Chapter 3
A Roadmap to Integrate Document Clustering in Information Retrieval

R. Subhashini
Sathyabama University, India

V. Jawahar Senthil Kumar
Anna University, India

ABSTRACT

The World Wide Web is a large distributed digital information space. The ability to search and retrieve information from the Web efficiently and effectively is an enabling technology for realizing its full potential. Information Retrieval (IR) plays an important role in search engines. Today’s most advanced engines use the keyword-based (“bag of words”) paradigm, which has inherent disadvantages. Organizing web search results into clusters facilitates the user’s quick browsing of search results. Traditional clustering techniques are inadequate because they do not generate clusters with highly readable names. This paper proposes an approach for web search results in clustering based on a phrase based clustering algorithm. It is an alternative to a single ordered result of search engines. This approach presents a list of clusters to the user. Experimental results verify the method’s feasibility and effectiveness.

INTRODUCTION

Since large information is available in an unstructured manner, retrieving out relevant documents containing the required information is the primary goal of research. This task is known as Information Retrieval. Most of the information retrieval systems are limited to the query processing based on keywords and key phrases. In Vector space model (VSM) (Khaled et al., 2004) a document is represented as a vector of index terms. In information retrieval system (Manning et al., 2008) the matching of the query against a set of text record is the core of the system. Any IR (Information Retrieval) system defines four basic elements:

- A collection profile,
- A document and query representation,
- A matching function,
- A ranking criteria.
Retrieval of the relevant natural language text document is a great challenge. The ability to search and retrieve information from the Web (Jansen, 2000; Page, & Brin, 1998) efficiently and effectively is an enabling technology for realizing its full potential.

The IR community has explored document clustering as an alternative method of organizing retrieval results (Branson & Greenberg, 2002). Document clustering is widely applicable in areas such as search engines, web mining, information retrieval, and topological analysis. Document clustering is an automatic grouping of text documents into clusters so that documents within a cluster have high similarity, but are dissimilar to documents in other clusters. Existing search engines such as Google, Yahoo and MSN often return a long list of search results, ranked by their relevancies to the given query. Most of the web search engines (Cutting et al., 1992) are also characterized by extremely low precision. Clustering algorithms, attempt to group documents together based on their similarities; thus documents relating to a certain topic will hopefully be placed in a single cluster. Clustering technique relies on four concepts: data representation model, similarity measure, clustering model and clustering algorithm that generates the clusters using the data model and the similarity measure. This can help users both in locating interesting documents more easily and in getting an overview of the retrieved document set. Moreover, most traditional clustering algorithms cannot be directly used for search result clustering, because of some practical issues. Zamir and Etzioni (1999) gave a good analysis on these issues. The algorithm should take the document snippets instead of the whole documents as input, because of the downloading time of original documents; the clustering algorithm should be fast enough for online calculation; and also the generated clusters should have readable descriptions for quick browsing by users, etc. We also follow these requirements to design our algorithm.

In this algorithm, several properties for each phrase such as phrase frequencies, document frequencies, phrase length, etc are considered and score of the phrase is calculated. The phrases are ranked according to the score, and the top-ranked phrases are taken as base clusters, which are further merged according to their corresponding documents. Our method is more suitable for Web search results clustering because we emphasize the efficiency of identifying relevant clusters for Web users. Furthermore, the clusters are ranked according to their scores, thus the more likely clusters required by users are ranked higher. It generates shorter and more readable cluster names, which enable users to quickly identify the topics of a specified cluster.

RELATED WORKS

The vector space model, VSM, provides a baseline for the purpose of ranking documents based on a query (Baeza-Yates & Ribeiro-Neto, 1999). The similarity between the terms or keywords of the document and the terms of the query are determined based on the frequency of the occurrence of the terms in the document. The higher the frequency of the occurrence of query terms, the higher the document will be ranked. This method has the advantage of not depending on the order of the keywords. Automatically extracting key phrases from documents is a task with many applications in information retrieval and natural language processing. Turney’s Extractor (1999) extracts a small number of key phrases from documents. The research on clustering search results has been investigated in a number of previous works. Agglomerative Hierarchical Clustering (AHC) algorithms are probably the most commonly used. These algorithms are typically slow when applied to large document collections. Single-link and group-average methods typically take $O(n^2)$ time, while complete-link methods typically
Related Content

Analysis the Typhoon Eyes of Megi from MTSAT Satellite Cloud Images with 3-D Profile Reconstruction
[www.igi-global.com/chapter/analysis-typhoon-eyes-megi-mtsat/73771?camid=4v1a](www.igi-global.com/chapter/analysis-typhoon-eyes-megi-mtsat/73771?camid=4v1a)

Events Automatic Extraction from Arabic Texts
[www.igi-global.com/article/events-automatic-extraction-from-arabic-texts/142824?camid=4v1a](www.igi-global.com/article/events-automatic-extraction-from-arabic-texts/142824?camid=4v1a)

MapReduce Based Information Retrieval Algorithms for Efficient Ranking of Webpages
[www.igi-global.com/article/mapreduce-based-information-retrieval-algorithms/68374?camid=4v1a](www.igi-global.com/article/mapreduce-based-information-retrieval-algorithms/68374?camid=4v1a)

Retrieval of Web Pages on Real-World Events related to Physical Objects
[www.igi-global.com/article/retrieval-web-pages-real-world/72706?camid=4v1a](www.igi-global.com/article/retrieval-web-pages-real-world/72706?camid=4v1a)