Find Me: uma Ferramenta para Consulta Estruturada a dados *wiki*

Poline Lottin, Carina F. Dorneles

Depto. de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Santa Catarina – Brazil, 88.049-900

{poline,dorneles}@inf.ufsc.br

Resumo. *Sistemas de busca na Web atualmente são voltados a consultas por palavras-chaves que permitem pouca expressividade do usuário. Este artigo apresenta o Find Me, uma ferramenta, que permite consultas estruturadas a dados na Web. A ferramenta utiliza um método que identifica objetos, atributos e seus relacionamentos dentro de uma página HTML e gera uma estrutura de índices adequada para a construção das consultas estruturadas.*

Abstract. *Web search engines currently support keywords search, which do not allow the user express a more expressive query condition. This paper presents Find Me, a tool that allows structured queries to data on the Web. This tool uses an approach that identifies objects, attributes and their relationships within an HTML page and generates an index structure suitable for the construction of structured queries.*

1. Introdução

A popularização da *Web* proporcionou o aumento, cada vez maior, de dados espalhados pelo mundo em meio digital, em sua maioria, descritos em arquivos textuais. Hoje, a forma mais popular de recuperar estes dados é a busca através de palavras-chave. Esta forma de acesso aos dados se mostra pouco expressiva, apresentando problemas como sinonímia (busca por "*idioma*" não retorna páginas com o termo "*língua*") e polissemia ("*São Paulo*" cidade, estado ou clube de futebol?), por exemplo. A necessidade de encontrar informações mais relevantes em milhares de documentos na *Web* estimula pesquisas de técnicas mais eficazes de consultas.

Atualmente o método de busca mais conhecido é o utilizado pelo *Google*. Este mecanismo coleta o maior número possível de recursos (URLs) usando softwares robôs e em seguida analisa cada página coletada a fim de indexá-la através dos termos encontrados nas páginas. Esta estrutura de índice permite aos usuários formularem consultas através de palavras-chave que são comparadas com os termos indexados. No momento da consulta, o sistema procura no índice o termo solicitado e fornece um conjunto de páginas que melhor combinam com o critério. Esta forma de indexação, no entanto, limita o usuário a apenas buscar páginas através de palavras-chaves, obrigando-o a deduzir quais termos devem ser inseridos na consulta para se obter a página desejada como resultado.

Nos mecanismos de consultas atuais, uma consulta do tipo "*cidades que possuem prefeito chamado José*" não possui valor semântico e os documentos

retornados serão aqueles que possuem maior similaridade com os termos da consulta, mesmo que o documento retornado esteja relacionado ao prefeito da cidade de São José, por exemplo. A alternativa de se oferecer mecanismos que possibilitem a construção de consultas mais sofisticadas é de bastante interesse na comunidade científica atualmente [MERGEN et. al., 2008, MERGEN et. al., 2010], visto a necessidade de compreensão de consultas como a citada acima.

Este artigo descreve a ferramenta *Find Me*, que possui basicamente duas contribuições principais: (i) disponibilização de uma interface amigável para o usuário realizar consultas estruturadas sobre documentos HTML, facilitando a consulta até mesmo a usuários que não tenham conhecimento da sintaxe de consultas a banco de dados; e (ii) execução de um processo de indexação das páginas que permite a construção de consultas estruturadas no estilo SQL. Para a construção da estrutura de índice foram definidas algumas regras heurísticas para identificação de objetos, atributos e seus relacionamentos dentro de uma página HTML. A ferramenta foi testada no contexto do Wikipedia, por ser uma fonte rica de informações, que se encontra em constante crescimento, tendo em vista que seu conteúdo é alimentado colaborativamente. Os dados desta wiki seguem alguns guias de estilo, ou seja, alguns padrões definidos pelo site. Um dos padrões da Wikipedia é uma caixa informativa, chamada infobox, que disponibiliza dados com informações de cada artigo publicado.

O artigo está organizado como segue. A Seção 2 descreve a ferramenta *Find Me*, enquanto os experimentos e resultados obtidos são apresentados na Seção 3. Na Seção 4 são apresentados alguns trabalhos relacionados que propõem o uso de consultas estruturadas para acesso a dados na Web. Finalmente, na Seção 5, são descritas as considerações finais e as direções futuras.

2. *Find Me*

Esta seção apresenta o *Find Me*, uma ferramenta de busca que permite consultas estruturadas em documentos HTML. A principal contribuição da ferramenta é permitir a construção de consultas mais expressivas sobre documentos em HTML na *Web*, utilizando princípios de consultas estruturadas com dados não necessariamente estruturados.

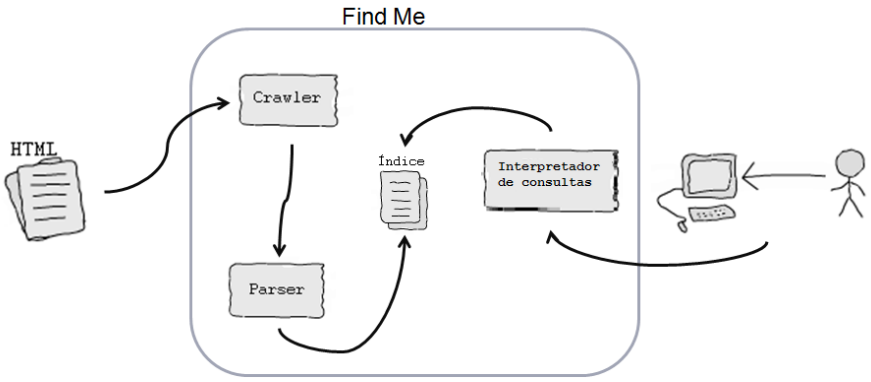


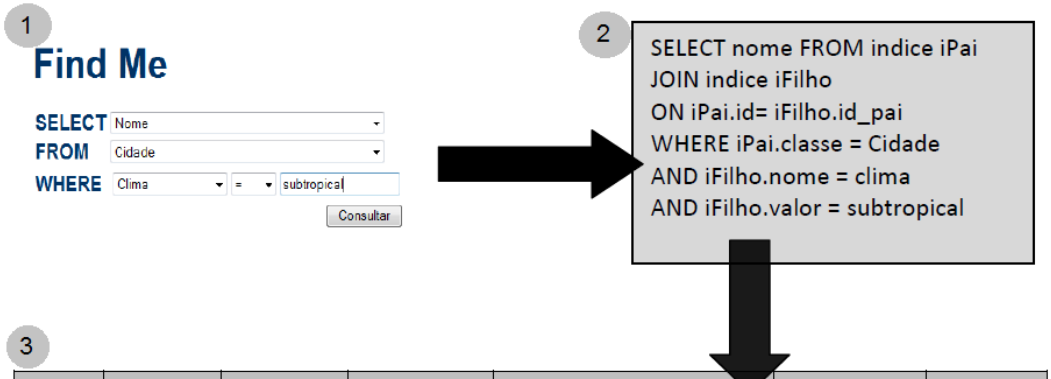
Figura 1 - Arquitetura da ferramenta *Find Me*

A arquitetura geral do *Find Me* é apresentada na Figura 1 e é composta por quatro componentes principais: (i) o ***crawler***, que é encarregado de percorrer a *Web* automaticamente, identificar documentos em HTML e encaminhar para o *parser* (o ponto inicial da varredura do *crawler* na *Web* é dado através de uma ou mais páginas *Web* que são denominadas sementes); (ii) o ***parser***, que recebe do *crawler* os documentos obtidos na Web, analisa cada um deles e identifica as características do documento, indexando-as de acordo com as regras heurísticas definidas na proposta (apresentadas na Seção 4); (iii) **índice**, que possui todas as informações sobre cada um dos documentos, estruturadas de acordo com as regras; e (iv) **interpretador de consultas**, que recebe a consulta construída na interface usada pelo usuário e a transforma em uma consulta compatível com os campos armazenados no índice.

2.1 Apresentação da Ferramenta

O *Find Me* foi desenvolvido em linguagem Java. Para a coleta das páginas foi utilizado um *Web* crawler de código aberto para linguagem Java, o *crawler4j*[1], e o parser *HtmlCleaner*[2], também de código aberto, para percorrer as páginas a fim de identificar os termos a serem indexados posteriormente. O desenvolvimento foi feito utilizando a linguagem Java SE, através da IDE Eclipse. A interface de consulta foi desenvolvida também em Java e JSPs.

Uma consulta precisa ser interpretada e adaptada antes de ser efetuada na tabela de índice. Na Figura 2 o usuário monta a consulta "*SELECT nome FROM cidade WHERE clima = subtropical*" através da interface principal (1), em seguida, esta consulta é interpretada de acordo com as cláusulas preenchidas (2) de forma que seja possível ser consultada na tabela de índice (3). Os atributos e objetos são indexados em uma mesma tabela, todo atributo possui um índice "pai" que referencia o objeto ao qual este atributo pertence. Para efetuar consultas as quais o usuário necessita comparar um atributo, é necessário efetuar um *join* entre duas linhas da tabela (objeto e atributo), conforme apresentado no item 2 da figura.



ID	ID_PAI	NOME	VALOR	NOME_URL	VALOR_URL	CLASSE
1		Erechim		http://wikipedia/.../Erechim		Cidade
2	1	Clima	Subtropical			

Figura 2 - Processo de consulta

[1] http://code.google.com/p/crawler4j/
[2] http://htmlcleaner.sourceforge.net/

2.2 Heurísticas para indexação

O modelo de dados do *Find Me* parte de premissas de que todo documento HTML é um objeto, todo objeto pertence a alguma classe e deve possuir atributos que devem possuir um valor. Os objetos se relacionam através de atributos que estão contidos em *links* que referenciam outro documento. Desta forma, um atributo pode ser atômico ou uma referência a outro objeto. Um exemplo da aplicação desta premissa é representado na Figura 3.

A definição de uma classe de objetos é feita através dos dados contidos na caixa informativa do documento. Caixa informativa, também chamada de *infobox*, é um padrão de todos os documentos *wiki*, onde são informados vários atributos sobre o documento. O título de um *infobox*, geralmente, inicia com o nome da classe do objeto. Por exemplo, o título *Município de Florianópolis*. A estrutura do *infobox*, normalmente, apresenta valores separados em uma tabela de duas colunas, onde a primeira coluna pode indicar o nome de um atributo e a segunda coluna o seu valor (Figura 3).

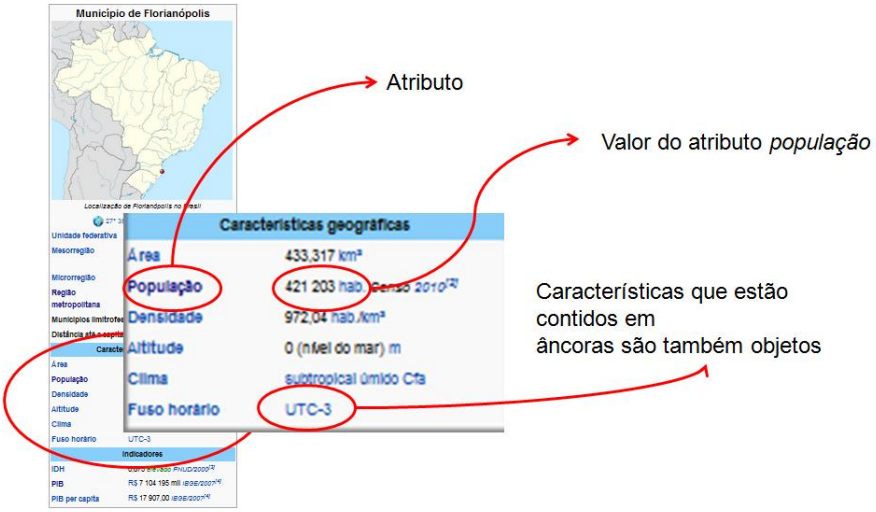


Figura 3 - Atributos do objeto *Florianópolis*.

3. Experimentos

Com base nos experimentos efetuados foi possível avaliar a confiabilidade dos resultados das consultas através de medidas de precisão, revocação e F-value [YATES e NETO, 1999]. Para a etapa de testes foram selecionados três domínios distintos, referentes a três classes: *Personalidade*, *Cidade* e *Filme*. O critério definido para escolha dos domínios foi o maior número de objetos classificados para determinada classe. Os testes foram realizados em uma base de dados com 22.640 índices, sendo que 2.162 índices correspondem a objetos e 20.478 atributos. Ao total 41 consultas foram efetuadas nos três domínios. Por exemplo, para o domínio *Cidade* um teste efetuado foi a execução da seguinte consulta: "*SELECT nome FROM Cidade WHERE fuso horário = UTC-4"*.

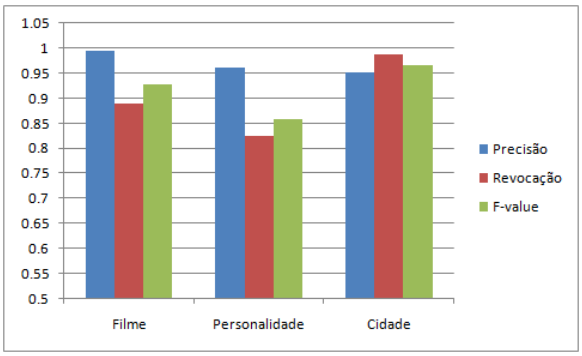


Figura 4 - Resultado dos experimentos

A Figura 4 apresenta a média das medidas de precisão, revocação e F-value obtidas nos experimentos nos três domínios escolhidos. Com base nos resultados, foi possível verificar que a média de precisão foi superior a 95%, porém, a média de revocação dos resultados foi pouco mais que 90%, isto significa que quase 10% dos objetos recuperados não foram relevantes à consulta.

Algumas limitações foram identificadas com base no resultado dos experimentos e a partir disto, novas melhorias foram incluídas aos trabalhos futuros. Por exemplo, no domínio *Personalidade* a média de revocação foi muito baixa com relação às demais medidas. Esta média deve-se a uma falha na indexação de atributos que possuem como valor uma *tag* com uma imagem antes da *tag* que representa o valor textual do atributo.

4. Trabalhos Relacionados

A ferramenta Mesa [MERGEN et. al., 2008] é um motor de busca que permite acesso aos dados estruturados em tabelas HTML de documentos *Web*. Mesa possibilita a consulta de dados previamente indexados tanto com palavras-chave quanto com *queries* formatadas em SQL. O resultado da consulta é exibido em um sumário com todas as tabelas resultantes e a URL de onde esta tabela foi extraída. Os autores utilizaram um *crawler* específico para acessar páginas que possuem tabelas em um contexto de filmes. A ferramenta indexa dados específicos de filmes armazenados no domínio do *Wikipedia*, portanto, como regra para a indexação das páginas, o documento deve possuir pelo menos uma tabela com as três colunas: *title*, *film* e *movie*. A necessidade de formulação de uma consulta em SQL foi apontada como limitação da ferramenta, visto que informar palavras-chave seria uma maneira muito mais fácil para o usuário, porém, pouco expressiva. Como trabalhos futuros, os autores sugerem a pesquisa de novas interfaces de consulta que equilibrem expressividade e simplicidade de uso.

Propostas de linguagens estruturadas para *Web* já surgiram, tais como Squeal [SPERTUS, STEIN, 2000] e SPARQL [PRUD'HOMMEAUX, SEABORNE, 2008]. A linguagem *Squeal* possui uma sintaxe similar à linguagem SQL, com construções efetuadas sobre os elementos HTML. Esta linguagem permite o acesso aos elementos e atributos das *tags* do HTML. Porém, não é possível efetuar consultas usando a semântica do domínio ao qual se está consultando. O *Squeal* difere da linguagem SPARQL (*SPARQL Protocol and RDF Query Language*) cujo objetivo principal é ser a linguagem padrão para a *Semantic Web*, e é considerada a peça chave nas tecnologias desenvolvidas na *Semantic Web*. Toda a linguagem foi projetada para ser executada

sobre documentos RDF. Portanto, o domínio a ser consultado deve estar descrito através de RDF para que consultas SPARQL possam ser executadas.

O presente trabalho difere dos trabalhos apresentados acima nos seguintes sentidos: ao contrário do *Mesa*, o *Find Me* está sendo desenvolvido para efetuar busca no documento HTML como um todo, não apenas em estruturas de elementos <*table*>. Em relação às linguagens estruturadas Squeal e SPARQL, a proposta apresentada neste artigo não constrói suas consultas sobre a estrutura dos elementos HTML, como a Squeal, nem faz uso de um descritor semântico como o RDF, como se propõe a SPARQL. O objetivo é efetuar consultas estruturadas sobre objetos, e atributos, encontrados em documentos HTML e que são definidos através de certas heurísticas.

5. Conclusões e Trabalhos Futuros

Este trabalho propõe um método para consultas *Web* que possa varrer documentos HTML e indexar dados semi-estruturados, tratando-os como objeto, atributo e valor de atributo dentro de um documento HTML. Como vantagem às demais técnicas de recuperação de informação atuais, foi também proposto ao trabalho, uma interface de consulta amigável, em que usuários mais leigos possam descrever suas consultas de forma mais expressivas. Esta interface deve possibilitar acesso a dados semi-estruturados através de consultas tipicamente formatadas para banco de dados relacionais, bem como o SQL. A idéia principal da proposta é considerar que todo documento HTML é um objeto, todo objeto pode possuir atributos e que todo atributo deve possuir algum valor. Além disso, os relacionamentos entre objetos são identificados através de links encontrados em seus atributos. Um dos principais trabalhos futuros identificados neste projeto é a implementação de *joins* entre objetos para permitir consultas como "*Personalidade de nacionalidade estaduniense que more na cidade de São Paulo*". Além disso, também proposta a aplicação de técnicas de classificação mais eficiente e inclusão de outros operadores de comparação.

6. Referências

Mergen, Sergio; FREIRE, Juliana; HEUSER, Carlos. Querying Structured Information Sources on the Web. In International Journal of Metadata, Semantic and Ontologies, 5:3, pp. 208-221, 2010

Mergen, Sergio; FREIRE, Juliana; HEUSER, Carlos. MESA: A Search Engine for QueryingWeb Tables. Salt Lake City - U.S [s.n.], 2008. 6 p.

Prud'hommeaux, E.; Seaborne, A. SPARQL Query Language for RDF. W3C Recommendation 15 January 2008.

Spertus, E.; Stein, L.A. Squeal: A structured query language for the web. In Proceedings of the 9th International World Wide WebConference (WWW9), pages 95– 103, Amsterdam, The Netherlands, 2000.

Yates, Ricardo B.; NETO Berthier R. Modern Information Retrieval. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.