Uma Proposta para Estruturação e Visualização Semântica de Resultados de Busca Exploratória

Lucas Pupulin Nanni
Departamento de Informática
Universidade Estadual de Maringá
Av. Colombo, 5790, zona 07, Maringá – PR, 87020-900
lucasnanni@gmail.com

Sérgio Roberto P. da Silva
Departamento de Informática
Universidade Estadual de Maringá
Av. Colombo, 5790, zona 07, Maringá – PR, 87020-900
+55 44 3011 4076
sergio.r.dasilva@gmail.com

RESUMO

A tarefa de identificar informações relevantes sobre um determinado assunto na *Web* tem sido dificultada devido ao grande volume de informação disponível atualmente, tornando-se particularmente mais complexa à medida que o assunto em questão não é de total conhecimento do usuário. O fato dos mecanismos de busca atuais não fornecerem uma organização semântica dos resultados recuperados acarreta dificuldades no entendimento e julgamento dos mesmos quando o usuário realiza buscas exploratórias. Desta forma, este trabalho tem como objetivo obter uma proposta para estruturação e visualização de resultados de busca exploratória baseada em grafos que leve em conta o agrupamento semântico dos resultados recuperados em mecanismos convencionais de busca, promovendo, assim, a facilitação da visualização e identificação dos resultados relevantes aos usuários.

Palavras-chave

Visualização, clusterização, busca exploratória.

1. INTRODUÇÃO

O grande volume de informação atualmente disponível para os usuários da *Web* tem dificultado a tarefa de identificar informações relevantes sobre um determinado assunto. Esta tarefa torna-se particularmente mais complexa à medida que o assunto em questão não é de total conhecimento do usuário, ou seja, quando ele está realizando uma busca exploratória no início de um projeto de pesquisa ou trabalho de graduação, por exemplo. Um problema enfrentado pelos usuários quando realizam buscas exploratórias nos mecanismos de busca atuais é decorrente do fato destes apresentarem seus resultados em forma de lista, empregando muitas vezes uma abordagem puramente léxica, não considerando a ordenação ou agrupamento semântico dos resultados. Esta desconsideração da semântica é particularmente problemática na busca exploratória, pois os usuários não têm conhecimento do jargão léxico mais adequado para a realização da busca, o que é essencial para se escolher as palavras-chave adequadas para a criação de uma *query* de qualidade, isto é, uma *query* que retorne documentos de alta relevância para o usuário. Em geral, a criação de uma *query* de qualidade vai além da escolha das palavras-chave, pois a forma como estas palavras são combinadas na *query* resulta em um número muito variável de resultados ordenados de forma diferenciada.

Um dos aspectos considerados importantes nos mecanismos de busca exploratória é a sua interação com os usuários [4], afetando a forma com que os mesmos constroem suas *queries* e avaliam os resultados da busca [1]. Uma das grandes apostas em técnicas de visualização da informação resultante de buscas é a utilização de grafos [2], já que os mesmos facilitam o agrupamento dos resultados por meio de elementos semânticos, auxiliando o usuário a relacionar e determinar a relevância dos documentos.

O objetivo deste trabalho é obter uma proposta para estruturação e visualização semântica de resultados de busca exploratória baseada em grafos, considerando o agrupamento dos resultados recuperados a partir de um sistema de buscas convencional, a fim de sugerir um protótipo de interface que possa ser avaliado em termos da facilitação da visualização e identificação dos resultados relevantes aos usuários.

Este artigo está organizado da seguinte forma. Na Seção 2 discutimos o problema de busca exploratória. Na Seção 3 falamos sobre o problema da representação visual. Na Seção 4 discutimos nossa proposta de visualização. Finalmente na Seção 5 apresentamos nossas conclusões e trabalhos futuros.

2. A BUSCA EXPLORATÓRIA

Segundo Marchionini [4], existem três tipos de atividades realizadas na busca pela informação. A primeira delas, conhecida como busca *lookup*, é a mais básica delas e tem sido o foco dos sistemas gerenciadores de banco de dados e dos motores de busca na *Web* atuais. Geralmente, as buscas *lookup* são adequadas em estratégias de busca analítica, na qual *queries* bem especificadas geram resultados precisos, sem a necessidade de verificação e comparação dos itens recuperados.

Entretanto, a *Web* tem se tornado fonte primária para aquisição de conhecimento exigindo que os sistemas de busca *lookup* sejam superados. Aliadas ao crescimento do volume de informação disponível surgiram duas outras atividades de busca, desta vez, focadas no aprendizado do usuário. A busca por aprendizagem e por investigação, como são conhecidas, envolvem diversas iterações de busca, requerendo esforço cognitivo desempenhado pelo usuário em análises, comparações e julgamento dos documentos recuperados. Esta atividade de busca e refinamento desenvolvida pelos usuários que se engajam em satisfazer sua necessidade de informação tem sido cunhada como "Busca Exploratória", a qual é caracterizada pela incerteza sobre o contexto da pesquisa e pela própria natureza do problema [4], [8].

O crescimento do interesse pela informação, aliada a grande disponibilidade da mesma, sugere que os problemas associados com a busca exploratória e seus usuários devam ser tratados com maior atenção. Deste modo, surgem ferramentas que, associadas à busca exploratória, visam auxiliar o usuário no entendimento do contexto da sua pesquisa, bem como na aquisição de conhecimento e habilidades. Ultimamente, a utilização de artefatos visuais e interfaces diferenciadas têm se tornado grandes apostas de apoio à busca exploratória [1], [3], [2].

3. A REPRESENTAÇÃO VISUAL

Inúmeras técnicas de visualização vêm sendo estudadas para propor melhorias na organização e apresentação dos resultados de buscas na *Web*. Quando aplicadas no contexto da busca exploratória, estas técnicas são ainda mais valorizadas por auxiliarem o usuário a identificar e relacionar os resultados, permitindo que o processo de compreensão do domínio da busca seja melhorado. Em seu trabalho, Sallaberry *et al*. [8] apresentam um sistema de visualização interativa para análise de conteúdo de resultados de busca. O sistema combina diversos algoritmos para apresentar um leiaute que auxilia os usuários a navegarem através de uma coleção de páginas *Web*.

Para este trabalho foi estudada e aplicada uma técnica de visualização por grafos que, como visto, é considerada uma das grandes apostas para exibição de resultados de buscas. A fim de gerar uma visualização de resultados que auxilie o usuário de busca exploratória a compreender o domínio de interesse, é importante estruturá-la de forma que forneça suporte natural a um componente visual. O problema do agrupamento dos resultados em núcleos (vértices) de acordo com sua semelhança pode ser visto como um problema de clusterização, que geralmente emprega aprendizagem de máquina não supervisionada sobre modelos espaciais para criar grupos (*clusters*) e decidir seus elementos (pontos) pertinentes. Segundo Zamir e Etzioni [7], este agrupamento auxilia os usuários a localizar documentos de interesse e obter uma visão geral do domínio dos documentos recuperados.

Em geral, as técnicas de clusterização de resultados se baseiam em duas abordagens. A primeira, e mais comum, consiste em aplicar um algoritmo clássico de clusterização sobre um conjunto de documentos e então "etiquetar" cada *cluster* gerado com o auxílio de técnicas de identificação de tópicos. Já a segunda utiliza o conceito inverso, descobrindo primeiramente termos representativos e, então, realizando a clusterização a partir deles. Algoritmos como *STC* [10], *Vivisimo* (*http://vivisimo.com*), SHOC [11] e *Lingo*© [6] aplicam esta abordagem utilizando diferentes técnicas de descoberta de termos e de agrupamento. Essa abordagem, empregada neste trabalho, se mostra mais promissora devido à criação dos *clusters* ser restringida aos termos representativos eleitos, gerando desta forma "*clusters* bem-descritos.

4. UMA PROPOSTA DE VISUALIZAÇÃO

O conceito base de nossa proposta de visualização é fornecer uma organização visual inicial dos resultados recuperados a partir de uma consulta realizada em um mecanismo convencional de busca. Para este trabalho, o mecanismo de busca considerado foi o *Google*®, devido a sua popularidade e qualidade reconhecidas. Entretanto, a proposta não é alterada ao ser considerado outro mecanismo, permitindo sua aplicação nas diversas ferramentas de busca disponíveis atualmente.

A organização visual fornecida é considerada inicial, pois sustentará o ponto de partida para uma visualização de resultados estendida que será abordada por um projeto futuro, no qual o usuário poderá externalizar seu modelo conceitual do domínio de busca ao adicionar, alterar e remover os elementos que estruturam a visualização gerada. Junto a esse conceito, outro que também poderá ser abordado é a aplicação da visualização de resultados estendida junto à visualização de histórico de busca, criando um ambiente visual ainda mais enriquecido. O projeto responsável pela visualização de histórico de buscas já está sendo desenvolvido pelo Grupo de Sistemas Inteligentes Iterativos (GSII) (*http://www.din.uem.br/gsii*) e possuirá implementação pareada com a proposta de extensão deste trabalho.

A construção do protótipo de visualização contou com a utilização do *framework* de clusterização *Carrot*^{2} (*http://project.carrot2.org*), da biblioteca de manipulação de documentos *D3.js* (*http://d3js.org*) e da *Google*® *Custom Search API* (*http://developers.google.com/custom-search*). Também foram criados *scripts*, utilizando a linguagem *Python* (*http://www.python.org*) para o processamento e redirecionamento do fluxo de dados entre as ferramentas utilizadas. O processamento dos resultados exigiu a utilização da interface à ontologia léxica *WordNet*© (*http://wordnet.princeton.edu*) fornecida pela ferramenta de processamento de linguagem natural *NLTK* (*http://nltk.org*). As ferramentas apresentadas estão disponibilizadas de forma gratuita e são mantidas atualizadas pelos seus desenvolvedores, o que favoreceu a escolha das mesmas.

A fim de esclarecer as atividades envolvidas no processo de visualização dos resultados, foi sugerida uma arquitetura geral cujo diagrama é ilustrado pela Figura 1.

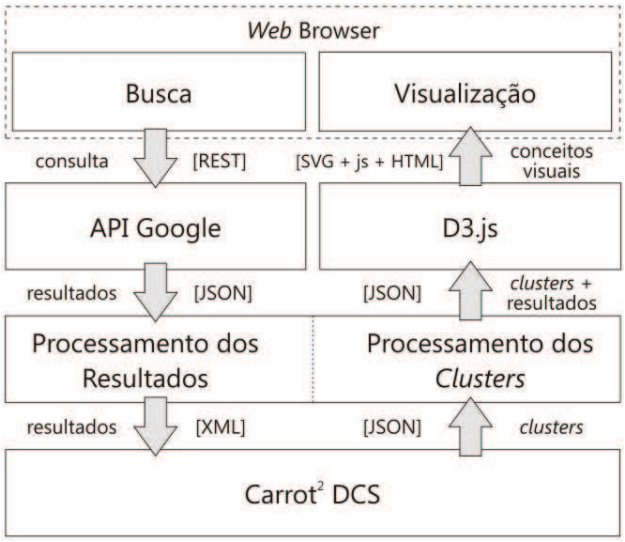

Figura 1 – Arquitetura da visualização proposta.

Por meio do diagrama apresentado é possível identificar que a busca do usuário é redirecionada à *API* de busca do *Google*® retornando um documento no formato *JSON* dos resultados recuperados. O documento retornado é processado e reestruturado a fim de ser submetido ao *Document Clustering Server* (*DCS*), um servidor de clusterização previamente configurado com o algoritmo *Lingo*© e disponibilizado pelo *framework Carrot*^{2}. Os *clusters* criados pelo *DCS* são registrados em um documento *JSON*, o qual especifica os resultados e anotações pertinentes a cada *cluster*. Por fim, o documento contendo o registro dos *clusters* é reestruturado para se adequar as especificações visuais estabelecidas. Com auxilio da biblioteca *D3.js* a visualização final dos resultados é criada junto a uma página *HTML* que será exibida ao usuário como resposta à busca realizada. As setas que envolvem as transações dos dados representam os *scripts* criados para o redirecionamento das entradas e saídas dos componentes utilizados.

4.1 Recuperação e processamento dos resultados

Atualmente existem duas técnicas principais para a recuperação dos resultados de busca na *Web*. A técnica mais primitiva é o emprego de *scraping* sobre as páginas *HTML* retornadas pelo mecanismo de busca, extraindo os resultados das mesmas. A outra técnica, utilizada neste trabalho, consiste em acessar diretamente os resultados por meio de *APIs* oferecidas pelos mecanismos de busca, como a *Custom Search API* do *Google*®, permitindo recuperar a estrutura completa dos resultados.

Após a recuperação dos resultados, estes são processados a fim de normalizá-los semanticamente e reestruturá-los de forma que possam ser clusterizados. A normalização semântica consistiu em inferir o sentido dos termos presentes no título e no resumo de cada resultado e, então, substituí-los pelo lema do sentido associado. O processo de inferir o sentido de um termo envolve buscá-lo em uma ontologia léxica, recuperar um conjunto de conceitos associados a ele e então aplicar alguma técnica de desambiguação de sentidos para determinar o conceito que deve ser associado ao termo. Para a construção e validação do protótipo foram utilizadas a ontologia léxica *WordNet*© e a técnica de desambiguação proposta por McCarthy *et al*. [5]. Com a substituição dos termos pelos lemas correspondentes, espera-se que termos lexicamente distintos, mas semanticamente similares, sejam fundidos e representados por um único termo, auxiliando o processo seguinte a clusterizar os resultados de forma mais concisa.

4.2 Configuração do serviço de clusterização

A escolha e configuração do algoritmo de clusterização foram baseadas nas características discutidas pelas técnicas de clusterização apresentadas. Dentre elas, foi definido que os *clusters* gerados deveriam estabelecer relações entre si, permitindo assim, serem associados visualmente. Além disso, os *clusters* deveriam ser anotados com termos significativos ao usuário, permitindo sua fácil identificação.

Este trabalho considerou a utilização do algoritmo *Lingo*© uma vez que o mesmo provê *clusters* com documentos compartilhados, permitindo estabelecer uma relação simples entre os grupos de documentos. Outro fator que corroborou para a escolha do *Lingo*© foi a possibilidade de configurá-lo de forma que os *clusters* pudessem ser anotados com termos compostos, ou frases, acrescendo na qualidade de descrição dos mesmos. Novamente, a escolha de outras técnicas de clusterização não é restringida pela proposta discutida, cabendo apenas a validação da mesma em relação às características exigidas para a construção da visualização.

Para que o *Lingo*© se adequasse às exigências estabelecidas, o mesmo sofreu alguns ajustes, possibilitando o aperfeiçoamento dos *clusters* gerados. Tais ajustes foram realizados com auxílio do *benchmark* disponibilizado pelo *framework Carrot*[2] e consistiu na modificação de alguns parâmetros como o *"Cluster count base"*, o qual permite ao algoritmo estimar o número aproximado de *clusters* que devem ser gerados. Neste caso, foi determinado que deveriam ser gerados em torno de 9 *clusters*, por este ser um número razoável de elementos passíveis de serem identificados com clareza pelo usuário. Outros parâmetros como *"Phrase Label Boost"* (9) e *"Phrase length penalty-start/stop"* (5) privilegiaram o uso de frases, ou palavras compostas, para a anotação dos *clusters*, fornecendo uma descrição mais aprimorada dos mesmos.

O recurso léxico de *stop words* utilizado pelo algoritmo também foi modificado com a inserção e remoção de regras léxicas, permitindo ao *Lingo*© processar documentos com termos mais abrangentes como *"information"*, e ao mesmo tempo ignorar frases pré-definidas, como *"Wikipedia, the free encyclopedia"*.

A configuração dos parâmetros do algoritmo visou, principalmente, reduzir o número de *clusters* gerados e priorizar a utilização de termos compostos para a anotação dos *clusters*, demandando menos esforço cognitivo por parte do usuário na identificação dos mesmos quando estruturados na visualização proposta.

4.3 Construção do protótipo de visualização inicial

Para a criação da visualização, primeiramente os *clusters* são recuperados fornecendo ao DCS os resultados já processados. Um documento *JSON* com o registro do conjunto de *clusters* é então gerado, no qual um *cluster C* pode ser definido pela tripla (*l*, *R*, *w*), em que *l* é a anotação gerada para o *cluster*, *R* é o conjunto de resultados contidos em *C*, e *w* é a pontuação associada ao *cluster*. O documento que especifica os *clusters* gerados é então processado para que componha um documento final que possa ser utilizado na produção da visualização. O processamento consiste em criar um documento *JSON* estendido que identifique as relações entre os *clusters* e que compreenda a estrutura original dos resultados, como metadados, formatações *HTML* e miniaturas. O relacionamento entre os *clusters* é estabelecido a partir da pertinência mútua dos resultados contidos neles. Um *cluster* C_i está relacionado a outro *cluster* C_j se existe um resultado *r* que pertença tanto ao *cluster* C_i quanto ao *cluster* C_j.

É requerido da representação visual que ela apresente um conceito visual simples e interativo, ao mesmo passo que permita ao usuário identificar de forma clara o significado dos conceitos aplicados. Para tanto, o protótipo de visualização foi construído utilizando o conceito de grafos não direcionados, no qual os vértices da estrutura consistem dos *clusters* gerados e as arestas representam as relações obtidas entre os *clusters*. Além do grafo, um painel lateral complementa a visualização com a listagem dos resultados contidos em um *cluster* selecionado pelo usuário.

Os vértices foram concebidos como círculos dimensionados de acordo com a pontuação do *cluster* associado e, então, coloridos aleatoriamente a partir de uma palheta de 15 cores distintas. Já as arestas, representando as relações entre os *clusters*, foram idealizadas como linhas retas que conectam os centros dos *clusters* relacionados. Para o desenho e interação da visualização foi utilizada a biblioteca *D3.js*, a qual permitiu, além de estruturar de forma adequada os componentes da visualização, manipular a estrutura da página *HTML* que a comportava. A Figura 2 ilustra a visualização gerada para a busca "*information retrieval"*.

A interação provida pela visualização permite ao usuário focalizar um grupo de *clusters* relacionados, bastando para isso selecionar qualquer *cluster* pertencente ao grupo. Uma vez focalizado um grupo de *clusters*, um *cluster* pertencente a ele pode ser analisado ao também ser selecionado. O foco é desfeito ao selecionar um *cluster* que não pertença ao grupo já focalizado, retornando a visualização ao seu estado inicial. O termo seleção foi empregado de forma genérica, não restringido o meio pelo qual se realiza a seleção, como um clique ou um toque. A Figura 3 ilustra a

situação em que o cluster *"Information Retrieval and Web"* foi selecionado e um grupo de 6 *clusters* relacionados foi focalizado.

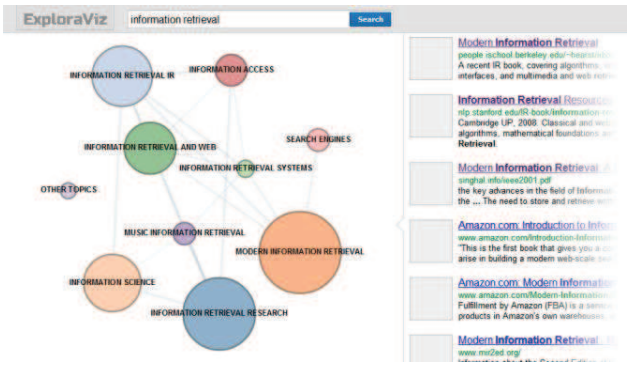

Figura 2 – Visualização gerada para a busca "*information retrieval*".

A visualização permite que os *clusters* sejam reposicionados espacialmente, fornecendo ao usuário a liberdade de organizar a informação apresentada. O posicionamento inicial dos *clusters* é automaticamente realizado pelo algoritmo de disposição de grafos dirigido por forças de atração e repulsão.

O painel de resultados lateral é inicialmente posicionado de forma recolhida, permitindo que o grafo ocupe a área central da visualização e forneça espaço suficiente para complementos visuais futuros, como expansão da proposta inicial. A utilização de miniaturas sugere, ainda que de forma parcial, a organização visual dos documentos recuperados pela busca, complementado a concepção do resultado textual.

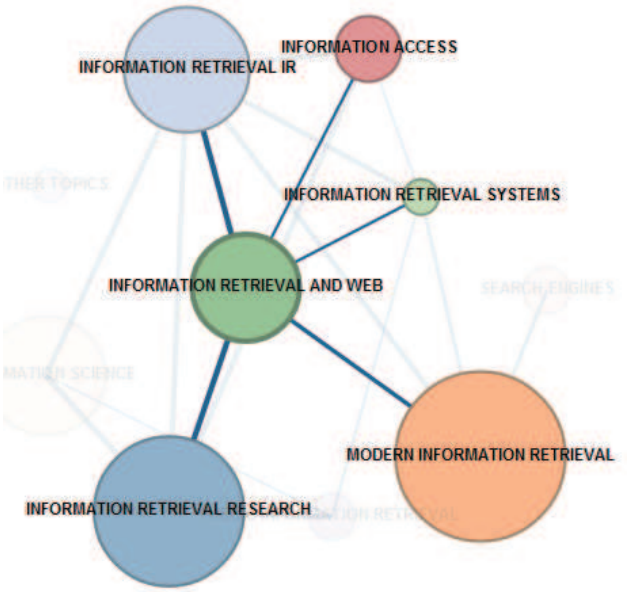

Figura 3 – Grupo de clusters focalizado após seleção do cluster *"Information Retrieval and Web"*.

5. RESULTADOS E CONCLUSÕES

A partir de testes iniciais realizados com o protótipo construído, foi verificado que o processo de clusterização se comportou de forma esperada, gerando um número adequado de *clusters* bem anotados. Acredita-se que o número reduzido de *clusters*, entre 8 e 10, permitirá ao usuário interpretar a visualização gerada com maior facilidade.

Em relação à visualização propriamente dita, esta se mostrou uma proposta viável para que futuramente novos atributos visuais e funcionalidades sejam adicionados a ela. Os resultados obtidos ainda são preliminares e necessitam ser complementados com avaliações qualitativas e quantitativas da proposta visual.

Como discutido anteriormente, este projeto condiciona trabalhos futuros a abordarem a extensão da visualização proposta a fim de permitir ao usuário modificar os elementos visuais apresentados, possibilitando, assim, a externalização de seu conceito de busca. Também poderá ser abordada, futuramente, a integração deste trabalho com a proposta de visualização de histórico de busca, a qual está sendo desenvolvida pelo GSII e estabelecerá um ambiente enriquecido de busca visual.

6. REFERÊNCIAS

[1] Alhenshiri, A. et al. 2010. Augmenting the visual presentation of Web search results. *2010 Fifth International Conference on Digital Information Management (ICDIM)* (Thunder Bay, Jul. 2010), 101–107.

[2] Angelaccio, M. et al. 2007. Graph Use to Visualize Web Search Results: MyWish 3.0. *Information Visualization 2007 IV 07 11th International Conference* (Roma, 2007), 245–250.

[3] Kajinmi, T. et al. 2008. Application of keyword map to decision support through exploratory search. *2008 IEEE International Conference on Systems, Man and Cybernetics* (Oct. 2008), 2177–2181.

[4] Marchionini, G. 2006. Exploratory search. *Communications of the ACM*. 49, 4 (Apr. 2006), 41.

[5] McCarthy, D. et al. 2007. Unsupervised Acquisition of Predominant Word Senses. *Computational Linguistics*. 33, 4 (Dec. 2007), 553–590.

[6] Osinsk, S. 2003. *An Algorithm for Clustering of Web Search Results*. Poznan University of Tecnology.

[7] Sallaberry, A. et al. 2010. Interactive visualization and navigation of web search results revealing community structures and bridges. *Proceedings of Graphics*. (2010), 105–112.

[8] White, R.W. et al. 2005. Exploratory Search Interfaces : Categorization , Clustering and Beyond. *SIGIR Forum*. 39, 2 (2005), 52–56.

[9] Zamir, O. and Etzioni, O. 1999. Grouper : A Dynamic Clustering Interface to Web Search Results. *Work*. 31, 11-16 (1999), 1361–1374.

[10] Zamir, O. and Etzioni, O. 1998. Web document clustering. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98* (New York, New York, USA, 1998), 46–54.

[11] Zhang, D. and Dong, Y. 2004. Semantic, hierarchical, online clustering of web search results. *Advanced Web Technologies and Applications*. 3007, (2004), 69–78.