominando o Google.** Rio de Janeiro: Brasport, 2005.

DUMAIS, S., CUTRELL, E., & CHEN, H. **Optimizing search by showing results in context.** In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Seattle, WA, p. 277-284. New York: ACM Press, 2001.

ECO, U. **Como se faz uma tese**. São Paulo: Perspectiva, 1999.

FRY, Benjamin J. **Computational Information Design.** Dissertação de doutorado defendida no Instituto de Tecnologia de Massachusetts, 2004.

GOMES FILHO, João. **Gestalt do objeto: sistema de leitura visual da forma.** 7. ed. São Paulo: Escrituras Editora, 2004.

HALVEY, Martin & KEANE, Mark. **An assessment of Tag Presentation Techniques.** In Proceedings of the WWW2007, Banff, Alberta, Canadá, 2007.

HEARST, Marti A. **Next generation web search: Setting our sites.** IEEE Data Engineering Bulletin: Special Issue on Next Generation Web Search, 2000.

HOSCHER, C. AND STRUBE, G. **Web search behavior of Internet experts and newbies.** In Proceedings of the 9th International World Wide Web Conference on Computer Networks. The International Journal of Computer and Telecommunications Networking. North-Holland Publishing Co., p. 337-346, 2000.

JANSEN, B. J., SPINK, A., & SARACEVIC, T. **Real life, real users, and real needs: A study and analysis of user queries on the Web.** Information Processing and Management, v. 36, n. 2, p. 207-227, 2000.

JANSEN, B. J., SPINK, A., & PEDERSEN, J. **A temporal comparison of AltaVista Web searching: Research articles.** Journal of the American Society for Information Science and Technology**,** v. 56, n. 6, p. 559-570, 2005.

JANSEN, B. J. & SPINK, A. **An analysis of Web searching by European AlltheWeb.com users.** Information Processing and Management: an International Journal**,** v. 41, n. 2, p. 361-381, 2005.

JORDAN, Patrick W. **An introduction to usability.** London: Taylor & Frances, 1998.

LUPTON, Ellen. **Pensar com tipos.** Cosac Naify, 2006.

MACEDO, J. **Recuperação de Informação Textual Distribuída por Fontes Autônomas com Sobreposição**. (Tese de Doutorado) Universidade do Ninho, Portugal, Julho 2001.

MARCHIONI, Gary. **Exploratory search: From finding to understanding.** Communications of the ACM, v. 49, n. 4, p. 41-46, 2006.

MARCONI, M. A.; LAKATOS, E. V. **Metodologia científica**. 3. ed. São Paulo: Atlas, 2000.

MEIRELLES, Junia Cristina J. P., MOURA, Mônica. **Web 2.0: novos paradigmas projetuais e informacionais.** Infodesign - Revista Brasileira de Design da Informação v. 4, n. 2, p. 12-19, 2007.

MONK, A.F., WRIGHT, P.C., DAVENPORT, L. & HABER, J. **Improving your human-computer interface: A practical technique.** Prentice Hall Practitioner series, 1993.

NIELSEN, Jakob. **Projetando Websites**. Rio de Janeiro: Campus, 2000.

PALTRIGE, Sam. **Mining and mapping web content.** INFO - The Journal of Policy, Regulation and Strategy for Telecommunications, Information and Media, v.1, n. 4, Camford Publishing, 1999.

PIROLLI, Peter & CARD, Stuart. **Information foraging in information access environments.** In Proceedings of the SIGCHI conference on Human factors in computing systems, p. 51-58, ACM Press/Addison-Wesley Publishing Co. New York, USA, 1995.

RIVADENEIRA, A. W, GRUEN, Daniel M., MULLER, Michael J. & MILLEN, David R., **Getting our head in the clouds: toward evaluation studies of tagclouds,** In Proceedings of the SIGCHI conference on Human factors in computing systems 2007, p. 995-998, San Jose, California, USA, 2007.

ROBERTSON, G., CARD, S. K. & J. D. MACKINLAY. **The cognitive coprocessor architecture for interactive user interfaces.** In Proceedings of the 2nd annual ACM SIGGRAPH symposium on User interface software and technology, p. 10-18, New York, USA, ACM Press, 1989.

ROSE, D. E., & LEVINSON, D. **Understanding user goals in web search.** In Proceedings of the 13th International Conference on World Wide Web, p. 13-19. New York, ACM Press, 2004.

SANTOS, Antonio Raimundo dos. **Metodologia científica: a construção do conhecimento.** 5. ed. Rio de Janeiro: DP&A, 2002.

SEBRECHTS, M.M, VASILAKIS, J., MILLER, M.S., CUGINI, J.V. & LASKOWSKI, S.J. **Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces**. 22nd Annual Int'l ACM-SIGIR Conference Research and Development in Information Retrieval, p. 3-10, ACM Press, New York, 1999.

SILVA, Ana A. **Representação gráfica e cartográfica da informação estatística.** Dissertação de mestrado defendida no Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa, 2003.

SILVERSTEIN, C., MARAIS, H., HENZINGER, M., & MORICZ, M. **Analysis of a very large Web search engine query log.** ACM SIGIR Forum, v. 33, n. 1, p. 6-12, 1999.

SMITH, Gene. **Tagging: People-powered Metadata for the Social Web.** Pearson Education, New Riders, Peachpit Press, 2007.

SPINK, A., WOLFRAM, D., JANSEN, M. B. J., & SARACEVIC, T. **Searching the Web: The public and their queries.** Journal of the American Society for Information Science and Technology, v. 52, n. 3, p. 226–234, 2001.

SPINK, A., JANSEN, B. J., WOLFRAM, D., & SARACEVIC, T. 2002. **From e-sex to e-commerce: Web search changes.** IEEE Computer Society, v. 35, n. 3, p. 107-109, 2002.

TEEVAN, Jaime, ALVARADO, Christine, ACKERMAN, Mark S. & KARGER, David R. **The Perfect Search Engine Is Not Enough: A Study of Orienteering Behavior in Directed Search.** Proceedings of the SIGCHI conference on Human factors in computing systems, p. 415-422, Vienna, Austria, 2004.

TUFTE, Edward R.**The Visual Display of Quantitative Information.** 2. ed. Graphics Press, Connecticut, 2001.

WARE, Colin. **Information Visualization: Perception for Design.** Morgan Kaufmann, 2000.

Supporting exploratory search. Communications of the ACM: Special Issue, v. 49, n. 4, p. 36-39, abril 2006.

WHITE, Ryen W., KULES, Bill, BEDERSON, Benjamin B. **Exploratory search interfaces: categorization, clustering and beyond.** Report on the XSI 2005 workshop at the Human-Computer Interaction Laboratory, University of Maryland. SIGIR Fórum, v. 9, n. 2, p. 52-56, 2005.

WIVES, L. K. **Tecnologia de Descobertas de Conhecimentos em Textos Aplicadas à Inteligência Competitiva**. (Exame de Qualificação), Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

WOLFRAM, D., SPINK, A., JANSEN, B. J., & SARACEVIC, T. 2001. **Vox Populi: The public searching of the Web.** Journal of the American Society for Information Science and Technology, v. 52, n. 12, p. 1073-1074, 2001.

ZIPF, G. K. **Selected studies of the principle of relative frequency in language.** Cambridge, MA: Harvard University Press, 1932.

ZAMIR, O. & ETZIONI, O. **Web Document Clustering: A Feasibility Demonstration.** Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval**,** Melbourne, Australia, p. 46-54, 1998.

Artigos na Internet

BONSIEPE, Gui. **Design as Tool for Cognitive Metabolism: from knowledge production to knowledge representation, 2000.** Disponível em: http://www.guibonsiepe.com/pdffiles/descogn.pdf

BLATTMAN, U. et al. **Recuperar a Informação Eletrônica pela Internet**. 2000. Disponível em: http://www.ced.ufsc.br/~ursula/papaers/buscanet.html

CENDÓN, B. V. **Ferramentas de Busca na Web.** Brasília, v. 30, n.1, p. 39 – 49, jan./abr. 2001. Disponível em: http://www.robotstxt.org/wc/threat-or-treat.html.

HEARST, Martin. **Informational visualization and presentation.** Disponível em: http://www.sims.berkeley.edu/courses/is247/s02/lectures/TextAndSearch.ppt

LEE, Rainie. **Online Activities & Pursuits: Tagging Report.** Pew Internet and American Life Project Report, 2007. Disponível em: http://www.pewinternet.org/PPF/r/201/report_display.asp

MOURA, Gevilacio Aguiar Coêlho de. **Sistemas de busca na web: diretórios e mecanismos de busca.** 2000. Disponível em: http://www.quatrocantos.com/tec_web/sist_busca/sb_sum.htm.

Nielsen/NetRatings. **Market shares de janeiro de 2007.** Disponível em: http://www.nielsen-netratings.com/pr/pr_070521.pdf. Acessado em junho de 2007.

QUINTARELLI, E. **Folksonomies: power to the people.** ISKO Italy-UniMIB meeting: Milan. Junho 24, 2005. Disponível em: http://www-dimat.unipv.it/biblio/isko/doc/folksonomies.htm

RIVADENEIRA, W. & BEDERSON, B. B. **A Study of Search Result Clustering Interfaces: Comparing Textual and Zoomable User Interfaces.** University of Maryland HCIL Technical Report HCIL-2003-36, 2003. Disponível em: http://hcil.cs.umd.edu/trs/2003-36/2003-36.pdf.

SPENCE, Ian. **William Playfair and the Psychology of Graphs.** ASA Section on Statistical Graphics. University of Toronto, 2006. Disponível em: http://www.psych.utoronto.ca/users/spence/Spence (2006).pdf

SULLIVAN, Danny (2004). Disponível em: http://blog.searchenginewatch.com/blog/041111-084221

Sites na Internet

AlltheWeb: http://www.alltheweb.com

Altavista: http://www.altavista.com

AltSearchEngines: http://www.altsearchengines.com

Apache Software Foundation, Lucene: http://lucene.apache.org

Clusty: http://www.clusty.com

Del.icio.us: http://www.delicious.com

Excite: http://www.excite.com

Fireball: http://www.fireball.de

Flickr: http://www.flickr.com

Globo.com: http://www.globo.com

Google: http://www.google.com

Grokker: http://www.grokker.com

Internet Systems Consortium: http://www.isc.org

KartOOVISU: http://beta.kvisu.com

Nielsen Netratings: http://www.nielsen-netratings.com

Open Directory Project: http://www.dmoz.org

Pollster: http://www.pollster.com

Quintura: http://www.quintura.com

The New York Times: http://www.nytimes.com

The Pew Internet & American Life Project: http://www.pewinternet.org

Wikipedia: http://www.wikipedia.com

World Wide Web Consortium: http:// www.w3.org

Yahoo: http://www.yahoo.com

Yahoo Developer Network: http://developer.yahoo.com

Anexos

Anexo I

Este primeiro anexo corresponde ao questionário inicial utilizado para a seleção dos participantes da pesquisa com o aplicativo. O questionário foi aplicado oralmente em forma de entrevista semi-estruturada. Os itens que constam nas respostas não foram colocados para os entrevistados, sendo de uso exclusivo do entrevistador.

A PARTE 1 do questionário identificou o perfil de uso da internet dos entrevistados. Através dessas perguntas foram eliminados os entrevistados que utilizavam ferramentas de busca avançada.

Na PARTE 2 foi avaliado o grau de conhecimento dos entrevistados com relação aos temas abordados na pesquisa.

Foram eliminados os entrevistados que detinham qualquer possível conhecimento a respeito dos assuntos abordados nas questões propostas como tarefas da pesquisa.

A partir desse questionário, os entrevistados que não foram eliminados foram divididos em dois grupos de constituição semelhante para a realização do experimento controlado. Também foram considerados para a elaboração dos grupos a idade, formação acadêmica e sexo dos participantes.

Universidade do Estado do Rio de Janeiro - UERJ
Centro de Tecnologia e Ciências - CTC
Escola Superior de Desenho Industrial - ESDI
Programa de Pós-Graduação em Design - PPD
Curso de Mestrado em Design - MDE

Questionário para uso em pesquisa de mestrado

Aluna: Márcia Severo Lunardi - Orientador: André Soares Monat

NOME: _____
SEXO: _____ IDADE:_____ FORMAÇÂO ACADÊMICA:_____
E-MAIL: _____ TELEFONE:_____

PARTE 1:

1- Quantas vezes por semana você costuma navegar na internet?

( ) diariamente ( ) nos fins de semana ( ) até 4 dias por semana

Obs: _____

2- Quanto tempo por dia você costuma navegar na internet?

( ) 2 horas ou menos ( ) entre 2 e 4 horas ( ) 4 horas ou mais

Obs: _____

3- Que tipo de atividade você costuma realizar pela internet?

( ) trabalho ( ) lazer ( ) transações bancárias ( ) compras

Obs: _____

4- Você utiliza os sistemas de busca com freqüência?

( ) sim ( ) não ( ) as vezes

Obs: _____

5- Você utiliza as ferramentas de busca avançada dos sistemas de busca?

( ) sim ( ) não ( ) as vezes

Obs: _____

PARTE 2:

1- Você trabalha ou já trabalhou na área de saúde?

( ) sim ( ) não

Obs: _____

2- Você trabalha ou já trabalhou em alguma atividade com crianças?

( ) sim ( ) não

Obs: _____

3- Tem filhos menores de 10 anos ou convive de perto com crianças dessa faixa etária?

( ) sim ( ) não

Obs: _____

4- Tirando o Rio de Janeiro, qual o seu grau de conhecimento a respeito da divisão interna dos estados brasileiros?

( ) profundo ( ) superficial ( ) depende do estado

Obs: _____

5- Qual o seu grau de conhecimento a respeito da culinária regional brasileira?

( ) conheço bastante ( ) superficial ( ) depende do estado

Obs: _____

6- É descendente de europeus?

( ) sim ( ) não **Caso positivo, de que país? _____**

Obs: _____

7- Você ou sua família é natural do Rio de Janeiro ou vem de outros estados?

( ) sim ( ) não **Caso negativo, que estados? _____**

Obs: _____

8 - Visitou algum outro estado nos últimos cinco anos?

( ) não ( ) sim **Caso positivo, quais? _____**

Obs: _____

Anexo II

Esse anexo corresponde às instruções por escrito que foram entregues aos participantes do experimento controlado. Nesse experimento não houve interferência do pesquisador. Todos os usuários tiveram a oportunidade de se familiarizar previamente com o funcionamento do aplicativo da nuvem de texto durante a avaliação cooperativa.

Universidade do Estado do Rio de Janeiro - UERJ
Centro de Tecnologia e Ciências - CTC
Escola Superior de Desenho Industrial - ESDI
Programa de Pós-Graduação em Design - PPD
Curso de Mestrado em Design - MDE

Instruções para uso em pesquisa de mestrado

Aluna: Márcia Severo Lunardi - Orientador: André Soares Monat

NOME: _____DATA: _____
SEXO: _____ IDADE:_____
E-MAIL: _____ TELEFONE:_____

Você irá realizar duas consultas seguidas, em duas versões diferentes do sistema de busca Yahoo. O objetivo é encontrar uma resposta para cada uma das duas perguntas.

As perguntas serão entregues em dois cartões separados. Para a primeira pergunta, que será a “Questão 3” ou a “Questão 4” conforme indicação do pesquisador, será utilizado o sistema de busca Yahoo, tal qual ele é conhecido.

Para a segunda pergunta, será utilizado o mesmo sistema, porém com uma nova funcionalidade. Essa nova funcionalidade é uma nuvem de texto. Ela aparece quando uma palavra ou mais palavras são inseridas no campo de busca, e a barra de espaço é pressionada.

A nuvem de texto que aparece é um resumo dos principais resultados do sistema. O tamanho das palavras é proporcional ao número de vezes que elas aparecem nos resultados. As palavras maiores são as mais freqüentes.

Clicando nas palavras da nuvem você poderá adicioná-las a sua consulta inicial antes de submetê-la, caso considere útil. Pressionando novamente a barra de espaço, uma nova nuvem será gerada. Clicando no botão buscar, a consulta é submetida.

Assim que encontrar a resposta para a questão informe ao pesquisador.

Procure agir naturalmente. Lembre-se que o que está sendo avaliado é o aplicativo e não seus conhecimentos ou suas habilidades.

Muito obrigado!

Sua participação foi de fundamental importância para esta pesquisa.

Anexo III

Esse anexo corresponde às questões submetidas aos participantes da **avaliação cooperativa**. Elas foram entregues em cartões separados imediatamente antes da realização de cada consulta conforme indicação do pesquisador. Nos cartões originais as respostas foram omitidas. Estas foram incluídas nesse anexo apenas para fins elucidativos da pesquisa.

Questão 1:

O pequi é uma fruta nativa brasileira típica de que vegetação?

Resposta: cerrado

Questão 2:

Toda criança precisa tomar uma série de vacinas ao longo da infância. A maioria das vacinas é oferecida pela rede pública de saúde, no entanto, algumas não fazem parte do calendário básico de vacinação.

A sociedade brasileira de pediatria recomenda outras vacinas além dessas oferecidas pela rede pública de saúde. Essas vacinas devem ser dadas em clínicas particulares de vacinação. Entre essas vacinas se encontram a contra varicela (catapora) e a hepatite A.

Qual a idade recomendada para a primeira e a segunda dose da vacina contra a hepatite A?

Resposta: 12 e 18 meses

Anexo IV

Esse anexo corresponde às questões submetidas aos participantes do **experimento controlado**. Elas foram entregues em cartões separados imediatamente antes da realização de cada consulta conforme indicação do pesquisador. Nos cartões originais as respostas foram omitidas. Estas foram incluídas nesse anexo apenas para fins elucidativos da pesquisa.

Questão 3:
Essa charmosa cidade foi habitada inicialmente por indígenas e até 1870 chamava-se Cruzinha. Um dos primeiros núcleos da imigração italiana no RS transformou-se, com o trabalho e dedicação de seus colonizadores, em uma das melhores regiões produtoras de vinho do Brasil. Situada na região do vale dos vinhedos, a cidade também é reconhecida como um dos maiores pólos moveleiros do sul do Brasil.

Que cidade é essa?

Resposta: Bento Gonçalves

Questão 4:
A culinária do Brasil é fruto de uma mistura de ingredientes europeus, indígenas e africanos. O tacacá é uma sopa, preparada com um caldo fino de cor amarelada chamado tucupi que é extraído da raiz da mandioca. A sopa contém também camarão. O tacacá é um dos pratos típicos de uma região do Brasil e de um estado em especial.

Que estado é esse?

Resposta: Pará

Anexo V

Esse anexo corresponde ao formulário utilizado para o controle dos pontos observados no experimento controlado. A partir desse formulário os dados foram tabulados e analisados. Obs: Na avaliação cooperativa os pontos foram anotados livremente em formulário corrido.

Universidade do Estado do Rio de Janeiro - UERJ
Centro de Tecnologia e Ciências - CTC
Escola Superior de Desenho Industrial - ESDI
Programa de Pós-Graduação em Design - PPD
Curso de Mestrado em Design - MDE

Controles para experimento controlado de pesquisa de mestrado

Aluna: Márcia Severo Lunardi - Orientador: André Soares Monat

NOME: _____ DATA:_____

YAHOO PADRÃO: Questão_____

Tempo total: _____ Número de páginas acessadas: _____

Obteve o resultado: ( ) sim ( ) não

Observações:

_____
_____
_____
_____

YAHOO NUVEM: Questão_____

Tempo total: _____ Número de páginas acessadas: _____

Utilizou palavras da nuvem de texto para refinar suas consultas: ( ) sim ( ) não

Caso positivo:

N° de palavras iniciais: _____ N° de palavras da nuvem: _____ N° de nuvens: _____

Obteve o resultado: ( ) sim ( ) não

Observações:

_____
_____
_____
_____

Anexo VI

Esse anexo corresponde ao questionário que foi distribuído aos participantes logo após a conclusão das tarefas no experimento controlado. O objetivo desse questionário foi coletar as principais impressões dos participantes a respeito do aplicativo visando complementar os pontos principais de avaliação do experimento controlado e da avaliação cooperativa.

Universidade do Estado do Rio de Janeiro - UERJ
Centro de Tecnologia e Ciências - CTC
Escola Superior de Desenho Industrial - ESDI
Programa de Pós-Graduação em Design - PPD
Curso de Mestrado em Design - MDE

Questionário para uso em pesquisa para dissertação de mestrado

Aluna: Márcia Severo Lunardi - Orientador: André Soares Monat

NOME: _____

SEXO: _____ IDADE:_____

E-MAIL: _____ TELEFONE:_____

1- Você encontrou alguma dificuldade na realização das tarefas propostas?

_____

_____

_____

_____

_____

_____

_____

2- Qual a sua impressão geral sobre a nuvem de texto dos resultados do sistema de busca?

_____

_____

_____

_____

_____

_____

_____

3- Você considera que a nuvem de texto ajudou de alguma forma na execução das consultas? Caso positivo, você conseguiria explicar brevemente o porquê?

_____

_____

_____

_____

_____

_____

_____

4- Se desejar faça outras observações abaixo:

_____

_____

_____

_____

_____

_____

_____

Muito obrigado!

Sua participação foi de fundamental importância para esta pesquisa.

Anexo VII

Esse anexo corresponde à tabela completa dos pontos observados no experimento controlado. Além dos dados transcritos nas tabelas comparativas do capítulo 8, nela podem ser vistos o número de nuvens geradas até a obtenção das respostas e também o número de palavras adicionadas às consultas a partir das nuvens.

Participante	Questão Yahoo Nuvem	Tempo	Nº de páginas acessadas*	Nº de nuvens geradas	Nº de palavras adicionadas à consulta a partir das nuvens	Questão Yahoo Padrão	Tempo	Nº de páginas acessadas
1	PARÁ	2:40 min	0	1	0	BENTO GONÇALVES	2 min	1
2	PARÁ	4:10 min	1	1	0	BENTO GONÇALVES	4:30 min	4
3	PARÁ	1:20 min	0	1	0	BENTO GONÇALVES	2:30 min	2
4	PARÁ	1:30 min	0	1	0	BENTO GONÇALVES	1:50 min	2
5	PARÁ	50 seg	0	1	0	BENTO GONÇALVES	3:20 min	4
1	BENTO GONÇALVES	4:10 min	0	1	0	PARÁ	2:20 min	1
2	BENTO GONÇALVES	19 min	1	6	2	PARÁ	1 min	1
3	BENTO GONÇALVES	5 min	0	3	2	PARÁ	50 seg	1
4	BENTO GONÇALVES	6:10 min	0	3	1	PARÁ	2:00 min	2
5	BENTO GONÇALVES	4:20 min	0	3	0	PARÁ	1:30 min	1

* Por número de páginas acessadas entende-se tanto páginas de resultados listados do Yahoo quanto *sites*.