ABSTRACT

Authors write headings for splitting a document into multiple semantic blocks of different topics. A block may include some other blocks, and the blocks in a document compose hierarchical heading structure. Information on hierarchical heading structure is very useful in document retrieval because the structure represents the topic structure, which is the most important factor in document retrieval. We can use the information on hierarchical heading structure in various ways for improving document retrieval systems. In this dissertation, we focus on the largest collection of digital documents, the World Wide Web. We first propose a method for extracting hierarchical heading structure in web pages, and also propose four methods that use the information on the extracted structure for improving web search systems. In the following, we explain the overview of these five methods.

Extracting Logical Hierarchical Structure of HTML Documents Based on Headings

We propose a method for extracting hierarchical heading structure in web pages in hypertext markup language (HTML). Human readers easily understand the structure by exploiting the following properties of headings: (1) headings appear at the beginning of the corresponding blocks, (2) headings are given prominent visual styles, (3) headings of the same level share the same visual style, and (4) headings of higher levels are given more prominent visual styles. Our method also exploits these properties. Our experiment shows our method outperforms existing methods.

Subtopic Ranking Based on Hierarchical Headings

We propose methods for generating diversified rankings of subtopics of keyword queries. Our methods are aware of hierarchical heading structure. Each heading concisely describes the topic of its corresponding block. Therefore, hierarchical headings in documents reflect the hierarchical topics referred to in the documents. Our methods score subtopics based on matching between the subtopics and hierarchical headings in documents. They give higher scores to subtopics matching hierarchical headings associated to more contents. Our methods generated significantly better subtopic rankings than query completion results by major commercial search engines.

Abstract

Heading-Aware Proximity Measure and Its Application to Web Search

Proximity of query keyword occurrences is one important evidence useful for effective document scoring. If a query keyword occurs close to another in a document, it suggests strong relevance of the document to the query. Most web pages contain hierarchical heading structure, and it affects logical proximity. For example, occurrences in headings describe the topics of the entire corresponding blocks, and the occurrences are strongly connected to any term occurrences in the blocks regardless of the simple distance between them. Based on such observations, we developed a heading-aware proximity measure, heading-aware semi-distance, applied the measure to three existing proximity-aware document scoring functions, and evaluated the existing and modified functions on the data sets of a well-known Text Retrieval Conference (TREC) web tracks. Our heading-aware Span method generated rankings significantly better than the existing Span method with simple distance.

Heading-Aware Block-Based Web Search

We propose a web search system that improves the ranking of search results by exploiting the hierarchical heading structure in web pages. Because a heading is a concise description of the topic of its associated block, query term occurrences in a heading indicate high relevance of the block to the query intent. In hierarchical heading structure, headings of ancestor blocks describe the topics of all their descendant blocks, although the ancestor headings may be located far from the descendant blocks. Based on these facts, we developed a search system which retrieves and scores blocks by taking into account their ancestor headings as well as their contents. We evaluated our system by using TREC data sets. The result indicates that our system can eliminate many irrelevant blocks with satisfactory recall and produce significantly better document ranking.

Heading-Aware Snippet Generation for Web Search

We propose new methods to generate search result snippets of web pages. To generate the snippets indicating the relevance of the pages, many existing methods extract sentences containing many query keywords. According to our observation, heading words are very often omitted from their associated blocks because readers can understand the topic of the block by reading its heading first. To score sentences considering the omitted heading words, our method counts keyword occurrences in their connected headings as well as in the sentences themselves. Finally, we evaluate and compare four methods, namely a naive method, the existing method, our method, and the combination of the existing method and our method. Our evaluation indicated that the combination generates the best snippets for the queries with clear intents and the queries containing four or more keywords.