Information Processing and Management 000 (2016) 1–20

Contents lists available at ScienceDirect

Information Processing and Management

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

PREFCA: A portal retrieval engine based on formal concept analysis

Eman Negm, Samir AbdelRahman*, Reem Bahgat

Department of Computer Science, Faculty of Computers and Information, Cairo University, Giza, Egypt

ARTICLE INFO

Article history:
Received 23 March 2015
Revised 14 August 2016
Accepted 20 August 2016
Available online xxx

Keywords:
Information retrieval
Formal concept analysis
Network analysis
Portal retrieval

ABSTRACT

The web is a network of linked sites whereby each site either forms a physical portal or a standalone page. In the former case, the portal presents an access point to its embedded web pages that coherently present a specific topic. In the latter case, there are millions of standalone web pages, that are scattered throughout the web, having the same topic and could be conceptually linked together to form virtual portals. Search engines have been developed to help users in reaching the adequate pages in an efficient and effective manner. All the known current search engine techniques rely on the web page as the basic atomic search unit. They ignore the conceptual links, that reveal the implicit web related meanings, among the retrieved pages. However, building a semantic model for the whole portal may contain more semantic information than a model of scattered individual pages. In addition, user queries can be poor and contain imprecise terms that do not reflect the real user intention. Consequently, retrieving the standalone individual pages that are directly related to the query may not satisfy the user's need. In this paper, we propose PREFCA, a Portal Retrieval Engine based on Formal Concept Analysis that relies on the portal as the main search unit. PREFCA consists of three phases: First, the information extraction phase that is concerned with extracting portal's semantic data. Second, the formal concept analysis phase that utilizes formal concept analysis to discover the conceptual links among portal and attributes. Finally, the information retrieval phase where we propose a portal ranking method to retrieve ranked pairs of portals and embedded pages. Additionally, we apply the network analysis rules to output some portal characteristics. We evaluated PREFCA using two data sets, namely the Forum for Information Retrieval Evaluation 2010 and ClueWeb09 (category B) test data, for physical and virtual portals respectively. PREFCA proves higher F-measure accuracy, better Mean Average Precision ranking and comparable network analysis and efficiency results than other search engine approaches, namely Term Frequency Inverse Document Frequency (TF-IDF), Latent Semantic Analysis (LSA), and BM25 techniques. As well, it gains high Mean Average Precision in comparison with learning to rank techniques. Moreover, PREFCA also gains better reach time than Carrot as a well-known topic-based search engine.

1. Introduction

The research in Information Retrieval (IR) field has been evolving rapidly to confront the flood of web information. The main challenge in the field of Web Information Retrieval is to retrieve relevant information from a gigantic number of

* Corresponding author.
E-mail addresses: e.negm@fci-cu.edu.eg (E. Negm), s.abdelrahman@fci-cu.edu.eg (S. AbdelRahman), r.bahgat@fci-cu.edu.eg (R. Bahgat).

http://dx.doi.org/10.1016/j.ipm.2016.08.002
0306-4573/© 2016 Elsevier Ltd. All rights reserved.

Table 1
Semantic search engines with their corresponding features.

Class	Search engine	Year	Live URL	Summarization	Predefined ontologies	Semantic tags	Clustering	Faceted search
Topic-based	*Carrot*^{2}	2002	http://search.carrot2.org/				✓	
	KartOO	2002	http://www.kartoo.com/				✓	
	SnakeT	2004	http://snaket.di.unipi.it/				✓	
	Sensebot	2007	http://www.sensebot.net/	✓				
	DuckDuckGo	2008	http://duckduckgo.com/	✓				
	Clusty	2009	http://www.clusty.com/				✓	
Knowledge-based	Hakia	2004	http://www.hakia.com/		✓			
	Swoogle	2004	http://swoogle.umbc.edu/			✓		
	Lexxe	2005	http://www.lexxe.com/			✓		
	Google (Knowledge graph)	2012	http://google.com/		✓			
	Google (Humming bird)	2013	http://google.com/		✓			
	Google (Images)		http://google.com/					✓
	Bing (Satori)	2013	http://www.bing.com/		✓			
	Yahoo (Knowledge graph)	2013	http://www.yahoo.com/		✓			
	Amazon		http://www.amazon.com/					✓
	DBLP	2008	http://dblp.uni-trier.de/					✓

web pages (documents) which often have dynamic, heterogeneous, and duplicate nature. Unfortunately, a typical user query usually consists of imprecise and insufficient terms (words). Web Search Engines (Croft, Metzler, & Strohman, 2010) are among the most important applications of Web IR. They are concerned with retrieving the nearest relevant documents that match the words in the user query. Traditional search engines, such as Salton and Buckley (1988) and Landauer, Foltz, and Laham (1998), represent the words of the user query and the web pages as a set of weights using some mathematical equations. Then, they retrieve the web pages having the words of the highest weights matching the words in the user query.

Recently, semantic search engines (Mangold, 2007; Negi & Kumar, 2014; Qureshi, Asma, & Khan, 2013) have been developed as extensions of these traditional search engines. Their aim is to enhance the traditional search results by analyzing the contextual meaning of the query words in the searchable network pages. The Semantic search engines depend on two main approaches. The first one is the topic-based approach, it is based on clustering or summarizing the search results into some web clues or topics. The second one is the knowledge-based approach, it is based on searching the query words in some predefined domain-specific ontologies, dictionaries, or annotated semantic tag corpora. The search engines that follow the topic-based approach do not constrain themselves to specific dictionaries or patterns as the knowledge-based search engines do. They basically collect the results of some other search engine techniques, such as TF-IDF and LSA. Then, they apply some semantic clustering or topic identification methods.

Table 1 classifies the semantic search engines according to their semantic features. Sensebot and DuckDuckGo are examples for the topic-based search engines, they apply summarization techniques on the search results returned from other search engines. *Carrot*^{2}, SnakeT, Clusty, and KartOO are Web clustering search engines (Carpineto, Osiński, Romano, & Weiss, 2009) that use the topic-based approach. Web clustering search engines display the final results to the user as meaningful groups (i.e. clusters), each group contains a title and a set of internal pages or other sub-groups. The objective of the Web clustering search engines is to give the user an overview on the main topics of the retrieved results. Therefore, the user can reach the relevant documents faster by selecting the related topic (Alam & Sadaf, 2013).

Faceted Search (Zheng, Zhang, & Feng, 2013) is another approach for filtering the search results to enhance the browsing process. It iteratively refines the search results based on the facets. The facets represent the aspects of a subject (such as title, field, publication year, and author could be considered as facets for the book). They could be hierarchically structured or totally independent of each other. The user uses the values of the facets to narrow down the search results (i.e. faceted browsing). Facet search achieves better values when the user is unfamiliar with the topic being searched. However, the main limitation of the faceted search is the cost and the time that are required to extract the facets and their values. This extraction is currently done manually by the domain expert. Google Images, Amazon, and DBLP Computer Science Bibliography are examples of search engines that implement faceted search. Faceted search could be easily implemented in PREFCA. In Section 5.4, we described a proof of concept that utilizes the domains of the portal as a facet to filter the search results. The domains of the portal are determined manually by the portal owner. Currently, we use data sets that are not prepared for faceted search as it does not contain extracted facets. We plan to utilize more data sets to support full faceted search.

Hakia, Lexxe, and Swoogle are examples for knowledge-based semantic search engines. They depend on predefined ontologies or semantic tags. Although Google, Yahoo, and Bing are considered as traditional search engines, they started to merge the knowledge-based approach with their traditional search approaches. Google rolled out multiple semantic updates like Knowledge Graph (Singhal, 2012) and Hummingbird. Microsoft announced Bing's semantic knowledge base named Satori. Yahoo engine is also building its own knowledge graph to enhance its search results.