Chapter 1

Introduction

1.1 GENERAL INTRODUCTION

The field of computer science emerging in analogous with database systems for several years is Information Retrieval (IR). It is concerned with processing of text databases. Thus, it deals with organization and retrieval of information from large document collection. Text databases comprises of huge collection of documents from various sources like news articles, research papers, books etc. (Singhal, 2001).

Information in text databanks are semi-structured i.e. it is neither structured nor unstructured. For instance, title, creator, date and category are structured data whereas abstract and information content are unstructured data in a document.

1.1.1 REPRESENTATION OF INFORMATION RETRIEVAL SYSTEM

The foremost intention of retrieval system is to examine the documents that are appropriate to the information requirement of the user from outsized document group. Every user has specific information requirement. User’s requirement will be the phrase to which answer is essential to execute certain task. The phrase given by the user has to be converted into the form appropriate for retrieval system. The information requirement of the user transformed into the form suitable for retrieval system is called query. Based on the user query, the retrieval of documents will take place from large document collection. The retrieved documents are ordered and presented to the user as an answer list. The general representation of the retrieval system is shown in Figure 1.1.
In this retrieval process, the performance of the system depends on how far the retrieved documents are relevant to the information need of the user.

1.1.2 APPLICATIONS

IR has been widely used in applications related to medical domain especially by the medical experts to practice Evidence-Based Medicine (EBM) (Frunza et al., 2011).

It is also enormously used in agricultural, spatial, scientific, industrial, governmental, engineering, business and in many other applications.

1.2 MEDICAL INFORMATION RETRIEVAL (MIR)

The IR system that deals with handling of vast measure of bio-medical documents is called MIR. It is totally centered around and connected to medical domain. It manages retrieval of medical information from vast bio-medical document accumulation.

People want to get reliable medical information from the available healthcare information systems to be aware of their health. The healthcare information systems satisfy the health requests of the people (Swartz, 2009).

1.2.1 PROBLEMS IN MIR

In a healthcare system, medical professionals need more reliable relevant information in a very timely manner. But the available sources of medical information...
fail to provide these due to abundant data in the medical domain. Not just the physicians but patients and their family members are increasingly using the net as a major source of advice regarding their illness, treatment options and procedures, dietary and health advice and disease prevention. However, little is known about the relevancy of medical advice obtained. Inaccurate or irrelevant medical information could potentially have a major detrimental impact. The reason can be due to the lack of knowledge of the user in understanding the information requirement or the interpretation of the user’s search request by the retrieval system. Figure 1.2 depicts the general problems that exist in MIR.

**Figure 1.2: General problems**

The general problems that lead to poor performance of MIR are as follows:

1. Improper search request by the user:
   - The factors leading to improper search request are as follows:
     i. Unawareness of required information
     ii. Inadequate formulation of user query
Due to lack of knowledge about the different keywords, they tend to input the symptoms, reactions instead of the disease itself and long lines of information need which lead to irrelevant information retrieval.

2. Misinterpretation of search request:
   Due to the misinterpretation of search request and huge availability of data, the retrieval system faces the following problems:
   i. Inadequate retrieval of information
   ii. Lack of relevant retrieved information

3. Medical errors:
   Due to insufficient relevant data, medical errors may occur. For example, it may lead to wrong disease diagnosis, treatment procedure and so on.

1.3 TASKS INVOLVED IN RETRIEVAL PROCESS

There are many tasks involved in the retrieval process. This section explains about these tasks.

1.3.1 KEYWORD EXTRACTION

The keyword extraction is an essential task involved in the retrieval process which determines the performance of the retrieval system. The process of recognizing the most suitable terms in the document is called keyword extraction. A good IR system must have a collection of beneficial keywords of corresponding field (Coursey et al., 2008). Annotation generation of keywords improve efficient retrieval of documents as they give auxiliary source of facts to retrieve the documents accurately (Frommholz, 2007; Matsuo and Ishizuka, 2004).
1.3.2 DOCUMENT INDEXING

Indexing is the process of associating documents in the corpus with various search terms (keywords). It shrinks the time spent in encumbered information and also creates the internal representation for documents and keywords.

This internal representation can have set of useful weighted keywords for each document as given below:

\[ \text{Doc}_1 \rightarrow \{(\text{term}_1, \text{wt}_1), (\text{term}_2, \text{wt}_2), \ldots \} \]
\[ \text{Doc}_2 \rightarrow \{(\text{term}_1, \text{wt}_1), (\text{term}_2, \text{wt}_2), \ldots \} \]
\[ \ldots \]
\[ \ldots \]
\[ \ldots \]
\[ \text{Doc}_n \rightarrow \{(\text{term}_1, \text{wt}_1), (\text{term}_2, \text{wt}_2), \ldots \} \]

Where,
‘Doc’ represents documents, ‘term’ represents keywords, ‘wt’ represents its weight and ‘n’ represents the total number of documents.

Example:

\[ \text{Doc}_1 \rightarrow \{(\text{heart}, 0.2), (\text{infection}, 0.3), \ldots \} \]
\[ \text{Doc}_2 \rightarrow \{(\text{kidney}, 0.6), (\text{blood}, 0.4), \ldots \} \]
\[ \text{Doc}_3 \rightarrow \{(\text{heart}, 0.4), (\text{kidney}, 0.3), \ldots \} \]
\[ \text{Doc}_4 \rightarrow \{(\text{heart}, 0.5), (\text{kidney}, 0.2), (\text{blood}, 0.7), \ldots \} \]

The internal representation can also be in the form of inverted file. This representation is used during the retrieval process for achieving better efficiency. For each keyword, it contains references to documents as given below:

\[ \text{term}_1 \rightarrow \{(\text{Doc}_1, \text{wt}_1), (\text{Doc}_2, \text{wt}_2), \ldots \} \]
\[ \text{term}_2 \rightarrow \{(\text{Doc}_1, \text{wt}_1), (\text{Doc}_2, \text{wt}_2), \ldots \} \]
\[ \ldots \]
\[ \ldots \]
\[ \ldots \]
\[ \text{term}_n \rightarrow \{(\text{Doc}_1, \text{wt}_1), (\text{Doc}_2, \text{wt}_2), \ldots \} \]
Here, ‘n’ represents the number of keywords.

Example:

\[
\begin{align*}
\text{heart} & \rightarrow \{(\text{Doc}_1,0.2), (\text{Doc}_3,0.4), (\text{Doc}_4,0.5), \ldots\} \\
\text{kidney} & \rightarrow \{(\text{Doc}_2,0.6), (\text{Doc}_3,0.3), (\text{Doc}_4,0.2), \ldots\} \\
\text{infection} & \rightarrow \{(\text{Doc}_1,0.3), \ldots\} \\
\text{blood} & \rightarrow \{(\text{Doc}_2,0.4), (\text{Doc}_4,0.7), \ldots\}
\end{align*}
\]

In the above examples, Doc\(_1\), Doc\(_2\), Doc\(_3\) and Doc\(_4\) represent the documents. The keywords are ‘heart’, ‘infection’, ‘kidney’, ‘blood’. Each keyword has its own weight with respect to the document.

1.3.3 DOCUMENT RANKING

Document ranking deals with ordering of all retrieved documents based on their relevance score to user query. In response to user query, ranked list of documents are returned by IR system. Ranking function is used to compute the relevance score between retrieved documents and user query. It is done by matching the keywords in the user query with those in the retrieved documents (Chou et al., 2008). Each retrieved document is scored and ranked based on how well it matches with the user query.

1.3.4 DOCUMENT SUMMARIZATION

Document summarization is an IR task that deals with constructing lessened form of the document (Gupta and Lehal, 2010). The summary made must safeguard the information content of the document and its meaning (Pimpalshende, 2013). The summarization method can be categorized based on the nature of text obtained, kind of detail delivered, content offered, quantity of input documents and linguistic (Munot and Govilkar, 2014; Mendoza et al., 2014). There are two types of summarizers under each category which is depicted in Table 1.1.
Table 1.1: Various categories and types of summarization methods

<table>
<thead>
<tr>
<th>Category</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nature of text obtained</td>
<td>Extractive</td>
<td>Summary is formed by mining key sentences from original document.</td>
</tr>
<tr>
<td></td>
<td>Abstractive</td>
<td>Summary is created by rephrasing the sentences in the original document. It is done by understanding the entire document content.</td>
</tr>
<tr>
<td>Kind of detail delivered</td>
<td>Indicative</td>
<td>Summary provides main idea of the original document.</td>
</tr>
<tr>
<td></td>
<td>Informative</td>
<td>Summary is produced by minimizing the length of the original document without changing its meaning.</td>
</tr>
<tr>
<td>Kind of content offered</td>
<td>Generic</td>
<td>Summary is not based on the user interest. It gives the same level of importance to all sentences when producing a summary.</td>
</tr>
<tr>
<td></td>
<td>Query-based</td>
<td>When producing a summary based on user interest or query, it gives importance to certain sentences.</td>
</tr>
<tr>
<td>Quantity of input documents</td>
<td>Single-document</td>
<td>Summary is produced for one document at a time.</td>
</tr>
<tr>
<td></td>
<td>Multi-document</td>
<td>Summary is produced for collection of related documents.</td>
</tr>
<tr>
<td>Linguistic</td>
<td>Mono-lingual</td>
<td>Summary is created for documents written in one specific language.</td>
</tr>
<tr>
<td></td>
<td>Multi-lingual</td>
<td>Summary is created for documents written in different language.</td>
</tr>
</tbody>
</table>
The summary can be genre-specific or domain-independent. In the former, documents belonging to one specific domain are considered and in the latter, documents belonging to any domain can be considered for summarization. These summarization approaches comprehends the bits of knowledge of information in the document, incase if a document does not contain the author written summary (Prakash et al., 2014; Shri and Subramaniyaswamy, 2015).

1.4 MEDICAL DOCUMENT CORPUS FOR IR EXPERIMENTS

The richest and most used source of information in medical domain is MEDLINE\(^1\) (MEDical literature analysis and retrieval system onLINE). It is the database of life science related articles. The process of identifying and disseminating relevant reliable information becomes a very difficult task as all research discoveries come and enter this repository at very high rate.

OHSUMED and PMC (PubMed Central) are the two subsets of MEDLINE that are commonly used for IR system performance evaluation (Hliaoutakis et al., 2009).

1.4.1 OHSUMED

OHSUMED is the standard corpus used for benchmarking IR systems evaluation which consists of collection of MEDLINE document abstracts from 270 medical journals from 1987 to 1991. It is the standard corpus for abstract indexing experiment.

1.4.2 PMC

PMC is the standard full document corpus used for benchmarking IR systems evaluation which consists of collection of MEDLINE documents. It is the standard corpus for full document IR indexing experiments.

\(^1\) www.nlm.nih.gov/bsd/pmresources.html
1.5 RESOURCES FOR SEMANTIC BASED IR

In this section, the various generic and medical domain-specific resources available for semantic based IR have been discussed. Among these, domain-specific ontologies are widely used in the literature.

1.5.1 WIKIPEDIA

The free online fact file maintained by outsized quantity of volunteers collaboratively is Wikipedia (Paci et al., 2010). It acts as a resource for IR in recent years. It contains inter-linked textual information, manually defined concepts and semantic relations. Therefore, the use of Wikipedia (Milne and Witten, 2013) can provide not only facts, but also exact semantic information. It is ever growing resource which is considered to be the potential source for Natural Language Processing (NLP), Knowledge Management (KM), Data Mining (DM) and many other research areas. It contains articles, redirects, categories and disambiguation pages for efficient retrieval of required information. The open source software system that is used by researchers and developers to integrate Wikipedia’s rich semantics into their own applications is Wikipedia Miner.

1.5.2 WORDNET

Ontologies are resources that allow researchers to extract semantic based information (Uschold and Gruninger, 1996). The general ontology developed at Princeton University is WordNet (Shubhangi, 2014). It is an electronic lexical database which is a resource for NLP and IR applications. It consists of set of synonyms called as Synset that represents one particular concept. Synset formation is done based on the synonymy and polysemy. Synonymy means many words having one meaning whereas polysemy means one word having many meanings.
1.5.3 UNIFIED MEDICAL LANGUAGE SYSTEM (UMLS)

In medical domain, various domain-specific ontologies are available which allows medical professionals and researchers to process bio-medical data from countless sources. Among those, the widely used medical domain-specific thesaurus is UMLS which is established by National Library of Medicine (NLM). It captures synonyms, variants, acronyms of specific medical term but does not look into the context.

UMLS integrates many subdomains related to medical domain such as clinical repositories, genetic knowledge bases, bio-medical literature, genome annotations, anatomy and many other subdomains which are depicted in Figure 1.3. MetaMap allows researchers and developers to use UMLS resource in their own applications (Milian et al., 2010).

\[\text{Figure 1.3: Various subdomains integrated in the UMLS}\]

\[\text{UMLS} \rightarrow \text{SNOMED} \rightarrow \text{Clinical Repositories} \]
\[\text{OMIM} \rightarrow \text{Genetic Knowledge Bases} \]
\[\text{MeSH} \rightarrow \text{Bio-medical Literature} \]
\[\text{GO} \rightarrow \text{Genome Annotations} \]
\[\text{UWDA} \rightarrow \text{Anatomy} \]
\[\ldots \ldots \rightarrow \text{Other subdomains}\]
1.5.4 GENERALISED ARCHITECTURE FOR LANGUAGES, ENCYCLOPAEDIAS AND NOMENCLATURES IN MEDICINE (GALEN)

The medical domain-specific ontology GALEN is the European Union project undertaken from 1992 – 1999 which provides reusable clinical terminology resources. It was designed before populating the ontology by defining the representation formalism and top level knowledge. OpenGALEN provides access to GALEN resource.

1.5.5 SYSTEMATIZED NOMENCLATURE OF MEDICINE CLINICAL TERMS (SNOMED CT)

SNOMED CT is another medical domain-specific dictionary established by the college of American Pathologists. It is the clinical repository which is now available as a part of UMLS. Therefore, it is widely used in medical information system.

1.6 MOTIVATION

The world population has increased by many folds. There have been a number of improvements in the various domains of medicine and clinical sciences. However, with the sophisticated standards of cure are knotted the highly resistant diseases, complexity of operations, the side effects and the malfunctions requiring big data to be processed.

Huge data available has made it impossible for a human to be aware of all the pros and cons however the need for the same cannot be sacrificed. Certain wide-spread diseases such as Ebola, Dengue Fever, Swine Flu whose symptoms are unnoticeable in the initial stages have caused havoc in the recent past.

In medical domain, EBM is practiced today (Frunza et al., 2011). Medical practitioners are judged based on their updated knowledge and not just their years of experience. Medical experts are expected to have updated knowledge about various treatments for a particular disease and all new discoveries related to that disease and treatment in order to identify the side effects and other malfunctions.
In this scenario, medical professionals like doctors, researchers and students demand faster and proper access to reliable relevant medical information. However, available medical information sources fail to provide these information in a timely manner, which result in wrong diagnosis or treatment in turn leading to many medical errors.

With the advent of EBM in the field of medicine, this thesis can prove to be a useful tool to design MIR system that enables effective retrieval of relevant reliable medical information from large medical document collection in order to aid the medical professionals to afford better service to the society.

1.7 AIM AND OBJECTIVES OF THE RESEARCH WORK

This research work is aimed at proposing and developing methods to achieve good performance improvement in the relevant reliable medical document retrieval by concentrating on the various tasks like keyword extraction, annotation of keywords, document indexing, document ranking and document summarization that are involved in the retrieval process.

The research work reported has the following objectives.

1. To generate context-aware keywords from medical documents using Wikipedia.
2. To annotate these context-aware keywords using Wikipedia and medical ontology and to use these annotated keywords as indexing keywords for document indexing.
3. To conduct detailed investigation on how to use Wikipedia for improving the performance effectiveness of MIR.
4. To design a new highly robust hybrid ranking function for the purpose of relevance score computation between the retrieved documents and user query.
5. To develop single document summarizer that exploits domain-specific knowledge using domain’s generic features in order to help the medical professionals rapidly understand the article.
6. To produce highly informative summaries for the retrieved documents which do not have author written summary or abstract.

1.8 ORGANIZATION OF THE THESIS

The thesis of the research work consists of seven chapters which presents the preliminaries and contributions.

Chapter 2 presents the literature review that reveals the existing works related to Wikipedia and ontology based information extraction and medical document indexing. It is also devoted to provide the prior works related to evolutionary algorithm based ranking functions and text document summarization. In addition to this, it also emphasizes the observations made from literature review for research problem identification.

Chapter 3 portrays the annotation based context-aware indexing for effective MIR. This work concentrates on generating annotated medical keywords using Wikipedia and medical ontology to address upon the synonymy and polysemy issues that exist in the IR.

Chapter 4 introduces a new hybrid neuro-fuzzy system based ranking function and its application to effective MIR. The hybrid neuro-fuzzy system is a highly robust uncertainty based model that is suitable for real time applications. This work concentrates on designing highly robust ranking function for relevance score computation and thereby captures the vagueness and uncertainty present in natural language text.

Chapter 5 presents the single medical document summarizer called Cue-Summarizer that exploits medical domain-specific information by utilizing cue-word feature of sentences. It assists the general population with defining whether the retrieved document is beneficial for detailed study.

Chapter 6 describes BSF-Summarizer, another single medical document summarizer. This summarizer exploits medical domain-specific information by using few best sentence features of medical domain and thereby produces highly acceptable quality summary for human use.
Finally, chapter 7 emphasizes the concluding remarks and scope of future work. Followed by conclusion and future work, list of references and publications have been presented.