INFORMATION RETRIEVAL USING NATURAL LANGUAGE PROCESSING

[1]Prof. S. A. Alatekar, [2]Prof. A. C. Suthar, [2]H. B. Bhadka
[1]Head, Department of Computer Engineering, ICRE, Gargoti, Dist. Kolhapur (MS).
salatekar@yahoo.co.in
[2]C. U. Shah College Of Engg. & Tech., Wadhwan. Gujarat (India).

ABSTRACT

People throughout society have to assimilate and manipulate increasing amounts of information, typically by interacting with sophisticated computer systems. It is essential that these systems have interfaces, which maximize the accessibility and usefulness of the underlying information.

There are three main areas in this field: the construction of integrated user interfaces using natural language for input and/or output to exact information from knowledge bases; the development of linguistically general but computationally tractable representations of meaning; and the mathematical study of the properties of linguistic notations and rule systems. The evolution of digital libraries and the Internet has dramatically transformed the processing, storage and retrieval of information. Even when there is no shortage of textual materials on a particular topic, procedures for indexing or extracting the knowledge or conceptual information contained in them can be lacking.

Recently developed information retrieval technologies are based on the concept of a vector space. Data are modeled as a matrix, and user's query of the database is represented as a vector. Relevant documents in the database are then identified via vector operations. Orthogonal factorizations of the matrix provide mechanisms for handling uncertainty in the database itself.

This paper deals with the concept of information retrieval to the web search to enhance the search process. The novel search approach interprets context in its natural setting. Searches are processed in the context of the query given, allowing more accurate and relevant search results. Natural language processing techniques have been used to assist an IR (Information Retrieval) system in selecting appropriate indexing of words and phrases, concepts and relations to sharpen the word based search.

1. Introduction

Web search engines work by storing information about a large number of web pages, which they retrieve from the WWW itself. These pages are retrieved by a web crawler, an automated web browser, which follows every link it sees. The contents of each page are then analyzed to determine how it should be indexed (for example, words are extracted from the titles, headings, or special fields called Meta tags). Data about web pages is stored in an index database for use in later queries. When a user comes to the search engine and makes a query, typically by giving key words, the engine looks up the index and provides a listing of best-matching web pages according to its criteria, usually with a short summary containing the document's title and sometimes parts of the text.

The proposed search enhancement tool has utilized the notion of *context*, making it one of the most abused terms in the field, referring to a diverse range of ideas from domain-specific search engines to personalization. A novel search approach interprets context in its most natural setting. The basic premise underlying the approach is that searches should be processed in the *context of the query given*, allowing more accurate and relevant search results that better reflect the user's actual intensions. Guiding user's search by the context surrounding the query eliminates possible semantic ambiguity and vagueness. NLP techniques have been used to assist an IR system in selecting appropriate indexing terms, both words and phrases that could be deemed to stand for actual entities, concepts and relations and therefore sharpen the word-based search. In other words, we are aiming at semantically motivated concept based representation.

2. Natural language processing

Given the constantly increasing information overflow of the digital age, the importance of information retrieval has become critical. Web search is today one of the most challenging problems of the Internet, striving at providing users with search results most relevant to their information needs. Current research efforts are on context based systems, which are aimed at increasing coverage and relevance. The goal of the Natural Language Processing (NLP) is to design and build software that will analyze, understand, and generate languages that humans use naturally and to be able to address the computer as though addressing another person. IR System is a system, which is used for Information Retrieval tasks. This is concerned with the use of storage of the documents about the subject and uses it to retrieve the document relevant to the users query. This is getting the document from the source and getting the query. The IR systems are usually known as Search Engines and then we use them for the web we get the better results. NLP in information retrieval consists of a software program that facilitates a user in finding the information the user needs, features and functions required to manipulate information items. It is capable of storage, retrieval and maintenance of information.

2.1 Steps in Information Retrieval

Document Processing

Documents are input to the system. Information retrieval systems build an inverted file, or list of words in alphabetical order. Stop words are left out of this list. NLP systems create and store a formal representation of each sentence including the role each word plays and its relationship to other words in the sentence.

Query Processing

When a query comes in, it must be interpreted for the system. NLP systems identify the terms to search for and it may look for stems and singular and plural forms. It may also assign weights to each term.

Query Matching

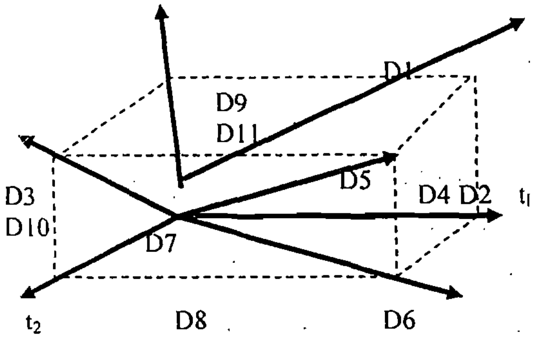
The interpreted query is matched against the inverted file and the knowledge base, if there is one.

Ranking and Sorting

Once all the candidate documents are selected that match the query, they are sorted by date, by field, or by how relevant the document is predicted to be to the query. NLP-based systems use the relevance ranking techniques. They rank the retrieved documents so that the top once returned match the query closest.

2.2 Vector Space Retrieval

Vector space is a mathematical model that is often used by Information Retrieval systems to determine how similar retrieved documents are to the user's query. In this, the document is represented as a vector of terms or keywords. We assume that there is no correlation between the keywords and they are independent of each other. For this we use the inverted index method to index the document in terms of keywords and their occurrences. We give the documents a unique ID so that every document, which has been indexed, has a unique identifier. We also remove the words which are in the document to indicate only grammatical information such as "a,an,the,is,was,were" etc. These are known as the stop words. We indicate the number of occurrences of the keywords along with the position of occurrence.



Then we create a matrix called Term – Document Matrix which gives the weight of the term in the document. This is accomplished by giving the values to the term document matrix by giving values according to some schemes.

(a) Boolean This just indicates that whether the term is found in the document or not. It gives the Boolean values to the matrix. It is easy to implement and can be used easily with Boolean queries like "and, or, not" with queries. This does not give the importance of the word and gives only presence and thus cannot distinguish between documents easily. We can use it as the most primitive weighting method.

(b) Term Frequency This indicates the occurrence frequency of the word. This gives the weight of the term as the number of times it appears in the document. This gives a very good local analysis method. It can retrieval has become critical. Web search is today one of the most challenging problems of the Internet, striving at providing users with search results most relevant to their information needs. Current research efforts are on context based systems, which are aimed at increasing coverage and relevance. The goal of the Natural Language Processing (NLP) is to design and build software that will analyze, understand, and generate languages that humans use naturally and to be able to address the computer as though addressing another person. IR System is a system, which is used for Information Retrieval tasks. This is concerned with the use of storage of the documents about the subject and uses it to retrieve the document relevant to the users query. This is getting the document from the source and getting the query. The IR systems are usually known as Search Engines and then we use them for the web we get the better results. NLP in information retrieval consists of a software program that facilitates a user in finding the information the user needs, features and functions required to manipulate information items. It is capable of storage, retrieval and maintenance of information.

2.1 Steps in Information Retrieval

Document Processing

Documents are input to the system. Information retrieval systems build an inverted file, or list of words in alphabetical order. Stop words are left out of this list. NLP systems create and store a formal representation of each sentence including the role each word plays and its relationship to other words in the sentence.

Query Processing

When a query comes in, it must be interpreted for the system. NLP systems identify the terms to search for and it may look for stems and singular and plural forms. It may also assign weights to each term.

Query Matching

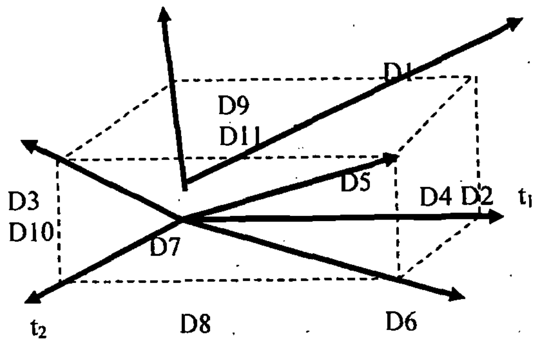
The interpreted query is matched against the inverted file and the knowledge base, if there is one.

Ranking and Sorting

Once all the candidate documents are selected that match the query, they are sorted by date, by field, or by how relevant the document is predicted to be to the query. NLP-based systems use the relevance ranking techniques. They rank the retrieved documents so that the top once returned match the query closest.

2.2 Vector Space Retrieval

Vector space is a mathematical model that is often used by Information Retrieval systems to determine how similar retrieved documents are to the user's query. In this, the document is represented as a vector of terms or keywords. We assume that there is no correlation between the keywords and they are independent of each other. For this we use the inverted index method to index the document in terms of keywords and their occurrences. We give the documents a unique ID so that every document, which has been indexed, has a unique identifier. We also remove the words which are in the document to indicate only grammatical information such as "a,an,the,is,was,were" etc. These are known as the stop words. We indicate the number of occurrences of the keywords along with the position of occurrence.



Then we create a matrix called Term – Document Matrix which gives the weight of the term in the document. This is accomplished by giving the values to the term document matrix by giving values according to some schemes.

(a) Boolean This just indicates that whether the term is found in the document or not. It gives the Boolean values to the matrix. It is easy to implement and can be used easily with Boolean queries like "and, or, not" with queries. This does not give the importance of the word and gives only presence and thus cannot distinguish between documents easily. We can use it as the most primitive weighting method.

(b) Term Frequency This indicates the occurrence frequency of the word. This gives the weight of the term as the number of times it appears in the document. This gives a very good local analysis method. It can distinguish between importances of terms in a document. This cannot find the importance of the word in the whole document collection.

(c) Inverse document Frequency This indicates the occurrence of the term with relation to the document collection. It is called inverted since it gives more weight to the rare word and gives less weight to the frequent words. It gives weight to the Distinguishing power of the word. This is mostly given as the value

Inverse Document Frequency (idf)=log10(n/dfj)

where n = No.r of documents in the collection.
Df = No of documents containing the keyword.

(d) TF * IDF:- In this we combine both the good properties of Tf and Idf by multiplying them to represent the weight in a better manner to the already available methods. We get the weight as the word, which occurs, frequently in a document and infrequently in the whole collection. This is the best method to give weights to vector representation of documents and this is the weight given to the terms in our Project.

$$W_{ik}=\frac{tf_{ik}\ \log(N/n_k)}{\sqrt{\sum^{t}_{k=1}(tf_{ik})^2[\log(N/n_k)]^2}}$$

3. New Search Model

3.1 Description

3.1.1 Crawler

The search engines use the indexing methods generally used for Information retrieval tasks. The process of getting the documents to index is known as the Crawling and the program is known as the crawler or spider.

3.1.2 Indexing

Then we use the indexing methods to make the set of documents to be easily retrievable by the keywords. There are a number or models to do the same.

3.1.3 Retrieval

We get the query from the user as a string and convert it into a document with the set of words and getting the weight as their Term Frequency of the document collection. Then we get the similarity between the document and the query as the Dot Product between the Query Vector and the Document vector. Find the minimal distance by getting the cosine of the angle between the two vectors. We can get this function for all the documents with the query and get the most similar documents to the query by sorting the results and getting the best N matches (In our case the value is 10). We can also get the concept matches by performing the query with the concepts given during the query.

$$\mathrm{Cos}(Q,D_2)=\frac{\sum_{j=1}^{t} w_{qj} \bullet w_{d\,ij}}{\sqrt{\sum_{j=1}^{t}(w_{qj})^2 \bullet \sum_{j=1}^{t}(w_{d\,ij})^2}}$$

3.2 Modular Description

This will provide an in-sight view of the various functions involved in this paper. This paper consists of the following modules.

- Searching
- Parsing
- Inverted Indexing
- Forming Term matrix
- Weight Document matrix (TF*IDF matrix)
- Finding Similarity.

3.2.1 Searching

After the query is given as input, the stop words from the query are removed. The result is used to search the HTML pages from the source based on Meta tags.

3.2.2 Parsing

Parsing is a process of removing the HTML TAGS from the source file. HTML TAGS begins with the symbol '<' and ends with '>'. In addition to those tags some other symbols namely delimiters like '!', '@', '#', '$' and '|' etc. are removed from the files.

3.2.3 Inverted Indexing

Inverted index is formed after parsing is performed. Inverted index consists of the term, the number of terms and the position of each term in every document. Inverted Indexes implemented by using the Linear Data Structure Arrays.

The Zeroth column consists of terms; the row consists of the number of words and the positions of each word in every document.

Eg: Emphasis 5 1.2 5.6 6.8 8.9 12.8

It denotes that the term "Emphasis" occurred in 5 positions and 1.2 denoted second position in the first document.

The output of this module is a file containing inverted index of the terms occurring in the HTML documents.

3.2.4 Forming Term Matrix

Term matrix gives the exact view of terms and the number of times that particular term occurs in each document. This matrix is also a two dimensional array having row as terms and column as documents.

3.2.5 Weight Document Matrix (tf*idf)

The Inverse Document Frequency of each term is found out by using the formula log (Ni/N), which is multiplied by Term Frequency Matrix. Each element in the Weight document matrix is indicating the weight of the each element in each document.

3.2.6 Finding Similarity

From the TF*IDF matrix the similarity between the query and the document is identified by taking query and document as two vectors and the Tf*IDF is considered as document vector and the invert document frequency of the query is taken as query vector.

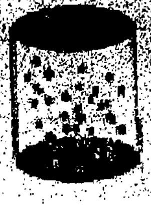
Finally the similarity values are sorted in the descending order and the higher value indicates that it has more similarity with that of the query. Each document has its own ID and files are identified by making use of these IDs. The files are displayed in the Similarity order.

4. Evaluation metrics

We can represent the evaluated measures as precision and recall. These are used to measure the IR systems to be better and get the precision and recall curves to get the maximum number of good documents.

$$\text{RECALL} = \frac{\text{\# relevant retrieved}}{\text{\# retrieved}}$$

$$\text{PRECISION} = \frac{\text{\# relevant retrieved}}{\text{\# retrieved in collection}}$$



The result is the retrieval of hyperlinks of the most relevant WebPages.

5. System Design

5.1 Input Design

The main objective of this is to provide a user friendly system. The user has to make minimum input as the whole process is automated. The design decision for handling input specifics how data are accepted for computer processing. User decides whether the data are entered directly or by a source document, such are the variable forms as how the data are transferred into the computer for processing. In the system the input is provided as a query for which the relevant documents are to be retrieved.

5.2 Output Design

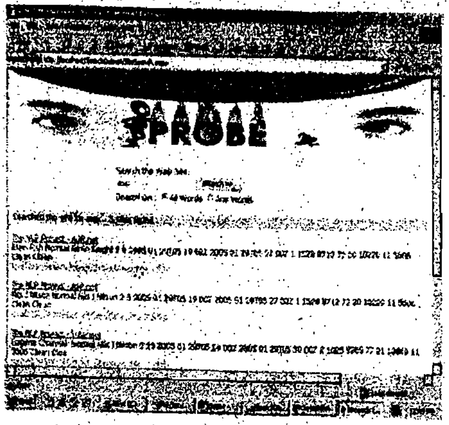
Output design of a system can be defined as the information being processed and then generated by the system in a specific format as specified by the user. In the system the output is a set of the most relevant documents.

6. System Implementation

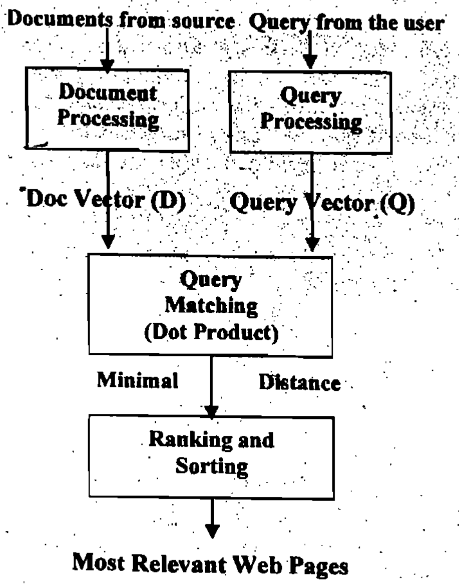
The implementation phase is one of the important phases in the evolution of the proposed system. If the implementation is not smooth and rise some problem, the user can doubt the usefulness of the system and reluctant to use it. Some of the issues that are to be dealt during the implementation are:-

- Drafting a plan of implementation.

- Checking the user site for a suitable implementation.
- Briefing the user on the advantages of the new system.



7. DATA FLOW DIAGRAM



8. Conclusion

NLP presents new tools for honing a search query so that it states our information need fully and then matches that query with an elaborate knowledge base built with NLP techniques. Techniques of NLP used here would work on the users query and retrieve highly relevant documents among the set of documents present in the source collection, to reduce users overhead and give specific related knowledge intact to the user.

The System does the job of finding the position of the keywords in the index. It can be improved to:

(i) Include the relative positions of the keywords to improve the Phrasal Matching of the keywords easily.

(ii) It can be found out that weight of the keyword can also be changed to indicate the part of speech of the sentence. Thus real information with very good precision is obtained.

This can be easily extended to information extraction, which can get the real information from the text from the content and the format of the web pages.

9. References

1. Strzalkowski, Tomek, Jose Perez – Carballo (1994). "Recent Developments in Natural Language Text Retrieval". Vol:7, Pg:123-136

2. Tomek Strzalkowski, Natural Language Information Retrieval, 1999.

3. Van RIJSBERGEN C.J, Information Retrieval, 2000.

4. Susan Gauch, Key Concept – "Incorporating Concept Matching in Search".

5. Robotstxt.com – "About web crawling".

6. Richard Vdovjak and Geert Jan Houben. Rdf-based architecture for semantic integration of heterogeneous information sources *International Workshop on Information Integration on the Web*, 2001.