WEB INFORMATION RETRIEVAL USING AUTOMATIC MULTI-DOCUMENT SUMMARIZATION

Rahul Shankarrao Khokale
Department of Computer Science & Engineering,
Priyadarshini Indira Gandhi College of Engineering,
NAGPUR (India)

Mohammd Atique
Post Graduate Department of Computer Science,
Sant Gadge Baba Amravati University,
AMRAVATI (India)

ABSTRACT

Today, internet has become the most important source of information. People are highly accustomed to the use of internet for acquiring information which they need. Many times, it is revealed that, the information seeker does not get relevant information very easily due to the presence of non-relevant web pages. This paper addresses the problem of effective information retrieval from the web. In this paper, the notion of Web Information Retrieval using Automatic Multi-document Summarization is presented. The proposed work is blend of Web technology and Natural Language Processing. When user will fire the query, the system tries to fetch web pages from different web servers, and they are indexed as per the order of relevance. The degree of relevance is not determined by the how many times the keywords of query is present in the document but it is determined on the basis of semantic content of the document and the user query.

Keywords: Web Information Retrieval, Automatic Multi-Document Summarizations, Web Technology, Information Retrieval and Natural Language Processing
I. INTRODUCTION

1.1 Information Retrieval

The Internet and the Web offer new opportunities and challenges to information retrieval researchers. With the information explosion and never ending increase of web pages as well as digital data, it is very hard to retrieve useful and reliable information from the Web. Materials from millions of web pages from organizations, institutions and personnel have been made public electronically accessible to millions of interested users. The Web uses an addressing system called Uniform Resource Locators (URLs) to represent links to documents on web servers. These URLs provide location information. Like titles of books in traditional libraries, no one can remember all URLs on the Web. Web search engines allow us to locate the internet resources through thousands of Web pages. It is almost impossible to get the right information as there is too much irrelevant and outdated information. Information retrieval systems provide useful information in libraries to researchers.

The Web can be viewed as a virtual library. Information retrieval is an important and major component of the Internet and the Web in the information age and should play an important role in knowledge discovery. General search engines such as, Google, AltaVista, Excite are considered as the powerful search engines so far. Most of the current search engines are based on words, not the concepts. When searching for certain information or knowledge with a search engine, one can only use a few key words to narrow down the search. The result of the search is tens or maybe hundreds of relevant and irrelevant links to various Web pages.

In spite of the voluminous studies in the field of intelligent retrieval systems, effective retrieving of information has been remained an important unsolved problem. Implementations of different conceptual knowledge in the information retrieval process such as ontology have been considered as a solution to enhance the quality of results. Furthermore, the conceptual formalism supported by typical ontology may not be sufficient to represent uncertainty information due to the lack of clear-cut boundaries between concepts of the domains [1] “Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.” (Salton, 1968).

1.2 Automatic Text Summarization

Automatic Text Summarization means the process of extraction and representation of most important content from the source document in the condensed form. This process involves Document Preprocessing, Feature Extraction, Sentence Ranking and Summary Generation [3]. Preprocessing is accomplished by Tokenization, Sentence Splitting, POS Tagging etc. Feature Extraction includes Word Frequency Extraction and Sentence Ordering. Weighing Sentences helps to score sentences required for Sentence Ranking. Summary Generation is the resulting phase of automatic text summarization.

1.3 Multi-Document Summarization

Multi-document text summarization deals with retrieval of salient information about a topic from various sources. The task of multi-document summarization is to identify a set of sentences, phrases or some generated semantically correct language units carrying some useful information. Then significant sentences are extracted from this set and re-organized them to get multi-documents’ summary [1]. Let D = {D_1, D_2, ..., D_n} be a set of documents. Let S = {S_1, S_2, ..., S_n} be a set of summary of each document respectively.
II. RELATED WORK

In recent years, the research focus in the domain of natural language processing and information retrieval has been shifted to the area of automatic document summarization. Automatic document summarization is of two types: *abstractive* and *extractive*.


III. PROPOSED WORK

This paper deals with the framework for “Web Information Retrieval based on Multi-Document Summarization”. The proposed framework is shown in Figure 2. However, in this paper, the emphasis is given on the multi-level document summarization. The basic purpose of this framework is to enhance the effectiveness of web information retrieval. As the indexing and ranking of the retrieved documents is supported by intelligent decision making system which is based on fuzzy inference rules, the degree of relevance can be increased.
3.1 Intelligent Query Processing
When a query is written by the user and submitted to the system, it is required to manipulate it and represent it in proper form. Intelligent Query Processing (IQP) module helps user to formulate his query. WordNet is used for identifying synonyms and thesaurus. This intelligent unit tries to understand the user need and accordingly it classifies the query into any of three types: Informational, Navigational and Transactional.

3.2 Search Engine
Search Engine uses web crawler to traverse the World Wide Web to fetch matching URLs.

3.3 Multi-Document Summarization
The n number of links/URLs is the input to the Multi-Document Summarization (MDS) unit. Each of these inputs can be HTML web page or a text/PDF file. MDS finds summary of each document separately and finally combines all of them to form single summary. It involves significant sentences identification, sentences reordering etc. The detail algorithm for this is discussed in section IV.

3.4 Indexing
The process of generating inverted indices of the retrieved documents is Indexing. It is also an important step in information retrieval.
3.5 Ranking

Ranking means determining the weight or rank of each retrieved web page or web document. Our page rank strategy is based on the summary generated by MDS unit. The Page Rank Algorithm will use fuzzy inference system to judge the relevance of the web page according to the user query.

IV. AUTOMATIC DOCUMENT SUMMARIZATION MODEL

To summarize a document or documents, a reader has to understand the document(s) and integrate information and make connections across sentences to form a coherent discourse representation. We designed and developed a new generic algorithm for automatic document summarization based on the analysis of human cognition and intelligence. In order make this model applicable for summarization we define the concept of ‘event’. ‘Event’ is a cognitive psychological concept, and can be either a story or a sentence in microstructure. We have treated each sentence as an event in this paper. The cognitive model is shown below.

![Cognitive Model](image)

**Figure 3: Cognitive Model**

As each event is representing sentence in the document, it is a combination of two parts: Subject and Predicate. e.g. Consider the example of an event:

```
Event : Tiger is a wild animal
```

This event can be represented in predicate form by using FOPL syntax

```
wild_animal(tiger)
```

The representation of events in predicate form is required to create the knowledge about the document. In addition to that, we have defined inference rules to understand the connection of one sentence to others. This helps us to decide the significance of sentence within the document and used for sentence re-ordering.
The Algorithm for Automatic text/document summarization is discussed below:

**Figure 4:** Algorithm for Text summarization

V. EXPERIMENTAL RESULTS

We have implemented the multi-document summarization system in JAVA.

5.1 **Test Data/Corpus**: We have used the standard CACM test data with 3204 test samples. Following are the documents chosen from CACM test data.

**Document**: CACM—0276.html

Program Organization and Record Keeping for Dynamic Storage Allocation

The material presented in this paper is part of the design plan of the core allocation portion of the ASCII-MATIC Programming System. Project ASCII-MATIC is concerned with the application of computer techniques to the activities of certain headquarters military intelligence operations of the U. Army.

CACM October, 1961

Holt, A. W.
5.2 Term Frequency Computation
The term frequency of all the terms in the sentence is calculated individually. The sum of all these frequencies will give the $T(s_i)$, term frequency of sentence $s_i$. This can be calculated by:

$$T(S_i) = \sum_{t=1}^{n} t * f(t_i)$$

Where, $t_i$ is the $i^{th}$ term in the sentence. And $t*f(t_i)$ is the term frequency of term $t_i$.

Therefore, $T(Si)$ for document 1

$$T(Si) = (1+1+1+1+1+1+2+5+1+1+1+1+1+3+1+1+1+1+2+2+1+1+1+1+1+1+1+1)$$

$$= 50$$

5.3 TF-IDF or term-frequency inverse-document-frequency
TF-IDF or term-frequency inverse-document-frequency is computed as the ratio of the quantity of terms in that document to the frequency of the quantity of documents containing that terms. For above document the TF_IDF is given below:

$$TF_{IDF} = \frac{41}{50}$$

$$= 0.82$$

V. CONCLUSION
In this paper, we have proposed a framework for web information retrieval using multi-document summarization. The emphasis was given on the multi-level document summarization. The basic purpose of this framework is to enhance the effectiveness of web information retrieval. As the indexing and ranking of the retrieved documents is supported by intelligent decision making system which is based on fuzzy inference rules, the degree of relevance can be increased. The experiment are carried out on the CACM test data, and it is found that information retrieval results can be improved after multi-document summarization process is performed.
REFERENCES


