Information Processing and Management 50 (2014) 416–425

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A review of ranking approaches for semantic search on Web

Vikas Jindal [a,*], Seema Bawa [b], Shalini Batra [b]

[a] School of Computational Sciences, Apeejay Stya University, Sohna 122103 Gurgaon, India
[b] Computer Science and Engineering Department, Thapar University, P.O. Box 32, Patiala 147004, India

ARTICLE INFO

Article history:
Received 3 April 2012
Received in revised form 14 August 2013
Accepted 18 October 2013
Available online 16 November 2013

Keywords:
Semantic search
Ranking
Ontology

ABSTRACT

With ever increasing information being available to the end users, search engines have become the most powerful tools for obtaining useful information scattered on the Web. However, it is very common that even most renowned search engines return result sets with not so useful pages to the user. Research on semantic search aims to improve traditional information search and retrieval methods where the basic relevance criteria rely primarily on the presence of query keywords within the returned pages. This work is an attempt to explore different relevancy ranking approaches based on semantics which are considered appropriate for the retrieval of relevant information. In this paper, various pilot projects and their corresponding outcomes have been investigated based on methodologies adopted and their most distinctive characteristics towards ranking. An overview of selected approaches and their comparison by means of the classification criteria has been presented. With the help of this comparison, some common concepts and outstanding features have been identified.

1. Introduction

Web search is a key application of the Web where present search technologies rely on link analysis techniques that exploit the structure of Web to determine important documents. At the same time, they rely on simple term statistics to identify documents that are most relevant to a query. Mark-up languages such as (X)HTML are primarily focused to documents whose content should be interpretable by human interpreters and hence focused on document structure and its presentation. Little efforts are paid to the representation of the semantics of the content itself.

The growing availability of structured information on the Web enables new opportunities for information access. Semantically oriented search engines and specifically that use ontologies as enabling technologies have gained considerable interest in the last decade. The ever growing amount of ontology-based semantic mark-up in the Web provides an opportunity to start working in the direction of a new generation of open intelligent applications (Motta & Sabou, 2006). Efficient search is one such major envisioned application of this next generation Web popularly known as Semantic Web (Burners-Lee, Hendler, & Lassila, 2001).

Current Web search techniques are not directly suited for indexing and retrieval of semantic mark-up. Document is treated as a *bag of words* where words or word variants are recognized as indexing terms. The existing semantic mark-up is either simply ignored by many search engines for indexing purposes or not processed in a way that allows the mark-up to be used distinguishably from other text during the search.

The upcoming Web search is no longer limited to matching keywords of the query against documents but instead complex information needs can be expressed in a structured way with precise and structured answers as results. The kind of

* Corresponding author. Tel.: +91 8295262540; fax: +91 0124 2013125.
E-mail addresses: jindal35@gmail.com (V. Jindal), seema@thapar.edu (S. Bawa), sbatra@thapar.edu (S. Batra).

0306-4573/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2013.10.004

search in which user's information needs are addressed by considering the meaning of user's query as well as available resources is referred to as *Semantic Search* (Tran, Haase, & Studer, 2009).

Due to the ever increasing move from data to knowledge and increasing popularity of the vision of Semantic Web, there is equally increasing interest and work in automatically extracting and representing the metadata as semantic annotation to the documents and services on the Web (Shah, Finin, Joshi, Cost, & Mayfield, 2002). It seems that each Web page would possess semantic annotation that record additional details concerning the page itself. Annotations are based on classes of concepts and relations among them. The "vocabulary" for the annotation is usually expressed by means of ontology. The information contained in such agreed upon ontology is quite valuable for determining the relevance of the retrieved documents based on the "known" facts, relationships or the other data. Table 1 shows a comparison of features of Traditional Keyword-based search and Semantic-based search based on various parameters.

The two elements of the ontology are quite significant from the "relevant information access" point of view. The first element is the *named entities* such as names of persons, objects, countries, places, research articles, artists, and museum. Available techniques had been developed for entity oriented search of documents (Aleman-Meza, Arpinar, Nural, & Sheth, 2010). The second element is the *relationships* which provide meaning to the entity. The value of such relationships relies on the fact that those are named relationships. Relationships play a vital role in the relevant information access as the Web evolves continuously (Sheth & Ramakrishnan, 2007).

2. Motivation for ranking

Many users try to analyze information either by browsing information space or using a search engine. Search engine based systems generally locate documents based on keywords. Although they do return documents involving keywords inputted by user, a lot of retrieved documents have very less to do with user's needs. The onus lies on the user to decide about the relevance of the retrieved documents using their mental model in order to obtain desired information. Efforts are consistently being made to extend or identify alternatives to traditional search mechanisms focused on finding documents based on keyword-based approaches. With the advent of the Semantic Web along with enabling technologies, a stage has been set which will facilitate in getting relevant documents from the massive data sources thereby assisting in information analysis.

The premise of search technologies today is primarily centered on enabling search for entities or other Semantic Web resources. Different from traditional text-based information retrieval systems which exclusively retrieve and rank documents, semantic search systems retrieve and rank entities of various types in response to user queries. Semantics of relationships among entities are defined in schema ontologies (e.g., through the domain and range constructs in RDF(S) or OWL languages). It is increasingly possible to analyze metadata extracted from Web to discover interesting relationships. Possibly, just as document ranking is a critical component in present search engines, the ranking of complex relationships is likely to be important component in the upcoming Semantic search engines. But it is very unlikely that ranking schemes for ranking entities (documents, resources etc.) may be applied for ranking complex relationships among entities. Furthermore, heterogeneous relationships existing among entities embedded into semantic annotations can be effectively exploited to define ranking strategies for semantically annotated Web pages.

2.1. Ranking for normal search

One of the most impressive and popular Ranking model for the ordering of retrieved documents is PageRank (Page, Brin, Motowani, & Winograd, 1998). It looks at the Internet as a big graph where pages are nodes and hyperlinks are edges. It has been successfully applied to distinguish the popularity of different Web documents through analyzing the link structure in the Web graph. It is obvious that in the Web graph, all the Web pages involving same keyword(s) are not equally popular. E.g. only some top conferences pertaining to a research field are highly important with high quality research papers. In order to help users to quickly locate their pages of interest, popularity of retrieved pages is required to be calculated. The more popular a Web page is, the more likely the user is interested in it and hence the more important that page is. PageRank algorithm facilitates to accurately approximate such global importance for a given page. It is based on the intuition that more the num-

Table 1
Comparison of features of Traditional Keyword-based search and Semantic-based search.

Parameter	Traditional Keyword-based search	Semantic-based search
Dataset	Documents	RDF triples, semantically annotated documents
Data organization	Unstructured	Semi- structured
Search orientation	Document – centric	Entity, relationship and semantic document centric
Collection	Bag of words	Bag of (RDF) assertions
Representation	Light weight syntax – centric models	Ontology based better expressive models
Domain of satisfaction	Work well for topical searches	Complex queries are satisfied, more precise answers
Query processing approach	Matching and filtering	Not just matching and filtering but also joining
Scalability	Web scale	Not scale to massive and heterogeneous Web environment