# A Framework for Automatic Document Understanding for Web Information Retrieval

Rahul S. Khokale
Priyadarshini Indira Gandhi College of Engineering,
Nagpur, (India)

Mohammad Atique, Ph.D
Associate Professor
SGB Amravati University, Amravati (India)

## ABSTRACT

Most of the web search engines use keyword based approach to search for needed information on the web. When a query is submitted by the user to the search engine, the web crawler tries to match the keywords with name of file, URL or the meta tags of the documents. Because of this, user may get many non-relevant documents along with relevant documents. It can lead to the frustration of information seekers. This problem can be alleviated, if the search is based on the contents and intents rather than only keywords. Automatic document understanding focuses on representation of a document in summarized form with its gist containing important contents and the intention of the author. This paper deals with the framework of a system for automatic document understanding for web information retrieval. The basic purpose of this work is to enhance the effectiveness of information search on the internet.

## General Terms

Algorithms, Natural Language Processing, Information Retrieval

## Keywords

Automatic Multi-document Summarization, Web Information Retrieval, Document Understanding

## 1. INTRODUCTION

Internet and 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. The location information of required web-pages is determined from these URLs. Like titles of books in traditional libraries, no one can remember all URLs on the internet. 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 out dated information. Information retrieval systems provide useful information in libraries to researchers.

The Web can be viewed as a virtual library. It is an enormous collection of documents. 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 [9] “Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.” (Salton, 1968).

### 1.1 Document Understanding

As more and more of our daily transactions involve computers, the volume of data and information on computers is enormous. Because of the Sophistication of requiring electronic documents (e.g. routing and retrieval) and the complexity of the documents themselves, it is not sufficient to simply scan and perform OCR (optical character recognition) on documents; deeper understanding of the document is s needed. Comprehensive document understanding involves the form (layout), as well as the function and the meaning of the document. as well as the function and the meaning of the document. Document understanding is thus a technology area which benefits greatly from the *integration* of text understanding. Text understanding is necessary to operate on the textual content of the document [11]

According to Sherif Yacoub,

> “*Document understanding is a field that is concerned with semantic analysis of documents to extract human understandable information and codify it into machine-readable form. Document understanding systems provide means to automatically extract meaningful information from a raster image of a document*”

Document understanding as a research endeavor consists of studying all processes involved in taking a document through various representations: from a scanned physical document to high-level semantic descriptions of the document. Some of the types of representation that are useful are: editable descriptions, descriptions that enable exact reproductions and high-level semantic descriptions about document content [15].

### 1.2 Information Retrieval

Information Retrieval (IR) is the science of searching for information within relational databases, documents, text, multimedia files, and the World Wide Web[13]. As the volume of information on the internet is increasing day by day so there is a challenge for website owner to provide proper and relevant information to the internet user[14]. 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 out dated information. Information retrieval systems provide useful information in libraries to researchers.

Definition

According to Salton (1968), information retrieval is defined as,

> *"Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information."*

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 [9]

## 2. 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.*

The research in this field began with Term Frequency based summarization. Following researchers used term frequency based approach for document summarization. G. Salton, 1989, Jun'ichi Fukumoto, 2004, You Ouyang, 2009 and Mr.Vikrant Gupta, 2012 [1], Inderjeet Mani, 1997, Rada Mihalcea, 2004, Junlin Zhang, 2005, Xiaojun Wan, 2008, Kokil Jaidka, 2010 [1] carried out research for document summarization using Graph-based approach. Kathleen McKeown, 1995, Xiaojun Wan, 2007 used Time-Based method for document summarization. Sentence Correlation method was implemented for document summarization by Shanmugasundaram Hariharan, 2012, Tiedan Zhu, 2012. Clustering-Based method for document summarization was proposed by Jade Goldstein, 2000. Vikrant Gupta el at [2], developed an auto-summarization tool using statistical techniques. The techniques involve finding the frequency of words, scoring the sentences, ranking the sentences etc. Yogan Jaya Kumar et al. [3] discussed Automatic Multi Document Summarization Approaches. Y. Surendranadha Reddy el at [4] presented a summarization system that produces a summary for a given web document based on sentence importance measures such as sentence ranking. Tiedan Zhu et al [5] proposed an improved approach to sentence ordering for Multi-document Summarization. Information Retrieval (IR) and Information Extraction (IE) play important role in the era of information technology. New trends are observed in IR and IE due to the rise of Web 2.0 paradigm [6]. It is required to review these trends and put them into context of what improvements and potential IR and IE have to offer to knowledge engineers, information workers, but also typical Internet users.

Efficient and intelligent information retrieval has been always a point of concern for the researchers. Various techniques and methodologies were tried by these intellectual fraternities to provide the best to the information seeker to satisfy their information needs. Information retrieval process is based on different models [12]. Support Vector Machines (SVM) is one of them. Monika Arora el at [8], presented the application of Support Vector Machine for designing the model for efficient and intelligent retrieval. Intelligent Information Retrieval (IIR) is also addressed by Vandana Dhingra et al [10]. The authors tried to focus on the areas where existing search engines are lacking and discuused the need of intelligent information retrieval. They have proposed the theoretical framework for IIR. The techonologies such as metadata, RDF, URI, XML, XMLS, Triples and Ontologies are discussed.

## 3. PROPOSED WORK

The objective of this research work is to develop user-friendly, interactive web based information retrieval system, which is based on the document understanding. The system focuses on effective retrieval of relevant information by minimizing the number of non-relevant documents. The mathematical model of our proposed system is described as below.

### 3.1 Document Model

A document can be represented in the form of vector using vector space model. i.e. d = {$t_1$, $t_2$, $t_3$, ………….,$t_m$ } be a document consisting of *m* terms or words, $t_1$,$t_2$,……$t_m$.

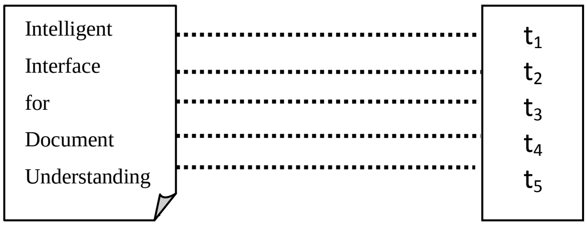

Figure 1 : Document vector space representation

q = { $q_1$, $q_2$, ...... ,$q_n$} be a set of queries. $L_{qi} = | q_i |$ is the length of query *qi* (number of keywords). $T_{qi}$ is the type of query *qi.*

$$T_{qi} = \begin{cases} 1 & \text{if qi is informational} \\ 0 & \text{if qi is navigational} \end{cases}$$

Let $f$ be a function from query vector to document vector, i.e. $f : Q \rightarrow D$. The matching between query and document is shown in following figure,

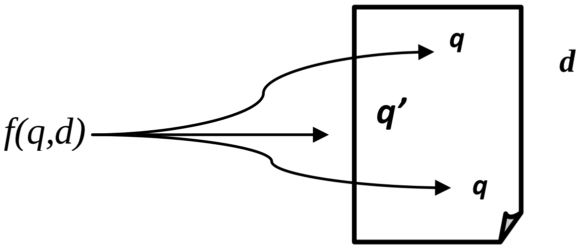

**Figure 2 : Matching between query and document**

The process of automatic document understanding is illustrated in figure 3

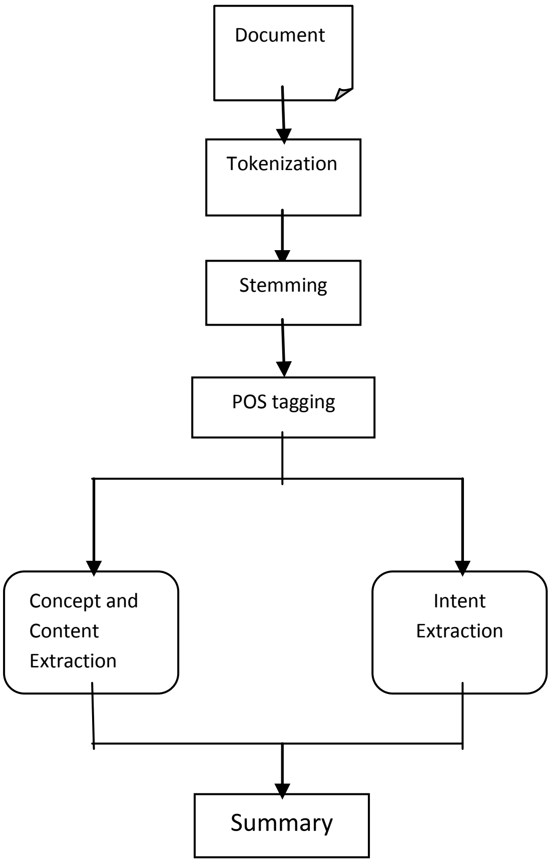

**Figure 3 : Document understanding process**

Tokenization is a primary step of machine translation. In this phase, the input sentence is decomposed into tokens. These tokens are give n to POS stagger function to tag the tokens with their respective type.

e.g. Sentence : "*India has won the match by six wickets*"

Tokens : "*India*" "*has*" "*won*" "*the*" "*match*" "*by*" "*six*" "*wickets*"

Stemmer is very useful to find stem i.e. root word of any word.

A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'.

Consider the input string , "*Boy is very good*"

The output of POS stagger for this sentence will be :

Boy → noun
is → verb
very → adjective
good→noun

Concept-based approach to text summarization can be effective means for document understanding. Soujanya Poria et al [16], presented a notion of concept extraction. They have devised an algorithm for concept extraction which is given below

**Algorithm (Concept Extraction)**

- **Input** : $S_{NL}$ (*Natural language sentence)*
- **Output**: $C = \{C_1, C_2, \ldots, C_N\}$ *List of concepts*
- **Method** :

```
Find the number of verbs in the sentence;
 for every clause do
        extract VerbPhrases and NounPhrases;
        stem VERB;
     for every NounPhrase with the associated verb do
        find possible forms of objects;
        link all objects to stemmed verb to get events;
     end
   repeat until no more
 end
```

The Web information retrieval system based on automatic document understanding is shown in figure 4. The proposed system is a blend of natural language processing and web technology. Given user query is analyzed and from it content and intent are extracted. These semantic features are compared with the semantic features extracted from the summary of each web page associated with the user query submitted to the search engine. Similarity distance metrics are used to determine the page rank of each of the document. The pages are indexed as per order of page rank and the most relevant pages are displayed to the user.

**Algorithm (Document Understanding)**

- **Input** : $d$ (Given document : PDF/HTML/txt/doc file)
- **Output** : $d_s$ (Document summary)
- **Method :**

```
Step 1) while i = 1 : EOF
            token(i) = get_token();        //tokenization
        end

Step 2) for j=1:NT                  // NT → No. of tokens
            stem_tok(j)=stemmer(j);
        end

Step 3) for k=1:NT
            POS_Tag(k)=POS_Tagger(k);
        End
```

```
Step 4) for l=1:NS                // NS → No. of sentences
            Concept{l} = Extract_Concept{l}
            Content(l) = Extract_Content{l}
            Intention{l} = Extract_Intention{l}
        end

Step 5) Determine most significant sentence. Rearrange
        the sentences.

Step 6) Generate summary
```

This algorithm includes get_token(), stemmer(), POS_Tagger(), Extract_Concept(), Extract_Content() and Extract_Intention() functions. Sentences are given the weights according to their concept, content and intent scores and they are rearranged accordingly. The least significant sentences are ignored and the summary of high significant sentences is constructed.

**Algorithm (Information Retrieval)**

```
• Input : q ϵ Q (user query)
• Output : D = {$d_1$,$d_2$,………..,$d_n$} set of relevant documents
• Method :
Step 1) Get the user query q, representation of q using NLP
        techniques.

Step2) Search Engines retrieves set of documents $d_1$, $d_2$,…..,$d_n$
Step 3) Perform Document Understanding for each of these
        documents. Obtain page score for each of these
        pages.

Step 4) Find inverted indexes for these documents according
        to the page_scores.

Step 5) Retrieve the most relevant documents
```

In the above algorithm, web based information retrieval is proposed. User will submit query in natural language form, which is processed and represented into Boolean form. The search engine will retrieve all possible links of documents satisfying the criterion of the query. The proposed system will perform document understanding for each of these documents.

The set of documents is $D$ = {$d_1$,$d_2$,….$d_N$} and the result of document understanding for the set $D$ is $DU$ = {$DU_1$, $DU_2$,………..,$DU_N$}. Each of the members of $DU$ represents some score or weight which is used for forming index and inverted indexes. The system displays the documents satisfying the threshold value of page score. As a result, it will show most relevant documents and removes various non-relevant pages.

To measure information retrieval effectiveness in the standard way, we need a test collection consisting of three things:

1. A document collection
2. A test suite of information needs, expressible as queries
3. A set of relevance judgments, standardly a binary assessment of either *relevant* or *non-relevant* for each query-document pair.

Relevance is assessed relative to an information need, *not* a query. For example, an information need might be:

> *Information on whether eating low calorie food is more effective at reducing your risk of heart attacks than high calorie food.*

This might be translated into a query such as:

> food AND low_calorie AND high_calorie AND heart AND attack AND effective

A document is relevant if it addresses the stated information need, not because it just happens to contain all the words in the query.

## 4. CONCLUSION

For web information retrieval many information retrieval systems are existing. These systems take keyword based queries. They have certain pros and cons. Though many of them are very popular but still they have to face challenges. In this paper, a framework for web information retrieval using document understanding is proposed. The emphasis was given on the document understanding using text summarization method. The basic purpose of this framework is to enhance the effectiveness of web information retrieval. After experimentation, it is found that information retrieval results can be improved after multi-document summarization process is performed.

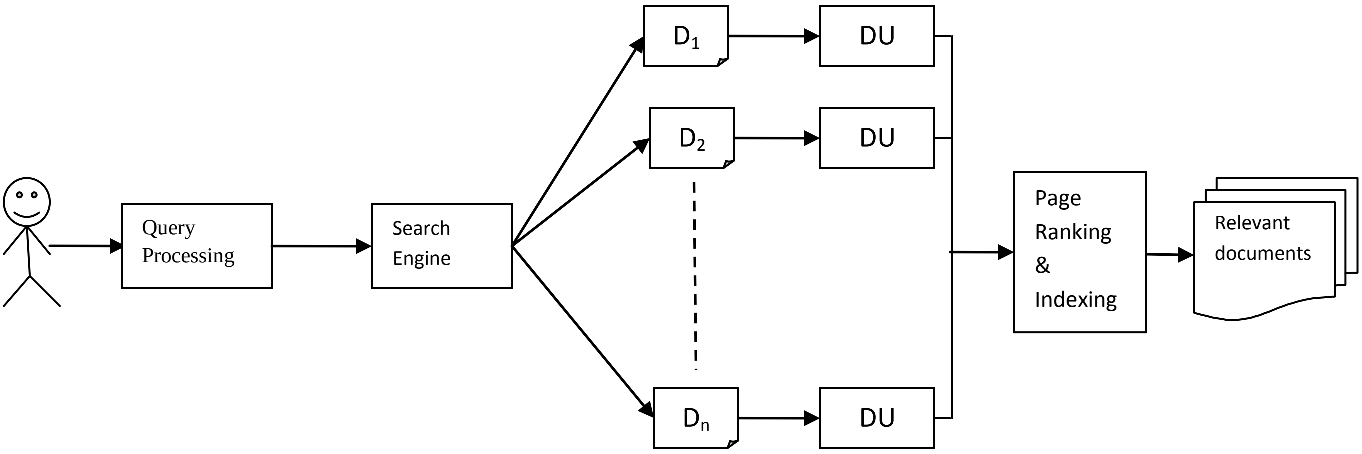

**Fig 4: Web Information retrieval using document understanding**

## 5. REFERENCES

[1] Md. Majharul, Suraiya Pervin and Zerina Begum 2013 Literature Review of Automatic Multiple Documents Text Summarization. International Journal of Innovation and Applied Studies, Vol. 3 No. 1 May 2013, pp. 121-129

[2] Vikrant Gupta, Priya Chauhan, Sohan Garg, Anita Borude, Shobha Krishnan, An Statistical Tool for Multi-Document summarization, International Journal of Scientific and Research Publications, Volume 2, Issue 5, May 2012

[3] Yogan Jaya Kumar and Naomie Salim, Automatic Multi Document Summarization Approaches, Journal of Computer Science 8 (1): 133-140,

[4] Y. Surendranadha Reddy and A.P. Siva Kumar, An Efficient Approach for Web document summarization by Sentence Ranking, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 7, July 2012

[5] Tiedan Zhu, Xinxin Zhao, An Improved Approach to Sentence Ordering For Multi-document Summarization, 2012 IACSIT Hong Kong Conferences, IPCSIT vol. 25 (2012) © (2012) IACSIT Press, Singapore

[6] Nikola Vlahovic, Information Retrieval and Information Extraction in Web 2.0 environment, International Journal Of Computers, Issue 1, Volume 5, 2011

[7] Yi Guo and George Stylios, An Intelligent Algorithm For Automatic Document Summarization

[8] Monika Arora, Uma Kanjilal, Dinesh Varshney, *"Efficient and Intelligent Information Retrieval using Support Vector machine (SVM)",* International Journal of Soft Computing and Engineering (IJSCE) , Volume-1, Issue-6, January 2012 pp 39-43

[9] Maryam Hourali and Gholam Ali Montazer, "An intelligent Information Retrieval Approach Based on Two Degrees of Uncertainty Fuzzy Ontology", Hindawi Publishing Corporation Advances in Fuzzy Systems Volume 2011, Article ID 683976, 11 pages

[10] Vandana Dhingra and Komal Kumar Bhatia, "Towards Intelligent Information Retrieval on Web", International Journal on Computer Science and Engineering (IJCSE), Apr 2011, Vol. 3 No. 4, pp 1721-1726

[11] Suzane, Liebowitz Taylor, Deborah A. Dahl, Mark Lipshutz, Carl Weir, Lewis M. Norton, Roslyn Nilson and Marciaa Linebarger, "Integrated Text and Image Understanding for Document Understanding"

[12] Djoerd Hiemstra, "Information Retrieval Models", Published in: Goker, A., and Davies, J. Information Retrieval: Searching in the 21st Century. John Wiley and Sons, Ltd., ISBN-13: 978-0470027622, November 2009

[13] Youssef Bassil, " A Survey on Information Retrieval,Text Categorization, and Web Crawling", Journal of Computer Science & Research (JCSCR) - ISSN 2227- 328X, Vol. 1, No. 6, December 2012, pp 1-11

[14] Dilip Kumar Sharma and A. K. Sharma, "A Comparative Analysis of Web Page Ranking Algorithms", International Journal on Computer Science and Engineering, Vol. 02, No. 08, 2010, pp 2670-2676

[15] Sargur Srihari, Stephen Lam, Venu Govindaraju, Rohini Srihari and Jonathan Hull, "Document Understanding: Research Directions", DARPA Document Understanding Workshop, Xerox PARC, Palo Alto, CA, May 6-8, 1992

[16] Soujanya Poria, Erik Cambria, Grégoire Winterstein, Guang-Bin Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis", Knowledge-based Systems Elsevier 2014