UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
NÚCLEO DE COMPUTAÇÃO ELETRÔNICA
INSTITUTO DE MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Renato Montaleão Brum Alves

Uma Análise Empírica Sobre o Uso Combinado de Técnicas de Busca e de Navegação na Recuperação de Informação

Orientador(es): Maria Luiza Machado Campos
Vanessa Braganholo Murta

Rio de Janeiro
2010

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
NÚCLEO DE COMPUTAÇÃO ELETRÔNICA
INSTITUTO DE MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Renato Montaleão Brum Alves

Uma Análise Empírica Sobre o Uso Combinado de Técnicas de Busca e de Navegação na Recuperação de Informação

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Instituto de Matemática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Informática.

Orientador(es): Maria Luiza Machado Campos
Vanessa Braganholo Murta

Rio de Janeiro
2010

A474 Alves, Renato Montaleão Brum.
Uma análise empírica sobre o uso combinado de técnicas de busca e de navegação na recuperação de informação. / Renato Montaleão Brum Alves. – 2010.
f.: il.

Dissertação (Mestrado em Informática) – Universidade Federal do Rio de Janeiro, Instituto de Matemática, Núcleo de Computação Eletrônica, Rio de Janeiro, 1988.

Orientadores: Maria Luiza Machado Campos
Vanessa Braganholo Murta

1. Uso Combinado de Técnicas de Busca – Teses. 2. Busca e Navegação – Teses. 3. Recuperação da Informação – Teses.
I. Maria Luiza Machado Campos (Orient.). II. Vanessa Braganholo Murta (Orient.).III. Universidade Federal do Rio de Janeiro. Instituto de Matemática. Núcleo de Computação Eletrônica. IV. Título

CDD

Renato Montaleão Brum Alves

Uma Análise Empírica Sobre o Uso Combinado de Técnicas de Busca e de Navegação na Recuperação de Informação

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Instituto de Matemática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Informática.

Aprovada em: Rio de Janeiro, 26 de fevereiro de 2010.

_____
Prof.ª Maria Luiza Machado Campos, Ph.D.

_____
Prof.ª Vanessa Braganholo Murta, D.Sc.

_____
Prof. Geraldo Bonorino Xexéo, D.Sc.

_____
Prof.ª Adriana Santarosa Vivacqua, D.Sc.

_____
Prof.ª Jonice de Oliveira Sampaio, D.Sc.

À Mônica Ferreira da Silva, com a minha admiração, carinho e inteira gratidão.

AGRADECIMENTOS

A Deus, por me oferecer a singular oportunidade de realização do presente trabalho; todos os meios necessários para tanto; e ainda permitir que eu compartilhe da convivência enriquecedora das pessoas aqui mencionadas.

Às minhas orientadoras, Professoras Vanessa Braganholo Murta e Maria Luiza Machado Campos, por, após terem me ofertado todas as chances possíveis, me darem mais uma. Pela paciência com as minhas dificuldades iniciais, e cuidado quando da gentil indicação dos meus erros. À Professora Maria Luiza, ainda, por apoiar e incentivar minha carreira profissional quando da minha aprovação em concurso público. Sua postura generosa tornou minhas decisões menos difíceis.

Aos Professores Geraldo Xexéo, Adriana Vivacqua e Jonice Sampaio pelas diversas sugestões e críticas construtivas quando da participação na banca de defesa da dissertação. A Professora Jonice Sampaio, ainda, e ao Professor Pedro Manuel, pelas muitas contribuições e apontamentos feitos durante a defesa do projeto, que permitiram ampliar meu entendimento e atingir as metas propostas.

À Professora, amiga e prima Mônica Ferreira da Silva por não me permitir desistir, por ir além do simples incentivo, me ensinando sobre metodologia de pesquisa e construção do conhecimento científico, tanto quanto sobre humildade, generosidade e compreensão. Obrigado pelo seu apoio incondicional, pela amizade sincera, incontáveis momentos de escuta paciente sobre diferentes assuntos, inúmeras horas de conselhos e orientações, inesgotável paciência e resiliente bom humor. Sua postura diante da vida pessoal e profissional é motivo da minha admiração, e insumo de constantes reflexões.

Aos funcionários do NCE, em especial à querida Tia Deise que, em diversos momentos, me auxiliou com prontidão e boa-vontade.

Ao Professor David Weiss da Poznan University of Technology, Polônia, e a Stanislaw Osinski, por terem gentil e prontamente cedido a utilização do algoritmo Lingo3G para a realização dos experimentos, bem como valiosa bibliografia para pesquisa.

À Professora Yi Zhang e ao doutorando Jonathan Koren da University of Califórnia, Estados Unidos, o meu pleno agradecimento pela parceria na árdua tarefa de desenvolver a coleção de testes utilizada nas sessões de experimentação.

Aos pesquisadores Daniel Tunkelang e Giovanni Maria Sacco por toda a consideração, incentivo, e diversas informações relevantes encaminhadas nos muitos e-mails trocados no decurso da pesquisa.

Ao Rafael Todor Rossini pela benevolência, apoio, interesse, e por todo o amparo na preparação de parte do ambiente de experimentação.

À Anna Maria Neville Morgado Nogueira por compreender minha necessidade de afastamento do ambiente de trabalho nos momentos em que o mestrado mais exigiu dedicação e empenho. Sem o seu apoio, e as condições que me proporcionou, a concretização do trabalho teria sido inviável. Obrigado ainda pelo seu estímulo constante, por ter incentivado, e continuar incentivando, a minha formação tanto acadêmica quanto profissional.

A meus queridos pais, Ney Alves e Eulália de Jesus Brum Alves, por terem sacrificado tantas vezes as próprias oportunidades para permitir que eu as tivesse em maior número e em melhor qualidade. Obrigado por estarem sempre dispostos a me auxiliar, pelo extenso esforço empregado na minha formação, e por em muitos momentos, silenciarem os próprios problemas para se ocuparem dos meus.

A minha amiga, fiel parceira e irmã Priscila de Jesus Brum Alves. Apoio e compreensão incondicionais. Obrigado por dividir comigo os momentos de sorrisos largados no rosto, e as inevitáveis e íntimas situações das lágrimas. Alma generosa, determinada e acolhedora, extremamente cuidadosa no apontamento das minhas dificuldades. A vida seria realmente sem graça sem você.

A sempre companheira, amiga e namorada, Maylle de Almeida Seabra Lo Feudo. Obrigado por compreender as muitas ocasiões em que precisei estar ausente, por acreditar em mim com tanta intensidade a ponto de me fazer continuar, pelas infindáveis e cansativas horas em que me ouviu atenciosa e pacientemente, e por me fazer estudar tanto só para poder te impressionar durante as disciplinas do mestrado. Aprendi muito assim! Obrigado, ainda, pela indispensável ajuda na preparação e análise da assustadora quantidade de informações geradas pelo experimento.

A minha família do coração, Carlos Seabra Lo Feudo e Ana Maria Barbosa de Almeida Lo Feudo pelo respeito, acolhimento, carinho e apoio demonstrado inúmeras vezes através de atitudes explícitas e por outras silenciosas e sinceras.

Ao amigo Carlos Rubini por me ajudar a entender e admitir que diante das nossas muitas limitações é indispensável aceitar, e até mesmo buscar, por ajuda. Ou muita ajuda. A extensa lista de agradecimentos é mostra de que, finalmente, aprendi o que tanto procurou me ensinar.

Aos colegas das várias turmas do NCE, e dos grupos de pesquisa DatAware e GRECO. Foi um prazer e um grande aprendizado estudar com vocês.

A todos os participantes do experimento: obrigado pela boa vontade, sugestões, opiniões, críticas e pela paciência notável durante as intermináveis sessões.

RESUMO

ALVES, Renato Montaleão Brum, **Uma análise empírica sobre o uso combinado de técnicas de busca e de navegação na recuperação de informação.** Rio de Janeiro, 2010. Dissertação (Mestrado em Informática). - Instituto de Matemática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2010.

Devido à grande produção de conhecimento presente nas últimas décadas, a área de Recuperação de Informação vem ganhando elevado papel de destaque. Diferentes abordagens de acesso a documentos vêm sendo utilizadas visando oferecer melhores ferramentas para os usuários de sistemas de Recuperação de Informação. A combinação dos tradicionais métodos de busca direta e navegação acenam com resultados promissores. Esta dissertação apresenta uma avaliação da eficiência dos seguintes métodos de Recuperação de Informação: (i) busca direta, (ii) navegação, (iii) agrupamento plano (*flat clustering)*, (iv) agrupamento hierárquico (*hierarchical clustering*) e (v) navegação facetada. A investigação foi desenvolvida através da aplicação de uma abordagem de pesquisa multimétodo, ou seja, recolhendo dados qualitativos e também quantitativos. Os dados quantitativos foram coletados utilizando a técnica de experimento com 16 participantes. O tempo utilizado para execução das tarefas, a taxa de erro, a quantidade de documentos relevantes encontrados por minuto e o *R-Precision* foram analisados para cada um dos métodos investigados. A técnica de entrevista por pauta proporcionou a coleta dos dados qualitativos. As percepções dos usuários, quanto à usabilidade e aos resultados da busca disponibilizados através de cada um dos métodos, foram recolhidas objetivando a triangularização com os dados quantitativos. A partir dos resultados empíricos foi possível obter evidências relevantes quanto à eficiência de cada método investigado, considerando as diferentes situações experimentadas. A pesquisa também possibilitou a análise da satisfação dos participantes, oferecendo material para melhor compreender os aspectos relacionados à eficiência das abordagens.

ABSTRACT

ALVES, Renato Montaleão Brum, **Uma análise empírica sobre o uso combinado de técnicas de busca e de navegação na recuperação de informação.** Rio de Janeiro, 2010. Dissertação (Mestrado em Informática). - Instituto de Matemática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2010.

Due to the large production of knowledge in recent decades, Information Retrieval has been gaining the spot lights. Different approaches for accessing documents have been used in order to offer better tools for users of Information Retrieval systems. The use of pure direct search or browsing does not suffice in several cases. In fact, the combination of these methods has been revealing promising results. However, there are still few works in literature that analyze the efficiency of the several possible combinations of these methods. This research work presents an efficiency assessment and comparison for the following information retrieval methods: (i) direct search, (ii) browsing, (iii) flat clustering, (iv) hierarchical clustering and (v) faceted navigation. By applying a multi-method approach, quantitative data were collected from an experiment conducted with 16 participants. The time taken to perform the tasks, the error rate, the amount of relevant documents found per minute and R-precision were analyzed for each of the investigated methods. Qualitative data were also collected through interviews, mainly. User's perceptions regarding the usability and search results provided by each of the investigated methods. From the empirical results it was possible to obtain relevant evidence regarding the effectiveness of each method, considering the different situations experienced. The results of our investigation also allowed the analysis of participant satisfaction, providing material to better understand the aspects related to the efficiency of the approaches.

Lista de Figuras

Lista de Quadros

Lista de Tabelas

Lista de Siglas

IR – Information Retrieval

RI – Recuperação da Informação.

URL – Uniform Resource Locator

WEB – World Wide Web

BBC – British Broadcasting Corporation

HP – Hewlett Packard

TREC – Text Retrieval Conference

IMDB – Internet Movie Database

CENADEM – Centro Nacional da Gestão da Informação

XML - eXtensible Markup Language

SUMÁRIO

1. INTRODUÇÃO

1.1. MOTIVAÇÃO

Há milhares de anos a humanidade vem registrando suas aquisições no campo do conhecimento com o objetivo de, posteriormente, recuperar essas informações em momento oportuno. Segundo o Centro Nacional da Gestão da Informação (CENADEM, 2009), nos últimos 50 anos foi gerada uma quantidade de informação equivalente aos 5 mil anos anteriores.

Nos últimos anos, com o contínuo avanço da tecnologia, as pessoas são diariamente sobrecarregadas com excessiva quantidade de informação. O número de publicações e documentos de diferentes tipos gerados tornou-se assustador. Conseguir a localização de uma informação desejada pode se traduzir, para os usuários de sistemas de informação, numa tarefa não apenas exaustiva, mas também frustrante (MARCHIONINI, 2006).

Segundo o Gartner (2009), uma pessoa comum perde até 6 semanas por ano procurando por documentos extraviados. O International Quality & Productivity Center (IQPC, 2009) corrobora com tal percepção, afirmando, por sua vez, que 30% do tempo dos funcionários de uma empresa é gasto procurando informações.

Desta maneira, tem sido atribuído à área de Recuperação da Informação (*Information Retrieval*) um grande destaque, em virtude de suas propostas para tentar equacionar o problema. Uma das principais questões é garantir que as informações relevantes às necessidades do usuário não serão desprezadas. Da mesma forma, também é preciso assegurar que o mínimo de conteúdo irrelevante será retornado.

As abordagens mais tradicionais de procura por informações consistem no uso dos métodos de busca direta e de navegação (BAEZA-YATES & RIBEIRO-NETO, 1999). Através da técnica de busca direta, um documento é retornado como resultado da consulta caso contenha alguma combinação dos termos utilizados pelo usuário. No entanto, muitos

documentos podem ser relevantes mesmo não possuindo os tais termos que foram especificados. Segundo Teevan et al. (2004), consultas simples digitadas em mecanismos de busca não são robustas o suficiente para atender todas as necessidades dos usuários.

Uma das maneiras de reduzir essa lacuna é aumentar o poder de consulta das linguagens utilizadas nos mecanismos de busca. Contudo, tal metodologia transfere para o usuário a responsabilidade e a incumbência de se adaptar ao sistema sendo utilizado.

A navegação é uma abordagem comumente empregada como alternativa ao uso da busca direta. Através de menus, que revelam uma cuidadosa categorização de conteúdos em acervos variados, o usuário é guiado por tópicos que guardam alguma relação semântica entre si, realizando o refinamento da sua pesquisa até a informação desejada.

No entanto, segundo Marchionini (2006), abordagens como a navegação não se revelam muito eficientes. Ele afirma que a "geração Internet" possui comportamento e expectativas diferentes; tratando-se de uma geração multitarefa, que espera encontrar informações dinâmicas, além de investigar e aprender através de interfaces mais flexíveis.

Por essa razão, o uso combinado das técnicas de navegação e busca direta vem sendo cada vez mais explorado (HEARST, 2009; MANNING, RAGHAVAN & SCHÜTZE, 2009). De acordo com Baeza-Yates & Ribeiro-Neto (1999), bibliotecas digitais modernas e a própria WEB devem procurar combinar as estratégias mencionadas, objetivando melhorar suas capacidades de recuperação. Tal combinação ainda não é uma abordagem totalmente estabelecida, não sendo ainda o paradigma predominante nos sistemas de recuperação de informação.

O presente trabalho proporciona um estudo experimental ao realizar análises por sobre diferentes métodos que usam, de forma combinada, as técnicas de busca direta e de navegação. O objetivo é verificar se as estratégias mescladas oferecem melhores resultados

do que ao serem utilizadas de forma isolada. Evidenciar indícios quanto à eficiência de tais abordagens também faz parte do escopo da pesquisa em questão.

1.2. OBJETIVO DA PESQUISA

A presente pesquisa possuiu como finalidade investigar as diferentes combinações de técnicas de busca e de navegação no processo de recuperação de informações. A partir da análise detalhada de dados gerados através da realização de um experimento, aliados às informações resultantes de entrevistas com os participantes, buscou-se avaliar a eficiência de um método em relação a outro.

Desta forma, os questionamentos que nortearam a realização da pesquisa foram:

1. **Os métodos que combinam navegação e busca, considerando agrupamento plano (*flat clustering*), agrupamento hierárquico (*hierarchical clustering*) e navegação facetada, são mais eficientes do que a navegação e a busca quando utilizados isoladamente?**
2. **Qual técnica de recuperação de informação oferece maior eficiência, dentre as técnicas de agrupamento plano, agrupamento hierárquico e navegação facetada?**

Segundo Manning, Raghavan e Schütze (2009), as medidas formais para avaliação dos métodos de recuperação de informação estão um tanto quanto distantes do objetivo final: a satisfação do usuário. Portanto, também figura como objetivo do presente trabalho analisar a percepção do usuário quanto à eficiência dos métodos averiguados. Tal análise visa proporcionar uma investigação mais abrangente da eficiência dos métodos de recuperação de informação.

1.3. DELIMITAÇÃO DO ESTUDO

Como delimitação do presente trabalho, o foco da pesquisa manteve-se restrito à análise das técnicas de busca, navegação, agrupamento plano, agrupamento hierárquico e navegação facetada.

Cabe ressaltar ainda que o objetivo da pesquisa foi direcionado a responder as perguntas apresentadas na seção anterior tendo em vista o âmbito de domínios homogêneos de conhecimento, a utilização de acervos expressivos em tamanho e a realização de buscas não-exploratórias.

1.4. ORGANIZAÇÃO DA DISSERTAÇÃO

A presente dissertação está organizada da seguinte forma. O capítulo 2 apresenta o referencial teórico que permeia a realização da pesquisa. São analisadas as técnicas de busca, navegação e suas combinações presentes na literatura consultada.

No capítulo 3, a metodologia da pesquisa é explicada, descrevendo seu caráter multimétodo e uso do método paralelo para coleta e análise de dados. O referido capítulo também contém a justificativa para a escolha dos métodos investigados. Por fim, são apresentadas as hipóteses formuladas para a pesquisa, além da forma como as mesmas foram operacionalizadas.

O capítulo 4 discorre sobre a pesquisa realizada, buscando definir uma coleção de testes seguida das etapas que envolveram a construção da mesma. São apresentadas também as diretrizes que serviram de norte para a construção das ferramentas, ou sua adaptação, com o objetivo da realização do experimento.

O capítulo 5 apresenta os resultados obtidos através da etapa piloto, que gerou insumo para aperfeiçoamento dos instrumentos da pesquisa. Em seguida, é apresentada uma análise dos dados quantitativos oriundos do experimento. E, depois, uma análise das informações qualitativas, obtidas através das entrevistas com os participantes. Por último, as hipóteses da pesquisa são revisitadas, objetivando sua validação.

No capítulo 6 são apresentadas as conclusões finais do trabalho. As principais contribuições efetuadas pela pesquisa são destacadas. Após, são apontadas oportunidades para realização de trabalhos futuros, visando lançar luz sobre questões relevantes e ainda não devidamente exploradas.

2. RECUPERAÇÃO DE INFORMAÇÃO

A área de Recuperação de Informação (RI, do inglês, *Information Retrieval*) trata da representação, armazenamento, organização e acesso aos artigos de conhecimento. Um sistema desse segmento tem como principal meta recuperar conteúdo útil, relevante sob a ótica do usuário. Trata-se justamente de um dos maiores desafios da área em análise: recuperar documentos que sejam relevantes para as necessidades do utilizador do sistema.

Segundo Baeza-Yates & Ribeiro-Neto (1999), o foco não é voltado para a recuperação de dados simplesmente, mas sim para a recuperação de informação. A primeira abordagem, recuperação de dados, consiste basicamente em determinar quais documentos atendem uma consulta fornecida, através das palavras-chaves passadas pelo usuário. Nesse cenário, a apresentação no resultado de um único objeto que não satisfaça a necessidade do usuário, é tida como uma falha total. A segunda abordagem, a recuperação de informações, preocupa-se mais com o resgate de conhecimento acerca de um assunto do que com dados que atendem a uma consulta. Desta forma, a recuperação de objetos pode ser imprecisa e pequenos erros podem ser tolerados.

Dois aspectos afetam a recuperação de informações relevantes: a tarefa do usuário e a visão lógica dos documentos adotada pelo sistema de RI (BAEZA-YATES & RIBEIRO-NETO, 1999):

- **Tarefa do usuário:** o utilizador do sistema precisa transformar a sua necessidade em uma consulta expressa em uma linguagem que possa ser compreendida pelo sistema. Significando que será preciso especificar um conjunto de palavras capazes de traduzir a semântica da informação que se deseja obter.

- **Visão lógica dos documentos:** documentos são normalmente representados através de palavras-chave ou por um conjunto de termos de índice.

O foco do presente trabalho de pesquisa encontra-se justamente na tarefa do usuário. Segundo Teevan et al. (2004), mesmo quando uma pessoa sabe exatamente o que deseja encontrar, a máquina de busca perfeita pode não ser suficiente. O motivo reside no fato de que seu sucesso está diretamente relacionado à capacidade de especificar sua necessidade de informação.

Máquinas de busca são empregadas no uso do método de busca direta. De forma resumida, a busca direta procura por termos definidos pelo usuário nos documentos de um dado acervo. Sua desvantagem mais evidente é tornar a tarefa do usuário mais difícil, na medida em que ele precisa se adequar à linguagem utilizada pela máquina de busca, além de não ter uma idéia clara de quais termos estão presentes nos documentos que deseja encontrar.

O método de navegação é normalmente encarado como uma alternativa à abordagem de busca direta. A navegação permite que uma pessoa explore o conteúdo desejado, fornecendo uma importante visão do todo e auxiliando a determinar a relevância de um dado documento ou tópico.

Em contrapartida, para acervos de grandes proporções, a navegação pode se traduzir numa atividade demasiadamente lenta, além de fazer com que o usuário se perca no conteúdo oferecido (BAEZA-YATES & RIBEIRO-NETO, 1999).

Do ponto de vista do usuário, navegar e buscar não são atividades contrastantes. As pessoas simplesmente querem obter a informação desejada. Portanto, a integração das duas visões pode fornecer um suporte mais adequado à forma como as pessoas realmente desejam buscar informação (KALBACH, 2007).

A seguir, os métodos de busca direta e navegação são brevemente apresentados, bem como as abordagens que exploram suas combinações, procurando oferecer aos usuários ferramentas mais completas e aderentes às suas necessidades.

2.1. BUSCA DIRETA

Método tradicionalmente empregado na recuperação de informações, a busca direta permite a localização de documentos indexados através da utilização de palavras-chave escolhidas para tal fim, ou até mesmo por meio da indexação de todas as palavras que compõem um texto existente nos documentos do acervo. Oferece como resposta uma lista ordenada de itens, segundo critério de relevância (BAEZA-YATES & RIBEIRO-NETO, 1999). O referido método vem se tornando cada vez mais popular nos sistemas de recuperação de informação, devido ao seu desempenho em acervos heterogêneos e de elevado número de documentos (HEARST, 2009).

No entanto, a recuperação de informação através do simples uso de palavras-chave oferece alguns inconvenientes. O primeiro deles é obrigar o usuário a conhecer a sintaxe utilizada pelo sistema para poder especificar sua consulta. Quanto maior o seu domínio sobre a linguagem de busca, tanto maior será a eficiência do resultado para as suas necessidades. Outra questão muito importante é que em algumas situações um dado documento pode não conter as palavras fornecidas pelo usuário na sua consulta, e ainda assim ser essencial para atender a demanda apresentada.

Apesar de apresentar desempenho superior ao da navegação quando utilizado em grandes acervos de informações (HEARST, 2009), o método de busca direta ainda possui problemas no referido quesito. Manning, Raghavan & Schütze (2009) afirmam que o usuário típico de RI quer convergir para um resultado de forma rápida, e não encontrar todas as possibilidades possíveis presentes numa coleção ou banco de dados. Dessa forma, as longas

ordenações de resultados da busca direta, com diversas páginas de documentos, fazem com que o usuário perca tempo em pesquisa ou esforço no sentido de descobrir a consulta que melhor se traduz em um conjunto de resultados mais adequados.

2.2. NAVEGAÇÃO

Uma maneira de diferenciar navegação de busca direta é ressaltar o fato de que a segunda tende a produzir coleções novas, com finalidade específica, e que nunca haviam sido reunidas anteriormente. Já no caso da navegação, o que existe é a seleção de *links* que leva o usuário de um ponto de vista a outro, em uma seqüência de operações de rastreamento e seleção (HEARST, 2009).

Estruturas de navegação se adéquam melhor a livros, coleções de informações, informações pessoais, sites da Web e a resultados de busca do que a domínios de conhecimento muito vastos como a Web. Tais estruturas são apoiadas por sistemas de categorias, que viabilizam a navegação em um sistema de informações, bem como a organização de resultados de busca.

Sistemas de categorias têm sua funcionalidade baseada no uso de rótulos que, aplicados aos itens de um conjunto, tornam possível ressaltar os tópicos relevantes daquele domínio. A categorização pré-definida auxilia na percepção da abrangência do referido campo de conhecimento, fornecendo ao especialista uma estrutura familiar de navegação. Ao mesmo tempo, pessoas que estão dando seus primeiros passos no domínio em pauta são beneficiadas por um arcabouço que as ajuda na compreensão de tais informações. Sistemas de categorias, usados na recuperação de informação, são normalmente planos, hierárquicos ou facetados (HEARST, 2009).

No entanto, Marchionini (2006) afirma que a “geração Internet” possui comportamento e expectativas diferentes; tratando-se de uma geração multitarefa, que espera

encontrar informações dinâmicas, além de investigar e aprender através de interfaces mais flexíveis.

Segundo Hearst (2009), o declínio da navegação utilizada isoladamente é acentuado, uma vez que a quantidade de informações disponíveis é cada vez maior, tornando a navegação confusa em sua estratégia de uso. Além disso, o método torna-se extremamente dispendioso em termos de tempo para utilização em grandes acervos, tornando-se uma experiência entediante e cansativa. A autora parece também concordar com Marchionini (2006) ao afirmar que as necessidades de se cruzar informações, ou suas categorias, são freqüentes, demandando interfaces mais aprimoradas.

Visando aliar as vantagens apresentadas por ambos os métodos, bem como mitigar suas vulnerabilidades, as abordagens que combinam busca direta e navegação estão sendo cada vez mais estudadas e experimentadas, conforme discutido nas próximas seções.

2.3. COMBINAÇÕES DE BUSCA DIRETA E NAVEGAÇÃO

Dumais et al. (2001) realizaram um experimento combinando os métodos de busca e de navegação. Como resultado, obtiveram o destaque de documentos importantes que ficaram a poucos cliques do usuário. Quando do uso da busca direta isoladamente, os mesmos documentos não apareciam entre os 20 primeiros da ordenação.

Zamir et al., (1999) corroboram com essa visão, afirmando que a busca direta não exibe a dimensão dos resultados relevantes. Em experimento realizado por Kules e Shneiderman (2008), alternando técnicas de busca direta e navegação, os usuários puderam ver o conjunto de itens relevantes de forma realçada e o consideraram mais organizado e eficiente.

A literatura fornece diversos tipos de combinações entre os métodos de busca direta e de navegação. Na seção 2.3.1, os métodos de agrupamento plano e agrupamento hierárquico

serão apresentados. A seção 2.3.2 introduz a abordagem de navegação facetada. E, por fim, a seção 2.4 discorre brevemente sobre as demais combinações conhecidas na literatura.

2.3.1. AGRUPAMENTO

A técnica de agrupamento (*clustering*) é proveniente da área de Mineração de Dados (Data Mining). Seu objetivo é realizar agrupamentos (clusters) automáticos de dados segundo algum critério de semelhança. A técnica também pode ser usada em arquivos de texto, utilizando algoritmos de Mineração de Texto (*Text Mining*). O objetivo do algoritmo é criar agrupamentos coerentes no seu conteúdo interno, mas claramente distintos entre si. Documentos dentro de um mesmo agrupamento devem ser o mais parecido possível, mas também devem ser o mais diferente possível daqueles que estão nos demais agrupamentos (MANNING, RAGHAVAN & SCHÜTZE, 2009).

A utilização de tal técnica na recuperação de informações deve-se à premissa de que se existe um documento proveniente de um agrupamento que é relevante para a consulta, então é provável que os outros documentos pertencentes àquele agrupamento também sejam relevantes. Tal hipótese está baseada no mecanismo de funcionamento do método.

Cabe ressaltar que o foco de utilização da técnica em questão, no presente estudo, é voltado para o agrupamento de resultados encontrados após a submissão de uma consulta. Outras aplicações da técnica na recuperação de informações são a detecção de duplicidade (YANG et al., 2006), a detecção de novidades (MANNING, RAGHAVAN & SCHÜTZE, 2009) e o descobrimento de metadados na web semântica (ALONSO et al., 2006).

Na técnica de agrupamento de documentos, o critério de semelhança é tipicamente calculado usando similaridade e associações entre palavras e frases. Suas principais vantagens são a completa automação do processo e a fácil aplicação a qualquer coleção de textos (HEARST, 2006). Por esse motivo, Manning, Raghavan e Schütze (2009) chegam a afirmar

tratar-se do método mais comum de aprendizado não supervisionado. O que significa não existir especialistas atribuindo documentos às suas classes correspondentes.

Uma importante distinção pode ser feita entre o agrupamento plano (*flat clustering*), ou agrupamento de otimização, e o agrupamento hierárquico (*hierarchical clustering*). No primeiro, um conjunto de agrupamentos é criado sem que haja nenhum tipo de relacionamento entre eles. Já no segundo, uma estrutura explícita de hierarquia é criada entre os agrupamentos. Outra importante distinção a ser feita é entre os algoritmos de agrupamento rígido (*hard clustering*) e o agrupamento flexível (*soft clustering*). Enquanto que no método de agrupamento rígido cada documento é membro de apenas um agrupamento, no agrupamento flexível um mesmo documento pode participar de vários agrupamentos simultaneamente (MANNING, RAGHAVAN & SCHÜTZE, 2009).

O método de agrupamento também pode revelar resultados interessantes e não esperados, ou algumas vezes indicar tendências dentro de uma temática. Exemplo: uma consulta sobre "*New Orleans*" feita em 16 de setembro de 2005 (data seguinte à devastação provocada pelo furacão Katrina), no Clusty.com[1], revelava um grupo denominado "Hurricane" no topo da lista de agrupamentos. Em seguida vinham os grupos mais esperados para a consulta, como "*hotels*", "*Louisiana*", "*University*" e "*Mardi Gras*" (HEARST, 2006).

As vantagens da técnica de agrupamento, destacadas por Hearst (2006) e Manning, Raghavan e Schütze (2009) são:

- Esclarece e conduz melhor uma consulta inicialmente vaga ao mostrar para o usuário os temas dominantes daquele resultado;
- Remove a ambigüidade de consultas, particularmente útil no caso de acrônimos.

[1] Ferramenta de agrupamento disponível na Web, apontado por Hearst (2006) como a melhor implementação da técnica até aquela data.

Desvantagens da técnica de agrupamento, destacadas também por Hearst (2006) e Manning, Raghavan e Schütze (2009):

- A união de muitas dimensões simultaneamente em apenas uma;
- Ausência de previsibilidade;
- Dificuldade em nomear os agrupamentos;
- Construção de agrupamentos com hierarquias poucos intuitivas.

A navegação através de agrupamentos é indicada como uma alternativa mais eficiente em relação à busca por palavra-chave, método tradicional de busca. Tal afirmação, segundo Manning, Raghavan e Schütze (2009) é especialmente verdadeira nos cenários em que o usuário não domina o tema em questão, e naturalmente possui dificuldades em escolher termos candidatos para a recuperação das informações.

No entanto, da mesma maneira que Hearst (2006), o autor afirma que a construção de menus de navegação, feita automaticamente pelos métodos de agrupamento, não possui a mesma clareza e efeito do que aqueles elaborados manualmente por especialistas do domínio.

A escolha do número de agrupamentos figura dentre as dificuldades da utilização do método. Também conhecido como a cardinalidade do agrupamento, esse número é freqüentemente uma estimativa baseada no conhecimento do domínio ou na experiência de utilização da técnica. Alguns algoritmos fazem uso de heurísticas para elicitar agrupamentos candidatos. Entretanto, encontrar um ponto de partida ideal para realizar as buscas é um problema que ainda aguarda solução (MANNING, RAGHAVAN & SCHÜTZE, 2009).

2.3.1.1. AGRUPAMENTO PLANO

O objetivo da técnica de agrupamento plano é agrupar os dados de um conjunto de forma ótima. Tal meta é obtida através da divisão iterativa do conjunto de dados em K-partições mutuamente exclusivas, as quais seguem um critério previamente estabelecido.

Dois algoritmos são comumente usados por tal técnica: *K-means* e EM (MANNING, RAGHAVAN & SCHÜTZE, 2009; CARPINETO et al., 2009). O *K-means* é o mais importante e também o mais largamente utilizado, dada a sua simplicidade e eficiência. O EM, por sua vez, pode ser aplicado a uma grande variedade de representações de documentos.

O ALGORITMO K-MEANS

O principal objetivo do *K-means* é aumentar a similaridade entre os documentos de um mesmo agrupamento, diminuindo a distância média euclidiana entre eles e o centro do agrupamento (também conhecido como centróide). Em contrapartida, essa atuação garante a maior diferenciação entre os documentos de agrupamentos distintos.

O algoritmo utiliza o modelo de vetores espaciais para representar os documentos. Uma medida de quão bem os centróides representam os membros do seu agrupamento é a Soma Residual dos Quadrados (SRQ, do termo original em inglês Residual Sum of Squares). Ela corresponde ao quadrado da distância de cada vetor do seu centróide, somado ao longo de todos os vetores.

O primeiro passo é informar ao algoritmo os centróides iniciais dos agrupamentos, que serão K documentos selecionados aleatoriamente. Tais documentos são conhecidos como as sementes (*seeds*). O algoritmo então se move em torno dos centróides no espaço a fim de minimizar os valores de SRQ (MANNING, RAGHAVAN & SCHÜTZE, 2009),

Para gerar os agrupamentos definitivos e determinar a classificação dos documentos, o algoritmo realiza tais ações repetidamente. Na medida em que as iterações do algoritmo vão acontecendo, cada documento é atribuído novamente ao agrupamento com centróide mais próximo. Em seguida, todos os centróides de cada um dos agrupamentos são recalculados (com base nos documentos pertences àquele grupo). Dessa forma, K agrupamentos são

gerados e os documentos classificados de acordo com suas distâncias em relação aos centróides finais.

A seguir, apresentamos o algoritmo *K-means* passo a passo (MANNING, RAGHAVAN & SCHÜTZE, 2009):

Passo 01: Atribuir valores iniciais dos centróides.

No primeiro passo, todos os K agrupamentos devem receber valores iniciais para os seus centróides, conforme apresenta a Figura 1. Os documentos (sementes) do conjunto a ser classificado, são escolhidos de forma randômica pelo algoritmo. É importante que todos os documentos estejam atribuídos a um agrupamento qualquer, para que o processamento continue corretamente.

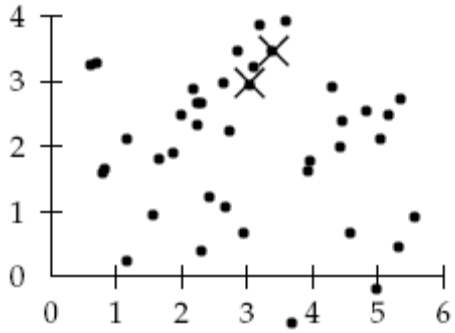


Figura 1 – Seleção das sementes.

Passo 02: Calcular a matriz de distância entre documentos e centróides.

Nesse momento, a distância entre cada documento e os centróides dos agrupamentos é calculada. Se houver N documentos e K agrupamentos, então será necessário calcular N x K distâncias. Por esse motivo, o presente passo é o que mais demanda recursos de tempo e cálculo. Ele é representado graficamente pela Figura 2.

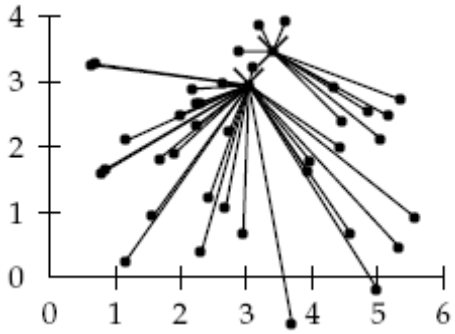


Figura 2 – Cálculo da distância entre documentos (iter. 1)

Passo 03: Classificar cada documento em seu agrupamento.

Nesse passo, os documentos são classificados de acordo com sua distância em relação aos centróides. A regra é simples: o documento vai pertencer ao grupo cujo centróide estiver mais perto, como apresenta a Figura 3. Vale salientar que o algoritmo termina quando nenhum documento for atribuído a um agrupamento diferente daquele em que estava anteriormente.

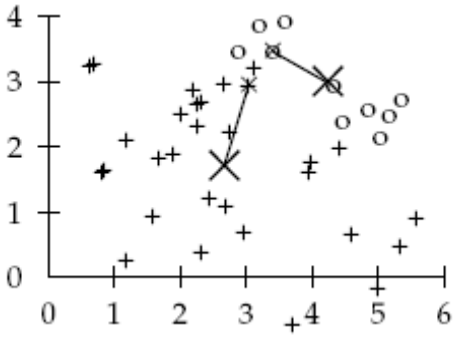


Figura 3 – Movimentação dos centróides (iter. 1)

Passo 04: Calcular novos centróides para cada agrupamento.

Para cada agrupamento que possui mais de um documento atribuído, o novo valor do centróide é calculado. Para tanto, é utilizado novamente o cálculo da distância média euclidiana entre cada documento pertencente àquele agrupamento e o centróide do agrupamento.

Passo 05: Repetir até o término.

Retornando para o passo 02, o algoritmo repete o ciclo em várias iterações, realizando assim um sucessivo refinamento da classificação. A Figura 4 exibe a movimentação dos centróides após realizadas nove iterações.

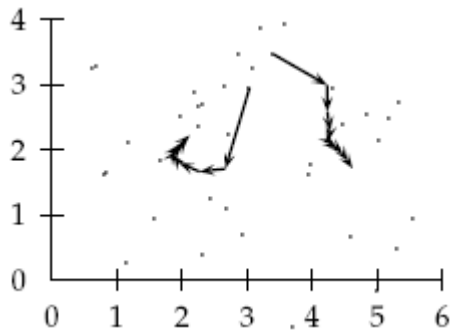


Figura 4 – Movimentação dos centróides após 9 iterações

Cabe ressaltar que o algoritmo K-means realiza uma classificação do tipo agrupamento rígido, pois cada documento, ao final das iterações, pertencerá a apenas um agrupamento. Algoritmos que trabalham com agrupamento flexível normalmente possuem uma métrica para mencionar o quanto cada documento pertence àquele agrupamento.

2.3.1.2. AGRUPAMENTO HIERÁRQUICO

Apesar da técnica de agrupamento plano ser eficiente e conceitualmente simples, ela possui algumas desvantagens em relação ao agrupamento hierárquico. Os algoritmos de agrupamento plano retornam um conjunto de agrupamentos não estruturado, requerem que o número de agrupamentos seja definido como entrada, além de serem não-determinísticos.

Já os algoritmos de agrupamento hierárquico produzem uma estrutura hierárquica de agrupamentos, sendo mais informativa do que o resultado plano da técnica anterior. Pré-definir o número de agrupamentos a ser usado não é necessário, e a maior parte dos algoritmos utilizados para a recuperação de informação são determinísticos. No entanto, tais vantagens do agrupamento hierárquico possuem um preço: a baixa eficiência. Os algoritmos

mais comumente empregados têm uma complexidade, no mínimo, quadrática em relação ao número de documentos, em comparação com a complexidade linear dos algoritmos K-means e EM (MANNING, RAGHAVAN & SCHÜTZE, 2009).

Diversos pesquisadores acreditam que o agrupamento hierárquico produz melhores resultados (JAIN & DUBES, 1988), (CUTTING et al., 1992), (LARSEN & AONE, 1999). No entanto, não há um consenso sobre essa questão. Recentemente alguns experimentos sugeriram o contrário, como pode ser observado em (ZHAO et al., 2002).

AGRUPAMENTO HIERÁRQUICO AGLOMERATIVO

De acordo com a abordagem, os algoritmos usados em agrupamento hierárquico podem ser classificados como aglomerativos ou divisivos. Os aglomerativos fazem uso da estratégia *bottom-up*, ou seja, primeiramente calculam os agrupamentos de um determinado conjunto de documentos. Em seguida, consolidam esse conjunto de agrupamentos, unindo os mais próximos para formar um subconjunto pai. Isso é feito sucessivamente, até que reste apenas um agrupamento na raiz, contendo todos os documentos da coleção.

Os algoritmos divisivos, adotam a estratégia *top-down*. Iniciando com a raiz da árvore hierárquica, o agrupamento inicial é dividido em *n* agrupamentos. A cada passo novos agrupamentos vão surgindo, até que se chegue aos documentos individualmente ou até que se atinja algum critério de parada. Algoritmos aglomerativos são mais utilizados na recuperação de informações (MANNING, RAGHAVAN & SCHÜTZE, 2009; ZHANG, 2007) e por isso também serão foco de estudo do presente trabalho.

Os métodos para criação do agrupamento hierárquico aglomerativo, seguem, de um modo geral, um mesmo algoritmo padrão. A diferença entre eles está presente na forma de cálculo da medida de similaridade entre os agrupamentos, objetivando a união dos mesmos.

A seguir, são apresentados os passos de um algoritmo padrão de agrupamento hierárquico aglomerativo segundo Manning, Raghavan e Schütze (2009), aonde *n* representa o número total de itens do conjunto:

1. Calcular a matriz C de similaridades N x N
2. Repetir N-1 vezes:
3. Unir o par de agrupamentos cuja similaridade fora a maior calculada
4. Atualizar a matriz C

Existem quatro tipos de algoritmos aglomerativos para criação do agrupamento hierárquico, são eles: *single-link, complete-link, group-average*, e *centroid similarity* descritos em detalhe no trabalho de Manning, Raghavan e Schütze (2009).

AGRUPAMENTO HIERÁRQUICO DIVISIVO

Um agrupamento hierárquico também pode ser gerado a partir de uma abordagem *top-down*, ou seja, iniciando-se pelo topo com todos os documentos em apenas um agrupamento. Em seguida, o agrupamento é dividido usando um algoritmo de agrupamento plano. Tal procedimento é aplicado então de forma recursiva até que todos os documentos estejam sozinhos em um agrupamento. Essa técnica é chamada de agrupamento hierárquico divisivo.

Conceitualmente a técnica de agrupamento hierárquico divisivo é mais complexa do que a de agrupamento hierárquico aglomerativo, uma vez que é necessária a utilização de um segundo algoritmo, o de agrupamento plano. Ele é utilizado como sendo uma sub-rotina para permitir as divisões. No entanto, a técnica possui a vantagem de ser mais eficiente caso não seja construída toda a estrutura até se chegar às folhas, isto é, agrupamentos com apenas um documento.

Para um número fixo de agrupamentos de topo, e fazendo uso de um algoritmo eficiente de agrupamento plano, como o K-means, a técnica de agrupamento hierárquico divisivo apresenta algoritmos lineares no número de documentos e de agrupamentos (MANNING, RAGHAVAN & SCHÜTZE, 2009).

2.3.2. NAVEGAÇÃO FACETADA

A classificação facetada se utiliza de um conjunto de categorias mutuamente exclusivas, aonde cada uma delas reflete, isoladamente, uma característica de um determinado grupo de itens. Conhecidas como facetas, essas categorias podem ser usadas de forma combinada para descrever completamente todos os objetos de um dado conjunto. Dessa maneira, as pessoas podem usá-la para navegação e busca de informações (DENTON, 2003).

A noção de faceta repousa na crença de que há mais de uma maneira de enxergar o mundo, e que mesmo classificações hoje vistas como estáveis são, de fato, provisórias e dinâmicas. O desafio está em construir classificações flexíveis, que possam absorver novas transformações do domínio, ainda, que sejam simples de manter e também de utilizar (KWASNICK, 1999).

Com o hipertexto e a Web, visualizar dinamicamente um conteúdo requer apenas alguns cliques. As facetas formam o esquema multidimensional, enquanto que os navegadores (*browsers*) são ferramentas fáceis e familiares de navegação em ambientes multidimensionais. Os benefícios proporcionados pela classificação facetada podem ser materializados pela Web.

2.3.2.1. DO USO, VANTAGENS E DESVANTAGENS

Segundo Denton (2003), o método de navegação facetada não deve ser usado quando o domínio em questão for classificado através de uma única perspectiva, uma vez que hierarquias e árvores são mais eficientes nesse aspecto. Quando o domínio considerar duas

perspectivas, Denton afirma que paradigmas similares a uma planilha eletrônica são mais adequados. No entanto, quando houver três ou mais aspectos mutuamente exclusivos a serem analisados, o método de navegação facetada é indicado.

As vantagens da navegação facetada destacadas por Kwasnick (1999) são:

- Não requer conhecimento completo das entidades ou dos seus relacionamentos.
- Flexibilidade (podem acomodar novas entidades com facilidade).
- Alto nível de expressividade.
- Podem ser especializadas ou livres em sua forma.
- Permitem diferentes perspectivas e abordagens na classificação.

Desvantagens destacadas também por Kwasnick (1999):

- Dificuldade em se escolher as facetas corretas.
- A falta de habilidade para se expressar o relacionamento entre elas.
- A dificuldade de se visualizar todas.

Escolher as facetas certas é um fator crucial, bem como possuir um bom conhecimento dos itens que serão classificados e dos usuários que farão uso da ferramenta. Mas também é verdade que esse é um problema comum a qualquer organização em categorias.

A pobreza na explicitação das relações semânticas entre as categorias é um problema real, mas que não será tratado no presente trabalho. Será assumido que os usuários poderão suprir essa deficiência através das inferências que poderão fazer usando seu próprio conhecimento a cerca dos itens e categorias envolvidos.

Quanto à dificuldade de visualização do conteúdo, a Web dispõe atualmente de ferramental iterativo e de visualizações multidimensionais que tornam possível contornar a

dificuldade apontada. Kwasnick havia previsto em 1999 que a tecnologia avançaria a ponto de preencher tal lacuna.

2.3.2.2. CRIANDO UMA CLASSIFICAÇÃO FACETADA

Após analisar o trabalho de Ranganathan (1962) e seu complexo conjunto de regras para criação de sistemas de classificação facetada, Spiteri (1998) propôs uma simplificação, estabelecendo seu próprio conjunto. O objetivo era oferecer uma ferramenta para estudantes da Ciência da Informação e projetistas de sistemas de classificação facetada.

Da mesma forma que Ranganathan, Spiteri divide a classificação em três partes: o Plano das Idéias, o Verbal e o da Notação. O Plano das Idéias se ocupada a avaliação do assunto em suas partes componentes. O Verbal, por sua vez, trata do processo que envolve a escolha da terminologia apropriada para expressar tais partes. O de Notação se destina a representar, através de algum mecanismo de notação, as mesmas partes.

Vickery (1960) propõe a criação de um esquema para classificação facetada em quatro passos. Em primeiro lugar, de acordo com Vickery (1960), a essência da análise por sobre as facetas está focada na organização dos termos dentro de um dado campo do conhecimento, de forma homogênea e mutuamente exclusiva. Cada uma delas deve ser derivada dentro desse domínio representando uma característica única de divisão conceitual.

O segundo passo seria definir uma ordenação, segundo a qual as facetas seriam usadas para se construir uma mistura de títulos de assuntos. O terceiro seria utilizar uma notação que permita adequar os termos selecionados em combinações completas e flexíveis, sem deixar de ser intuitiva para quem está selecionando a lista de títulos. O quarto passo seria tornar possível a utilização do esquema facetado de tal forma que fosse viável o uso através de uma referência específica, ou ainda no intuito de uma pesquisa com o requerido grau de inespecificidade.

2.4. OUTRAS COMBINAÇÕES EXISTENTES NA LITERATURA

2.4.1. MÉTODO GROUPING

Mecanismos de busca mais sofisticados realizam um agrupamento automático dos itens similares por assunto em subgrupos. O agrupamento pode ser exibido ao lado da página de resultados, permitindo assim a navegação pela lista recém criada. Ao clicar em algum subgrupo o resultado da busca é então filtrado, e um novo inventário é exibido para o visitante. Um dos métodos empregados para a construção automática dos subgrupos é o *grouping*[2]. Apesar de haver uma confusão comum com o método de agrupamento plano, existe uma diferença entre as técnicas utilizadas.

No *grouping*, os termos já existem previamente e estão armazenados de alguma forma, como numa taxonomia, por exemplo. A lista de resultados sofrerá variações de acordo com a busca efetuada. Os termos que intitulam os subgrupos, ao contrário, serão sempre os mesmos. O sistema determina quais são os tópicos daqueles agrupamentos ou então cada item de resultado já possui um termo associado a ele, indicando o seu subgrupo. Em seguida essas categorias são utilizadas para navegação (KALBACH, 2007).

Já no caso do agrupamento plano, as categorias são construídas de forma automática e os termos que irão nomeá-las são derivados do conjunto de resultados apresentado. O algoritmo encontra os grupos que possuem similaridade e, em seguida, classifica os itens segundo a categorização de termos definidos (HEARST, 2009).

Como exemplo de utilização bem sucedida da técnica de *grouping*, podemos citar o endereço do site da Sun (2009) na Web. Na página inicial do site é possível visualizar as abas de navegação dispostas no topo da página, de forma horizontal (Figura 5). Algumas dessas

[2] Mantido no original, em inglês, para evitar confusão com o termo agrupamento, já utilizado na tradução de *clustering*.

abas possuem até um segundo nível de agrupamento. O estilo de uso das abas, e apresentação das mesmas, remete ao caráter tradicional de navegação. No entanto, trata-se de uma estrutura gerada no momento do acesso, baseado em termos pré-definidos de categorias.

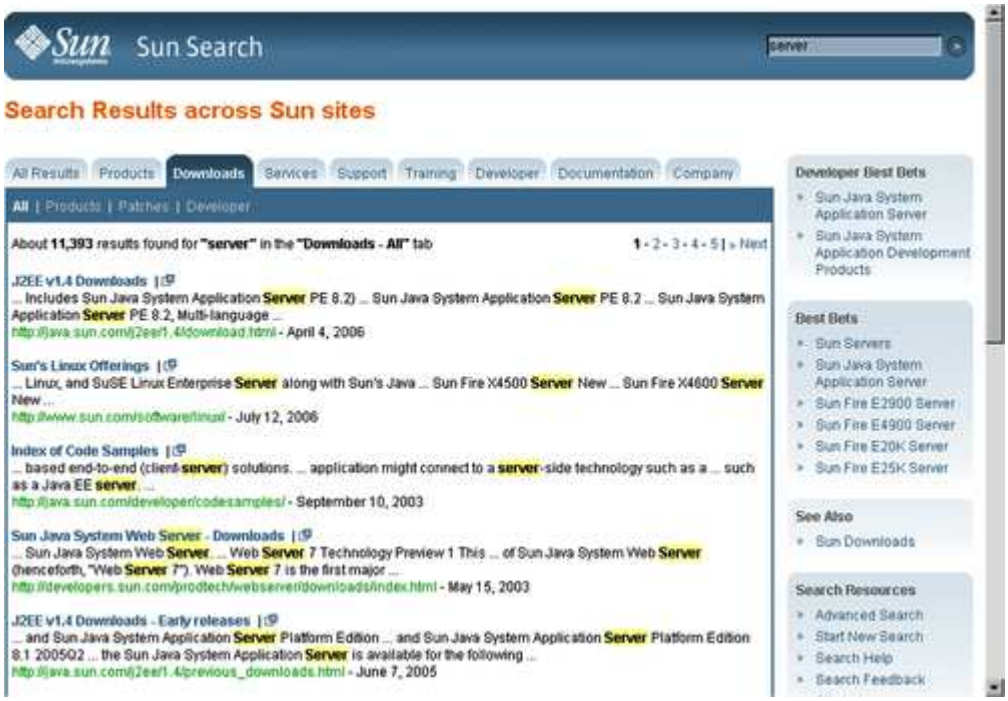


Figura 5 – Sun Web Site, emprego do grouping

Segundo Kalbach (2007), esse é um exemplo de integração bem sucedida das abordagens de busca e navegação. Nitidamente trata-se de resultados de uma busca, porém o agrupamento se assemelha em termos de interface, e também se comporta, como uma navegação tradicional.

2.4.2. MÉTODO TOPIC LINKS

Em geral os resultados de uma busca contêm detalhes sobre cada página ou item encontrado, bem como o tópico de cada um deles. Se os objetos dessa coleção estiverem catalogados através dos tópicos existentes no acervo, então ao clicar na categoria de um dos resultados seria possível ter acesso a todos daquela mesma tipificação (KALBACH, 2007).

A vantagem é oferecer ao usuário mais material sobre o mesmo ponto de interesse, a mesma temática. Pode ser citado, como exemplo de uso dessa técnica, o portal de notícias Digg (2009). O portal possui 35 tópicos, ou categorias, de assuntos. Os resultados de busca incluem um *link* de tópico em cada um dos itens apresentados como resposta, como pode ser observado na Figura 6. Ao clicar sobre qualquer um dos links, uma nova página de resultados é apresentada, desta vez com todos os itens daquela mesma categoria.

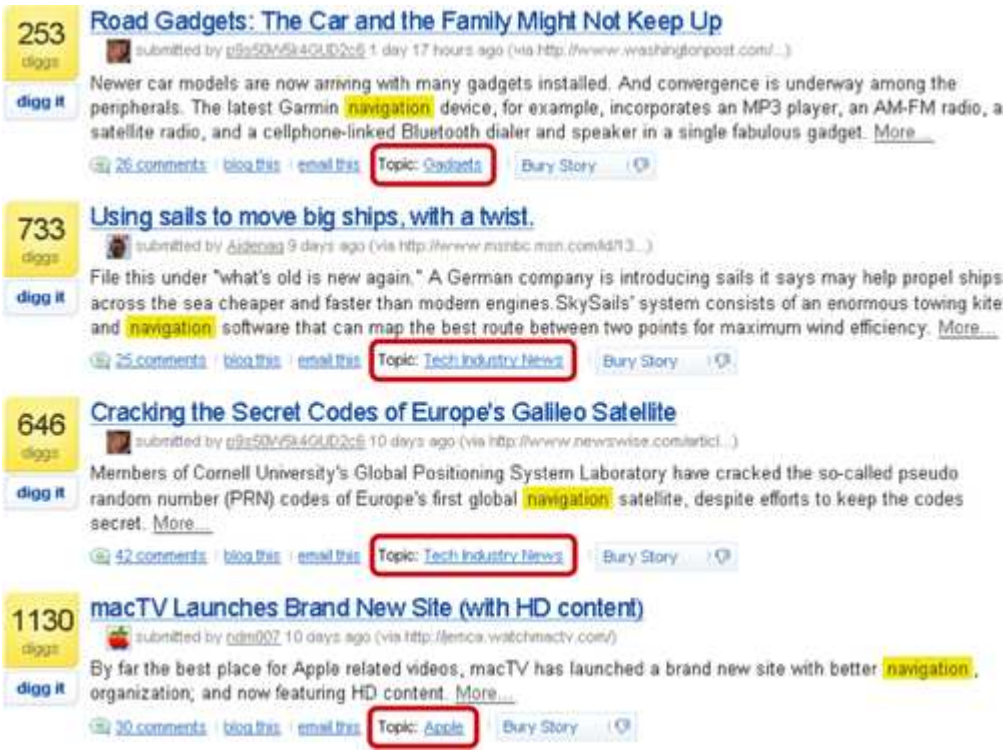


Figura 6 – Digg.com e o uso de Topic Links

2.4.3. MÉTODO SUGESTÕES

Outro recurso também utilizado pelos mecanismos de busca é o de Sugestões (*Suggestions*). O objetivo de tal estratégia é oferecer ao usuário termos alternativos para serem utilizados após a execução da consulta. A idéia é melhorar os resultados obtidos no passo anterior ou simplesmente ajudar o visitante a refinar as informações apresentadas na busca (HEARST, 2009).

Geralmente uma lista de palavras é formada com base no conteúdo do conjunto de resultados encontrados. Existem diferentes formas de realizar a sugestão de termos, e também diversas maneiras de se realizar a geração automática da lista de sugestões (KALBACH, 2007).

Na Figura 7 é possível acompanhar um exemplo de uso da técnica em questão. Trata-se do site de leilões online eBay (2009). Na figura são apresentadas as propostas de termos exibidas ao se realizar uma busca usando a palavra "*speakers*". As opções sugeridas no topo do quadro de alternativas se assemelham às relações associativas. Elas conduzem a opções de conteúdo correlacionado. A parte inferior do quadro apresentado, no entanto, fornece acesso às alternativas de refinamento da busca efetuada.

Figura 7 – eBay.com e a sugestão de termos

2.4.3.1. MÉTODO SUGESTÕES - *FEEDBACK* DE RELEVÂNCIA

Segundo Baeza-Yates & Ribeiro-Neto (1999), o conceito tradicional de *feedback* de relevância (*relevance feedback*) se refere a ciclos de interações nas quais o usuário seleciona um pequeno conjunto de documentos que aparentam ser relevantes para a consulta. Em seguida, o sistema utiliza características derivadas das escolhas feitas para revisar a consulta usada inicialmente.

De acordo com Kalbach (2007), se alguém explicitamente indicar que alguns documentos são mais relevantes do que outros, o sistema pode incorporar essa informação para melhorar os resultados apresentados. Ainda de acordo com Kalbach (2007), o escopo da técnica de *feedback* de relevância foi ampliado. Mecanismos de busca baseados na Web vêm adotando estratégias extremamente simples, como botões "*More like this*". De um modo geral, essa abordagem se refere a qualquer técnica na qual o usuário possa oferecer retroalimentação sobre a relevância dos documentos encontrados, com o objetivo de melhorar ou estender os resultados.

O princípio maior por trás da idéia é que o processo de busca é iterativo. Informações encontradas em um primeiro momento podem redefinir a estratégia de busca de uma pessoa. O problema encontrado é que os usuários dificilmente querem participar desse processo de retroalimentação, a menos que seja extremamente simples, como clicar em algumas estrelas, atribuindo uma nota a um dado resultado da busca.

Pseudo-relevance feedback, também conhecido como *local feedback* ou *blind feedback*, é uma variante automatizada do processo de retroalimentação pelo usuário. A idéia consiste, basicamente, em extrair os termos originários dos documentos classificados nos primeiros lugares dos conjuntos de resultados de busca. Esses termos então são utilizados na execução de uma segunda consulta, de forma automática ou manual.

Um exemplo do uso de *pseudo-relevance feedback* é o mecanismo de consultas Ask.com (2008), apresentado na Figura 8. O sistema extrai os termos de cada conjunto de consultas e os apresenta como *links* no lado direito da página. Dessa forma, um resumo semântico é oferecido por sobre o conteúdo do site.

Caso o usuário clique em algum dos links exibidos, uma nova consulta é executada para o termo ou frase selecionado. Novamente o *pseudo-relevance feedback*

é acionado. Os termos são extraídos dos documentos classificados em primeiro lugar no conjunto de resultados e, em seguida, exibidos na listagem do lado direito do site. O processo continua a se repetir enquanto o usuário assim o desejar.

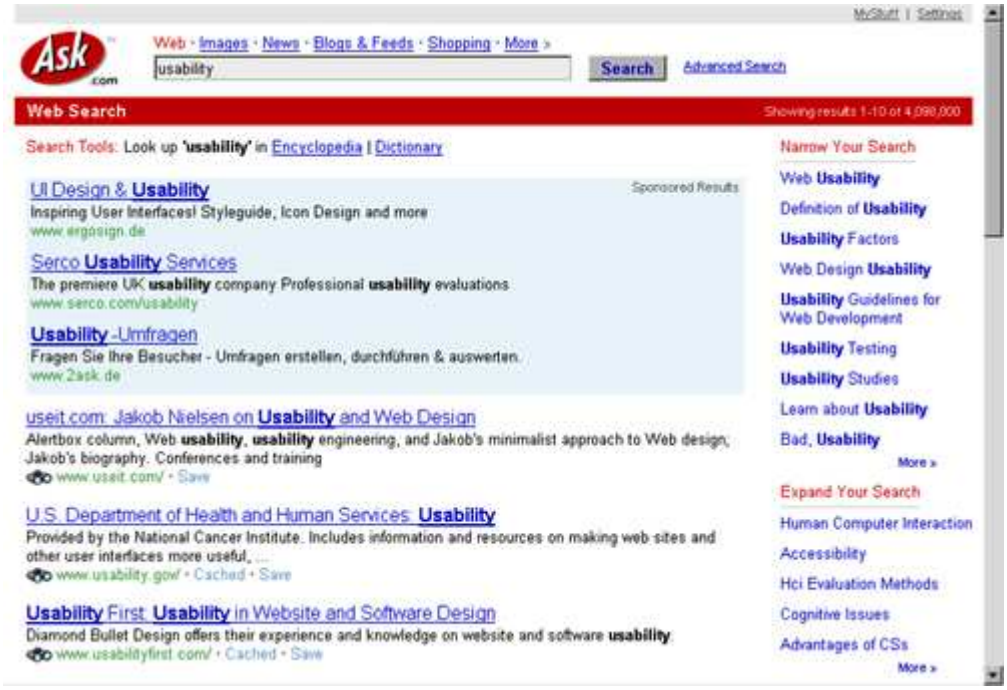


Figura 8 – Ask.com e a técnica de *pseudo-relevance feedback*.

2.4.3.2. MÉTODO SUGESTÕES - *BEST BETS*

Outra abordagem para realizar sugestões é predeterminar as páginas que serão incluídas nos resultados quando determinadas palavras-chaves são utilizadas. Tal abordagem recebe o nome de *Best bets* (KALBACH, 2007).

As recomendações manuais que aparecem no lado direito do site da Sun, apresentada pela Figura 5, por exemplo, são divididas em duas categorias de *Best bets*. O site da BBC (2009) também faz uso da técnica em questão. As recomendações aparecem no topo da lista de resultados. Na Figura 9, uma busca pela palavra *Eastenders* (um programa de televisão popular na Inglaterra), produziu três links marcados com o rótulo “BBC Best Link”.

Figura 9 – Website da BBC, abordagem de Best Bets utilizada.

A técnica de recomendação manual vem sendo largamente utilizada para patrocínio de sites e serviços na Web. Um dos exemplos mais presentes atualmente é o mecanismo de buscas Google (2009). Ao realizar uma busca pela palavra “carros”, por exemplo, a ferramenta de busca exibe os anunciantes na seção “Links Patrocinados”, como pode ser observado na Figura 10. A posição de destaque, acima dos resultados da busca, confere maior visibilidade ao anúncio. Além disso, a interface pode induzir o usuário a acreditar que são os primeiros itens do resultado da sua busca. Dado que o Google organiza os itens do conjunto de resultado por ordem decrescente de relevância, o visitante pode ser levado a julgar que são os itens mais importantes do conjunto da página de resultados.

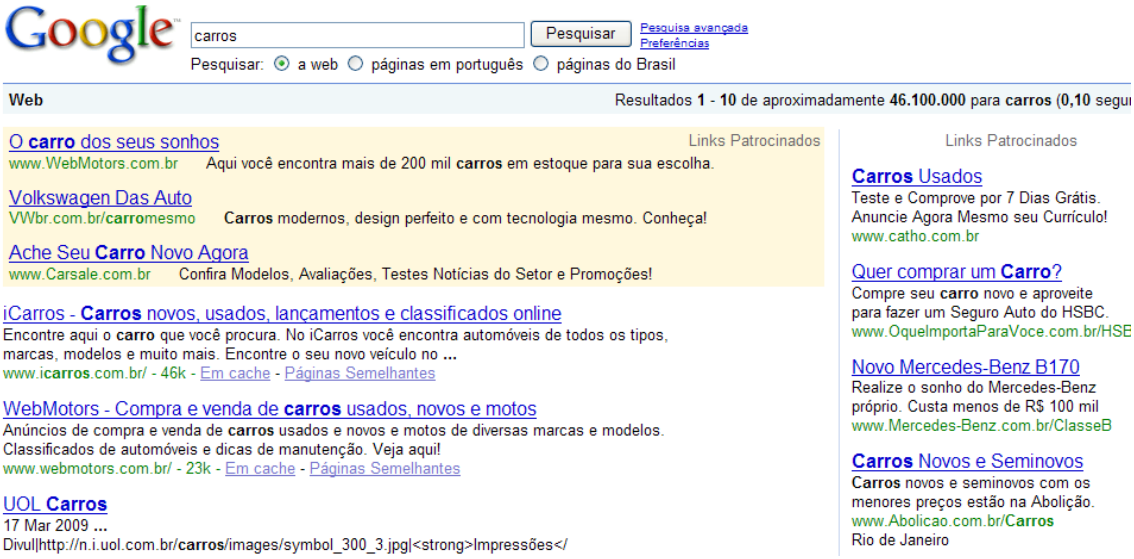


Figura 10 – Google e o uso de Best Bets

2.4.3.3. MÉTODO SUGESTÕES – NENHUM RESULTADO

No lugar de exibir uma simples e desagradável mensagem informando que nenhum item foi encontrado, os mecanismos de busca agora oferecem alternativas para essa situação. Usando a navegação, os resultados são expandidos, oferecendo aos usuários opções, mesmo na ausência de resultados específicos.

Uma estratégica consiste em confrontar os termos utilizados na busca com um índice interno e oferecer, por exemplo, alternativas de ortografia, para casos em que houve distração do visitante. Várias máquinas de busca na Web oferecem hoje esse tipo de serviço.

Outra estratégia é exibir categorias para navegação ou uma parte do mapa do site. A premissa é que, se o sistema não pode detectar um erro na ortografia, então talvez a navegação, em termos inicialmente ausentes da consulta, possa auxiliar o usuário. Além disso, navegar por palavras ou opções não previstas antes pelo visitante pode oferecer uma idéia do conteúdo de um site.

A página de resultados do site da empresa HP (2009) fornece as pesquisas mais comuns entre os visitantes da página. As sugestões podem ser encontradas ao final da página

contendo o resultado da busca (Figura 11). A idéia é transformar o que seria o ponto de término do uso do site no início de uma nova busca e navegação.

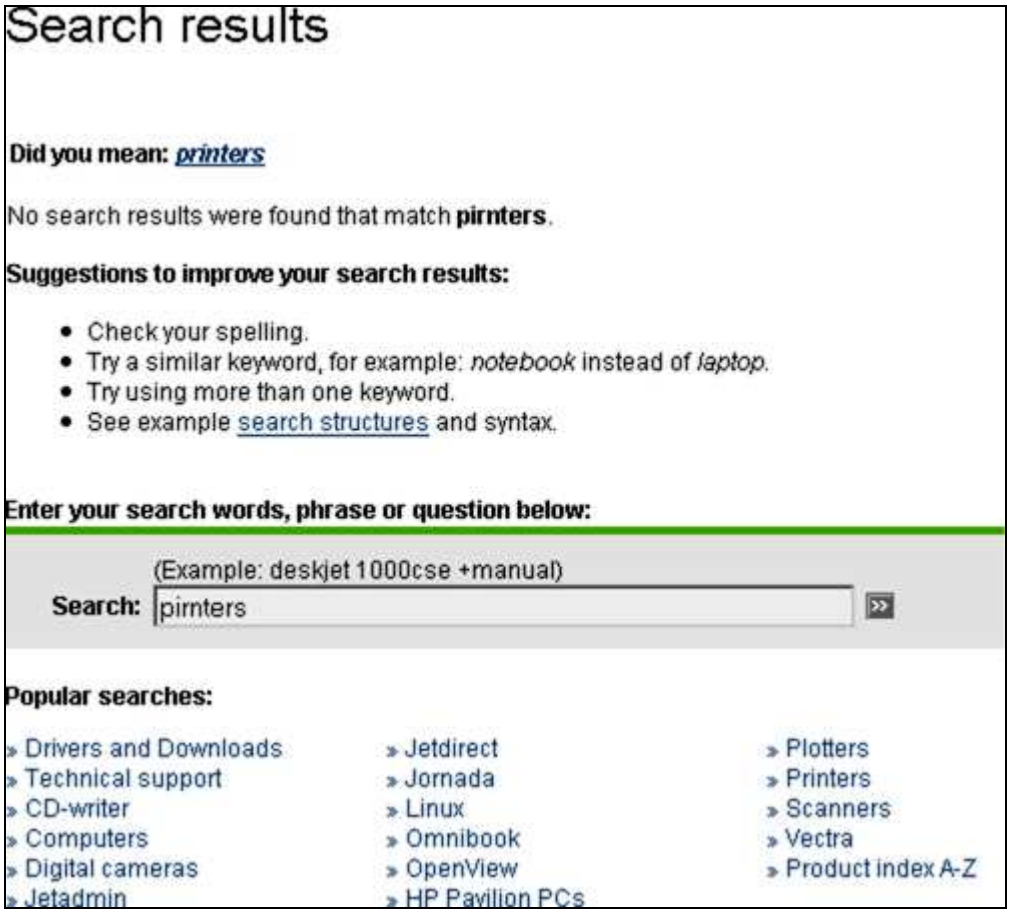


Figura 11 – HP: Página de sugestões de pesquisas populares

2.4.4. MÉTODO SCOPING SEARCH

Existem várias maneiras em potencial de se realizar a integração de mecanismos de navegação dentro do processo de busca. É possível incluir opções no próprio formulário de busca tais como menus de *Scoping Search* ou *Word Wheel*, que permitem ao usuário se concentrar no foco da sua busca, delimitando o seu escopo de atuação através da seleção de pré-filtros do conteúdo disponibilizado. Tais estratégias permitem ainda exibir alguns termos candidatos, auxiliando a formulação da consulta (KALBACH, 2007).

Alguns formulários utilizados para busca apresentam uma opção para limitar a abrangência da pesquisa antes de executar a consulta propriamente dita. Freqüentemente essa

opção é implementada através de um menu do tipo *drop-down* que contem os principais tópicos daquele acervo. O visitante então pode escolher uma dada seção e limitar o escopo da sua busca apenas àquela área. Essa técnica é especialmente indicada para sites cujo conteúdo seja muito abrangente em termos de assuntos, ou então aqueles que possuam elevado número de documentos.

Bom exemplo de uso da abordagem mencionada é o site norte americano de vendas de roupas para práticas esportivas Eddie Bauer (2008). Na página principal é oferecido ao visitante um pequeno menu *drop-down* que permite delimitar o escopo da intenção da pesquisa a ser realizada. As categorias disponíveis para seleção refletem a maior parte do conteúdo existente, sendo claramente diferenciadas, o que permite um conjunto de resultados mais próximos do esperado, como apresentado pela Figura 12.

Para ilustrar uma típica situação de uso, um homem que estivesse visitando o site à procura de sapatos, poderia escolher a opção “homem” no menu. Dessa maneira veria apenas resultados relativos a sapatos masculinos.

Figura 12 – Exemplo da técnica *Scoping Search*

2.4.5. MÉTODO WORD WHEEL

O recurso *Word Wheel* consiste em exibir uma pequena lista de termos previamente utilizados na indexação dos documentos, funcionando como palavras-chave (KALBACH, 2007). Por exemplo, ao digitar a letra N uma lista contendo cinco ou dez termos iniciados por essa letra é exibida para seleção. Ao continuar a digitação, dessa vez escolhendo a letra A, uma nova lista de itens é exibida, dessa vez com termos iniciando com a sílaba NA. E assim por diante.

Da mesma maneira ao se reduzir o número de letras informadas no campo de busca, a lista retorna para o conjunto anterior de opções. A qualquer momento um dos itens oferecidos na lista pode ser selecionado para sua utilização na busca.

O site Answers.com (2008) faz uso da abordagem *Word Wheel* (Figura 13). Um pequeno atraso na produção da lista de termos é percebido ao se digitar as letras, mas não inviabiliza ou sequer prejudica o potencial apresentado pelo recurso.

Figura 13 – Exemplo de uso de Word Wheel

Cabe ressaltar que essa abordagem não possui relação com o recurso utilizado em alguns navegadores, aonde é exibida uma lista de palavras ou frases previamente digitadas e armazenadas de forma a aguardar um histórico do uso. Trata-se de uma lista de termos associados às páginas ou a outros documentos.

2.4.5.1. MÉTODO *WORD WHEEL - PATTERN MATCHING*

Uma estratégia um pouco mais sofisticada de uso do *Word Wheel* é chamada de *Word Wheel Pattern Matching*. Através dessa técnica é possível realizar a comparação não somente com as primeiras letras de uma palavra, mas sim com qualquer ocorrência daquela string dentro dos termos pré-indexados (KALBACH, 2007).

O PubMed (2009) é uma grande base de dados online para pesquisas médicas. O mecanismo de busca para artigos e revistas nessa coleção é feito através da abordagem de pattern matching.Após a entrada de pelo menos dois caracteres, o word wheel busca todas as ocorrências daquela string em qualquer posição dos termos armazenados. Na Figura 14, é possível acompanhar o resultado da digitação da expressão “cardio”. Como resultado é obtido não apenas a *The American Journal of Cardiology*, mas também a *Cardiology Research*, dentre outras.

Figura 14 – Exemplo de Word Wheel Pattern Matching

A vantagem do uso dessa abordagem consiste na diminuição dos problemas tão comumente encontrados nas ferramentas de busca com os sinônimos ou mesmo as variantes de algum termo ou expressão. Além disso, erros de ortografia ficam evidenciados, reduzindo o tempo gasto com buscas equivocadas.

2.4.5.2. MÉTODO *CANNED SEARCHES*

A exemplo de um recurso largamente utilizado em mecanismos de busca, a opção “Encontre mais resultados”, os links que fazem uso da opção *Canned Search* contêm um

atalho para a execução de uma busca. *Links* que implementam esse tipo de estratégia são facilmente reconhecidos pela sua URL, que possui os termos como parâmetros de uma *string* (KALBACH, 2007).

Particularmente úteis para direcionar o visitante para um conteúdo desconhecido, ou sobre o qual não se possui a gestão, a estratégia em questão é largamente utilizada na ligação entre sites de conteúdo igual ou similar. Outro benefício reside no fato de que a página de resultado da busca pode ser organizada de forma automática, o que evitaria a atualização manual do conteúdo de um site de vendas, por exemplo. A desvantagem estaria na falta de personalização que alguns usuários poderiam exigir de determinados tipos de conteúdo.

Exemplo do uso de *canned search*, o site Wine.com (2008) exibe uma página de resultados de busca, após o usuário clicar na opção "*Wine Collections*". Não se trata de uma página de galeria escondida, mas sim do resultado de uma busca, incluindo o número de resultados encontrados. Além disso, os visitantes podem continuar sua navegação por sobre os resultados, fazendo uso do menu "*narrow search by*" presente à esquerda da página (Figura 15).

Figura 15 – Exemplo de uso de *Canned Search*

2.5. CONSIDERAÇÕES FINAIS

No presente capítulo foram apresentados os métodos de busca direta, navegação e suas abordagens combinadas existentes na literatura. Em particular, os métodos de agrupamento, tanto o plano quanto o hierárquico, e a navegação facetada foram detalhados por estarem presentes nas hipóteses do presente trabalho de pesquisa. O detalhamento das hipóteses, bem como suas operacionalizações, são temas do próximo capítulo.

3. METODOLOGIA DE PESQUISA

Neste capítulo, é abordada a metodologia de pesquisa que guiou esta investigação, realçando-se seu caráter multimétodo, ou seja, com uma etapa qualitativa utilizando o método de estudo de caso e outra etapa quantitativa apoiada em um experimento.

Inicialmente, o caráter científico da investigação é discutido assim como sua classificação segundo critérios definidos na literatura. Em seguida, a seleção dos métodos de recuperação de informação é delineada, e as questões da pesquisa são apresentadas.

A seguir, a abordagem de avaliação das hipóteses é discutida, através do universo investigado e da amostra utilizada no experimento, bem como da coleta de dados e de sua análise posterior.

3.1. CLASSIFICAÇÃO DA PESQUISA

A condução de uma pesquisa segundo princípios científicos pressupõe a eleição do paradigma que servirá de guia para o processo de investigação. Creswell (1994, 1998) distingue os principais paradigmas de pesquisa existentes: qualitativo, quantitativo e multimétodo:

- Uma pesquisa quantitativa visa entender os problemas sociais ou humanos a partir de testes da teoria existente. Para tanto, faz uso de variáveis medidas por números e analisadas com procedimentos estatísticos. Esse paradigma é também chamado de tradicional, positivista, pós-positivista, experimental, ou ainda de empirista.
- Uma pesquisa qualitativa baseia suas contribuições em uma visão holística e complexa formada por palavras que relatam as interpretações dos entrevistados. O paradigma qualitativo é conhecido também como construtivista, naturalista ou interpretativo.
- Uma pesquisa multimétodo, ou mista (do inglês *mixed*), caracteriza-se pela mescla das abordagens qualitativa e quantitativa.

Existem três estratégias associadas à abordagem multimétodo:

- Seqüencial: visa aprofundar ou expandir os resultados encontrados em uma primeira fase. Pode iniciar por uma fase quantitativa e depois seguir uma fase qualitativa, ou vice-versa.
- Paralela: efetua-se a coleta dos dois tipos de dados no mesmo momento e, em seguida, é feita uma análise de forma distinta, esperando convergir os resultados.
- Transformativa: ocorre a coleta dos dois tipos de dados, porém obedecendo a diferentes tópicos de interesse. Parte da pesquisa é qualitativa e parte quantitativa.

A presente dissertação faz uso do paradigma de investigação multimétodo, como ilustra a Figura 16. A etapa quantitativa tem por objetivo coletar as medidas relativas à utilização de cada método de recuperação de informação investigado, viabilizando a avaliação dos mesmos quanto a sua eficiência. A etapa qualitativa, por sua vez, busca aprofundar a análise da eficiência sob a ótica da percepção do usuário de ferramentas de RI. Como observado por Manning, Raghavan e Schütze (2009), as medidas quantitativas formais estão freqüentemente a certa distância do objetivo final dos sistemas de RI: a satisfação do usuário.

A estratégia utilizada é a paralela (CRESWELL, 1998), ou seja, com a coleta de dois tipos de dados seguida de análise distinta, de caráter confirmatório, objetivando a triangularização e a convergência dos resultados.

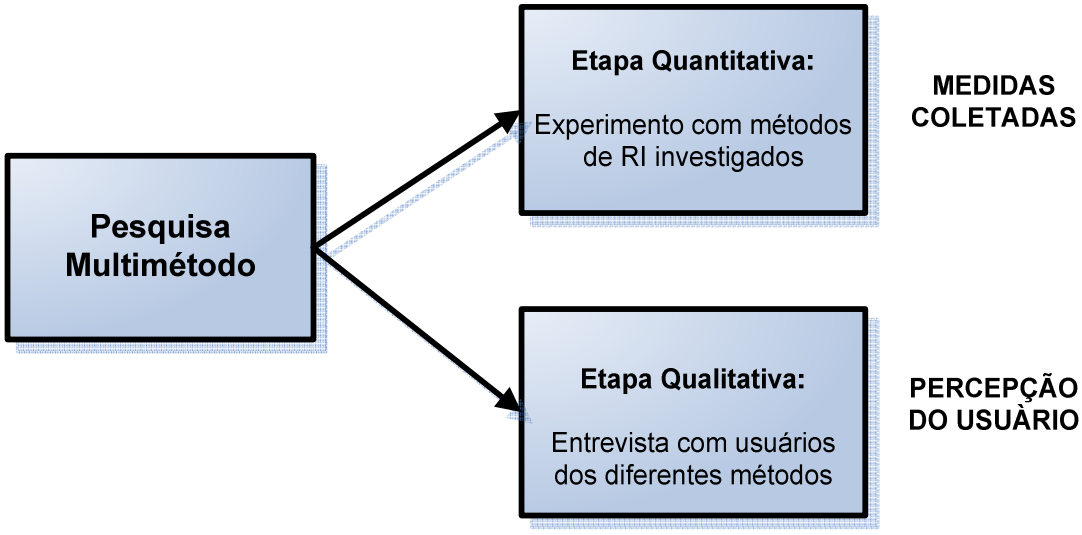


Figura 16 – Etapas do trabalho de pesquisa

Na etapa quantitativa, o método utilizado é o experimental, por ser adequado aos questionamentos que o estudo coloca. Um experimento tem como base a observação de um fenômeno em um ambiente controlado, no qual o pesquisador elabora um cenário de pesquisa e define as variáveis a serem observadas (DIAS & SILVA, 2010). Segundo Vergara (2009), um experimento (em campo ou em laboratório) é uma investigação empírica na qual o pesquisador manipula e controla diretamente as variáveis independentes e observa as variações das variáveis dependentes.

Neste estudo temos como variáveis independentes os métodos de recuperação da informação, o perfil dos participantes, o número de tarefas a serem executadas, bem como o tipo das mesmas. As variáveis dependentes são caracterizadas pela precisão do resultado de busca, pela cobertura do mesmo resultado, pela taxa de erro decorrente da realização das atividades propostas, pelo número de documentos relevantes encontrados por minuto, e ainda, pelo esforço empregado pelo usuário, a se refletir no tempo necessário para realização das tarefas.

Já na etapa qualitativa, utiliza-se o método de entrevista por pauta, pelo seu caráter exploratório, no qual o entrevistador agenda vários pontos para serem averiguados com o

entrevistado, sendo o método de entrevista que permite maior profundidade na abordagem dos temas (VERGARA, 2009).

Segundo Vergara (2009) e Yin (1994), quanto aos objetivos, esta dissertação pode ser classificada como:

- Exploratória, já que versa sobre um fenômeno relativamente recente (o da recuperação de informações com métodos combinados de busca direta e de navegação); e
- Descritiva, uma vez que visa explicitar as características e diferenças entre os métodos analisados.

A natureza empírica da presente investigação é confirmada por Demo (2000), pois lida com dados reais provenientes de um experimento. Caracteriza-se, na categoria mencionada, também pela observação do fenômeno na prática, contrapondo-se a pesquisas de cunho somente teórico.

3.2. SELEÇÃO DAS ABORDAGENS DE RECUPERAÇÃO DE INFORMAÇÃO

A presente investigação teve como foco o estudo comparativo das seguintes abordagens de recuperação de informação: (i) busca direta, (ii) navegação, (iii) agrupamento plano, (iv) agrupamento hierárquico e (v) navegação facetada. Buscou-se verificar a eficiência de tais métodos, bem como a satisfação dos usuários associada aos mesmos.

A navegação e a busca direta, ambas utilizadas de forma isolada, foram selecionadas devido a sua utilização amplamente difundida, uma vez que se tratam dos métodos de recuperação de informação mais tradicionais (MANNING, RAGHAVAN & SCHÜTZE, 2009; HEARST, 2009). Aliado ao fato mencionado, também pela aderência aos questionamentos trazidos no presente trabalho, tais métodos de recuperação de informação não poderiam estar ausentes do escopo das experimentações.

Dentre os métodos apontados pela literatura como sendo os mais promissores, frutos da combinação das características pertencentes à navegação e à busca direta, o método de agrupamento recebe papel de destaque (CARPINETO et al., 2009; KALBACH, 2007; HEARST, 2006, 2009). Diversos pesquisadores debruçaram-se sobre o tema, o que, no entanto, não garante homogeneidade nas conclusões. Muitos acreditam que o método de agrupamento hierárquico produz melhores resultados (JAIN et al., 1988), (CUTTING et al., 1992), (LARSEN & AONE, 1999). No entanto, não há um consenso sobre essa questão. Recentemente alguns experimentos sugeriram o contrário, como pode ser observado em (ZHAO et al., 2002). Visando proporcionar alguma contribuição para a questão acima, e também pela evidência de bons resultados apontada pela literatura, ambos os métodos de agrupamento (plano e hierárquico), foram escolhidos como parte da presente investigação.

Destacada como a combinação mais holística dentre as técnicas de navegação e busca, a navegação facetada tem sido tema cada vez mais explorado tanto pelas pesquisas acadêmicas quanto por aplicações comerciais. No âmbito acadêmico, o projeto Flamenco, conduzido por Marti Hearst (HEARST, 2006a), representa mais de uma década de desenvolvimento de sistemas de navegação facetada e aplicação de estudos de usabilidade. O projeto Relation Browser++, liderado por Gary Marchionini (ZHANG, J. & MARCHIONINI, 2005) é um dos esforços mais conhecidos na mesma área, apresentando igualmente bons resultados. Dentre as iniciativas comerciais, o eBay lançou, em 2006, um projeto denominado Ebay Express (EBAY, 2006), um site comercial que empregava as técnicas de navegação facetada. Antes dessa iniciativa, a Amazon, em 2002, também fez uso da mesma técnica no seu projeto Ruby (COX, 2002).

Aliadas às razões acima explicitadas, os métodos selecionados para aprofundamento na investigação apresentavam quantidade suficiente de ferramental disponível para concretização do trabalho.

3.3. HIPÓTESES DA PESQUISA

O objetivo da pesquisa é responder as seguintes perguntas, tendo em vista o âmbito dos domínios homogêneos de conhecimento, a utilização de acervos expressivos em tamanho, e a realização de buscas não-exploratórias:

1. Os métodos que combinam navegação e busca, considerando agrupamento plano, agrupamento hierárquico e navegação facetada, são mais eficientes do que a navegação e a busca utilizados isoladamente?
2. Qual técnica de recuperação de informação oferece maior eficiência, dentre agrupamento plano, agrupamento hierárquico e navegação facetada?

Segundo o levantamento bibliográfico realizado, foram construídas as hipóteses apresentadas pelo Quadro 1, de acordo com os trabalhos dos autores correlacionados na coluna “Autores”.

Hipótese	Autores
H1: a busca direta é menos eficiente que o agrupamento plano.	(DUMAIS et al., 2001) (ZAMIR et al., 1999) (KAKI, 2005) (KULES & SHNEIDERMAN., 2008) (MANNING, RAGHAVAN & SCHÜTZE, 2009)
H2: a busca direta é menos eficiente que o agrupamento hierárquico.	
H3: a busca direta é menos eficiente que a navegação facetada.	
H4: a navegação é menos eficiente que o agrupamento plano.	(BAEZA-YATES & RIBEIRO-NETO,1999)

H5: a navegação é menos eficiente que o agrupamento hierárquico	(MARCHIONINI, 2006) (KALBACH, 2007)
H6: a navegação é menos eficiente que a navegação facetada.	
H7: o agrupamento plano é menos eficiente que o agrupamento hierárquico.	(JAIN, A., DUBES, R., 1988) (CUTTING et al., 1992) (LARSEN & AONE, 1999)
H8: o agrupamento hierárquico é menos eficiente que a navegação facetada.	(HEARST, 2006b, 2009) (MANNING, RAGHAVAN & SCHÜTZE, 2009)

Quadro 1 – Lista de Hipóteses

As hipóteses H1, H2, H3, H4, H5 e H6 visam responder a primeira pergunta da pesquisa, ou seja, quanto aos métodos que fazem uso de busca e de navegação de maneira combinada possuírem eficiência superior aos que os utilizam de forma isolada. As hipóteses H7 e H8 visam verificar a eficiência dos métodos investigados conforme a segunda pergunta da pesquisa.

3.4. OPERACIONALIZAÇÃO DAS HIPÓTESES

3.4.1. COLETA DOS DADOS

Visando a captação de informações para subsidiar a validação das hipóteses, os dados quantitativos foram coletados no decorrer do experimento com os cinco métodos de recuperação de informação. O processo foi constituído de dez tarefas e de uma entrevista com os participantes ao final da sessão. Para a realização do experimento foi utilizado um acervo

de informações referente a uma coleção de filmes (IMDB, 2009), que será detalhada no capítulo seguinte.

Houve uma grande preocupação com a definição das tarefas a serem realizadas pelos participantes, uma vez que poderiam impactar consideravelmente nos resultados da pesquisa. Segundo Nielsen (1993), a regra básica a ser seguida é a de que as atividades devem ser escolhidas visando, ao máximo possível, alcançar a representatividade do uso que será feito do sistema em questão. Somada a essa observação, as tarefas também foram projetadas para serem pequenas o suficiente a fim de caber no tempo estipulado para o experimento, mas não tão simples a ponto de se tornarem demasiadamente triviais. Por exemplo, atividades do tipo "*Encontre o produtor do filme The Matrix*" ou "*Quem foi o compositor de 2001: A Space Odyssey?*" não foram utilizadas.

Segundo Marti Hearst (2009), torna-se necessário ainda levar em conta que os participantes podem cansar, e por isso também é de se esperar que seja realizado apenas um pequeno número de consultas durante a sessão de testes. Além disso, procurou-se evitar a introdução de tarefas que possam gerar resultados irrelevantes, tendo um efeito nulo nas análises futuras.

Pelos motivos expostos, duas tarefas foram destinadas para cada um dos 5 métodos em investigação. Divididas em dois tipos, elas buscaram representar usos reais das ferramentas, bem como permitir olhares distintos do ponto de vista da adequação dos métodos aos propósitos de recuperação de informação relevante. Um tipo de tarefa, denominada para efeitos de referência como Tarefa 1, representava a necessidade de se efetuar uma busca segundo dois critérios presentes na coleção de filmes utilizada. Já o outro modelo de tarefa, permitia o uso exploratório, de caráter livre, da ferramenta, para buscar filmes sobre um tema fornecido.

Aspecto igualmente importante do experimento, a motivação do participante procurou ser alimentada através da busca por temas atraentes, cujo número de títulos disponíveis pudesse ser elevado, e que provocasse interesse como “máfia italiana” ou “nazismo”. A motivação também norteou a delineação da Tarefa 1, levando então a buscas por gêneros específicos de diretores famosos como “Francis Ford Coppola” e “Martin Scorsese”. No entanto, todas as 10 tarefas foram elaboradas de tal forma que os participantes pouco provavelmente pudessem antecipar suas respostas, levando-os, dessa forma, a utilizar os métodos disponíveis.

Uma mesma base de dados, de conhecimento homogêneo, foi utilizada para o experimento realizado com todas as técnicas a fim de permitir a comparabilidade dos resultados. Durante o experimento foram capturadas então as medidas de: tempo para execução das tarefas, respostas para cada tarefa (a fim de avaliar a taxa de erro) e as consultas utilizadas em cada método, para cada tarefa, o que foi feito através de gravação das sessões. As gravações permitiram o cálculo da média da precisão e cobertura dos resultados para cada método.

As informações qualitativas, por sua vez, foram capturadas através de entrevista com os participantes ao final de cada sessão. O roteiro das entrevistas pode ser observado em detalhes no ANEXO D - Roteiro da Entrevista com Usuário. As perguntas presentes no referido roteiro possuem correlação com as medidas quantitativas coletadas, segundo mostra o ANEXO C - Itens de Medida para cada Experimento. O objetivo foi viabilizar o caráter confirmatório, presente na pesquisa de natureza multimétodo, usando o procedimento paralelo proposto por Creswell (1998), que permite coletar dados quantitativos e qualitativos ao mesmo tempo, para posterior análise e triangularização dos dados.

Todas as sessões de experimentação seguiram a cartilha definida no ANEXO A - Roteiro do Experimento para o Pesquisador. Uma breve explicação foi feita a cada

participante sobre o objetivo da pesquisa e a coleção de informações disponíveis para o seu uso. Em seguida, a introdução aos recursos de interface do primeiro método era realizada. Um exemplo de tarefa similar ao que ele deveria executar era demonstrado, e logo depois o participante poderia usar o tempo que julgasse conveniente para se ambientar com o programa. Depois a tarefa referente ao método da vez era entregue, o tempo então passava a ser contado, e a captura da tela do computador utilizado era iniciada. Para cara uma das ferramentas disponíveis os passos descritos eram repetidos.

Com o objetivo de reduzir algum possível viés na execução do experimento, as duas tarefas definidas para cada ferramenta foram diferentes, porém equiparáveis na sua dificuldade de execução e também no número de documentos relevantes como resposta. Ao todo, então, foram elaboradas 10 diferentes tarefas, como pode ser observado no ANEXO B – Roteiro de Tarefas.

Ainda para evitar a introdução de qualquer tipo de viés, com relação às medidas de tempo utilizadas na execução das tarefas, alguns cuidados foram tomados. Em primeiro lugar, os participantes tiveram acesso a todas as cinco ferramentas em um servidor localmente instalado, o que evitou problemas com tráfego de dados em rede. As aplicações estavam igualmente disponibilizadas na máquina de utilização do participante. Um segundo cuidado foi tomado com relação à visualização do histórico de consultas feitas por participantes anteriores, nos menus *drop-down* do navegador. Ao fim de cada turno de experimentação, todo o *cache* de histórico foi devidamente apagado.

3.4.2. UNIVERSO E AMOSTRA

Com relação à escolha dos participantes é importante ressaltar que *"as duas questões mais importantes na recuperação de informações e usabilidade são as tarefas dos usuários, e as suas diferenças e características individuais" (NIELSEN, 1993).* Em um de seus estudos,

Nielsen (1989) avaliou 30 publicações sobre avaliação de sistemas de RI, e verificou que a diferença entre indivíduos respondia por 4 dos 10 aspectos que mais influenciavam o desempenho dos métodos, e 2 deles eram referentes à diferença entre as tarefas.

O experimento foi realizado com 16 participantes, sendo 8 analistas de sistemas de uma grande empresa multinacional, e 8 profissionais de profissões diversas (professor, advogado, dentista, biólogo, psicólogo, farmacêutico, militar). O objetivo era considerar os dados do conjunto com 16 participantes, mas procurar alguma evidência, ainda que pequena, da diferença de comportamento ou resultado entre os dois grupos.

No estudo conduzido por Lindgaard e Chattratichart (2007), não foi encontrada nenhuma correlação entre o número de participantes e o número de problemas detectados pelo uso de diferentes interfaces de busca. No entanto, o trabalho revelou que há uma clara correspondência entre o número de tarefas distintas e a quantidade de problemas verificados. Essas conclusões sugerem que a diferença presente nas tarefas revela distintos aspectos com os métodos utilizados. Por tais razões, as tarefas devem pertencer a tipos muito diferentes entre si, prática que foi adotada na presente pesquisa.

Ainda sobre o perfil dos participantes, o experimento foi conduzido com pessoas entre 24 e 33 anos, possuindo, no mínimo, o ensino superior completo. Todos faziam uso regular de sistemas de busca ou navegação, e detinham conhecimento avançado ou fluente na língua inglesa, o que era relevante uma vez que a coleção de testes utilizada encontrava-se no referido idioma.

3.4.3. ANÁLISE DOS DADOS

Em sistemas cujo objetivo seja prover a recuperação de dados, o tempo de resposta e o espaço necessário são usualmente as duas métricas utilizadas para avaliação (BAEZA-YATES & RIBEIRO-NETO, 1999). Vários aspectos são então analisados através dessa

perspectiva. No entanto, em sistemas voltados para promoção da recuperação de informação, outras métricas são também relevantes.

Uma vez que a consulta do usuário é inerentemente imprecisa, os documentos recuperados não constituem uma resposta acurada, exata. Por esse motivo, tais documentos precisam ser ranqueados de acordo com a sua relevância em relação à consulta efetuada. A necessidade do ranking, inexistente na recuperação de dados, exerce um papel fundamental na recuperação de informação, que demanda uma análise da qualidade do conjunto resposta fornecido pelo sistema.

Tal avaliação é comumente baseada em uma coleção de teste, utilizada como referência, e em métricas de avaliação dos resultados. Uma coleção de teste consiste em um conjunto de documentos, alguns exemplos de consultas a serem empregadas, e um grupo de documentos relevante para cada uma dessas consultas (normalmente preparado por especialistas daquele domínio de conhecimento).

Após a execução da consulta, os documentos recuperados são então comparados com aqueles apontados pelos especialistas como sendo os itens relevantes da coleção. Em seguida, as métricas de avaliação são aplicadas por sobre o resultado encontrado.

As medidas mais utilizadas na análise de desempenho dos sistemas de recuperação de informação são precisão e cobertura, que podem ser definidas da seguinte forma ((BAEZA-YATES & RIBEIRO-NETO, 1999):

- **Cobertura**: fração dos documentos relevantes que foi recuperada.
- **Precisão**: fração dos documentos recuperados que é relevante para as necessidades de informação representadas pela consulta.

Uma análise mais minuciosa, no entanto, revela alguns problemas com as duas medidas. Segundo Manning, Raghavan e Schütze (2009) as métricas de precisão e cobertura servem apenas para os métodos que não fazem uso de classificação de resultados. Quando o

objetivo é avaliar o conjunto ordenado de documentos retornados como resposta à consulta efetuada, os primeiros documentos são naturalmente o objeto de estudo.

Hearst (2009) e Manning, Raghavan e Schütze (2009) concordam que o usuário típico de RI quer convergir para um resultado de forma rápida, e não encontrar todas as possibilidades possíveis presente numa coleção ou banco de dados (toda a cobertura). Por isso o método *Precision at K* ficou mais famoso (HEARST, 2009). A grande vantagem do método é isentar o pesquisador de conhecer toda a cobertura disponível, tarefa impraticável para grandes coleções de documentos. No entanto, o uso de uma constante possui vulnerabilidades, uma vez que a dependência do valor de K se torna demasiadamente elevada.

Como alternativa ao problema, a métrica utilizada nesta dissertação é o *R-Precision* (MANNING, RAGHAVAN & SCHÜTZE, 2009). A abordagem requer o conhecimento do número de documentos relevantes para uma dada consulta. Se existirem |*Rel*| itens relevantes, então são examinados os primeiros |*Rel*| elementos do conjunto ordenado de resultado. Se forem encontrados *r* documentos relevantes nas primeiras |*Rel*| posições, então diz-se que a precisão e a cobertura são iguais a *r*/|Rel|.

A indicação do *R-Precision* reside justamente no fato de que ele utiliza como base de análise a cobertura conhecida, no entanto, não aplica a verificação da cobertura em toda a extensão da coleção, nem mesmo da precisão. A referida abordagem foi, inclusive, adotada como métrica oficial de avaliação na TREC HARD track (ALLAN, 2005).

Para determinar o conjunto de documentos relevantes das tarefas do tipo 1, ou seja, as coberturas referentes a cada tarefa, foi utilizado o método de *Pooling*. Usando tal metodologia, os documentos são pesquisados e definidos como relevantes através da utilização de todas as ferramentas que sejam alvo do estudo em questão, e mais nenhuma outra além dessas. Ou seja, a definição da cobertura, para o presente trabalho, foi realizada

por meio das ferramentas de busca direta, navegação, agrupamento plano, agrupamento hierárquico e pela navegação facetada (MANNING, RAGHAVAN & SCHÜTZE, 2009).

No caso das tarefas do tipo 2, a definição da cobertura ficou a cargo da utilização das cinco ferramentas por todos os participantes do experimento. A proposta tem origem também no método de *Pooling*. Dado ao tamanho da coleção de documentos utilizada, cerca de 115 mil, seria inviável definir de forma exaustiva a cobertura de tais assuntos, a exemplo do que foi feito para as tarefas de tipo 1.

Para determinar os erros e acertos nas respostas fornecidas pelos participantes, uma folha de respostas foi fornecida a cada participante para preenchimento durante a utilização das diferentes ferramentas. Após a execução do experimento, as respostas fornecidas foram confrontadas individualmente, de forma manual, com a cobertura previamente definida. De tal maneira, também foram determinados a taxa de erro por ferramenta e o número de documentos relevantes encontrados por minuto.

Em relação ao cálculo do *R-Precision*, seriam necessárias, no melhor caso, 160 análises individuais, uma vez que foram aplicadas 2 tarefas, para cada um dos 5 métodos, gerando 10 resultados, por 16 participantes diferentes. Como cada participante, assim como na vida real, pode executar mais de uma consulta para cada tarefa, no melhor caso, em termos de análises, a conta chegaria a 160. Tal esforço de avaliação não é compatível com o tempo de uma dissertação, visto que cada sessão teve uma duração média de 90 minutos. No caso do agrupamento, por exemplo, se considerado todos os cliques em diversos grupos, ainda que numa mesma consulta, o número de análises cresceria enormemente.

Portanto, para calcular o valor do *R-Precision* foram selecionadas as melhores consultas de cada par (tarefa, método) realizadas no experimento, levando em conta todos os participantes envolvidos. Para definir a melhor consulta foi utilizado o critério de maior

número de acertos por minuto, ou seja, as consultas que retornaram o maior número de documentos relevantes em menor tempo.

A combinação dos parâmetros obtidos através do experimento e da entrevista com o usuário viabilizou a análise das hipóteses: precisão, cobertura, tempo utilizado para execução das tarefas, número de documentos relevantes encontrados por minuto, e a taxa de erro. Para o presente trabalho, a eficiência foi considerada como sendo a combinação do maior R-Precision, com o maior número de documentos relevantes encontrados por minuto, fazendo uso do menor tempo, e apresentando menor taxa de erro. A Figura 16 ilustra o caráter experimental da pesquisa, bem como o relacionamento entre as variáveis componentes da mesma.

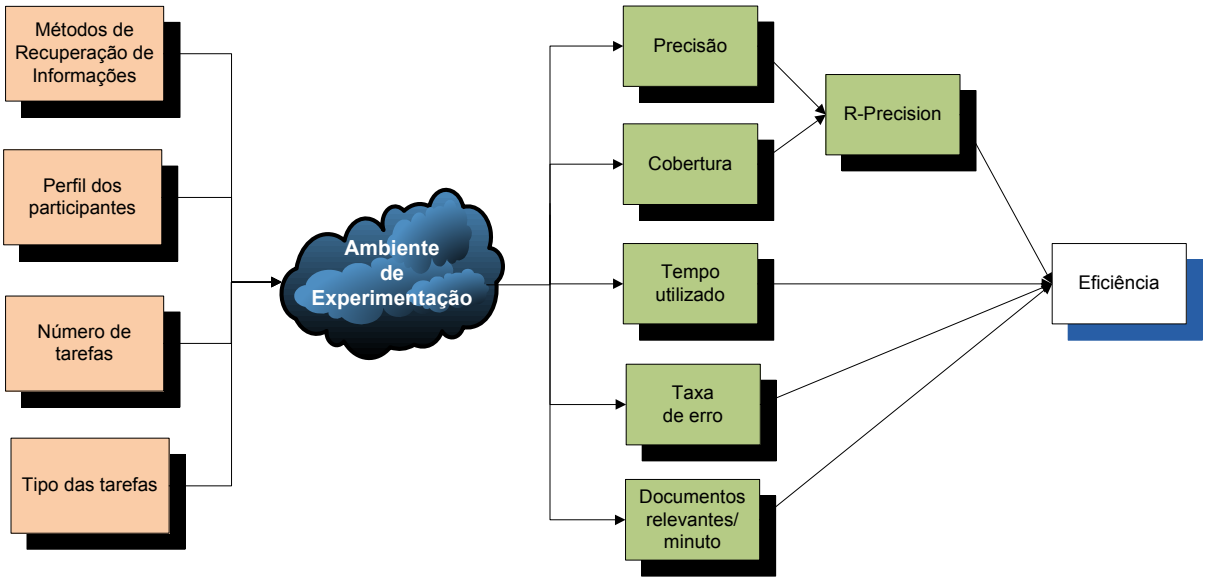


Figura 17 – Relacionamento entre variáveis independentes e dependentes.

Na Figura 17 é apresentado o relacionamento entre as variáveis independentes e dependentes da pesquisa. As variáveis independentes são representadas pelos métodos de recuperação de informação, pelo perfil dos participantes, pelo número de tarefas a serem executadas, bem como pelo tipo das mesmas. As variáveis dependentes são caracterizadas

pela precisão do resultado de busca, pela cobertura do mesmo resultado, pelo número de documentos relevantes encontrados por minuto, pela taxa de erro decorrente da realização das atividades propostas, e ainda, pelo esforço empregado pelo usuário, a se refletir no tempo necessário para a realização das tarefas.

4. O AMBIENTE DE EXPERIMENTAÇÃO

Neste capítulo, é apresentado o trabalho realizado referente à preparação do ambiente de experimentação. Na primeira seção, a pesquisa por uma coleção de testes disponível, adequada aos objetivos da investigação, é exposta. Segue-se com uma descrição do processo que envolveu a criação de uma coleção própria, que viabilizasse tanto o uso de métodos de agrupamento, quanto as necessidades inerentes da navegação por facetas. Na última seção, é apresentada uma análise das três ferramentas inteiramente desenvolvidas, bem como de duas outras que demandaram ajustes no seu desenvolvimento para atender aos requisitos da pesquisa. A seção inclui ainda a exposição dos preparativos concernentes à máquina de busca utilizada por todas as cinco ferramentas envolvidas na presente investigação.

4.1. COLEÇÃO DE TESTES UTILIZADA

Uma vez que, no âmbito da recuperação de informação, a consulta do usuário é inerentemente imprecisa, os documentos recuperados não constituem uma resposta exata. Por esse motivo, tais documentos precisam ser ranqueados de acordo com a sua relevância em relação à consulta efetuada. A necessidade do ranking, inexistente na recuperação de dados, demanda uma análise da qualidade do conjunto resposta fornecido pelo sistema.

Tal avaliação é comumente baseada em uma coleção de teste, utilizada como referência pelos pesquisadores do tema, e em métricas de avaliação dos resultados. Uma coleção de teste consiste em um conjunto de documentos estruturados, alguns exemplos de consultas a serem empregadas, e um grupo definido de documentos relevantes para cada uma dessas consultas (normalmente preparado por especialistas daquele domínio de conhecimento).

Em um primeiro momento, a busca por uma coleção de testes disponível foi realizada dentre as bases formais já reconhecidas e utilizadas pelo meio acadêmico. O Quadro

2 relaciona as lista de coleções analisadas, seus respectivos assuntos, bem como aponta as referências utilizadas na análise.

COLEÇÃO	ASSUNTO	REFERÊNCIA
20 NEWSGROUPS	Artigos do newsgroup da Usenet (20 categorias diferentes)	(MANNING, 2009)
ADI	Documentos sobre a ciência da informação.	(YATES & NETO, 1999)
CACM	Abstracts de artigos da computação, eletrônica e física.	(YATES & NETO, 1999)
CLEF	Idiomas da Europa para recuperação cross-language	(MANNING, 2009)
CRANFIELD	Abstracts de artigos sobre aerodinâmica	(MANNING, 2009; HEARST, 2009)
Cystic Fibrosis DB	Aspectos da Fibrose Cística	(YATES & NETO, 1999)
GOV2	Páginas da web coletadas pelo NIST	(MANNING, 2009)
INSPEC	Abstracts de artigos da computação, eletrônica e física.	(YATES & NETO, 1999)
ISI / CISI	Artigos médicos.	(YATES & NETO, 1999)
LISA	Biblioteconomia e Ciencia da Informação.	(YATES & NETO, 1999)
Medlars	Artigos médicos.	(YATES & NETO, 1999)
NPL	Engenharia Elétrica.	(YATES & NETO, 1999)
NTCIR	Idioma do Leste Asiático para recuperação cross-language.	(MANNING, 2009)
Reuters-21578	Artigos de revistas/jornais online.	(MANNING, 2009)
Reuters-RCV1	Artigos de revistas/jornais online.	(MANNING, 2009)
TIME	Artigos diversos.	(YATES & NETO, 1999)
TREC	Text Retrieval Conference (TREC) - Artigos diversos	(MANNING, 2009; HEARST, 2009)
TREC 6-8	Subconjuntos da TREC.	(MANNING, 2009)

Quadro 2 – Coleções de teste formais.

No entanto, apesar da extensa avaliação realizada nas bases mencionadas, nenhuma correspondia aos requisitos necessários para o presente estudo. Um dos métodos mais exigentes, nesse aspecto, é a navegação facetada. Para viabilizar sua utilização, o tema da coleção necessariamente deve permitir sua segmentação sob a ótica das facetas, ou seja, o domínio não pode ser descrito através de uma única perspectiva. A classificação facetada é um conjunto de categorias mutuamente exclusivas, aonde cada uma delas reflete, isoladamente, uma característica de um determinado grupo de itens (DENTON, 2003).

Uma base de testes que vise o uso da navegação facetada deve, portanto, apresentar as facetas definidas, bem como os tópicos dentro de cada uma delas também pré-determinado, além da indexação dos seus documentos disposta segundo essa estrutura. Além dos requisitos já expostos, deveria também apresentar: grande volume de documentos, conjunto de consultas de testes definido, julgamento de relevância conhecido para cada par (consulta, documentos), e também possuir domínio de conhecimento acessível, visando facilitar tanto a construção correta das facetas e tópicos, quanto à captação de participantes para o experimento.

Diante da aparente indisponibilidade de uma coleção de testes formal, contemplando todos os requisitos necessários, tornou-se aparente a necessidade da construção de uma coleção própria, o que levou a nova pesquisa. As bases pré-selecionadas encontram-se listadas no Quadro 3.

COLEÇÃO	ASSUNTO	REFERÊNCIA
Chave	Notícias do Folha de São Paulo e do PUBLICO	(Coleção CHAVE, 2009)
DBLP	Artigos da Ciência da Computação.	(DBLP, 2009a; 2009b)
DMOZ	Diretório navegável contendo diversos assuntos.	(DMOZ, 2009)
IMDB	Informações sobre filmes, atores diretores, etc.	(IMDB, 2009)
MusicBrainz Database	Dados de artistas, releases, trilhas, e gravadoras.	(MusicBrainz Database, 2009)
New York Times Annotated Corpus	Artigos de assuntos diversos.	(New York Times Annotated Corpus, 2009)
Wikipedia Dumps	Variados, classificados em categorias.	(Wikipedia Dumps, 2009)

Quadro 3 – Coleções não formais.

Dentre as coleções listadas, a pertencente à Internet Movie Database (IMDB) demonstrou ser a mais atrativa, uma vez que possuía: facetas e tópicos já definidos, documentos categorizados de forma manual, grande volume de dados, palavras-chave já definidas por filme, domínio de conhecimento popular, além de ser gratuita e disponível para utilização. Outro aspecto interessante é o de que um pequeno subconjunto da IMDB já foi utilizado em pesquisa anterior, sobre a navegação facetada, com sucesso (KOREN & ZHANG, 2008).

A desvantagem referente à base selecionada, é que a mesma disponibiliza seu acervo apenas de forma não estruturada, em arquivos de texto, com tamanhos de até 500Mb. Outra desvantagem é a de que, embora contenha todas as facetas e tópicos dentro das mesmas, definidos previamente, o acervo não explicita o relacionamento entre eles. A Figura 18 revela um pequeno fragmento de um dos arquivos disponíveis, que trata da questão da classificação dos filmes quanto ao gênero.

```
 1 Adam & Evelyn (2008)                    Short
 2 Adam & Evelyn (2008)                    Comedy
 3 Adam & Evelyn (2008)                    Fantasy
 4 Adam & Evelyn (2008)                    Musical
 5 Adam & Evelyn (2008)                    Romance
 6 Adam & Evil (2004)                    Horror
 7 Adam Funn (2009) (V)                    Short
 8 Adam Funn (2009) (V)                    Comedy
 9 Adam Funn (2009) (V)                    Drama
10 Adam Funn (2009) (V)                    Sci-Fi
11 Adam Had Four Sons (1941)                 Drama
12 Adam Had Four Sons (1941)                 Romance
13 Adam Hart i Sahara (1990) (V)                 Fantasy
14 Adam Hart i Sahara (1990) (V)                 Adventure
15 Adam Hart i Sahara (1990) (V)                 Short
16 Adam Hart i Sahara (1990) (V)                 Mystery
17 Adam: His Song Continues (1986) (TV)              Drama
18 Adam Hunter: Dysfunctional (2007) (V)             Comedy
19 Adamianta sevda (1984)                    Comedy
20 Adam i Eva (1963) (TV)                    Drama
21 Adam i Eva (1969) (TV)                    Drama
22 Adam i Eva 66 (1966)                    Short
23 "Adam i Ewa" (2000)                   Drama
24 "Adam i Ewa" (2000)                   Romance
25 Adam i Heva (1969)                    Comedy
26 Adamini bul (1975)                    Comedy
27 Adamini bul (1975)                    Romance
```

Figura 18 – Fragmento da IMDB

Foram criados e utilizados diversos scripts na linguagem Pearl, usando o shell padrão do sistema operacional Unix, para a construção dos, aproximadamente, 676 mil documentos relativos aos filmes disponíveis da IMDB. O formato escolhido foi o XML por permitir conteúdo estruturado, o que seria utilizado posteriormente para a indexação dos arquivos na máquina de busca. Cada arquivo XML continha as seguintes informações a cerca de um filme:

- Ano
- Países de origem
- Atores
- Atrizes
- Gênero
- Cor
- Idioma
- Tipo de som
- Diretores
- Escritores

- Compositores
- Classificação – por país
- Tempo de execução
- Local(is) de filmagem – cidades e país
- Editor
- Data de lançamento - por país
- Produtores
- Tipo (filme, série de TV, etc.)

Outra exigência da investigação proposta é a de que os documentos contivessem texto suficiente para utilização dos algoritmos de agrupamento. Então, após a construção da coleção em XML, novos scripts foram criados para percorrer a coleção e incorporar a sinopse de cada filme. No entanto, dos 676 mil documentos criados, apenas a sinopse de aproximadamente 115 mil filmes estavam disponíveis, pela IMDB, no momento da criação da base. Todos os experimentos foram conduzidos então com essa última quantidade de documentos.

4.2. FERRAMENTAS DESENVOLVIDAS E UTILIZADAS

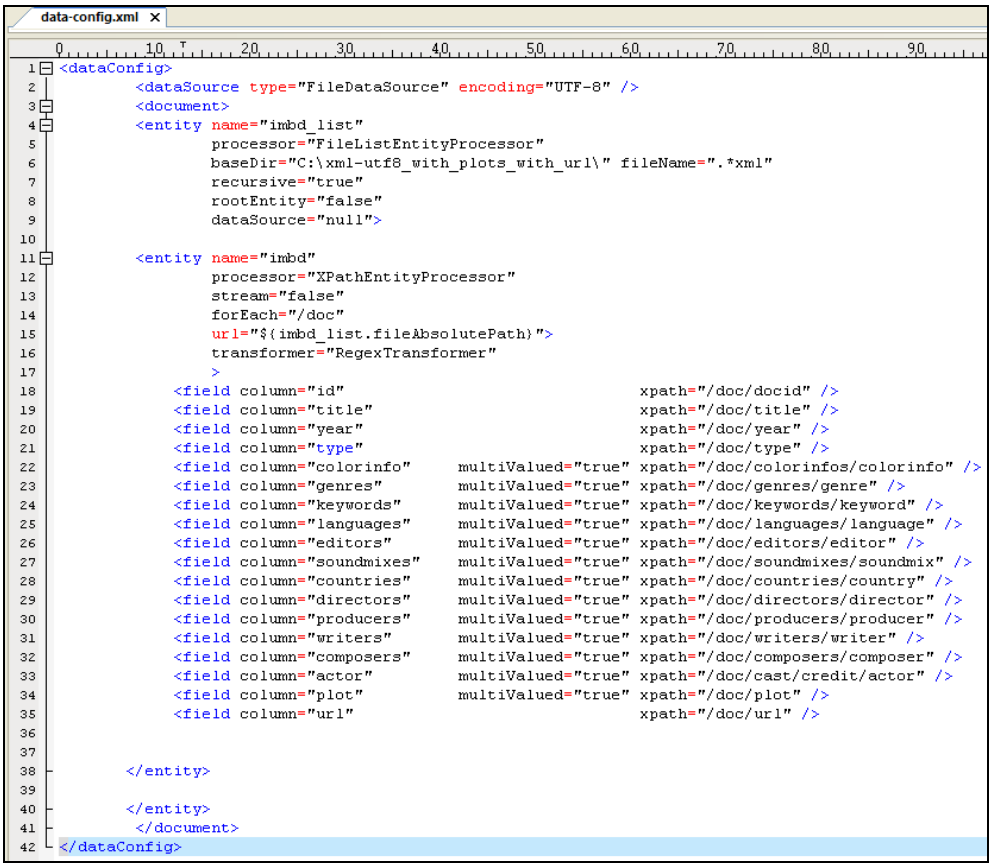
Para atuar na busca e na ordenação dos conjuntos de resultados, servindo a todas as ferramentas de recuperação de informação a serem investigadas, foi utilizada a máquina de busca Apache SOLR, na sua versão 1.4 (SOLR, 2009). Trata-se de uma reconhecida plataforma de busca *open source*, que faz uso das bibliotecas de indexação desenvolvidas pelo projeto Lucene (LUCENE, 2009).

O SOLR é uma máquina de busca consolidada, cuja utilização é exercida por inúmeras instituições presentes na Internet, dentre elas: AT&T, Apple, NASA, Disney, Goldman Sachs, dentre outras (SOLR, 2010).

Embora o SOLR traga inúmeros aprimoramentos à plataforma disponibilizada pelo projeto Lucene, o principal deles é permitir a construção de navegação facetada. Na sua

versão 1.4, possibilita ainda a integração com as ferramentas de agrupamento Carrot² (2009) e Lingo3G (2009), também utilizadas pelo presente trabalho de pesquisa.

Após a configuração do SOLR, com a definição detalhada do esquema dos dados e a parametrização dos seus respectivos tipos a serem utilizados, foi dado início ao processo de indexação da coleção de testes. Para realizar a indexação dos 114 mil documentos XML no Apache SOLR, foi preparado e testado um *DataImportHandler* específico para tal tarefa. O código utilizado para realizar o procedimento pode ser visto na Figura 19 a seguir. Todos os campos disponíveis na coleção, ou seja, título do filme, ano, diretores, sinopse, etc., foram indexados e armazenados para conferir agilidade na resposta das ferramentas. As consultas efetuadas pelos participantes, portanto, eram realizadas por sobre todas as informações disponíveis na coleção de testes. Finalizada a etapa de indexação, as configurações relacionadas à viabilização das facetas foram realizadas. Parte dos procedimentos seguidos para preparação do SOLR foi feita segundo os direcionamentos presentes em (SMILEY & PUGH, 2009).



data-config.xml

```
<dataConfig>
        <dataSource type="FileDataSource" encoding="UTF-8" />
        <document>
        <entity name="imbd_list"
                processor="FileListEntityProcessor"
                baseDir="C:\xml-utf8_with_plots_with_url\" fileName=".*xml"
                recursive="true"
                rootEntity="false"
                dataSource="null">

        <entity name="imbd"
                processor="XPathEntityProcessor"
                stream="false"
                forEach="/doc"
                url="${imbd_list.fileAbsolutePath}">
                transformer="RegexTransformer"
                >
            <field column="id"                                  xpath="/doc/docid" />
            <field column="title"                               xpath="/doc/title" />
            <field column="year"                                xpath="/doc/year" />
            <field column="type"                                xpath="/doc/type" />
            <field column="colorinfo"     multiValued="true" xpath="/doc/colorinfos/colorinfo" />
            <field column="genres"        multiValued="true" xpath="/doc/genres/genre" />
            <field column="keywords"      multiValued="true" xpath="/doc/keywords/keyword" />
            <field column="languages"     multiValued="true" xpath="/doc/languages/language" />
            <field column="editors"       multiValued="true" xpath="/doc/editors/editor" />
            <field column="soundmixes"    multiValued="true" xpath="/doc/soundmixes/soundmix" />
            <field column="countries"     multiValued="true" xpath="/doc/countries/country" />
            <field column="directors"     multiValued="true" xpath="/doc/directors/director" />
            <field column="producers"     multiValued="true" xpath="/doc/producers/producer" />
            <field column="writers"       multiValued="true" xpath="/doc/writers/writer" />
            <field column="composers"     multiValued="true" xpath="/doc/composers/composer" />
            <field column="actor"         multiValued="true" xpath="/doc/cast/credit/actor" />
            <field column="plot"          multiValued="true" xpath="/doc/plot" />
            <field column="url"                                 xpath="/doc/url" />


        </entity>

        </entity>
         </document>
</dataConfig>
```

Figura 19 – DataImportHandler utilizado na indexação com SOLR.

Encerrada a preparação da máquina de busca, a criação da busca direta teve o seu início. É importante ressaltar que o desenvolvimento, ou adaptação, de todas as cinco ferramentas (busca direta, navegação, agrupamento plano, agrupamento hierárquico e navegação facetada) compartilharam da mesma identidade visual. Trabalhos anteriores constataram que a satisfação do usuário está relacionada, com freqüência, às questões de *design* da interface da ferramenta, que são claramente independentes da qualidade dos resultados retornados em si (MANNING, RAGHAVAN & SCHÜTZE, 2009; LINDGAARD & DUDEK, 2003; HASSENZAHL, 2004).

Ainda em relação às questões de *design* e usabilidade, houve grande preocupação e esforço no sentido de prover, ao máximo possível, os mesmos recursos e as mesmas formas de exibição de resultados em todas as cinco ferramentas. Características como paginação do conjunto retornado, apresentação dos resultados de dez em dez registros, exibição do título do

filme e um pequeno fragmento da sua sinopse nos itens de resultado, e apresentação dos detalhes de cada filme, foram igualmente implementadas em todas as ferramentas.

O sistema de busca direta foi desenvolvido utilizando a recuperação através de palavras ou expressões. Nos casos em que mais de uma palavra é fornecida, a busca é realizada através do operador AND. Quando duas ou mais palavras entre aspas são usadas como entrada, então é realizada uma procura pela expressão exata. O tratamento de *stop words* (MANNING, RAGHAVAN & SCHÜTZE, 2009) foi realizado aproveitando as funcionalidades nativas do Apache SOLR. A ferramenta de busca direta foi concluída com seu comportamento similar ao padrão das máquinas de busca disponibilizadas na Web atualmente. A interface da ferramenta pode ser observada na Figura 20, que representa uma busca realizada usando a consulta “Francis Ford Coppola”, sem as aspas.

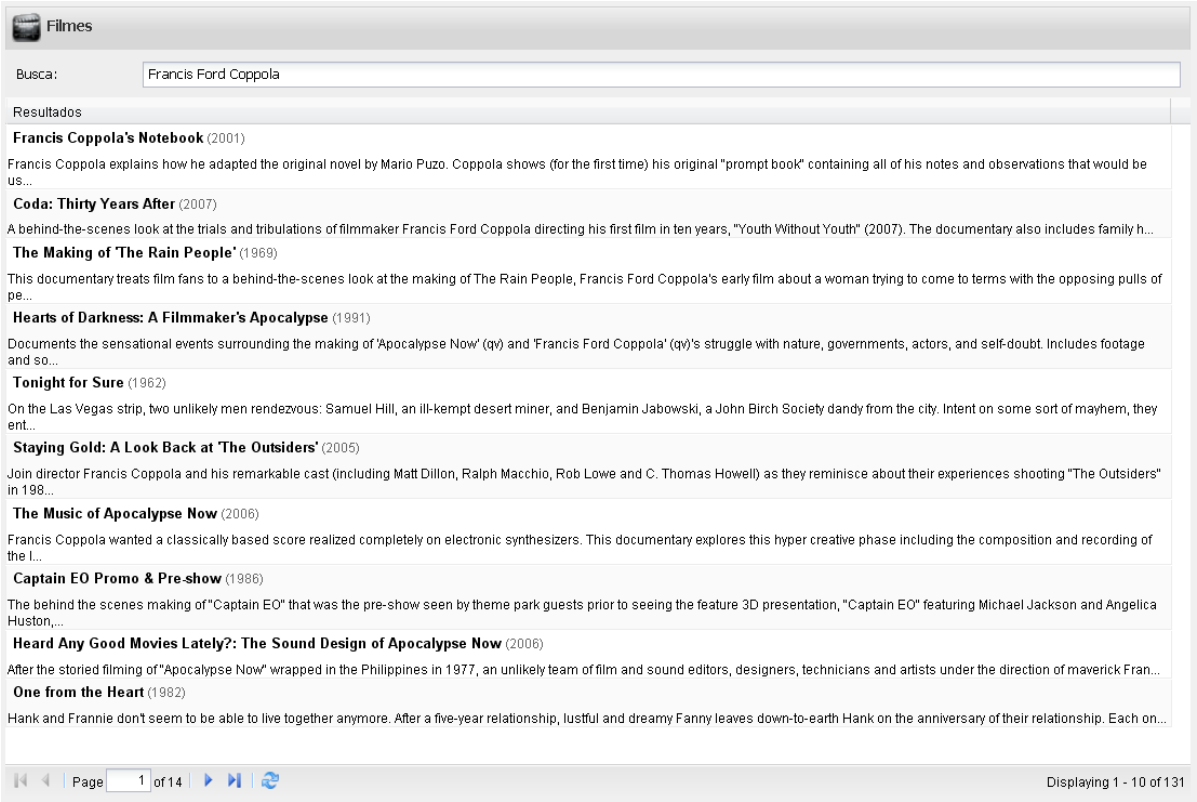


Figura 20 – Tela da ferramenta de Busca Direta

Ao clicar em algum filme presente na lista de resultados, uma ficha técnica contendo todos os detalhes do mesmo é exibida para o usuário, sem que seja necessário sair da tela

atual. Tal funcionalidade foi obtida através da combinação de HTML e JavaScript. Cabe ressaltar, no entanto, que o mesmo recurso foi disponibilizado para as demais ferramentas. A Figura 21 demonstra a ficha técnica do filme "*Tonight for Sure*", exibida quando da seleção do quinto filme retornado para a consulta anterior.

Figura 21 – Ficha Técnica com detalhes dos resultados

O desenvolvimento da ferramenta de navegação procurou seguir, de igual forma, as melhores práticas existentes atualmente para a Web (LEAVITT & SHNELDERMAN, 2006). Um menu com todas as categorias existentes na coleção de testes foi disponibilizado para o usuário no topo da página: cor, idioma, país, gênero, atores, compositores, editor, tipo, escritor, tipo de som (*sound mixer*), diretores, produtores. O menu também foi disposto horizontalmente, concordando com os estudos de usabilidade atuais. Cada categoria, ao ser

clicada, apresentava seu conteúdo ordenado de forma alfabética, ou numérica crescente, conforme o caso. O menu pode ser observado na Figura 22.

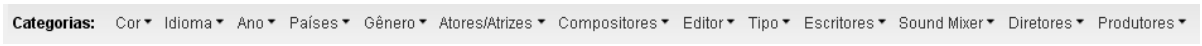


Figura 22 – Menu de categorias da ferramenta de navegação.

Nos casos em que uma categoria possuía elevada quantidade de itens que poderiam tornar a navegação mais difícil ou ainda prejudicar a sua usabilidade, um menu hierárquico foi desenvolvido. A Figura 23 representa uma consulta realizada por um determinado diretor, mais especificamente por filmes de Alfred Hitchcock. Trata-se de um exemplo de hierarquia criada para facilitar a navegação, e conseqüentemente, procurar extrair o melhor que o método poderia oferecer.

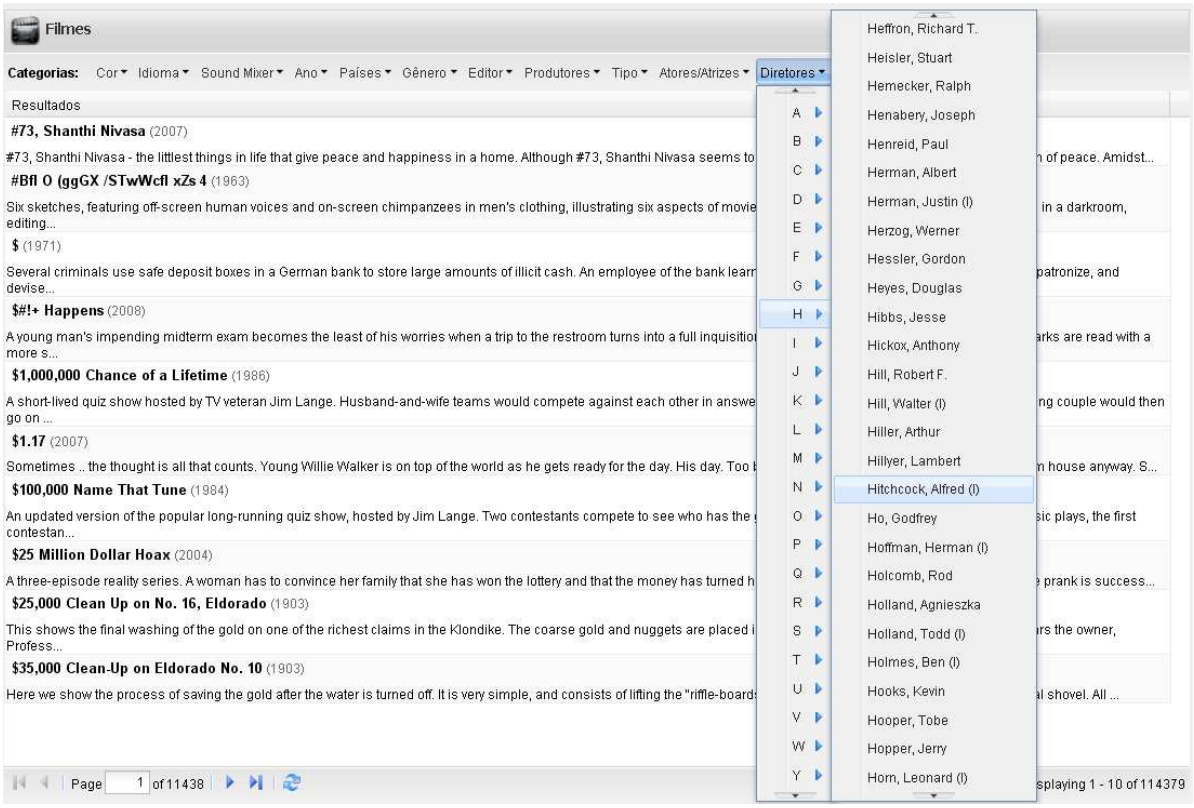


Figura 23 – Consulta por diretor na navegação.

No caso dos sistemas que fazem uso dos algoritmos de agrupamento, não foram criadas ferramentas inteiramente novas. Houve, na realidade, um desenvolvimento com o foco

na adaptação de programas existentes. Para a execução das tarefas de agrupamento plano, a ferramenta Carrot² (2009), na sua versão 3.2, foi selecionada devido a sua integração com o Apache SOLR, versão 1.4. Sua escolha também levou em consideração os bons resultados apresentados em Carpineto et al. (2009), o fato de ser uma ferramenta *open source*, e ainda, de disponibilizar um ambiente completo para ajustes no funcionamento do seu algoritmo Lingo.

Após a realização dos ajustes necessários para que o Carrot² pudesse acessar a máquina de busca, ou seja, a coleção de testes indexada, a forma de exibição dos resultados foi adaptada para se assemelhar às ferramentas anteriores. Os títulos dos filmes presentes no conjunto de resultados, e os fragmentos de suas respectivas sinopses são exibidos logo após a execução da consulta, ou da seleção de algum dos agrupamentos. Também foi desenvolvida a mesma ficha técnica para exibição, caso o usuário procurasse por mais detalhes de um determinado filme. A Figura 24 apresenta o resultado da busca para uma consulta por “mafia”, utilizando o agrupamento plano. O conjunto resposta pode ser refinado através da navegação pelos diversos agrupamentos exibidos à esquerda da lista ordenada de filmes. Cada agrupamento indica, entre parênteses, o número de documentos ali contido, ou seja, classificados. Cabe ainda ressaltar que o Lingo é um algoritmo de agrupamento flexível.

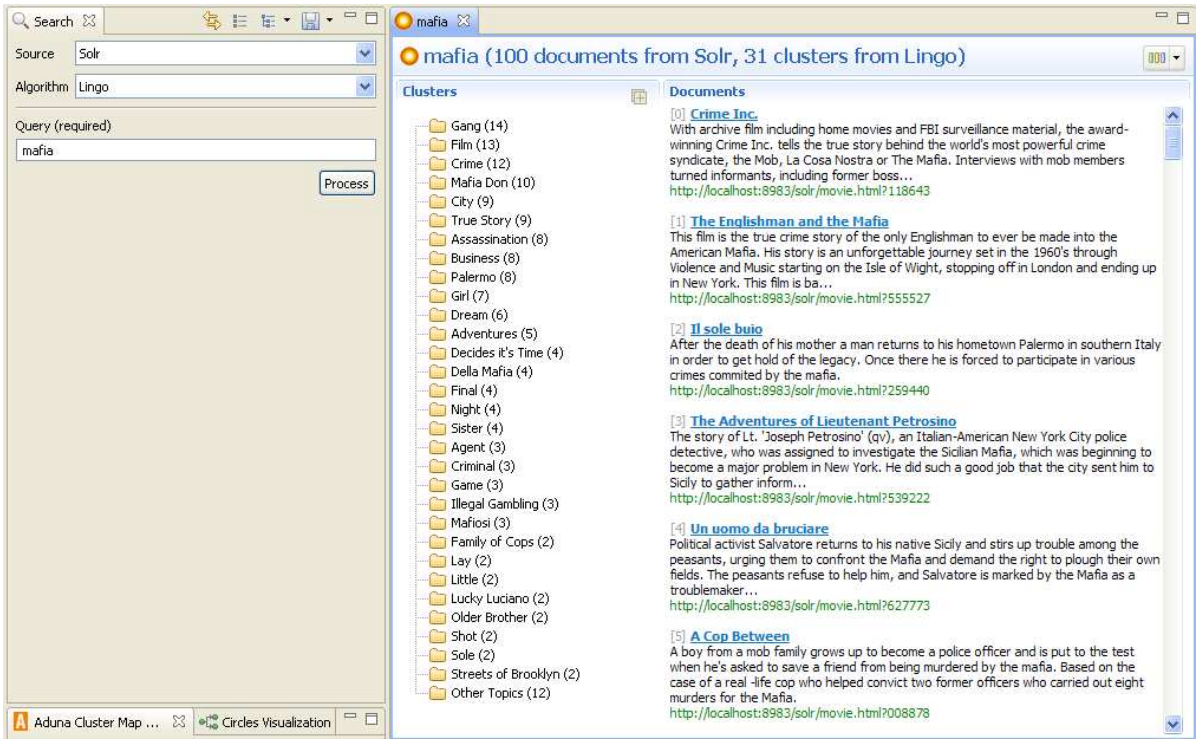


Figura 24 – Consulta por tema na ferramenta de agrupamento plano.

Para a avaliação do método de agrupamento hierárquico houve também o desenvolvimento com o foco na adaptação de um programa existente. De excelente desempenho, como mostrado em Carpineto et al. (2009), o Lingo3G (2009), que utiliza um algoritmo homônimo, possui também, em sua versão 1.3.0, integração com o Apache SOLR. No entanto, trata-se de uma ferramenta comercial, não disponível para livre uso ou cópia na Internet. Sua utilização no presente trabalho deve-se à licença gentilmente cedida pelos proprietários da empresa Carrot Search (2009) para este fim.

Todos os ajustes, assim como as mesmas adaptações, realizados para a ferramenta Carrot² também foram empregados objetivando o uso do Lingo3G. A imagem a seguir ilustra uma consulta realizada sobre o tema anterior (máfia), porém demonstra a hierarquia atribuída automaticamente pelo algoritmo. A seleção do cluster “Sicilian”, filho do cluster de nome “Mafia Boss”, exibe à direita os quatro filmes classificados segundo essa mesma hierarquia.

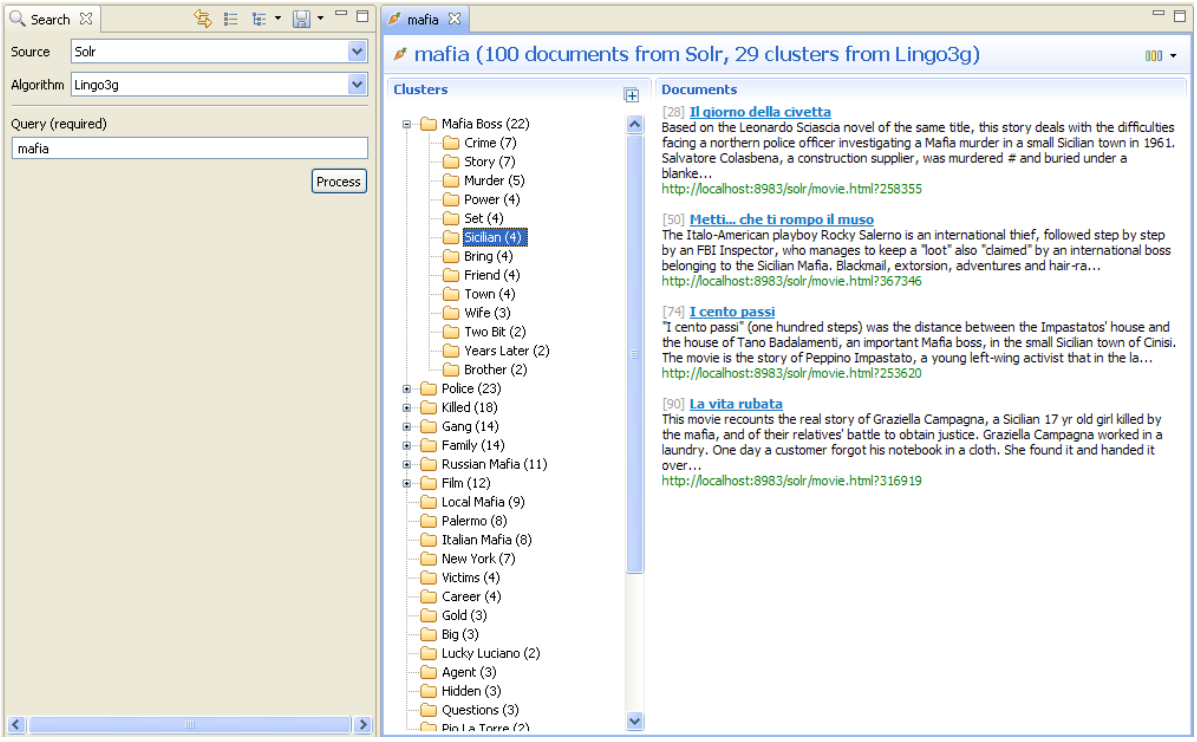


Figura 25 – Consulta por tema na ferramenta de agrupamento hierárquico.

Por fim, o sistema de navegação facetada foi construído exclusivamente para a presente pesquisa. O desenvolvimento atendeu a todas as recomendações realizadas por William Denton, (2003), Marti Hearst (2006a e 2008) e Daniel Turkelang (2009), visando extrair o melhor do método durante o experimento.

Devido ao grande número de facetas envolvidas no domínio, totalizando 13 delas, e também à elevada quantidade de tópicos, ou seja, de conteúdo presente na maioria das facetas, a interface exibia todas listadas à esquerda, permitindo ao usuário navegar em cada uma separadamente.

No primeiro momento do acesso ao programa, uma caixa de busca era apresentada, bem como todos os 114 mil resultados possíveis, e todas as treze facetas exibidas, com suas respectivas categorias listadas. Para cada categoria, dentro de cada uma das facetas, era exibido um número ao lado do seu nome, representando a quantidade de documentos, pertencentes àquela classificação, retornados como resultado da consulta. A ordenação de

todas as categorias seguia a ordem decrescente numérica, no caso dos anos, e a ordenação alfabética para as demais categorias, dado que diversos estudos de navegação em interfaces sugeriram que os usuários preferem ordenações previsíveis e conhecidas (PRATT et al., 1999), e que não são alteradas durante o seu uso, ou seja, ordenações estáticas (TUNKELANG, D. 2009).

Ao clicar em uma categoria, todas as demais eram atualizadas para refletir a seleção feita. A contagem de resultados em cada categoria também sofria atualização. O mesmo comportamento ocorria quando do uso de busca através de palavras ou expressões. A lista ordenada dos resultados também era atualizada a cada interação do usuário. Sua exibição obedecia as mesmas regras e usabilidade presentes nas demais ferramentas: paginação do conjunto de resultados, exibição de dez resultados por página, exibição do número de documentos encontrados para a dada consulta, e, para cada filme, o título do mesmo e o fragmento de sua sinopse eram exibidos.

Para permitir ao usuário saber aonde ele se encontrava exatamente durante o seu processo de busca, o recurso conhecido como *breadcrumb*, ou “migalha-de-pão”, foi disponibilizado na ferramenta. Desta forma, todas as categorias selecionadas de cada uma das diferentes facetas eram incluídas em uma espécie de caminho percorrido. A fim de permitir uma perfeita integração entre busca e navegação, o *breadcrumb* foi adaptado para também armazenar as palavras utilizadas na busca como parte do histórico. Desta forma, várias palavras poderiam ser incluídas na pesquisa, alternadas com as seleções das categorias. Tal procedimento é apontado como primordial para o sucesso da navegação facetada (HEARST, M., 2009; TUNKELANG, D. 2009). A imagem a seguir ilustra a situação aonde uma consulta foi realizada pela navegação e seleção das categorias “USA”, “Documentary” aliadas à utilização da busca pela palavra “NASA”.

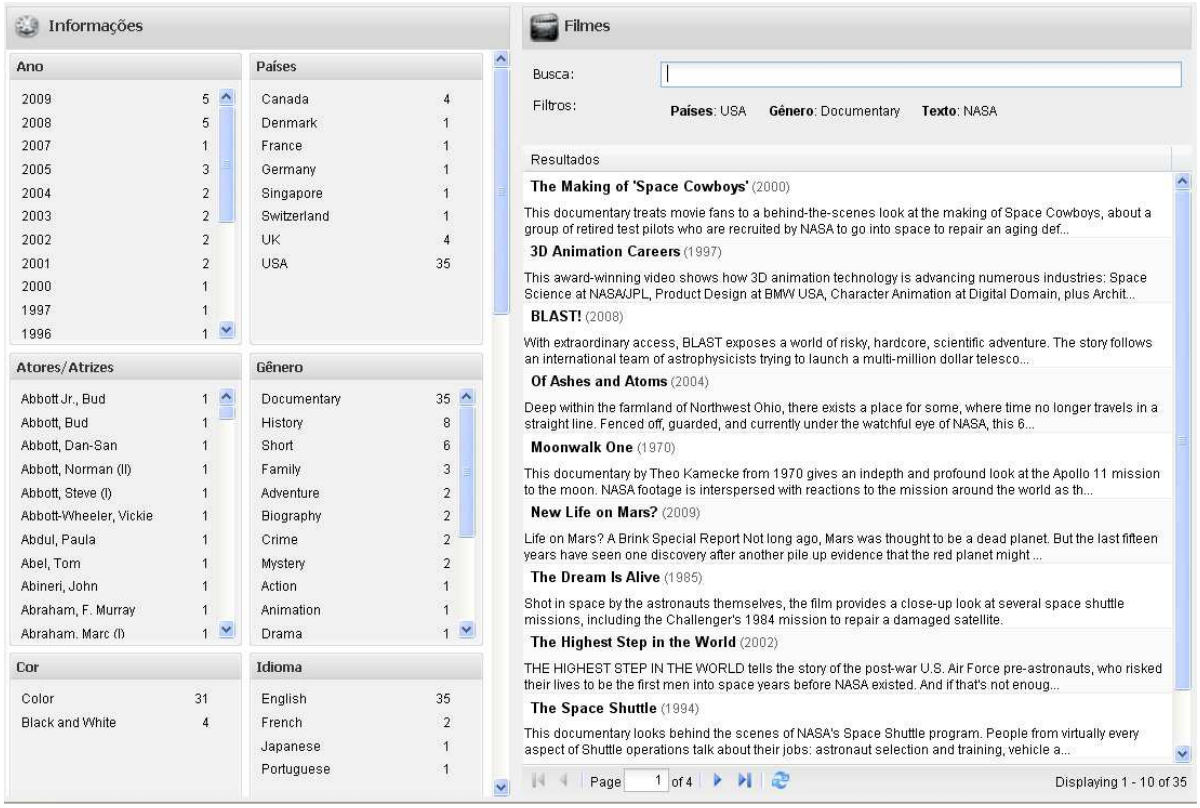


Figura 26 – Navegação facetada: seleção de categorias e palavra-chave.

A ordem de exibição dos critérios no *breadcrumb* refletia a ordem da seleção dos mesmos durante a consulta. Essa abordagem demonstrou ser a mais apropriada às necessidades dos usuários (HEARST, 2006a). Ainda respeitando os recentes trabalhos encontrados na literatura, mesmo quando uma faceta não possuía nenhuma categoria relevante para a consulta, seu nome continuava sendo apresentado na interface.

A qualquer momento da navegação, o usuário poderia optar por remover um critério da sua busca. Para tanto, bastava clicar em cima do mesmo, na listagem exibida pelo *breadcrumb*. A ordem de remoção não precisava obedecer a ordem de entrada do critério no seu histórico de utilização.

Outro aspecto comportamental importante da ferramenta é que para facetas nas quais mais de uma de suas categorias eram freqüentemente associadas a um mesmo filme, por exemplo, gênero, pois um mesmo filme pode pertencer à categoria de suspense e também a de terror, a navegação facetada permitia um comportamento diferente. Em tais situações era

possível realizar uma conjunção de disjunções, ou seja, era possível selecionar mais de uma categoria dentro da mesma faceta, como demonstra a Figura 27. Tal comportamento é destacado como uma das melhores práticas em navegação facetada por (TUNKELANG, D. 2009).

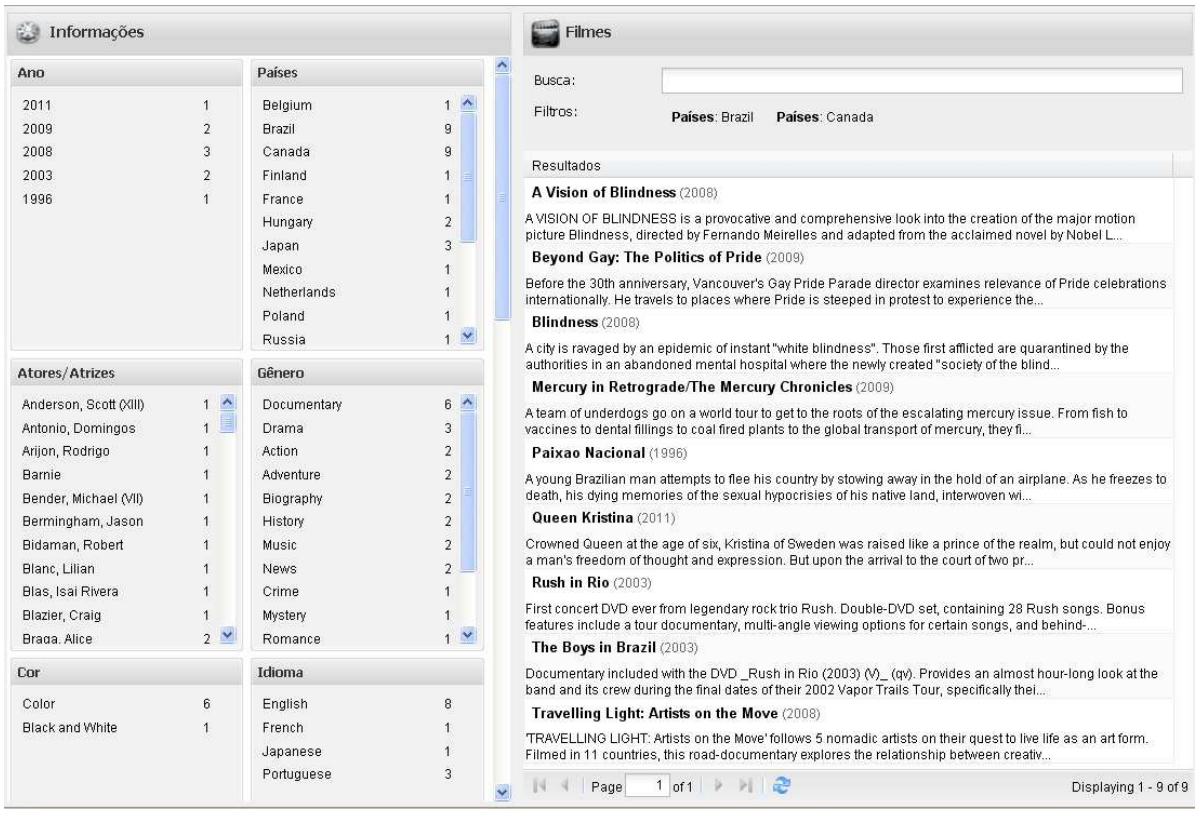


Figura 27 – Navegação facetada: conjunção de disjunções

Importante ressaltar que todas as ferramentas, em suas versões finais para o experimento, possuíam a mesma identidade visual, com as idênticas formas de apresentação dos dados de resultados. Utilizavam também a mesma máquina de busca, realizando a busca pelos mesmos campos indexados, e ainda com tempos similares de resposta. Tal requisito representa as melhores práticas de avaliação quando do envolvimento de participantes. O cuidado incluiu desde a apresentação da lista de resultados, cores e imagens utilizadas, até a apresentação de detalhes dos filmes.

5. EXPERIMENTO, ENTREVISTAS E ANÁLISE DE RESULTADOS

Neste capítulo, são apresentados os resultados obtidos na apreciação dos dados coletados do experimento e das entrevistas. Na primeira seção, é apresentada a análise da etapa piloto. A seguir, é realizada a descrição dos dados obtidos no experimento, permeados pela análise feita com base na teoria compulsada. Segue-se com a descrição e a análise dos dados coletados durante as entrevistas com os participantes, realçando alguns aspectos encontrados, bem como o seu caráter confirmatório em relação ao experimento. Na última seção, o modelo de referência da pesquisa é testado a fim de verificar as hipóteses deste trabalho.

5.1. REALIZAÇÃO DO PILOTO

Objetivando validar e aperfeiçoar os procedimentos de aplicação do experimento, e assegurar que as entrevistas e métricas quantitativas possibilitem medir as variáveis que se deseja realmente medir, ou seja, verificar a aderência do instrumento de medida com a pesquisa, um piloto foi realizado antes da realização das atividades do experimento em si.

Segundo Gil (2007), é necessário que tal fase seja feita com população similar à que será estudada, sem requerer, todavia, uma amostra rigorosamente representativa dessa população. Portanto, o piloto foi realizado com dois participantes de nível superior, representando os dois diferentes grupos do experimento: um analista de sistemas e uma psicóloga. Ambos se enquadravam dentro dos demais quesitos do universo da amostra, sendo a faixa etária, o domínio do idioma inglês e a familiarização com sistemas de busca.

O resultado do piloto permitiu perceber que as tarefas pertencentes ao tipo Tarefa 1, possuíam cobertura muito pequena, ou seja, a quantidade de documentos relevantes era 2 ou 3 filmes. O número restrito visava tornar a cobertura mais fácil de ser conhecida, assim como simplificar a análise necessária para o cálculo do *R-Precision*. No entanto, a precisão

mostrou-se nitidamente afetada, tornando sua medição pouco atraente e menos significativa. Todas as tarefas pertencentes ao tipo Tarefa 1 foram então substituídas por outras novas que, apesar de obedecerem ao mesmo tipo, possuíam o dobro da cobertura.

Outro aspecto observado foi em relação ao tempo necessário para execução das dez tarefas presentes no plano de experimentação. As sessões do piloto duraram, em média, cerca de 90 minutos, fazendo com que os participantes chegassem ao final já demonstrando sinais de cansaço. Esse fato poderia prejudicar os últimos métodos a serem utilizados pelo participante, em especial a navegação facetada. Contudo, o experimento já havia sido projetado de forma parcimoniosa, contendo apenas duas tarefas para cada um dos cinco métodos investigados. A ordem de utilização das ferramentas também não poderia sofrer alteração a cada experimento visando mitigar o problema encontrado, pois, conforme mostra o ANEXO A, o roteiro foi elaborado a fim de fazer com que o participante utilizasse os métodos seguindo a ordem com que estes revelam o conteúdo. Assim, o experimento inicia com a busca direta e termina com a navegação facetada, a qual revela o conteúdo de forma mais explícita.

O piloto também possibilitou validar as soluções propostas para captação dos dados e alimentação das variáveis dependentes em suas medidas quantitativas (tempo, taxa de erro e *R-Precision*) e também no seu aspecto qualitativo (respostas às perguntas da entrevista). A triangularização das informações obtidas, através do caráter multimétodo da pesquisa e sua coleta de dados paralela, permitiu a execução do seu papel confirmatório.

No aspecto mais operacional, a validação da ferramenta *freeware* responsável pela captura das telas e geração dos vídeos com as consultas efetuadas pelos participantes, para posterior geração do *R-Precision*, mostrou-se ineficiente durante o piloto. Sua substituição por uma ferramenta comercial do mesmo segmento, utilizada em seu período *trial*, permitiu gerar gravações com qualidade superior e tamanho vinte vezes menor. Além disso, a nova

ferramenta não causava impacto no desempenho do computador durante a utilização da máquina de busca e das ferramentas de recuperação de informação.

Após a conclusão da fase piloto, e dos ajustes que se fizeram necessários, todos os dados provenientes dessa etapa foram descartados, não sendo, portanto, considerados para a análise de resultados da presente pesquisa. Os envolvidos na etapa piloto não participaram do experimento, conforme sugerido por Vergara (2009) e Gil (2007).

5.2. RESULTADOS

5.2.1. Descrição e Análise dos Resultados Quantitativos

Quanto aos dados relativos às variáveis dependentes, a

Tabela 1 apresenta o cálculo do tempo médio gasto por cada um dos participantes na realização das consultas definidas para o experimento. O maior tempo utilizado é interpretado como o método que demanda maior esforço do usuário para que o mesmo possa alcançar seus objetivos.

Tabela 1 – Média do tempo gasto na execução das tarefas.

	Busca Direta	Agrupamento Plano	Agrupamento Hierárquico	Navegação	Navegação Facetada
Tarefa 1	04:18	08:22	05:03	05:30	01:58
Tarefa 2	05:34	05:56	05:23	06:39	05:14
Média Total	**04:56**	**07:09**	**05:13**	**06:05**	**03:36**
Desvio Padrão	**02:33**	**03:05**	**02:16**	**01:58**	**02:09**

Conforme pode ser observado, o método de navegação facetada apresentou o menor tempo dentre todos os cinco investigados, particularmente na Tarefa 1, que demandava uma análise sob duas perspectivas distintas da entidade central da coleção: o diretor do filme e o gênero ao qual ele pertencia. Pela facilidade que o método oferece para esse tipo de consulta, era de se esperar que se destacasse por diminuir o esforço do usuário durante sua utilização.

Outro ponto que chama atenção, e que não era previsto, foi o elevado tempo empregado na execução da Tarefa 1 usando o agrupamento plano. Uma causa provável para esse resultado é a pouca previsibilidade dos agrupamentos oferecidos através desse tipo de método, agravada no caso do agrupamento plano por não oferecer a mesma organização semântica oferecida no agrupamento hierárquico. Tal fenômeno também foi percebido por Hearst (2006b; 2009).

A seguir, a Tabela 2 exibe a taxa de erro calculada a partir das respostas oferecidas para as tarefas. Como pode ser verificado, novamente a navegação facetada se destacou, permitindo uma considerável redução na taxa de erro em comparação com os demais métodos. Para a Tarefa 1 ela foi, inclusive, a única com taxa de erro igual a zero. É importante ressaltar o desempenho do agrupamento hierárquico, que vem logo em seguida, com uma taxa de erro de apenas 1 em 16 dos casos, para a mesma tarefa.

Tabela 2 – Taxa de erro na execução das tarefas.

	Busca Direta	**Agrupamento Plano**	**Agrupamento Hierárquico**	**Navegação**	**Navegação Facetada**
Tarefa 1	8/16	7/16	1/16	7/16	0/16
Tarefa 2	5/16	13/16	11/16	7/9	6/16
Total	**13/32**	**20/32**	**12/32**	**14/25**	**6/32**

Quando da análise da Tarefa 2, a busca direta e a navegação facetada se sobressaem oferecendo baixas taxas de erro. Em relação aos métodos de agrupamento, que permitem uma classificação adicional ao conteúdo dos resultados da busca, esperava-se uma baixa taxa de erro. Entretanto, a taxa de erro se mostrou muito elevada. Tal fenômeno pode ser atribuído ao caráter subjetivo da Tarefa 2, que versava sobre a busca de filmes em temas amplos: “máfia italiana” e “furacões”. Alguns filmes que não discorriam sobre os furacões, não documentavam seus efeitos na sociedade, ou ainda onde os furacões nem faziam parte da trama central, eram apontados como resposta. Por exemplo, “*A Concert for Hurricane Relief*” foi um título utilizado por vários participantes como sendo uma resposta para a consulta,

sendo, no entanto, apenas um festival de música para angariar fundos destinados aos sobreviventes do furacão Katrina, em Nova Orleans.

Outro aspecto fundamentalmente importante, revelado pela Tabela 2, é o comportamento dos participantes na Tarefa 2, durante o uso da navegação. O objetivo era utilizá-la para encontrar filmes sobre o tema “nazismo”. Diante do grau de dificuldade, 7 dos 16 participantes desistiram após algum tempo de tentativas. Foram os únicos casos de desistência em todo o experimento.

No resultado total, no entanto, o agrupamento hierárquico demonstrou taxa de erro próxima à busca direta. A navegação facetada apresentou novamente o melhor resultado.

Apesar das análises de tempo empregadas na realização das consultas e das observações acerca da taxa de erro dos métodos terem oferecido importantes reflexões, a avaliação de ambos torna-se mais atraente quando realizada de forma concomitante, como visto a seguir.

O usuário pode gastar muito tempo em uma ferramenta, mas encontrar poucos resultados relevantes, e vice-versa. Por esse motivo, a Tabela 3 fornece dados que merecem relevo no contexto investigativo da pesquisa. Tal tabela apresenta a quantidade de documentos relevantes encontrados por minuto, em média, para cada um dos métodos e em cada uma das tarefas.

Tabela 3 – Documentos relevantes encontrados por minuto.

	Busca Direta	Agrupamento Plano	Agrupamento Hierárquico	Navegação	Navegação Facetada
Tarefa 1	1,37	0,39	1,08	0,80	2,07
Tarefa 2	1,42	0,79	0,95	0,22	0,98
Média Total	1,40	0,59	1,02	0,51	1,53
Desvio Padrão	0,49	0,34	0,44	0,40	0,67

Embora a tabela apresente apenas ligeira vantagem para a navegação facetada em relação à busca direta, uma observação se faz necessária. O método de navegação facetada

apresenta as facilidades da busca direta acrescidas das oriundas da navegação. Por esse motivo, é de se estranhar a menor quantidade de itens relevantes encontrados em comparação com a busca direta quando da execução da Tarefa 2. Uma explicação para tal comportamento pode ser atribuída ao cansaço dos participantes ao final do experimento, como relatado na descrição de resultados da etapa piloto. Uma vez que a navegação facetada foi o último método a ser utilizado pelos participantes, após mais de uma hora de experimento, era natural esperar algum tipo de queda no rendimento dos usuários.

Quanto ao número de documentos relevantes encontrados por minuto, cabe ressaltar ainda que a navegação e o agrupamento plano apresentaram resultados muito inferiores, comparativamente aos demais métodos pesquisados.

A lógica leva a pressupor que os métodos de agrupamento, por viabilizarem uma organização do conjunto de resultados e ainda permitirem a navegação por esse mesmo conjunto de forma mais prática, conseguiriam expor mais documentos relevantes em menor tempo quando comparados à navegação e à busca direta. De fato, para a navegação isso foi observado. No entanto, o mesmo não se verificou para o método de busca direta.

Uma possível explicação para o aparente contra-senso é baseada na forma de utilização desse tipo de ferramenta pelos participantes. Por serem usuários acostumados com as ferramentas de busca da Web, como o Google, os participantes efetuaram consultas que pouco favoreciam os algoritmos de agrupamento, utilizando grandes conjuntos de palavras na recuperação. Dos 16 participantes do experimento, apenas 4 perceberam que conseguiam resultados melhores, ou seja, agrupamentos mais significativos, quando usavam apenas uma ou duas palavras, e também quando essas eram mais representativas diante da sua necessidade de informação. Mesmo os participantes do grupo dos analistas de sistemas não utilizaram os agrupamentos da melhor forma. Alguns tiveram resultado ainda pior ao tentar combinar várias

palavras com expressões através do uso de aspas. Apenas 2 dos 8 analistas perceberam que não podiam usar a mesma estratégia da busca direta nas ferramentas de agrupamento.

Outra característica interessante a ser observada, e que denota a subutilização dos métodos de agrupamento, é o fato de os participantes não terem usado, em sua maioria, os conceitos revelados pelos nomes dos agrupamentos para realimentar as consultas gerando novas entradas. Por exemplo, na pesquisa por “máfia italiana”, o conceito “*mob*”, sinônimo de "máfia italiana" apresentado como um agrupamento na resposta, não era utilizado em nova busca. No caso da pesquisa por “furacões”, apesar da palavra “Katrina” aparecer em mais de um agrupamento, os participantes não experimentavam usá-la como novo parâmetro da busca. Nesse ponto as pessoas pareciam repetir o comportamento apresentado diversas vezes quando do uso da busca direta pelos participantes que não atuavam como analistas de sistemas: a realização de poucas consultas.

Por fim, a Tabela 4 revela o resultado para as variáveis dependentes precisão e cobertura, aferidas através do método *R-Precision*. Novamente a navegação facetada mantém o destaque já alcançado nas métricas anteriores, sendo seguida de perto pelo agrupamento hierárquico. Aqui o desempenho da navegação facetada também pode ter sido prejudicado pelo cansaço dos participantes, uma vez que o *R-Precision* da Tarefa 2 é de apenas 0,68 para tal método, e de 0,80 para o método de busca direta.

Tabela 4 – *R-Precision* calculado no experimento.

	Busca Direta	Agrupamento Plano	Agrupamento Hierárquico	Navegação	Navegação Facetada
Tarefa 1	0,67	0,83	0,83	0,00	1,00
Tarefa 2	0,80	0,63	0,81	0,11	0,68
Total	**0,73**	**0,73**	**0,82**	**0,06**	**0,84**

A navegação ficou com um resultado muito ruim provavelmente devido ao tamanho da coleção utilizada no experimento. Aparentemente, esse método não é indicado para

volumes expressivos de documentos, como no caso em questão, sob pena de deixar a precisão e a cobertura em níveis inaceitáveis de uso.

Interessante observar também que, apesar de os participantes não terem utilizado as ferramentas de agrupamento da melhor forma, o valor total do *R-Precision* do agrupamento plano foi idêntico ao da busca direta. Esse dado endossa a idéia de que, caso os participantes fizessem um melhor uso das ferramentas de agrupamento, possivelmente os resultados obtidos nas métricas anteriores teriam sido mais promissores para esses métodos.

5.2.2. Descrição e Análise dos Resultados Qualitativos

Quanto aos dados demográficos, não foi percebida diferença entre participantes de gêneros ou idades distintos. No entanto, a área de atuação dos participantes parece influenciar a forma pela qual as consultas são realizadas. Como visto na seção 3.4.2, havia 8 analistas de sistemas dentre os 16 participantes.

Os analistas geralmente não utilizavam os dados obtidos através da primeira consulta. Ao perceberem que o conjunto de resultados não lhes era favorável, imediatamente inseriam uma nova consulta para tentar obter um conjunto de resposta menor, e mais próximo do esperado. Realizavam então muitas consultas e analisavam menos os resultados retornados.

Os participantes que atuam em outras áreas demonstraram comportamento diferente. As consultas não eram descartadas apenas pela apresentação da lista de filmes, ou pela quantidade de itens retornados. Não havia estratégia para refinar os resultados. Além disso, o conjunto de resposta era extensamente analisado em busca de um item relevante. Poucas consultas eram efetuadas e os resultados retornados eram mais analisados.

Segundo Manning, Raghavan e Schütze (2009), as medidas formais para avaliação dos métodos de recuperação de informação estão, de certa forma, distantes do objetivo final das medições: a satisfação do usuário, que é independente da qualidade dos resultados

retornados pois inclui outros fatores como a usabilidade, o tempo de resposta do sistema e questões de leiaute.

Visando proporcionar uma investigação mais abrangente da eficiência dos métodos de recuperação de informação em análise, após o encerramento das atividades planejadas na etapa de experimento, os participantes foram entrevistados seguindo a pauta presente no ANEXO D - Roteiro da Entrevista com Usuário.

A seguir é feita uma análise de cada uma das questões quanto à usabilidade das ferramentas e quanto aos resultados das buscas efetuadas através das mesmas.

Usabilidade

Quando perguntados sobre qual o método mais agradável para uso, 15 dos 16 participantes responderam ser a navegação facetada, conforme pode ser observado na Figura 28. Os motivos mais comuns apontados para a escolha foram: a praticidade na utilização; a combinação dos recursos de navegação com a possibilidade de usar a busca por palavras ou expressões; a possibilidade de cruzar as categorias existentes na coleção; e a funcionalidade de refinar os resultados de uma consulta já efetuada. Apenas um participante indicou a busca direta como sendo o método mais agradável para uso, devido à maior familiaridade com o mesmo.

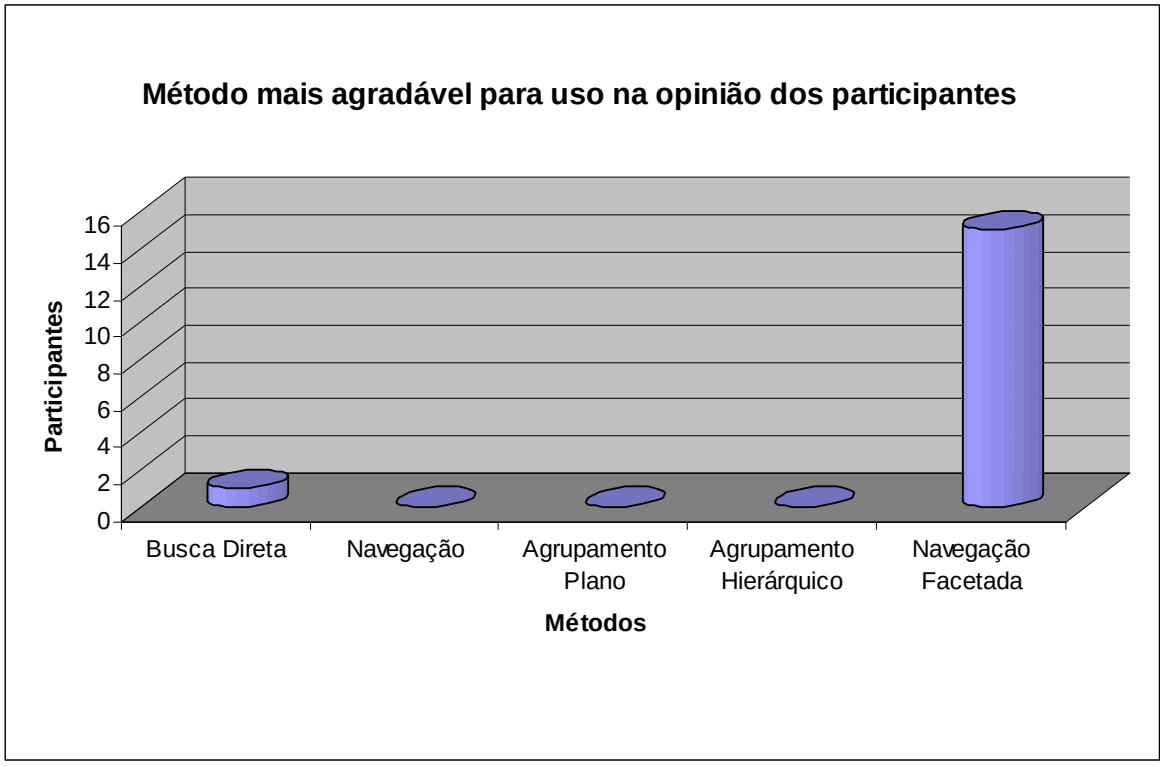


Figura 28 - Método mais agradável para uso na opinião dos participantes

Em relação ao método que se mostrou menos agradável para utilização, as opiniões ficaram divididas conforme ilustra a Figura 29. As razões para o desagrado em relação ao método de agrupamento plano foram: a pouca previsibilidade dos resultados; os nomes dos agrupamentos não conferirem com a necessidade da informação desejada, sendo, portanto pouco representativos; e a dificuldade de elaborar uma consulta que consiga extrair bons resultados. Quanto à navegação, as reclamações foram unânimes: a impossibilidade de cruzar as categorias e a ausência de busca direta.

Ainda com relação ao método menos agradável, um participante apontou a busca direta, por retornar uma quantidade muito grande de filmes, além de não permitir quaisquer filtros para refinar o resultado da consulta. Outro participante julgou ser o agrupamento hierárquico por não o considerar de fácil compreensão.

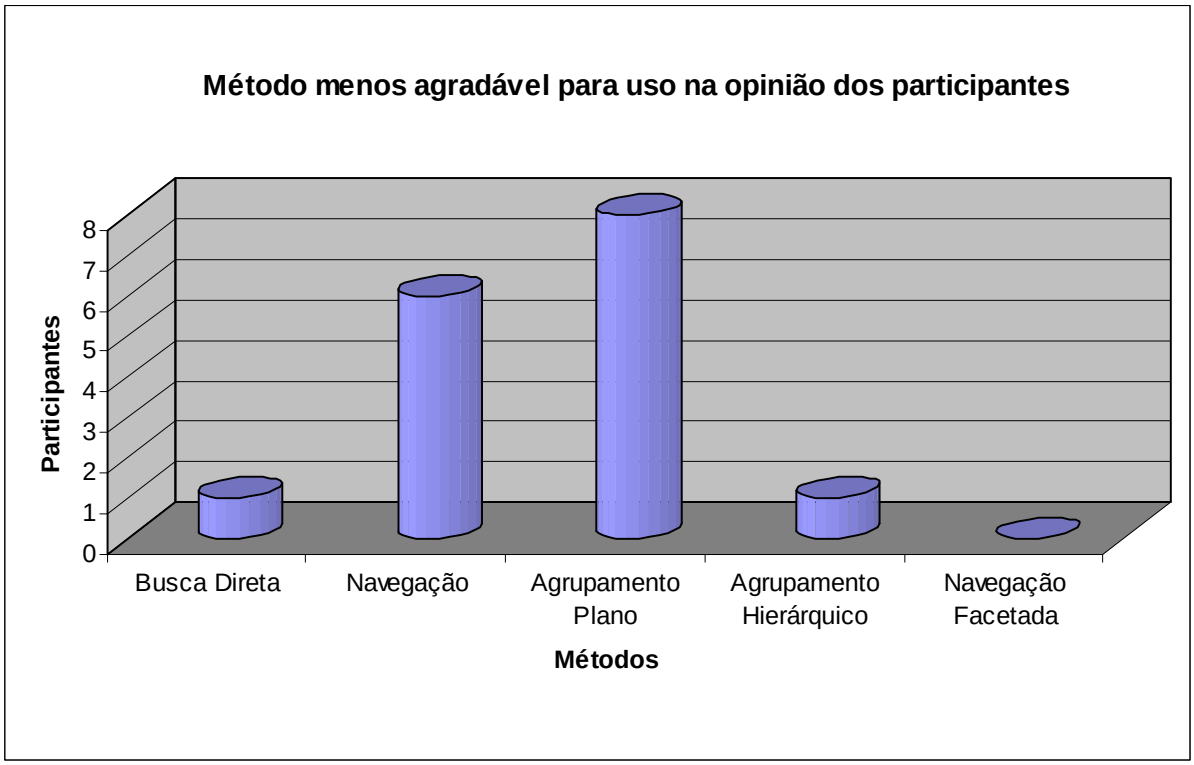


Figura 29 – Método menos agradável para uso.

Quando questionados sobre qual método permitiu conhecer melhor os tipos de informação disponíveis na coleção de filmes, as respostas ficaram divididas basicamente entre a navegação facetada e o agrupamento hierárquico. Nesse ponto, uma observação interessante foi a de que metade do grupo de analistas de sistemas acreditou que o agrupamento hierárquico auxiliava de forma mais eficiente a conhecer o que era possível extrair de informações. No entanto, dentre o grupo pertencente aos participantes de outras profissões, a escolha pela navegação facetada foi unânime. A Figura 30 apresenta a distribuição das opiniões no quesito em análise.

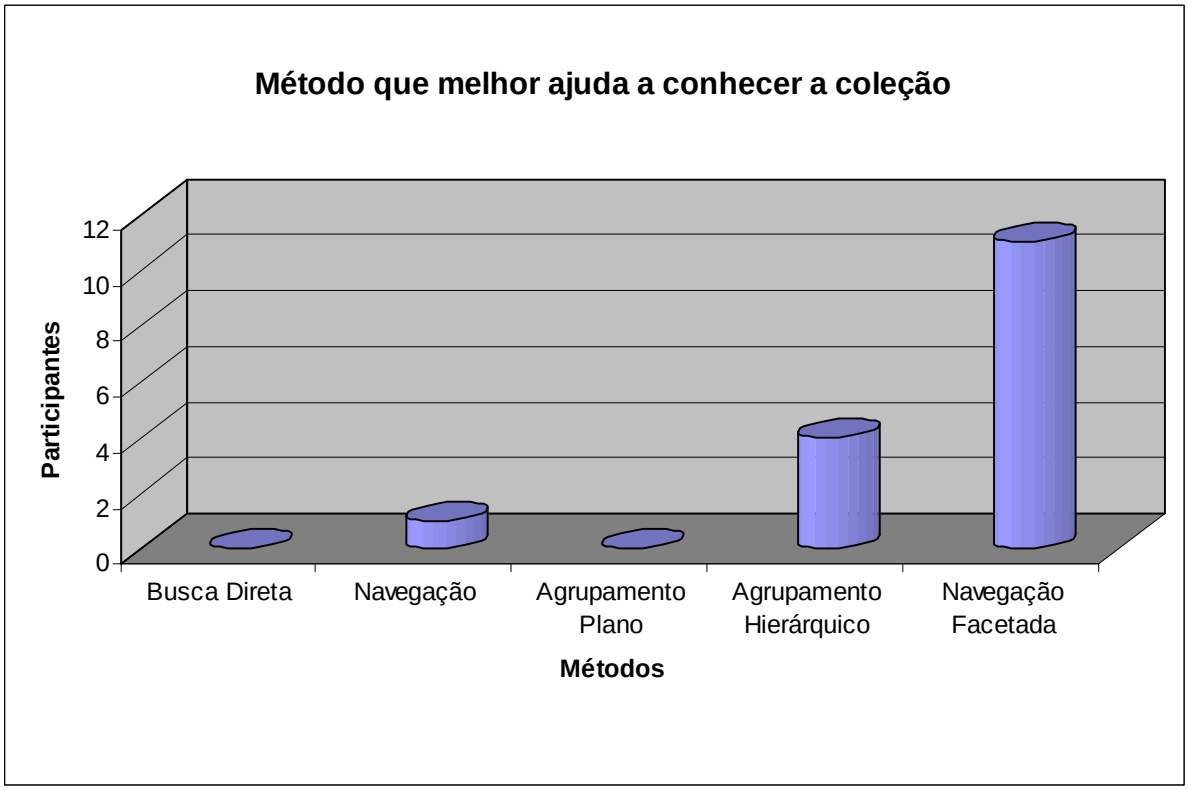


Figura 30 – Método que melhor ajuda a conhecer a coleção.

Ao serem questionados sobre qual método escolheriam caso todos os cinco estivessem disponíveis para realização das tarefas do experimento a resposta foi única dentre todos os 16 participantes: navegação facetada. As justificativas se repetiram de acordo com o que foi destacado na pergunta 1.1, que discorria sobre o método mais agradável para uso, ou seja, foram apontadas: a praticidade no uso; a combinação dos recursos de navegação com busca direta; a possibilidade de cruzar as categorias predefinidas da coleção; e a funcionalidade de refinar os resultados de uma consulta já efetuada anteriormente. Outro ponto a ser observado é que mesmo o participante que havia elegido a busca direta como o método mais agradável para uso, acabou por escolher a navegação facetada para a execução de todas as tarefas caso lhe fosse possível optar por apenas uma das ferramentas.

Resultados das buscas

O conjunto de perguntas sobre o desempenho dos métodos em relação aos objetivos e aos resultados que precisavam ser alcançados, possuía forte correlação com as medidas quantitativas colhidas no decorrer do experimento. Tal correspondência pode ser observada segundo o ANEXO C – Correlação dos Itens Quantitativos com os Qualitativos.

Questionados sobre a precisão dos métodos, sendo pedido para que elegessem o mais preciso dentre os cinco analisados, dos 16 participantes, 15 escolheram a navegação facetada como sendo o mais preciso, conforme apresentado na Figura 31. Um participante optou pelo agrupamento hierárquico. Quando a eleição possuía como objetivo apontar o método menos preciso, as opiniões se mostraram distribuídas segundo apresentado na Figura 32. Um aspecto interessante é o de que, de uma forma geral, para 5 dos 8 participantes analistas de sistemas, o método menos preciso é o de navegação, enquanto que, para o outro grupo, a resposta que mais se sobressaiu foi a busca direta, também para 5 dos 8 profissionais envolvidos no experimento. É necessário ressaltar ainda que tanto a eleição do método mais preciso, como também a do menos preciso, exercem um papel confirmatório daquilo que foi aferido segundo medidas quantitativas, conforme exibe a Tabela 4. De fato, o cálculo do *R-Precision* apontou a navegação facetada e o agrupamento hierárquico como sendo os mais precisos, e a navegação e a busca direta como sendo os menos precisos.

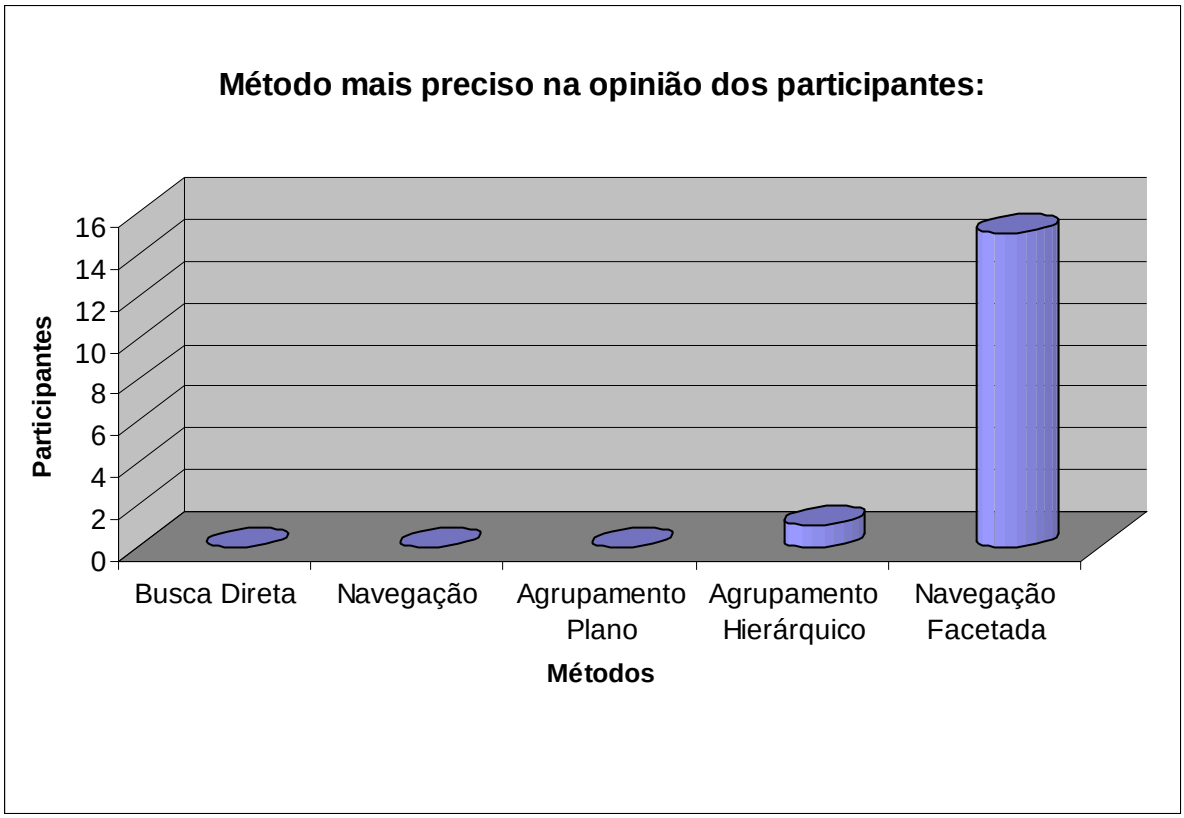


Figura 31 – Método mais preciso na opinião dos participantes.

Na questão qualitativa referente à cobertura, os participantes foram perguntados se alguma informação que eles esperavam encontrar havia ficado de fora por algum método, e, em caso afirmativo, qual havia sido o método. A maioria dos participantes, 12 deles, respondeu que não percebeu nenhuma informação ausente. No entanto, duas pessoas apontaram problemas de cobertura na navegação, uma na busca direta e outra no agrupamento plano, o que de fato reafirma o que havia sido constatado pelos dados quantitativos, como pode ser observado na Figura 32.

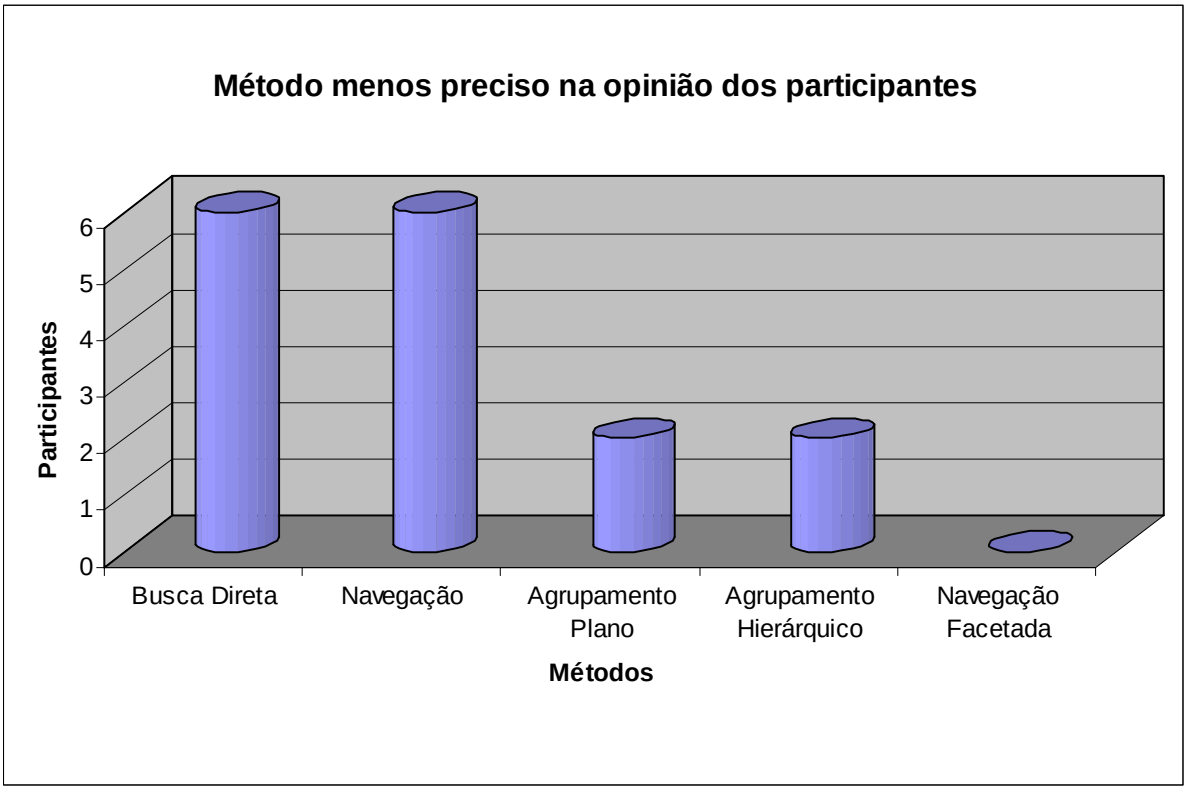


Figura 32 – Método menos preciso na opinião dos participantes.

A pergunta seguinte dizia respeito ao método que levava aos documentos relevantes de forma mais rápida, com menor esforço. Em outras palavras, qual método gastava menos tempo até se obter os filmes desejados. Novamente todos os 16 participantes responderam que a navegação facetada conduzia até as respostas corretas em menor tempo. A Figura 33 ilustra a distribuição das respostas obtidas. Quando questionados sobre qual método demandava mais tempo, as opiniões ficaram concentradas na navegação e no agrupamento plano. Um participante também concluiu que a busca direta o fez gastar mais tempo. A Figura 34 apresenta os resultados da pergunta em análise revelando a distribuição de suas respostas.

Novamente os dados qualitativos, oriundos da percepção dos participantes, confirmam os resultados obtidos através dos dados quantitativos do experimento. Conforme pode ser observado na Figura 34, a navegação e o agrupamento plano foram revelados pelos números do experimento como sendo os que demandam mais esforço do usuário, consumindo mais do seu tempo na tarefa.

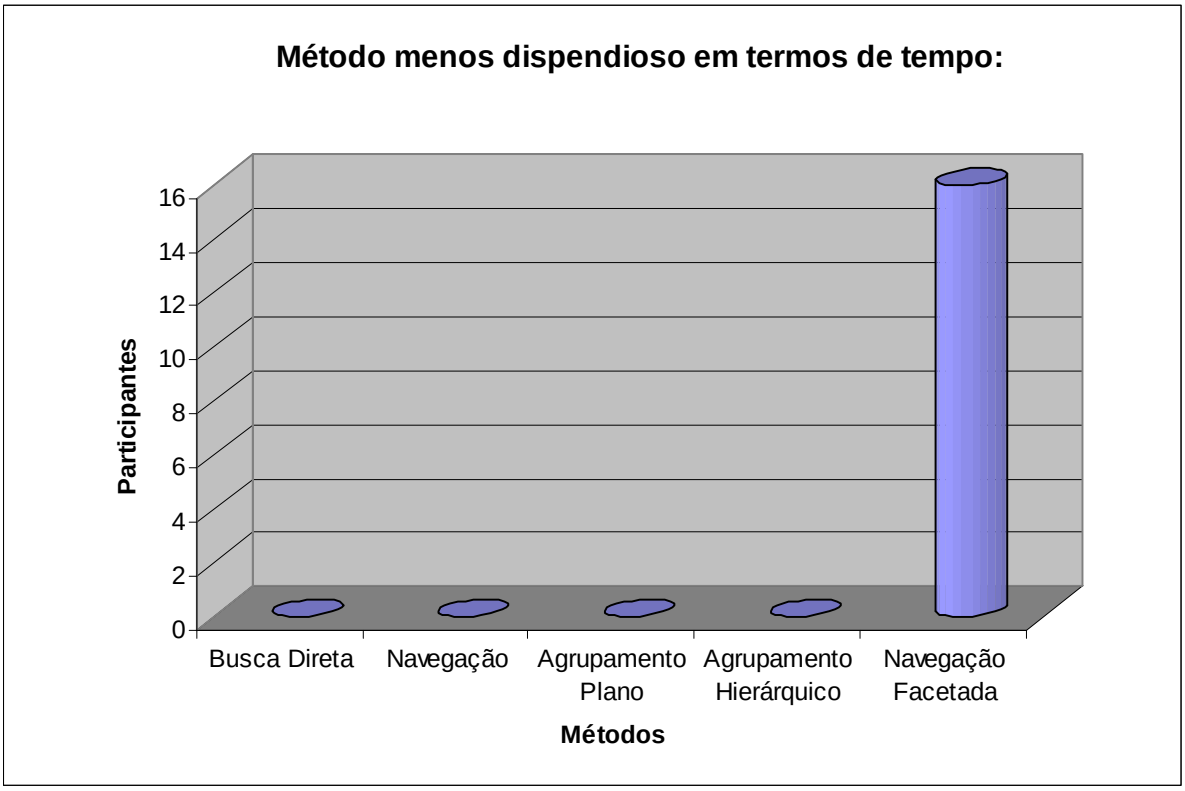


Figura 33 - Método menos dispendioso em termos de tempo.

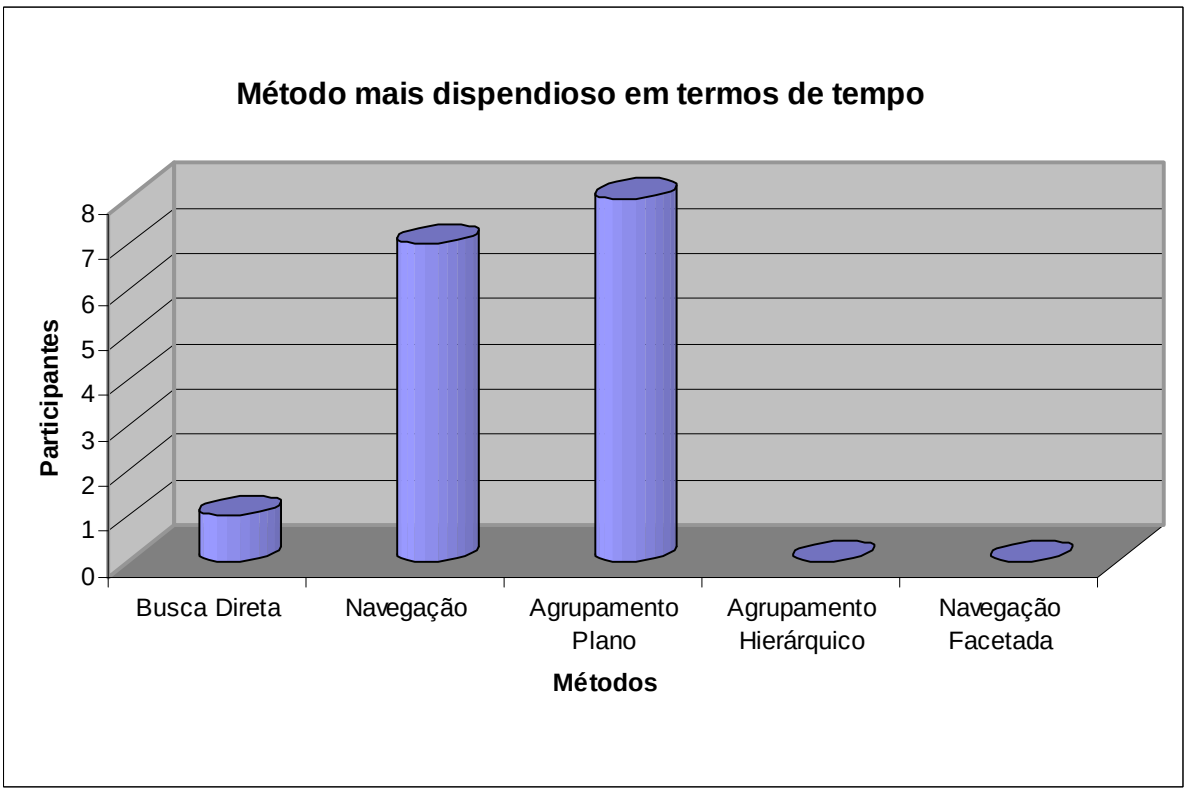


Figura 34 – Método mais dispendioso em termos de tempo.

Por último, quando questionados sobre o método que lhes proporcionou maior certeza no momento de oferecer as respostas, a navegação facetada foi unanimidade dentre os 16 participantes, conforme ilustra a Figura 35. Já na escolha do método que proporcionava menos segurança, as opiniões ficaram divididas segundo a Figura 36. Busca direta e agrupamento plano ficaram empatados em primeiro lugar, seguidos pela navegação.

Tal resultado, quando comparado à aferição quantitativa da taxa de erro por método, apresentada na Tabela 2, possui pequenas variações. De fato, a confirmação de que o agrupamento plano possui uma das mais elevadas taxas de erro, condiz com a análise qualitativa. No entanto, apesar de a busca direta apresentar baixa taxa de erro, foi percebida como um dos métodos que oferece menos segurança. Talvez isso se deva ao fato de a busca direta não associar os resultados com uma categorização explícita, limitando a ordenar os itens do conjunto resposta.

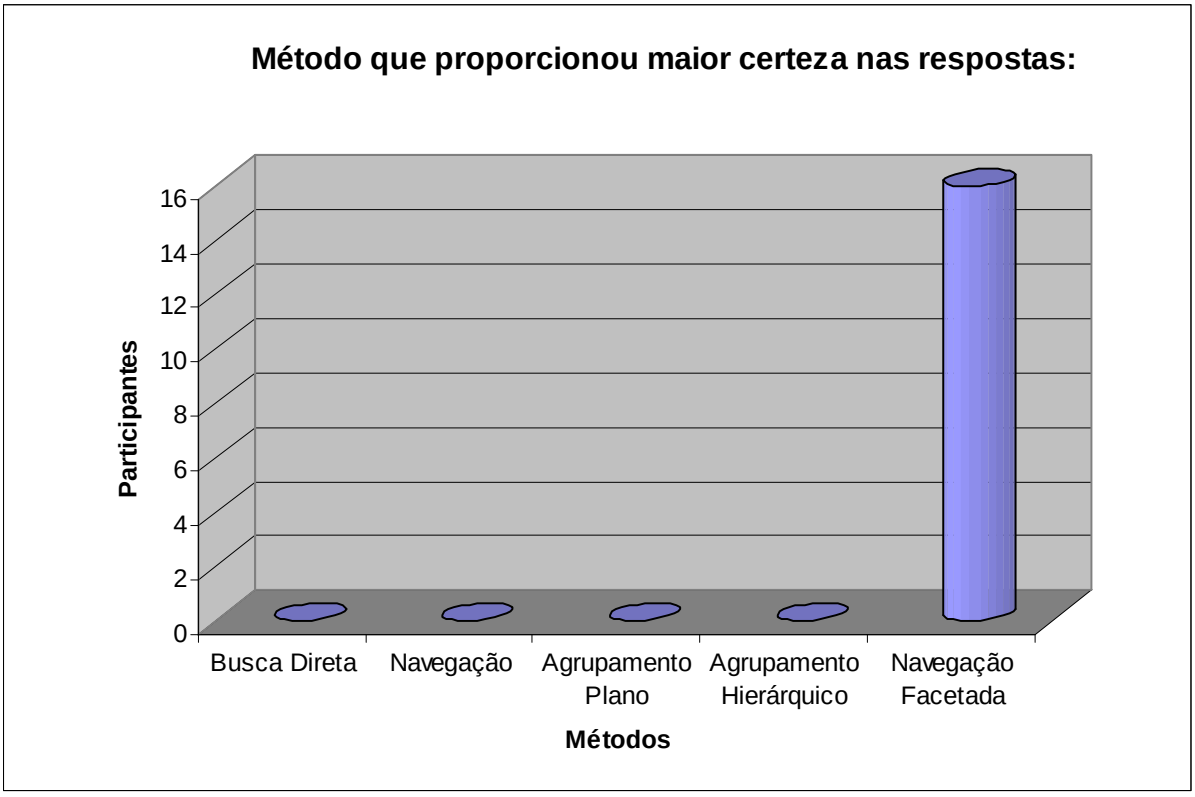


Figura 35 - Método que proporcionou maior certeza nas respostas.

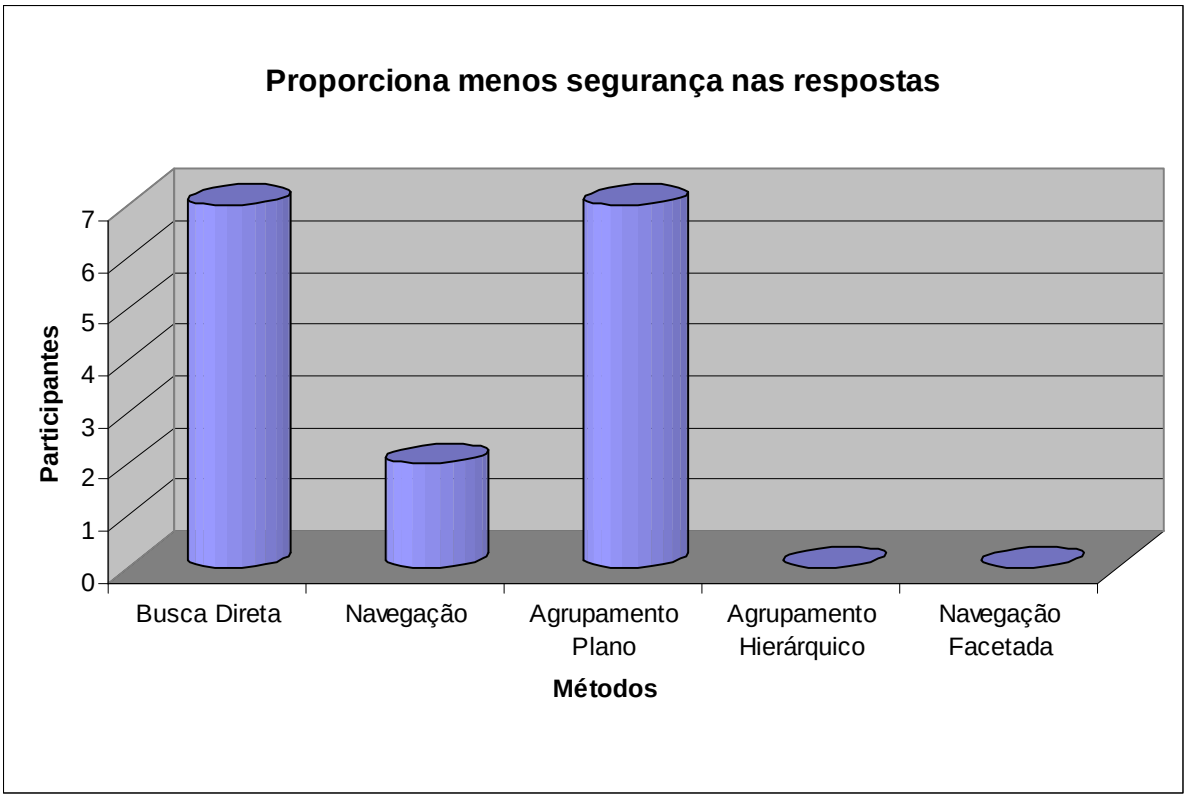


Figura 36 – Métodos que proporcionam menos segurança nas respostas.

5.3. REVISÃO DAS HIPÓTESES

O Quadro 4 apresenta a revisão das hipóteses de acordo com a análise dos dados do experimento. Para validar as hipóteses foram utilizados todos os quesitos levantados na pesquisa, ou seja, para verificar a eficiência de cada método em relação a outro, foram utilizados o tempo de execução das tarefas, a taxa de erro, o número de documentos relevantes encontrados por minuto, a precisão e a cobertura calculados através do *R-Precision*.

Hipótese	**Tempo de Execução** (ver Tabela 1)	**Taxa de Erro** (ver Tabela 2)	**Documentos Relevantes/Min.** (ver Tabela 3)	***R-Precision*** (ver Tabela 4)	**Validação**
H1: a busca direta é menos eficiente que o agrupamento plano.	Não	Não	Não	Inconclusivo	–
H2: a busca direta é menos eficiente que o agrupamento hierárquico.	Inconclusivo	Inconclusivo	Não	Sim	–
H3: a busca direta é menos eficiente que a navegação facetada.	Sim	Sim	Sim	Sim	Aceita.
H4: a navegação é menos eficiente que o agrupamento plano.	Não	Não	Inconclusivo	Sim	–
H5: a navegação é menos eficiente que o agrupamento hierárquico.	Sim	Sim	Sim	Sim	Aceita.
H6: a navegação é menos eficiente que a navegação facetada.	Sim	Sim	Sim	Sim	Aceita.
H7: o agrupamento plano é menos eficiente que o agrupamento hierárquico.	Sim	Sim	Sim	Sim	Aceita.

H8: o agrupamento hierárquico é menos eficiente que a navegação facetada.	Sim	Sim	Sim	Sim	Aceita.

Quadro 4 – Revisão das hipóteses.

Como pode ser observado no Quadro 4, as hipóteses H3, H5, H6, H7 e H8, formuladas pela presente pesquisa, foram confirmadas segundo os dados proporcionados pelo experimento. No entanto, quanto à validação das hipóteses H1, H2 e H4, não se pôde chegar a um resultado conclusivo, uma vez que alguns dados coletados não apontam claramente quanto à eficiência de um método em relação a outro.

No caso da hipótese H1, o fato do *R-Precision* da busca direta se mostrar muito próximo ao do agrupamento plano sugere que uma melhor compreensão do método pelos participantes, a ser refletida no uso mais adequado da ferramenta, possa gerar resultados mais favoráveis ao agrupamento plano e também mais confiáveis.

O uso incorreto dos métodos de agrupamento pode ter trazido conseqüências também para a validação da hipótese H2, uma vez que as medidas de cobertura e precisão, representadas pelo *R-Precision*, indicaram a superioridade do agrupamento hierárquico em relação à busca direta. Cabe ressaltar ainda que as entrevistas apontam a busca direta como um dos métodos que mais oferecem insegurança no momento das respostas, o que está relacionado à taxa de erro de um método. Apesar dos dados qualitativos darem clara vantagem ao agrupamento hierárquico, os dados quantitativos não se mostraram suficientemente claros quanto à eficiência dele em relação à busca direta.

As observações realizadas para a hipótese H1 são aplicáveis à hipótese H4, uma vez que o *R-Precision* do agrupamento plano se mostrou nitidamente superior ao da navegação.

6. CONCLUSÕES

A presente pesquisa analisou a eficiência dentre os seguintes métodos de recuperação da informação: (i) busca direta, (ii) navegação, (iii) agrupamento plano (*flat clustering)*, (iv) agrupamento hierárquico (*hierarchical clustering*) e (v) navegação facetada.

Através da aplicação de uma abordagem multimétodo, os dados quantitativos foram coletados utilizando a técnica de experimento com 16 participantes. O tempo utilizado para execução das tarefas, a taxa de erro, a quantidade de documentos relevantes encontrados por minuto e o *R-Precision* foram analisados para cada um dos métodos investigados.

A técnica de entrevista por pauta proporcionou a coleta dos dados qualitativos. As percepções dos usuários, quanto à usabilidade e aos resultados da busca disponibilizados através de cada um dos métodos, foram recolhidas objetivando a triangularização com os dados quantitativos.

Neste capítulo, as principais contribuições do presente estudo são realçadas e novas pesquisas são propostas para futuras investigações.

6.1. PRINCIPAIS CONTRIBUIÇÕES

Dentre as contribuições do presente trabalho de pesquisa, destacam-se as seguintes.

A realização do levantamento das coleções formais de testes disponíveis atualmente no âmbito acadêmico, com a correlação de seus respectivos assuntos e referências. Da mesma forma, também se pode ressaltar o resultado da pesquisa por coleções não formais que, embora não se caracterizem atualmente como acervos de referência, podem sofrer adequações, a fim de atender a trabalhos investigativos com necessidades específicas.

Ainda no quesito das coleções de teste, o presente trabalho oferece como contribuição a construção de uma grande coleção de filmes, rica em informações e com aproximadamente 676 mil documentos. A nova coleção poderá proporcionar outras investigações a cerca do uso

do método de navegação facetada. Este subsídio é particularmente relevante, uma vez que não há disponibilidade por tal tipo de coleção na Web.

Outro destaque, dentre as contribuições, são as ferramentas de busca inteiramente desenvolvidas assim como aquelas que foram adaptadas, totalizando cinco programas específicos, todos seguindo as melhores práticas existentes para cada um dos métodos investigados. A configuração da máquina de busca, com toda a indexação do acervo utilizado, aproximadamente 114 mil documentos, também estará disponível para uso.

Quanto ao tema da pesquisa, relevante contribuição do trabalho foi realizada através da utilização de metodologia multimétodo, segundo a abordagem paralela (CRESWELL, 1994, 1998), para a coleta e análise de dados não apenas quantitativos, mas também qualitativos. Segundo Manning, Raghavan e Schütze (2009), a percepção do usuário é a mais importante métrica a ser analisada, e que normalmente se encontra ausente dos estudos desenvolvidos na área de Recuperação de Informações, levando-se em conta apenas as métricas formais.

O experimento realizado durante o presente trabalho investigativo apontou o método de navegação como sendo menos eficiente que o agrupamento hierárquico nas situações testadas. Esta análise foi reafirmada através da averiguação das respostas fornecidas pelos participantes no momento da entrevista.

A pesquisa ofereceu também mais um indício de que o agrupamento hierárquico possui melhor eficiência quando comparado ao agrupamento plano, conforme apontado por trabalhos anteriores (JAIN, A., DUBES, R., 1988), (CUTTING et al., 1992), (LARSEN, B., AONE, C., 1999).

Cabe ressaltar ainda que os resultados da presente pesquisa permitiram observar que a navegação facetada se mostrou como o mais eficiente dentre os cinco métodos analisados, mesmo havendo o viés do cansaço do participante contra ela, por ter sido o último método

executado em cada sessão do experimento. Tal constatação foi inclusive confirmada através das medições feitas utilizando a percepção do usuário.

Dentre as contribuições oriundas da análise da etapa qualitativa, pode-se destacar a indicação dos métodos de navegação e busca direta como sendo os menos precisos na opinião dos participantes. E, ainda, a percepção de que a busca direta e o agrupamento plano foram os métodos que menos ofereceram segurança quanto à relevância dos itens encontrados.

6.2. TRABALHOS FUTUROS

A realização de um novo experimento que busque a comparação da eficiência entre busca direta e agrupamento hierárquico, porém utilizando uma amostragem maior de participantes, poderia responder ao questionamento sobre qual dentre tais métodos é o mais eficiente. A coleta e análise dos dados poderiam ser feitas de forma quantitativa. Nos resultados do presente trabalho, os dados qualitativos apontaram maior eficiência do agrupamento hierárquico, porém os dados quantitativos não apresentaram clareza suficiente para se chegar a uma conclusão.

Na atual investigação sobre os métodos de recuperação da informação, os participantes do experimento não tiraram o melhor proveito da capacidade das ferramentas de agrupamento. A realização de um treinamento mais demorado, seguido de novos experimentos, poderia verificar se tanto o agrupamento plano quanto o hierárquico apresentariam melhores resultados.

Novos estudos poderiam investigar as diferentes técnicas que promovem melhoria da nomeação automática dos agrupamentos (POPESCUL, A., UNGAR, L., 2000; GLOVER et al., 2002; STEIN, B., ZU EISSEN, S. 2004), visto que tal característica foi responsável por boa parte das reclamações dos participantes. Pela atribuição de nomes pouco expressivos ou que não tinham correlação com a busca efetuada, os participantes apontaram o agrupamento

plano como sendo o método menos agradável para uso, bem como o que mais despendia tempo de utilização durante a execução das tarefas.

Além de novos experimentos com os métodos de agrupamento, precedidos de treinamento, e da utilização de técnicas para melhoria da nomeação dos agrupamentos, a realização de novas pesquisas sobre a eficiência dos métodos, efetuando a comparação dos mesmos em duplas, poderia complementar os resultados encontrados na presente pesquisa. Ao diminuir o número de métodos em análise, a quantidade de tarefas utilizadas no experimento poderia ser maior, variando tanto o seu tipo quanto o número. Nesse sentido, uma pesquisa futura poderia investigar o mais eficiente dentre os métodos de busca direta e agrupamento hierárquico, e dentre os métodos de navegação e agrupamento plano.

Visando o caráter confirmatório existente na produção do conhecimento cientifico, novas pesquisas, similares à realizada pelo presente trabalho, poderiam ser conduzidas utilizando os métodos de busca analisados e outra coleção de conhecimento homogêneo.

REFERÊNCIAS BIBLIOGRÁFICAS

ALLAN, J. HARD track overview in TREC 2005: high accuracy retrieval from documents. In: TEXT RETRIEVAL CONFERENCE, 14., 2005, Gaithersburg, **Proceedings …** Gaithersburg: NIST, 2005. p. 174-521.

ALONSO, O. ; BANERJEE, S. ; DRAKE, M. GIO: a semantic web application using the information grid framework. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 15., 2006, Edinburgh. **Proceedings …** New York: ACM, 2006. p. 857–858.

ANSWERS.com. Disponível em: http://answers.com. Acesso em: 15 dez. 2008.

ASK.com. Disponível em: http://ask.com Acesso em: 27 dez. 2008.

BAEZA-YATES, R. ; RIBEIRO-NETO, B. **Modern information retrieval**. Boston: Addison Wesley.1999.

BBC. Disponível em: http://www.bbc.co.uk Acesso em: 29 jan. 2009.

BERKHIN, P. **Survey of clustering data mining techniques**. San Jose: Accrue Software, 2002. (Technical Report).

BRIN, S. ; PAGE, L. **The Anatomy of a large-scale hypertextual web search engine.**. Stanford: Computer Science Department, Stanford University, 1998.

CARPINETO, C.,et al.. A survey of web clustering engines.**ACM Computing Surveys**, New York, v. 41, n. 3. Ju.l. 2009.

CARROT². 2009. Disponível em: http://search.carrot2.org Acesso em: 05 set. 2009.

CARROT² Search. 2009. Disponível em: http://company.carrot-search.com/ Acesso em: 05 set. 2009.

CENADEM. Centro Nacional da Gestão da Informação 2009. Disponível em: http://www.cenadem.com.br Acesso em: 05 abr. 2009.

COLEÇÃO chave. 2009. Disponível em: http://www.linguateca.pt/chave/ Acesso em: 04 ago. 2009.

COX, B. It´s official, it´s called ruby, it´s in beta. **Ecommerce-guide.com**. November 1, 2002.

CRESWELL, J. W. **Research design: qualitative & quantitative approaches**. Thousand Oaks: Sage, 1994.

_____. Data anaslysis and representation In: _____. **Qualitative inquiry and research design: choosing among five traditions**. Thousand Oaks: Sage, 1998. chap. 8, p. 139-165.

CROFT, W. B. Combining approaches to information retrieval. In: _____. **Advances information retrieval**; recent researcn from the center for intelligent information retrieval. Norwell: Kluwer Academic Publishers. 2000. cap 1. p.1–36.

CUTTING, D.et al.. Scatter/Gather: a cluster-based approach to browsing large document collections. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 15. 1992, Copenhagen **Proceedings …** New York: ACM Press, 1992. p. 318–329.

DBLP. **Digital bibliography & library project**. 2009a. Disponível em: http://www.informatik.uni-trier.de/~ley/db/ Acesso em: 14 ago. 2009.

DBLP. **Digital bibliography & library project**. 2009b. Disponível em: http://dblp.l3s.de/dblp++.php Acesso em: 14 ago. 2009.

DEMO, P. **Metodologia do conhecimento científico**. São Paulo: Atlas. 2000.

DENTON, W. **How to make a faceted classification and put it on the web**. Nov. 2003. Disponível em: http://www.miskatonic.org/library/facet-web-howto.html Acesso em: 12 jun. 2009.

DIAS, D. ; SILVA, M.; **Como fazer uma monografia**. Rio de Janeiro: Atlas, 2010.

DIGG. 2009. Disponível em: http://www.digg.com. Acesso em: 05 fev. 2009.

DIJCK, P. **XFML Core eXchangeable faceted metadata language**. 2003. Disponível em: http://www.xfml.org/spec/1.0.html Acesso em: 2009.

DMOZ. **Open directory project**. 2009. Disponível em: http://www.dmoz.org/ Acesso em: 05 ago. 2009.

DUMAIS, S. ; CUTRELL, E. ; CHEN, H. Optimizing search by showing results in context. In: SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 2001, Seattle. **Proceedings …** New York: ACM, 2001. p. 277–284.

EBAY. 2009. Disponível em: http://www.ebay.com Acesso em: 10 abr.2009.

EBAY´S express to take on Amazon. **The Sunday Times,** London, October 1, 2006.

EDDIE Bauer sport wears. 2009. Disponível em: http://www.eddiebauer.com/. Acesso em: 07 jan. 2009.

ENGLISH, J. et al.. **Examining the usability of web site search**. 2001. Disponível em: http://flamenco.berkley.edu/papers/epicurious-study.pdf. Acesso em: 12 jun. 2009.

FRAKES, W. ; YATES, R. **Information retrieval**: data structures & algorithms. New Jersey: Prentice Hall, 2000.

GARTNER. **Gartner technology business research**. 2009. Disponível em: http://www.gartner.com Acesso em: 07 abr. 2009.

GIL, A. C. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2007.

GLOVER, E. et al. Using web structure for classifying and describing web pages. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 11., 2002, Honolulu. **Proceedings …** New York: ACM, 2002. p. 562-569.

GOOGLE. Disponível em: http://www.google.com.br Acesso em: 05 fev. 2009.

GREENGRASS, E. **Information retrieval:** a survey. November, 2000.

HASSENZAHL, M. The Interplay of beauty, goodness, and usability in interactive products. **Human-Computer Interaction**. Hillsdale, v. 19, n. 4, p. 319-349, Dec. 2004.

HEARST, M. A. Clustering versus faceted categories for information exploration. **Communications of the ACM, N**ew York, v. 49, n. 4, Apr. 2006.

_____. Design recommendations for hierarchical faceted search interfaces. **ACM SIGIR Workshop on Faceted Search.** August, 2006. Disponível em: http://flamenco.berkeley.edu/papers/faceted-workshop06.pdf. Acesso em: 2009.

_____. UIs for faceted navigation: recent advances and remaining open problems. In: WORKSHOP ON HUMAN-COMPUTER INTERACTION AND INFORMATION RETRIEVAL, 2., 2008. Redmond, WA. **Proceedings …** Regmond: Microsoft Research, 2008.

_____. **Search user interfaces**. Cambridge: Cambridge University Press. 2009

HEARST, M. ; PEDERSEN J. Reexamining the cluster hypothesis. In ANNUAL INTERNATIONAL SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT, IN INFORMATION RETRIEVAL, 19., 1996, Zurich. **Proceedings …** New York, ACM, 1996. p. 76–84.

HEARST, M.A. et al. Finding the flow in web site search. **Communications of the ACM**, New York, v. 45, n. 9, p. 42-49, Sept. 2002.

HEARST, M. A. et al. **Flexible search and navigation using faceted metadata**. Berkeley: University of California, Berkeley. Search Engines Meeting, April 2002. (Technical Report).

HP. 2009. Disponível em: http://www.hp.com Acesso em: 8 abr. 2009.

IMDB. 2009. Disponível em: http://www.imdb.com/interfaces/ Acesso em: 05 set. 2009.

IQPC. International quality & productivity center. 2009. Disponível em: http://www.iqpc.com Acesso em: 07 abr. 2009.

JAIN, A. ; DUBES, R. **Algorithms for clustering data**. Upper Saddle River: Prentice Hall. 1988.

KAKI, M. Findex: search result categories help users when document rankings fail. In: SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS. 2005, Portland. **Proceedings …** .New York: ACM, 2005.

KALBACH, J. **Designing web navigation: optimizing the user experience**. Beijing: O'Reilly, 2007.

KOREN, J. ; ZHANG, Y. Personalized interactive faceted search. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 17., 2008, Beijing. **Proceedings …** New York: ACM, 2008. p. 477-486.

KULES, B. ; SHNEIDERMAN, B. Users can change their web search tactics: design guidelines for categorized overviews. **Information Processing and Management**, Elmsford, v. 44, n. 2, p. 463–484, Mar. 2008.

KWASNICK, B. H. The role of classification in knowledge representation and discovery. **Library Trends**, Champaing, v. 48, n. 1, p. 22-47, 1999.

LARSEN, B. ; AONE, C. Fast and effective text mining using linear-time document clustering. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 5., 1999, San Diego. **Proceedings …** New York: ACM, 1999. p. 16–22.

LEAVITT, M. ; SHNELDERMAN, B. **Research-based web design and usability guidelines.** U.S. Department of Health and Human Services. 2006. Disponível em: http://www.usability.gov/pdfs/guidelines.html Acesso em: 16 fev. 1999.

LINDGAARD, G. ; CHATTRATICHART, J. Usability testing: what have we overlooked? In: SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 2007, San Jose. **Proceedings …** New York: ACM, 2007.p. 1415-1424.

LINDGAARD, G. ; DUDEK, C. What is this evasive beast we call user satisfaction? **Interacting with Computers**, London, v. 15, n. 3, p. 429-452, Jun. 2003.

LINGO 3G. 2009. Disponível em: http://company.carrot-search.com/lingo-applications.html Acesso em: 05 set. 2009.

LUCENE. apache. 2009. Disponível em: http://lucene.apache.org/ Acesso em: 10 set. 2009.

MACKINLAY, J. ; ZELLWEGER. P. Panel: browsing vs. search: can we find a synergy?. In: CONFERENCE COMPANION ON HUMAN FACTORS IN COMPUTING SYSTEMS, 1995, Denver. **Proceedings …** New York: ACM, 1995. p. 179-180.

MANBER, U. ; SMITH, M. ; GOPAL, B. WebGlimpse: combining browsing and searching. In: ANNUAL CONFERENCE ON USENIX ANNUAL TECHNICAL CONFERENCE, 1997, Anaheim. **Proceedings …**Berkeley: USENIX Association, 1997. p.15-15.

MANNING, C. ; RAGHAVAN, P. ; SCHÜTZE, H. **Introduction to information retrieval**. Cambridge University Press. 2009. Disponível em: http://nlp.stanford.edu/IR-book/information-retrieval-book.html Acesso em: 15 jun. 2009.

MARCHIONINI, G. Exploratory search: from finding to understanding. **Communications of the ACM**, New York, v. 49, n. 4, p. 41-46, Apr. 2006.

MORVILLE, P. ; ROSENFELD, L. **Information architecture for the world wide web**. Beijing: O´Reilly. 2006.

MUSICBRAINZ database. 2009. Disponível em: http://musicbrainz.org/doc/Database_Download Acesso em: 01 set. 2009.

NEW York Times annotated corpus. **New York Times**. 2009. Disponível em: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T19 Acesso em: 16 ago. 2009.

NIELSEN, J. **Usability engineering**. San Diego: Academic Press, 1993.

_____. The matters that really matter for hypertext usability. In: ANNUAL ACM CONFERENCE ON HYPERTEXT, 2.,1989. Pittsburgh. **Proceedings …** New York: ACM, 1989. p. 239–248.

POPESCUL, A. ; UNGAR, L. **Automatic labeling of document clusters**. U. Pennsylvania. 2000. Disponível em: http://www.cis.upenn.edu/ popescul/Publications/popescul00labeling.pdf. Acesso em: 16 ago. 2009.

PRATT, W. ; HEARST, M. ; FAGAN, L. A knowledge-based approach to organizing retrieved documents. In: National Conference on Artificial Intelligence, 16., 1999. Orlando. **Proceedings …** Menlo Park: American Association for Artificial Intelligence, 1999. p. 80-85.

PUBMED.gov. Disponível em: http://www.ncbi.nlm.nih.gov. Acesso em: 10/01/2009.

RANGANATHAN, S. R. **Elements of library classification**. New York: Asia Publishing House. 1962.

SACCO, G. M. Research results in dynamic taxonomy and faceted search systems. In: INTERNATIONAL WORKSHOP ON DYNAMIC TAXONOMIES AND FACETED SEARCH, 2007, Regensburg. **Proceedings …** Washington: IEEE, 2007.

SALOMON, D. V. **Como fazer uma monografia**. 11 ed. São Paulo: Martins Fontes, 2004.

SMILEY, D. ; PUGH, E. **Solr 1.4 enterprise search server**. [S.l]. Packt Publishing. 2009.

SOLR. apache. 2009. Disponível em: http://lucene.apache.org/solr/ Acesso em: 10 set. 2009.

SPITERI, L. A simplified model for facet analysis: Ranganathan 101. **Canadian Journal of Information and Library Science,** Ontario, v., 23. n. 1-2, p. 1-30, Apr./Jul.1998.

STEIN, B. ; ZU EISSEN, S. Topic identification: framework and application. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE MANAGEMENT, 13., 2004, Washington. **Proceedings …** New York, ACM, 2004.

SUN. 2009. Disponível em: http://www.sun.com Acesso em: 10 jan. 2009.

TEEVAN, J. et al. The perfect search engine is not enough: a study of orienteering behavior in directed search. In: Proc. of the SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEM, 2004, Vienna. **Proceedings …** New York: ACM, 2004. p. 415-422.

TODA, H., KATAOKA, R. A search result clustering method using informatively named entities. In: INTERNATIONALWORKSHOP ONWEB INFORMATION AND DATA MANAGEMENT, 7., 2005. Bremen. **Proceedings …** New York: ACM, 2005. p. 81–86.

TOMBROS, A. ; VILLA, R. ; VAN RIJSBERGEN, C. J. The effectiveness of query-specific hierarchic clustering in information retrieval. **Information Processing & Management**, Elmsford, v. 38, n. 4, p. 559-582, Jul. 2002

TUNKELANG, D. **Faceted search**. San Rafael: Morgan & Claypool. 2009.

VERGARA, S. **Projetos e relatórios de pesquisa em administração**. Rio de Janeiro: Atlas, 2009.

VICKERY, B. C. **Faceted classification**: a guide to construction and use of special schemes. London: Aslib, 1960.

WINE.com. 2008. Disponível em: http://www.wine.com. Acesso em: 20 dez. 2008.

HUI, Y. : CALLAN, J. Near-duplicate detection by instance level constrained clustering. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 29., 2006, Seattle. **Proceedings …** New York, ACM, 2006. p. 421–428.

YEE, K. P. et al. Faceted metadata for image search and browsing. In: SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS. 2003, Ft. Lauderdale. **Proceedings …** New York, ACM, p. 401–408.

YIN, R. K. **Estudo de caso: planejamento e métodos**. Porto Alegre: Bookman, 2005.

_____.**Applications of case study research**. Newbury Park: Sage, 1994.

ZAMIR, O. ; ETZIONI, O. Grouper: A dynamic clustering interface to web search results. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 8., 1999. Toronto **Proceedings …** New York: Elsevier North-Holland, 1999. p. 1361-1374.

ZHANG, J. **Visualization for information retrieval**. Berlin: Springer. 2007.

ZHANG, J. ; MARCHIONINI, G. Evaluation and evolution of a browse and search interface: relation browser++,. In: 2005. NATIONAL CONFERENCE DIGITAL GOVERNMENT RESEARCH, 2005, Atlanta. **Proceedings …** Chapel Hill: Digital Government Society of North America, 2005.p. 179-188.

ZHAO, Y. ; KARYPIS, G. Evaluation of hierarchical clustering algorithms for document datasets. In: INTERNATIONAL CONFERENCEOF ON INFORMATION AND KNOWLEDGE MANAGEMENT, 11., 2002, McLean. **Proceedings …** New York: ACM, 2002. p. 515-524.

ANEXOS

ANEXO A

ROTEIRO DO EXPERIMENTO PARA O PESQUISADOR

1. Introdução aos recursos de interface da ferramenta.
2. Demonstração de uso com um exemplo de tarefa.
3. Tempo para adaptação do participante com a interface.
4. Início da captura de tela do computador a ser utilizado.
5. Execução das atividades (entrega do formulário com as tarefas).
6. Repetir os passos acima para cada tipo de ferramenta/método.
7. Entrevista com participante.

Observação:

- Permitir o abandono de uma atividade caso o participante considere a sua execução muito difícil e/ou muito demorada.

Duração da sessão:

Início do experimento: _____
Término do experimento: _____

Medição dos tempos:

Ferramenta	Tarefa 1	Tarefa 2
Busca		
Clustering Flat		
Clustering Hierárquico		
Navegação		
Nav. Facetada		

Fim do experimento:

1. Limpar *cache* do histórico de navegação para evitar que futuros experimentos utilizem as consultas já formuladas anteriormente por outro participante.

2. Remover vestígios de uso também das ferramentas de clustering.

ANEXO B

ROTEIRO DE TAREFAS

- Realizar as seguintes tarefas utilizando a ferramenta de **busca direta**:
  - Quais filmes de suspense ("*mystery*") foram dirigidos por POLANSKI, Roman?
  - Encontre filmes disponíveis na base de dados sobre homossexualismo (*homosexuality*).

- Realizar as seguintes tarefas utilizando a ferramenta de **clustering flat**:
  - Quais comédias ("*comedy*") foram dirigidas por COPPOLA, Francis Ford?
  - Encontre filmes disponíveis na base de dados sobre furacões (*hurricane*).

- Realizar as seguintes tarefas utilizando a ferramenta de **clustering hierárquico**:
  - Quais filmes de suspense ("*thriller*") dirigidos por KUBRICK, Stanley?
  - Encontre filmes disponíveis na base de dados sobre a Máfia Italiana (*italian mafia*).

- Realizar as seguintes tarefas utilizando a ferramenta de **navegação**:
  - Quais biografias ("*biography*") foram dirigidas por SCORSESE, Martin?
  - Encontre filmes disponíveis na base de dados sobre o Nazismo (*nazism*).

- Realizar as seguintes tarefas utilizando a ferramenta de **navegação facetada**:
  - Quais filmes de suspense ("*mystery*") foram dirigidos por ALLEN, Woody ?
  - Encontre filmes disponíveis na base de dados sobre a NASA e/ou suas missões (*missions*).

ANEXO C

CORRELAÇÃO DOS ITENS QUANTITATIVOS COM OS QUALITATIVOS

Itens	Numeração do item no questionário da entrevista
R-Precision	2.1 e 2.2
Esforço (Tempo utilizado)	2.3
Taxa de Erro	2.4

ANEXO D

ROTEIRO DA ENTREVISTA COM USUÁRIO

Por favor, informe os dados abaixo:

Profissão: _____

Escolaridade: _____

Sexo: Masculino____ Feminino ____

Idade: _____

Com que freqüência usa sistemas de navegação/busca?

___ Diária ___ Semanal ___ Quinzenal ___ Mensal

Conhecimento do idioma inglês:

___ Pouco ___ Regular ___ Avançado ___ Fluente

1. Quanto à usabilidade das diferentes ferramentas:
   1.1. Qual método lhe pareceu mais agradável para o uso? Por quê?
   1.2. Qual método apresentou maior dificuldade de utilização? Por quê?
   1.3. Que método ajudou a conhecer melhor os tipos de informação disponíveis no banco de dados?
   1.4. Se todos os cinco métodos estivessem disponíveis, qual você escolheria para uso? Por quê?

2. Quanto ao resultado dos diferentes métodos:
   2.1. Sendo precisão a quantidade dos documentos recuperados que é relevante, indique o método mais preciso e o menos preciso, em sua opinião.
   2.2. Nos resultados exibidos, você esperava encontrar alguma informação que não estava presente? Se sim, em qual(is) dos cinco métodos?
   2.3. Qual método fez você convergir mais rapidamente para o resultado? E qual fez você gastar mais tempo?
   2.4. Qual método lhe proporcionou maior certeza nas respostas? E qual lhe deixou mais inseguro quanto às respostas?

3. Espaço livre para comentários.