Enhancing Web Search through Web Structure Mining

Ji-Rong Wen
Microsoft Research Asia, China

INTRODUCTION

The Web is an open and free environment for people to publish and get information. Everyone on the Web can be either an author, a reader, or both. The language of the Web, *HTML (Hypertext Markup Language),* is mainly designed for information display, not for semantic representation. Therefore, current Web search engines usually treat Web pages as unstructured documents, and traditional information retrieval (IR) technologies are employed for Web page parsing, indexing, and searching. The unstructured essence of Web pages seriously blocks more accurate search and advanced applications on the Web. For example, many sites contain structured information about various products. Extracting and integrating product information from multiple Web sites could lead to powerful search functions, such as comparison shopping and business intelligence. However, these structured data are embedded in Web pages, and there are no proper traditional methods to extract and integrate them. Another example is the link structure of the Web. If used properly, information hidden in the links could be taken advantage of to effectively improve search performance and make Web search go beyond traditional information retrieval (Page, Brin, Motwani, & Winograd, 1998, Kleinberg, 1998).

Although *XML (Extensible Markup Language)* is an effort to structuralize Web data by introducing semantics into tags, it is unlikely that common users are willing to compose Web pages using XML due to its complication and the lack of standard schema definitions. Even if XML is extensively adopted, a huge amount of pages are still written in the HTML format and remain unstructured. Web structure mining is the class of methods to automatically discover structured data and information from the Web. Because the Web is dynamic, massive and heterogeneous, automated Web structure mining calls for novel technologies and tools that may take advantage of state-of-the-art technologies from various areas, including machine learning, data mining, information retrieval, and databases and natural language processing.

BACKGROUND

Web structure mining can be further divided into three categories based on the kind of structured data used.

- **Web graph mining:** Compared to a traditional document set in which documents are independent, the Web provides additional information about how different documents are connected to each other via hyperlinks. The Web can be viewed as a (directed) graph whose nodes are the Web pages and whose edges are the hyperlinks between them. There has been a significant body of work on analyzing the properties of the Web graph and mining useful structures from it (Page et al., 1998; Kleinberg, 1998; Bharat & Henzinger, 1998; Gibson, Kleinberg, & Raghavan, 1998). Because the Web graph structure is across multiple Web pages, it is also called *interpage structure.*
- **Web information extraction (Web IE):** In addition, although the documents in a traditional information retrieval setting are treated as plain texts with no or few structures, the content within a Web page does have inherent structures based on the various HTML and XML tags within the page. While Web content mining pays more attention to the content of Web pages, Web information extraction has focused on automatically extracting structures with various accuracy and granularity out of Web pages. Web content structure is a kind of structure embedded in a single Web page and is also called *intrapage structure.*
- **Deep Web mining:** Besides Web pages that are accessible or crawlable by following the hyperlinks, the Web also contains a vast amount of noncrawlable content. This hidden part of the

Web, referred to as the *deep Web* or the *hidden Web* (Florescu, Levy, & Mendelzon, 1998), comprises a large number of online Web databases. Compared to the static surface Web, the deep Web contains a much larger amount of high-quality structured information (Chang, He, Li, & Zhang, 2003). Automatically discovering the structures of Web databases and matching semantically related attributes between them is critical to understanding the structures and semantics of the deep Web sites and to facilitating advanced search and other applications.

MAIN THRUST

Web Graph Mining

Mining the Web graph has attracted a lot of attention in the last decade. Some important algorithms have been proposed and have shown great potential in improving the performance of Web search. Most of these mining algorithms are based on two assumptions. (a) Hyperlinks convey human endorsement. If there exists a link from page A to page B, and these two pages are authored by different people, then the first author found the second page valuable. Thus the importance of a page can be propagated to those pages it links to. (b) Pages that are co-cited by a certain page are likely related to the same topic. Therefore, the popularity or importance of a page is correlated to the number of incoming links to some extendt, and related pages tend to be clustered together through dense linkages among them.

Hub and Authority

In the Web graph, a *hub* is defined as a page containing pointers to many other pages, and an *authority* is defined as a page pointed to by many other pages. An authority is usually viewed as a good page containing useful information about one topic, and a hub is usually a good source to locate information related to one topic. Moreover, a *good* hub should contain pointers to many good authorities, and a *good* authority should be pointed to by many good hubs. Such a mutual reinforcement relationship between hubs and authorities is taken advantage of by an iterative algorithm called HITS (Kleinberg, 1998). HITS computes authority scores and hub scores for Web pages in a subgraph of the Web, which is obtained from the (subset of) search results of a query with some predecessor and successor pages.

Bharat and Henzinger (1998) addressed three problems in the original HITS algorithm: mutually reinforced relationships between hosts (where certain documents "conspire" to dominate the computation), automatically generated links (where no human's opinion is expressed by the link), and irrelevant documents (where the graph contains documents irrelevant to the query topic). They assign each edge of the graph an authority weight and a hub weight to solve the first problem and combine connectivity and content analysis to solve the latter two. Chakrabarti, Joshi, and Tawde (2001) addressed another problem with HITS: regarding the whole page as a hub is not suitable, because a page always contains multiple regions in which the hyperlinks point to different topics. They proposed to disaggregate hubs into coherent regions by segmenting the DOM (document object model) tree of an HTML page.

PageRank

The main drawback of the HITS algorithm is that the hubs and authority score must be computed iteratively from the query result on the fly, which does not meet the real-time constraints of an online search engine. To overcome this difficulty, Page et al. (1998) suggested using a random surfing model to describe the probability that a page is visited and taking the probability as the importance measurement of the page. They approximated this probability with the famous PageRank algorithm, which computes the probability scores in an iterative manner. The main advantage of the PageRank algorithm over the HITS algorithm is that the importance values of all pages are computed off-line and can be directly incorporated into ranking functions of search engines.

Noisy link and topic drifting are two main problems in the classic Web graph mining algorithms. Some links, such as banners, navigation panels, and advertisements, can be viewed as noise with respect to the query topic and do not carry human editorial endorsement. Also, hubs may be mixed, which means that only a portion of the hub content may be relevant to the query. Most link analysis algorithms treat each Web page as an atomic, indivisible unit with no internal structure. This leads to false reinforcements of hub/authority and importance calculation. Cai, He, Wen, and Ma (2004)