Chapter 13
Exploring Semantic Characteristics of Socially Constructed Knowledge Repository to Optimize Web Search

Dengya Zhu
Curtin University, Australia

Heinz Dreher
Curtin University, Australia

ABSTRACT

Short-term queries preferred by most users often result in a list of Web search results with low precision from a user perspective. The purpose of this research is to improve the relevance of Web search results via search-term disambiguation and ontological filtering of search results based on socially constructed search concepts. A Special Search Browser (SSB) is developed where semantic characteristics of the socially constructed knowledge repository are extracted to form a category-document set. kNN is employed with the extracted category-documents as training data to classify Web results. Users' selected categories are employed to present the search results. Experimental results based on five experts' judgments over 250 hits from Yahoo! API demonstrate that utilizing the socially constructed search concepts to categorize and filter search results can improve precision by 23.5%, from Yahoo's 41.7% to 65.2% of SSB based on the results of five selected ambiguous search-terms.

DOI: 10.4018/978-1-4666-2494-8.ch013

INTRODUCTION

The introduction and subsequent explosion of the Web has dramatically changed our approach to access and use of information. Internet is becoming a part of life for most people in the world. However, as indicated by Baeza-Yates and Ribeiro-Neto (1999), most users have difficulties in expressing their information needs in search-term format: they prefer short queries instead of the Boolean expressions (Jansen & Spink, 2006). To address this issue, most search engines encourage users to enter very short search terms as queries, and then return a list of search results which are ranked by technologies such as traditional information retrieval models and PageRank (Page, et al., 1998), according to the relevance degree of the results with respect to a given query. However, as the volume of information on the Web is becoming unbelievably huge, short search terms based Web search usually leads to search engines return a list of thousands, even millions of search results. Searchers are thus frustrated when facing such a long list of results especially when half of the search results are irrelevant to their information needs (Gauch, Chaffee, & Pretschner, 2003). It is now commonly recognized that information search services are far from perfect. The challenges of search engines are summarized in Table 1.

The first challenge for search engines is the *Information overload* (C1). What does "1-10 of 55,400,000 for jaguar" mean? Can we really access 55,400,000 information items about jaguar? Are all of the items are relevant to my information need? Oh my God, how can manage to read all of them. It seems that search engines prefer to present a number of search results as huge as this for short queries. However, as research indicated (Jansen & Spink, 2006), the tendency is fewer results pages are browsed. Therefore, millions of search results are a kind of information overload for users.

The second challenge of search engines is *Mismatching hits* (C2). The performance of an information retrieval system is usually measured by precision and recall. Precision is an evaluation of how retrieved results of the information retrieval system are relevant; whereas the recall is an evaluation of how the relevant results are retrieved (Baeza-Yates & Ribeiro-Neto, 1999). While millions of search results being retrieved, meaning that the recall is very high, precision remains low, because most the retrieved results are irrelevant. For example, the top ten search results of a Chinese name "Wei Liu" are about ten different persons (Zhu, 2007), and you are lucky if the person you are searching for is among the top ten results; or on the other hand, the information you are looking for is located on the tenth page.

The third issue of search engines is that search results are presented in the form of a *flat list* (C3). The flat list of results is suitable for a small amount of results because it provides an easy and quick way to locate a relevant information item. However, when there are hundreds or even thousands of retrieved results returned, users have to do re-search amongst the returned items; they need to go through page by page to pick up useful information items. Finding relevant information

Table 1. Challenges of web search engines (Zhu, 2007)

	Challenges	Phenomena
C1	Information overload	Millions of Web hits
C2	Mismatching hits	High recall, low precision, many irrelevant results
C3	Flat list of results	Results are presented in a flat list, users have to pick up useful items among the list, like finding a needle in haystack
C4	Mismatching mental model	Automatically formed hierarchy used to re-organize Web hits usually mismatches human mental model
C5	Homogeneity	Search engines present "the same for all" hits, not personalized
C6	Low recall of Web navigation	Web navigation is more accurate, but the recall is very low