Semantic Information Retrieval System

Dr A R Nadira Banu Kamal,
Head of Computer Science,
Thassim Beevi Abdul Kader College for Women, Kilakarai.

M Manimekala,
Assistant Professor of Computer Science,
ThassimBeevi Abdul Kader College for Women, Kilakarai.

Abstract— Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources.In computer science and information science ontology is a formal representation of set of concepts within a domain and the relationships between those concepts. This paper explains how to search a document using keyword.

Keywords – Information Retrieval, Information Retrieval System.

I. INTRODUCTION

Information retrieval, concerns the retrieving of relevant information from databases. It is basically concerned with facilitating the user's access to large amounts of (predominantly textual) information.Information Retrieval is a process of searching some collection ofdocuments, using the term document in its widest sense, in order toidentify those documents which deal with a particular subject.Information retrieval isthe technique and process of searching, recovering, and interpretinginformation from large amounts of stored data. In formal search there is a chance to produce irrelevant pages as results for search using ontology the unrelated pages are avoided and the relevant to the search are provided to the user for reference.

II. BASIC CONCEPTS OF IR MODELS

A conceptual model of IR when dealing with text documents, one should distinguish different views on these documents. Several subfields of computer science and related fields deal with documents, where most of the fields focus on one or two views and ignore the others. Here I try to present an integration of the different perspectives of information retrieval system. For this purpose, I propose to extraction of relevant keywords of a text document. Standard information retrieval results is that automatic indexing in which algorithms do statistical word counting and indexingleads to better performance than systems in which people do manual indexing.
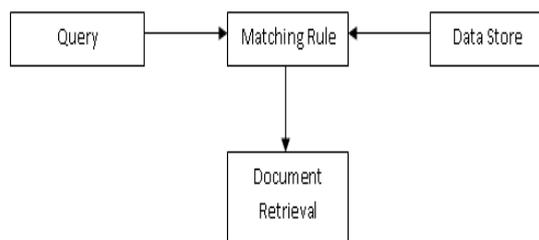
Fig. 1 General model of information retrieval systems

III. CONCEPTUAL MODELS OF IR

The most general facet in the scheme is *conceptual model*. An IR conceptual model is a general approach to IR systems. Several taxonomies for IR conceptual models have been proposed.

External attributes comprise data that is not contained within the document, i.e. a user looking at the document only may not see these values. External attributes contain information that is needed for certain types of processing the document, e.g. the name of the creator of the document, access rights, or publication information. In digital libraries, this type of data often is called metadata.

Logical structure: The media data that is contained within the document, and its internal structure comprise the logical structure of a document. Usually, documents have a hierarchical structure (e.g. a book divided into chapters, chapters containing sections, consisting of subsections, which comprise paragraphs, images

and tables). In this tree structure, the data is located in the leaves, where a leaf contains single media data only (e.g. text, graphics, images, audio, video, animation, 3D). Hypermedia links allow for nonhierarchical structures.

Layout structure: In order to show a document to a user, it must be presented at some kind of output media (e.g. when a document is printed, we have a sequence of pages). Based on a so-called style sheet, the layout process maps the logical structure onto the output media. The layout structure describes the spatial distribution of the data over the output media, e.g. the sequence of pages, which in turn are subdivided into rectangular areas (e.g. page header, footer, columns). This concept can be extended to time-dependent media (e.g. audio, video), where the layout structure describes the temporal and spatial distribution on an appropriate output device

Content deals with the meaning of a document The content is derived from the logical structure, in most cases by an automatic process. The content representation may have an internal structure, too, but often rather simple schemes are used. For example, in text retrieval, content mostly is represented as a set of concepts.

IV. EXTRACTING RELEVANT IDENTIFIERS OF A DOCUMENT

The application will extract keywordsfrom the documents using preprocessing method. In preprocessing, the system eliminates the stop words, numeric and punctuation marks from the

documents. The preprocess parsing application break the documents into individual words, after that count the occurrences of each and every words,then list those words and occurrences. The system will negotiate the words those occurrence value is equal or less than 3. After that Documents are automatically parsed and translated by the machine translation systems and the parse and output informationis stored as relevant identifiers of the document. This separation of the pre-processing stage from the actual information retrieval stage allows querying to be done at high speed.

The figure 2 illustrate the extraction of relevant identifiers from a document.
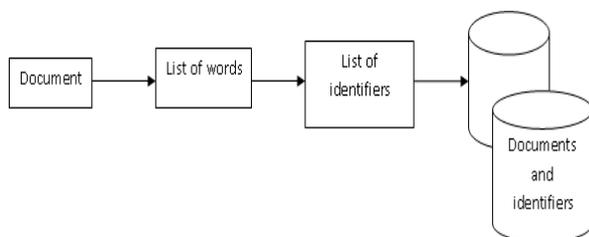
Fig. 2 Extraction of Identifiers of a Document

V. RETRIEVAL OF RELEVANT DOCUMENT

In information system the user can provide phrase to search. The application eliminates the stopwords from the phrase and extract the remaining words as keywords. The application will search those keywords in the database for keyword appearance in the repository. The system retrieves those related documents from the database as result

of provided phrase. The most common and well known application of information retrieval is the retrieval of text documents from the internet. With its recent growth, the internet is fast becoming the main media of communications for business and academic information. Thus it is essential to be able to tap the right document from this vast of information. This is in fact, one of the main pushing force for the development of information retrieval. To date, many relatively successful systems have been developed.
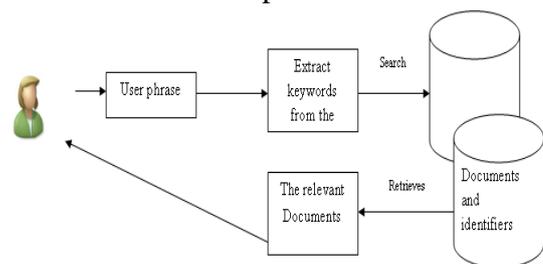
Fig. 3 Retrieval of Relevant Document from the Database

VI. CONCLUSION

In information science there are lot of information are available for reference. To find the exact one for our perusal is our problem, the semantic keyword search concept will reduce our manual work for search. The main advantage of this application is to avoid the irrelevant results and save our precious time.

VII. REFERENCES

[1] W. B. FRAKES "INTRODUCTION TO INFORMATION STORAGE AND RETRIEVAL SYSTEMS," SOFTWARE ENGINEERING GUILD, STERLING, VA 22170.

[2] Christopher D. Manning, PrabhakarRaghavan, HinrichSchutze, "Introduction to Information Retrieval," Cambridge University Press 978-0-521-86571-5.

[3]Norbert Fuhr, "Models in Information Retrieval," University of Dortmund, Germany.