An Intelligent Web Search Using Multi-Document Summarization

Sheetal A. Takale, Vidya Pratishthan’s College of Engineering, Baramati, India
Prakash J. Kulkarni, Walchand College of Engineering, Sangli, India
Sahil K. Shah, Vidya Pratishthan’s College of Engineering, Baramati, India

ABSTRACT

Information available on the internet is huge, diverse and dynamic. Current Search Engine is doing the task of intelligent help to the users of the internet. For a query, it provides a listing of best matching or relevant web pages. However, information for the query is often spread across multiple pages which are returned by the search engine. This degrades the quality of search results. So, the search engines are drowning in information, but starving for knowledge. Here, we present a query focused extractive summarization of search engine results. We propose a two level summarization process: identification of relevant theme clusters, and selection of top ranking sentences to form summarized result for user query. A new approach to semantic similarity computation using semantic roles and semantic meaning is proposed. Document clustering is effectively achieved by application of MDL principle and sentence clustering and ranking is done by using SNMF. Experiments conducted demonstrate the effectiveness of system in semantic text understanding, document clustering and summarization.

KEYWORDS

Document Clustering, Multi-Document Summarization, Semantic Role, Semantic Similarity, Web Search Result Summarization

1. INTRODUCTION

Information available on the Internet is huge, diverse and dynamic. Current Search Engine is doing the task of intelligent help to the users of the Internet. Without search engines, the Internet would be a huge, disorganized and unmanageable information that would certainly be interesting but perhaps not very useful. The search engines are drowning in information, but starving for knowledge. For a user keyword query, current web search engines return a list of ranked individual web pages with respect to the query. The information for such query is scattered among all the ranked web pages. Hence user finds it cumbersome to locate the desired information. For a multi-topic query, occurrence frequency of individual query keyword is more as compared to the co-occurrence frequency of query keywords. For such multi-topic queries, information is often distributed among multiple physical pages. Thus search engines face the result representation challenge: search engine returns a list of documents ranked by their relevance to the request, and users have to go through the list and examine the documents one by one to identify their desired cases. To address the challenges faced by present web search engines, we propose an intelligent web search using multi-document summarization. It automatically generates a summarized solution for the user query using the search results from the web using search engines like Google, Yahoo, etc.
Multi-document summarization (Bollegala, 2012; Bollegala, 2010; Xiaoyan, 2012) is a challenging problem. It has to cover single document summarization as well as it should be able to select or extract the best information from multiple documents. Basically, an automatic document summarization is a process of reducing the size of documents provided, it preserves the semantic content. The task of automatic summarization with human intelligence is not to simply extract and concatenate sentences. But, it satisfies coherence, pronoun resolution, semantic content, grammar, etc. Our aim of multi document summarization is query focused extractive summarization with a focus on semantic content extraction. The aim of our work is to represent the information drawn by search engine in the form of knowledge.

A preliminary version of this paper (Shah, Takale, 2013) is presented in IEEE ICARET-13. It focuses on iAssist as an Intelligent Help Desk System.

The main contributions of the paper are:

- The method proposed for summarization follows the principle of: “A sentence is ranked higher if it is contained in the theme cluster which is more relevant to the given query while a theme cluster is ranked higher if it contains many sentences which are more relevant to the given query”;
- An outline of the new approach to the problem of search result summarization considering the semantics and context information of query and search result is presented (section 3);
- A typical approach to text preprocessing generally involves, stop word removal, TF-IDF ranking and suffix removal steps. The problem of text preprocessing is addressed with a new approach. Preprocessing of search engine results and role labeling is discussed (section 3);
- A new approach to sentence level semantic similarity computation is making use of semantic role labeling and semantic dictionary. Document ranking is done using the semantic score of each document. Sentence level semantic similarity computation and document ranking are used to identify top relevant search results (section 4);
- The most intuitive approach to multi document summarization involves clustering of similar documents as the important and fundamental step. Clustering of top ranking search results is achieved using the Minimum Description Length MDL principle (section 5);
- Once the topics or themes are identified by clustering, next step in summarization process is sentence clustering. It is done using symmetric non-negative matrix factorization (SNMF) (section 6);
- Sentence selection from Sentence cluster (summarization of sentence cluster) is based on the score of the sentence. Sentence score is computed using its internal similarity score and external similarity score (section 7);
- An analysis of results of search result summarization for user queries is presented (section 8).

See Table 1 for abbreviations used in this article.

2. LITERATURE SURVEY

Currently the web is a huge collection of information. Search engines are helping us to make effective utilization of information available on the internet. However, for a given user query, search engines don’t answer the query. They merely provide a long list containing the titles and snippets of each of the retrieved documents. Such lists are useful only for navigational queries but not for informational queries. The problem arises when the information for the query is vague, too broad or ill-defined. In such situation the range of covered topics becomes unmanageably large, confusing and doubtful to the user. To solve this problem, we address the task of generating a set of query focused summaries of search results.
Query Recommendation for Improving Search Engine Results
www.igi-global.com/article/query-recommendation-improving-search-engine/53126?camid=4v1a