Interface para Busca Semântica: Um Estudo sobre o Estado da Arte

Paulo R. Nietto, Rodrigo Costa Camargos

Programa de Mestrado em Ciências da Computação – Faculdade de Campo Limpo Paulista (Faccamp) – Campo Limpo Paulista, SP – Brasil

pnietto@gmail.com, rodrigocamargos@hotmail.com

Abstract. *Conventional searching mechanisms can present results that oppose to the meaning of sentences expressed on queries. In order to solve this divergence between the query and the result, interfaces are being implemented for semantic search. This paper focus on presenting, analyzing and discussing recent researches about Semantic Search Interfaces. In order to execute this research, scientific papers published in specifics journals and conferences are used as foundations for this paper.*

Resumo. *Os mecanismos de busca convencionais podem apresentar resultados que não condizem com o significado da sentença expressada na consulta. Para solucionar essa divergência entre a consulta e o resultado, estão sendo implementadas interfaces para a busca semântica. Este artigo tem o objetivo de apresentar, analisar e discutir as recentes pesquisas sobre o tema de Interface para Busca Semântica. Para a realização desse artigo são utilizados como base, artigos científicos publicados em revistas e conferencias.*

1. Introdução

As consultas nos mecanismos de busca convencionais podem retornar resultados que não condizem com o significado ou intenção da pesquisa, uma vez que mais de uma palavra-chave pode fazer parte da mesma sentença, mas com semântica diferente. A busca semântica tenta solucionar esse problema, disponibilizando um conteúdo mais ajustado e possibilitando buscas mais complexas.

A busca semântica é a forma de se consultar objetos em uma base de conhecimento e obter como resultado vários atributos e conceitos relacionados que são interligados. Essa técnica permite explorar melhor o significado e intenção do que está sendo consultado.

A ontologia é uma técnica de organização de informações e representação de conceitos dentro de um domínio, que tem como propósito permitir o compartilhamento e reutilização do conhecimento. A estrutura da ontologia é baseada na descrição dos conceitos e relacionamentos semânticos entre eles.

As interfaces de busca semântica são construídas utilizando diversas técnicas, dentre elas estão as que implementam linguagem natural, busca em grafos, uma linguagem formal própria ou uma linguagem formal já estabelecida, como a OWL ou SPARQL.

Este artigo apresenta algumas das recentes pesquisas sobre as interfaces de busca semântica, apresentando seus objetivos e resultados. Também são discutidas as limitações das pesquisas que podem ser futuramente exploradas.

Para a realização desse artigo foram pesquisados diversos artigos sobre interface de busca semântica em revistas científicas e conferências, considerando as diversas áreas de pesquisas que são implementadas nas interfaces, a contribuição do artigo e ano de publicação.

O restante do artigo está organizado da seguinte forma: A metodologia utilizada para a realização desse artigo é descrita na Seção 2. Na Seção 3 são apresentados algumas das recentes pesquisas sobre interface para busca semântica. Na Seção 4 são discutidos os objetivos, resultados e limitações das pesquisas apresentadas.

2. Metodologia

Para a realização desse artigo, foram pesquisados 22 artigos como limite máximo. O critério de escolha foi que o ano das publicações deveriam estar entre 2009 e 2014, sendo esses artigos publicados em universidades ou em revistas e conferências dos portais IEEE (2015), ACM (2015), Springer (2015) e ScienceDirect (2015). Os 4 portais foram selecionados por apresentarem grande parte das mais recentes pesquisas sobre o assunto de interface de busca semântica. As palavras-chave utilizadas foram *semantic search*, *semantic web*, *ontology* e *semantic interface*.

Dos 22 artigos pesquisados, 8 foram selecionados utilizando como critério primeiramente as métricas utilizadas nas pesquisas, em segundo lugar a variação de assunto, para que não sejam todos os artigos sobre o mesmo foco de pesquisa em busca semântica (ex. todos artigos sobre busca semântica utilizando busca em linguagem natural), e por último a data de publicação. Os detalhes dos artigos selecionados são descritos na próxima seção.

3. Visão Geral dos Artigos Recentes

Nessa seção são apresentados os objetivos, métodos utilizados para a realização das pesquisas, resultados e limitações de 8 artigos que pesquisam sobre as diversas áreas dos tipos de interface para busca semântica. Os resultados sobre cada objetivo e as limitações encontradas em cada artigo possuem foco principalmente na usabilidade, interface amigável ao usuário e experiência obtida pelos usuários.

O artigo de Price et al. (2009) realiza uma comparação entre um sistema que utiliza indexação de palavras-chave e textos inteiros, chamado de sistema 1, com um sistema que utiliza componentes semânticos junto com a forma de indexação existente, chamado de sistema 2. O estudo envolveu usuários e documentos do portal nacional de saúde dinamarquês (sundhed.dk). Essa pesquisa tem como objetivo verificar se componentes semânticos melhoram a precisão das buscas na perspectiva do sistema e se na perspectiva dos usuários, a utilização de componentes semânticos em um sistema de busca resulta em uma experiência mais agradável.

Para a realização da pesquisa, Price et al. (2009), utilizou 24712 documentos copiados do portal sundhed.dk, o mecanismo de busca Ultraseek na versão 5.6 (o mesmo mecanismo de busca do portal sundhed.dk), tanto para sistema 1 quanto para o

sistema 2. Também foram convidados 30 médicos que utilizam constantemente o portal sundhed.dk. A interface de busca para o sistema 1, utiliza campos semelhantes aos utilizados no portal sundhed.dk, enquanto a interface utilizada no sistema 2 utiliza todos os campos do sistema 1 junto a campos específicos para a busca usando componentes semânticos. Para o sistema 2, não é necessária a utilização dos campos específicos para a busca semântica, pois nos resultados estão inclusas informações sobre os componentes semânticos independente do fato de utilizá-los. Para a classificação dos componentes semânticos, foram consultados funcionários da sundhed.dk envolvidos no processo de indexação que mantém contato constante com o usuário. Foram desenvolvidos 4 cenários baseados na área médica para que os usuários façam buscas de documentos.

Os resultados obtidos no artigo de Price et al. (2009) foram que de 60 consultas realizadas em cada sistema, o sistema 1 resultou em 45 buscas bem-sucedidas enquanto o sistema 2 resultou em 48 buscas bem-sucedidas na opinião dos usuários. A média de tempo gasto na busca de documentos utilizando o sistema 2 foi de 1 minuto e 22 segundos a mais que no sistema 1. Sendo 1 = muito satisfeito; e 5 = muito insatisfeito, a média da satisfação dos usuários em relação aos resultados das buscas foi de 2.3 no sistema 2 e 2.1 no sistema 1. A média da facilidade de expressar a intenção da pesquisa no sistema 2 foi de 2.1 e 2.0 no sistema 1. As limitações encontradas na pesquisam foram que nem todos usuários utilizaram os campos específicos para a utilização dos componentes semânticos, os usuários que participaram do experimento já conheciam o sistema 1 e são de uma mesma categoria (médicos), dificultando a medida precisa de tempo gasto e a variação de opiniões.

Outro artigo selecionado é o de Lee et al. (2012), que propõe um método de pesquisa semântica que se baseia em recuperar todos os conceitos que estão relacionados a determinada palavra-chave mesmo que a palavra-chave não apareça no documento, ou seja, localiza resultados relevantes que são relacionados semanticamente à consulta efetuada pelo usuário. Os resultados das buscas são ordenados pelos pesos das propriedades e recursos que são baseados em especificidade e generalidade. Para que seja possível a busca entre os conceitos relacionados é utilizada uma ontologia que produz as descrições dos conceitos e seus relacionamentos.

O método utilizado por Lee et al. (2012), para a realização de sua pesquisa foi a construção de uma ontologia de domínio de eletrônicos, que utilizou as informações que foram publicadas em diferentes fontes de dados coreanos, extraindo conceitos como título, autor, palavras-chave, URL e data de publicação. Foram aplicados e normalizados os pesos de generalidade e especificidade na base de conhecimento utilizando a linguagem OWL (*Web Ontology Language*) e o algoritmo de busca utilizado foi o de ativação e propagação (*spreading activation algorithm*). Para a redução do tempo de pesquisa, os pesos, propriedades e instâncias são calculadas durante o pré-processamento. Para a avaliação do experimento, foi comparado o mecanismo de busca proposto com as metodologias SemRank e RSS. Também foram convidadas 10 pessoas que possuem o título de doutor ou mestre, com especialidade em web semântica ou engenharia de ontologias.

A pesquisa de Lee et al. (2012), apresentou como resultado a média da comparação dos resultados da busca com base na intenção do usuário, os valores de 0.60 para o sistema utilizando o peso de especificidade, 0.67 para o sistema utilizando o peso

de generalidade, 0.65 para o sistema RSS e 0.61 para o sistema SemRank, sendo o valor 0 como muito irrelevante e 1 como muito relevante. As limitações encontradas foram que poucos cenários são utilizados para a comparação entre as metodologias, o domínio utilizado para a base de conhecimento é muito limitado e a quantidade de pessoas no experimento é pequena e específica.

O terceiro artigo selecionado é realizado por Tablan et al. (2014), que apresenta um framework para busca semântica chamado Mímir, criado para tarefas de descoberta de informação, que suporta tanto estruturas de consulta complexa utilizando operadores, como busca por palavra-chave através de tokens. O foco do artigo é apresentar o framework Mímir, testar seu desempenho e avaliar a usabilidade com base nos usuários.

Para a realização de sua pesquisa, Tablan et al. (2014) utilizou o framework Mímir, a linguagem de consulta SPARQL, 9.5 milhões de documentos dos repositórios semânticos da GeoNames e DBpedia, especialistas em imunologia treinados no uso do framework para avaliar as pesquisas semânticas complexas e 23 participantes que participaram do workshop sobre Mímir para avaliar sua usabilidade e interface.

Tablan et al. (2014), obteve como resultado que somente 13 dos 23 usuários concluíram as 4 consultas no tempo estabelecido de 30 minutos. Entre os resultados obtidos da consulta com palavras-chave e busca semântica, os usuários preferiram os resultados apresentados pela busca semântica. Como avaliação da interface de busca semântica, 56.3% dos usuários ficaram satisfeitos com a usabilidade, 87% dos usuários reprovaram a complexidade da interface e 93,75% dos usuários aprenderam a utilizar o sistema em 10 minutos, sem a necessidade de mais tempo de aprendizado. As limitações encontradas são que não foram comparados os resultados com nenhum sistema semelhante (como o Brocolli ou o KIM), a quantidade de consultas é pequena, dificultando a comparação por parte dos usuários entre a busca por palavra-chave com a busca semântica, a quantidade de usuários que testaram o sistema é pequena e não é possível estimar precisamente o tempo que os usuários gastam nas buscas, pois é somente estimado um tempo máximo.

A quarta pesquisa selecionada é realizada por Kaufmann e Bernstein (2010), que apresenta um comparativo da usabilidade sob perspectiva de usuários casuais e ocasionais entre as interfaces de linguagem natural de consulta com uma interface de linguagem de consulta formal. Essa pesquisa investiga o quanto são utilizáveis as linguagens de consulta natural para encontrar dados em bases de conhecimento de Web Semântica.

Para a realização da pesquisa, Kaufmann e Bernstein (2010) usaram as interfaces de consulta Ginseng, NLP-Reduce e a Querix, que utilizam linguagem natural e a Semantic Crystal, que utiliza linguagem formal. Para analisar a usabilidade foram convidadas 48 pessoas de diversas áreas, que formularam 4 questões a serem consultadas em uma base de conhecimento de informações geográficas e responderam ao questionário escala de usabilidade do sistema (SUS – System Usability Scale).

A pesquisa de Kaufmann e Bernstein (2010) obteve o resultado da média do tempo gasto pelos usuários nas consultas, sendo o NLP-Reduce com 2,39 minutos, o Querix com 4,11 minutos, o Ginseng com 6,06 minutos e Semantic Crystal com 9,43 minutos. A quantidade de consultas para encontrar as respostas das 4 questões para o

NLP-Reduce foi de 7.94, para o Querix foi de 7.75, para Ginseng foi de 11.06 e para o Semantic Crystal foi de 7.02. O questionário de escala de usabilidade do sistema (SUS) resultou em 66.67% dos usuários preferiram a interface do Querix e 60,42% dos usuários não gostaram da interface do Semantic Crystal. As limitações encontradas são que somente uma linguagem de consulta formal foi utilizada contra três linguagens de consulta natural, a quantidade dos tipos de consulta e o domínio utilizado impede uma análise mais detalhada de qual ferramenta é melhor para o uso.

O quinto artigo selecionado foi realizado por Hahn e Diaz (2013), onde é apresentada a ferramenta de busca de interface semântica Deneb 1.0, que fornece recomendações de assuntos com base na informação total contida nos metadados desse assunto. Nesse artigo é modelado uma nova interface de apresentação de resultados obtidos das pesquisas, chamado de Deneb 2.0, com base nas preferências dos usuários.

Para a realização da pesquisa, Hahn e Diaz (2013), convidaram 8 usuários estudantes da universidade de Ilinóis para avaliação da interface do resultado das buscas, a ferramenta Deneb na versão 1.0, a base de dados da universidade de Ilinóis e o framework Bootstrap para a criação da interface do Deneb 2.0.

A pesquisa de Hahn e Diaz (2013) obteve como resultado da avaliação dos usuários sob a interface dos resultados das pesquisas utilizando o Deneb 1.0, que não estava clara a diferença entre os títulos e as sugestões dos assuntos nos resultados obtidos, a lista de recomendações exibida nos resultados é útil, mas nem sempre condiz com a intenção da pesquisa, os resultados por título não resultam em uma lista de títulos similar e não possui uma breve descrição sobre do que se trata. A partir das informações qualitativas fornecidas pelos usuários foi construído o Deneb 2.0. Essa pesquisa teve como limitação a pequena quantidade de usuários utilizados no experimento e a falta de outra ferramenta para comparação da interface de resultados e resultados obtidos nas buscas.

O sexto artigo é a pesquisa de Styperek et al. (2014), que apresenta um mecanismo de busca semântica (SSE), que fornece aos usuários um construtor de consultas em grafos. Nesse artigo é realizada uma comparação com outras ferramentas de pesquisa que utilizam linguagem natural, formal ou palavra-chave para descobrir quais apresentam as melhores respostas. A pesquisa também compara os sistemas de busca baseados em grafos para descobrir quais possuem interface para formulação de consulta mais amigável aos usuários.

A pesquisa de Styperek et al. (2014), utilizou na base de conhecimento da DBPedia as ferramentas SSE, GoR e NAGA de busca semântica baseada em grafos e o PowerAqua de linguagem natural. Também foi utilizado o Google utilizando linguagem natural e palavra-chave. Para avaliar qual interface é mais amigável aos usuários, foi utilizado como medida, quanto menor quantidade de elementos para construção da consulta, mais amigável é a interface.

Styperek et al. (2014), obteve como resultado da avaliação de qual interface de formulação de consultas é mais amigável aos usuários, sendo em primeiro lugar a ferramenta SSE, que utilizou a média de 2,786 palavras por consulta, em segundo lugar ficou a ferramenta Naga, com média de 5,714 palavras por consulta e em terceiro lugar a ferramenta GoR, que utilizou uma média de 5,714 palavras com consulta. As limitações

encontradas no artigo foram que nenhuma das interfaces de busca utilizando linguagem natural foi avaliada para verificar se possuem interface de formulação de consultas amigável aos usuários e não utilizaram como critério a opinião de usuários.

A sétima pesquisa, realizada por Dos Reis et al. (2013), efetuou uma comparação entre o mecanismo de busca ISM (*inclusive search mechanism*), com um mecanismo de busca sintático. O estudo envolveu usuários que utilizaram os dois mecanismos e o sistema VilanaRede. O foco do artigo é a produção de mecanismos de busca que possam ser utilizados por pessoas que possuem pouca proficiência com tecnologia.

Para a realização da pesquisa, Dos Reis et al. (2013) utilizou o sistema VilanaRede, o método de análise semântica (SAM), para extração e representação de significados em um modelo semântico, a ferramenta SONAR para a construção de diagramas a partir de outro diagrama, o mecanismo de busca ISM, o mecanismo de busca sintático e foram convidados 25 usuários para realizar buscas nos dois mecanismos, onde 16 usuários já conheciam o sistema VilanaRede e 9 não conheciam. Os usuários foram divididos em 5 grupos e para cada grupo foram elaborados 3 cenários de busca específicos.

Os resultados obtidos por Dos Reis et al. (2013) foram que, segundos usuários, o sistema ISM apresentou a maior quantidade de resultados e mais resultados relevantes. Tanto o sistema ISM quanto o sistema de busca sintático necessitaram de mais de uma tentativa para obter o resultado desejado, porém, o mecanismo ISM obteve uma leve vantagem a partir da segunda tentativa. As limitações encontradas no artigo foi que o domínio e quantidade de cenários para as buscas são pequenos e somente um mecanismo foi comparado ao mecanismo ISM.

O oitavo e último artigo analisado foi realizado por Batzios e Mitkas (2012), que apresenta um protótipo de um mecanismo de busca semântica chamado de WebOWL, que possui dois tipos de interface, uma utilizando consultas OWL e outra possui linguagem formal para a consulta. O foco do WebOWL é identificar e propor soluções dos maiores desafios das tecnologias de busca em dados semânticos. Esses desafios vão desde as técnicas de rastreamento, indexação, classificação e consultas até questões relacionadas aos usuários, como as interfaces de usuário.

Para a realização da pesquisa, Batzios e Mitkas (2012) utilizaram a base de dados db4OWL0, um rastreador que reconhece ontologias disponíveis na Web chamado BioCrawler, o mecanismo WebOWL e consultas de usuário nas classes de animais herbívoros na ontologia pessoas+animais de estimação.

As limitações encontradas no artigo de Batzios e Mitkas (2012) foram que a interface de consulta que utiliza linguagem formal, que é descrita pelos autores como mais amigável ao usuário, não foi testada por usuários comuns e apresenta somente um resultado em cada consulta, enquanto a interface que utiliza consultas OWL pode apresentar diversos resultados. Também não foram comparados outros mecanismos de busca para medir o desempenho do WebOWL.

4. Discussões sobre os Artigos Analisados e Trabalhos Futuros

Os artigos analisados apresentaram diversos resultados e limitações que nessa seção são discutidos levando em consideração a proposta dos mecanismos de busca dos artigos

apresentados. Na tabela 1 são apresentados de forma simplificada uma meta-analise dos objetivos, resultados e limitações das pesquisas analisadas.

Tabela 1. Objetivos, resultados e limitações gerais das pesquisas

Objetivo	Resultados	Limitações
Satisfação dos usuários em relação aos resultados das buscas.	Os resultados sugerem que 60% dos usuários experientes em buscas por palavra-chave mostram menos satisfação às buscas semânticas. 85% dos usuários que utilizam com pouca frequência os mecanismos de busca preferem a busca semântica.	Não foram todos usuários experientes que utilizaram os campos específicos da busca semântica. Utilizar usuários de uma única categoria dificulta a variação opiniões. Os domínios utilizados nas bases de conhecimento são limitados. Poucos cenários são utilizados para comparar os mecanismos de busca. A quantidade de pessoas nos experimentos, de forma geral, é pequena.
Aumentar a precisão da busca utilizando como base a intenção do usuário.	74% de resultados relevantes é obtido utilizando os mecanismos de busca semântica. 38% de resultados relevantes é obtido utilizando os mecanismos de busca sintático.	
Facilidade em expressar a pesquisa.	87% dos usuários experientes em buscas sintáticas mostram mais dificuldade nas buscas semânticas. 96% dos usuários que utilizam menos os mecanismos de busca preferiram a busca semântica.	

As pesquisas realizadas demonstram que cada sistema analisado opera com ontologias diferentes, onde essas ontologias possuem variação de dimensão e complexidade. Os artigos sugerem que no momento da indexação, a dificuldade em classificar as ontologias é similar à classificação de bases de dados que utilizam palavra-chave, no entanto, a quantidade de bases de dados que utilizam ontologia é muito menor do que a quantidade de bases de dados que utilizam palavra-chave.

Para as consultas realizadas em cada pesquisa, é subjetivo afirmar que os resultados apresentados estão corretos ou errados, que a quantidade de resultados apresentados é suficiente ou insuficiente ou que a quantidade de tentativas para se obter o resultado desejado é aceitável. Todos esses fatores estão fortemente relacionados aos conceitos e intenções utilizadas nas pesquisas, consultas e em alguns casos, os resultados apresentados são exibidos com base na proposta do mecanismo de busca.

Os artigos sugerem que alguns dos maiores desafios das interfaces dos mecanismos de busca semântica são mensurar sua usabilidade, realizar os experimentos com um número grande de pessoas e tornar a interface de usuário mais simples e intuitiva. Os resultados obtidos pelos artigos analisados demonstram que a busca semântica é um campo que possui futuro promissor e ainda pode ser muito explorado.

Para trabalhos futuros, serão estudadas novas pesquisas relacionadas a interfaces para busca semântica, para que seja possível futuramente se aprofundar mais ao tema e efetuar uma contribuição científica.

5. Referências

ACM (2015). Disponível em< https://www.acm.org >. Acessado em 17 de abril de 2015.

Batzios, A., and Mitkas, P. a. (2012). "WebOWL: A Semantic Web search engine development experiment". In *Expert Systems with Applications*, *39*, pages 5052–5060. Elsevier.

Dos Reis, J. C., Bonacin, R., and Baranauskas, M. C. C. (2013). "Addressing universal access in social networks: an inclusive search mechanism". *Universal Access in the Information Society*, pages 1–21. Springer-Verlag Berlin Heidelberg.

Hahn, J., and Diaz, C. (2013). "Formative Evaluation of Near-Semantic Search Interfaces". *Internet Reference Services Quarterly*, *18*, pages 175–188. University of Iowa Libraries Staff Publications.

IEEE (2015). Disponível em< https://www.ieee.org/index.html >. Acessado em 16 de abril de 2015.

Kaufmann, E., and Bernstein, A. (2010). "Evaluating the usability of natural language query languages and interfaces to Semantic Web knowledge bases". *Journal of Web Semantics*, *8*, pages 377–393. Elsevier.

Lee, M., Kim, W., and Park, S. (2012). "Searching and ranking method of relevant resources by user intention on the Semantic Web". *Expert Systems with Applications*, *39*, pages 4111–4121. Elsevier.

Price, S. L., Lykke Nielsen, M., Delcambre, L. M. L., Vedsted, P., and Steinhauer, J. (2009). "Using semantic components to search for domain-specific documents: An evaluation from the system perspective and the user perspective". *Information Systems*, *34*, pages 778–806. Elsevier.

ScienceDirect (2015). Disponível em< http://www.sciencedirect.com >. Acessado em 16 de abril de 2015.

Springer (2015). Disponível em< http://www.springer.com >. Acessado em 17 de abril de 2015.

Styperek, A., Ciesielczyk, M., and Szwabe, A. (2014). "Semantic search engine with an intuitive user interface", *23rd international conference on World wide web companion,* pages 383–384. ACM Digital Library.

Tablan, V., Bontcheva, K., Roberts, I., and Cunningham, H. (2014). "Mímir: An open-source semantic search framework for interactive information seeking and discovery". *Web Semantics: Science, Services and Agents on the World Wide Web*, *30*, pages 52–68. Elsevier.