19

Semantic Structure Analysis of Web Documents

Rupesh R. Mehta, Harish Karnick, and Pabitra Mitra

19.1 Introduction

Today the web has become the largest information source for a large section of the world population. The web contains documents that are highly volatile, distributed and heterogeneous. The content of a web page is usually much more diverse when compared with traditional plain-text documents and may encompass multiple regions/segments with unrelated topics. Currently, most information retrieval systems on the web considers a web page as the smallest and indivisible unit. However, often it is not appropriate to represent a whole web page as a single semantic entity as web documents are heterogeneous and contain multiple topics that are not necessarily related to each other. Moreover, for the purpose of browsing and publication, non-content materials, such as navigation bars, decoration items, interaction forms, copyrights and contact information, are usually embedded in web pages. Considering the web page not as an individable unit but as having an underlying semantic structure with topically coherent segments as atoms, relevant and wealth of information contents on the web page (excluding noisy and irrelevant contents) can be found out and a better performance of web information retrieval systems can be achieved.

Many web applications can exploit the semantic structure of web documents to its benefit. For example, in query expansion system, relevant words are added to the query to increase the information content. The quality of expansion terms is highly affected by top-ranked documents. Two major negative factors for query expansion systems are noisy content and multiple topics embedded in the document. With the use of semantic structure of a web page, semantically homogeneous web page segments can be easily obtained. As the term correlation within a segment will be much higher than those in other parts of web page, high-quality expansion terms can be

extracted from the segments and used to make more specific and relevant query. This helps in improving the information retrieval performance. Recently, link analysis has received more importance. Generally, web pages are treated as single semantic and hence all links in a web page get equal importance. Traditionally, all the links in a web page are treated equally. The basic assumption of link analysis in information retrieval systems is that multiple citations from a single web page are likely to cite semantically related web pages, i.e. if there is a link between two pages, there is some relationship between the two whole pages. In short, the relevance of a web page is a reasonable indicator of the relevance of its neighbours. But in most cases, a link from page X to page Y indicates that there might be some relationship between a certain part of page X (containing the link) and a certain part of page Y but not necessarily between the whole web pages X and Y. Also, it has been shown that the main reason for topic drift problem in the HITS (hyperlink induced topic search) algorithm [1, 2] is existence of a large amount of noisy information. The same observation can be made from recent works on topic distillation [3, 4] and focused crawling [5]. However, these works are based on a DOM (document object model) tree of the web page that does not have sufficient power to semantically segment the web page. Furthermore, due to the small size of screen of handheld devices, it is better to show the list of categories to which web page belongs and later display user-interested segment only, rather than the whole web page. This efficient browsing of large web pages on small handheld devices also necessitates semantic analysis of web pages [6].

19.2 Related Work

Document passage (segment) retrieval is a research topic with a long history in the information retrieval (IR) community that addresses the shortcomings of whole-document ranking. Previous work reveals that it is sometimes beneficial to apply retrieval algorithms to portions of a document, particularly when documents contain multiple drifting subjects or have varying lengths [7, 8].

In traditional passage retrieval, passages can be categorized mainly into three classes: discourse, semantic and window. Discourse passages rely on the logical structure of the documents marked by punctuation, such as sentences, paragraphs and sections [7–9]. Semantic passages rely on semantic structure of document to partition it into topics or sub-topics [8, 10, 11], whereas windows-based passage retrieval approach partition the document based on the fixed number of words (per passage) [7, 12, 13].

We cannot directly adopt these plain-text document passage retrieval techniques for partitioning web pages. Some research [14] on web page segmentation and its applications has been done using traditional passage retrieval methods, but the results are not encouraging. This indicates that