Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 4, Issue 4; July-September, 2017, pp. 258-262
© Krishi Sanskriti Publications
http://www.krishisanskriti.org

Analysis and Applications of Text Retrieval

Sindhu S Pandya

Laxmi Institute of Commerce & Computer Application (BCA), Sarigam
Laxmi Vidyapeeth, Sarigam
E-mail: sindhubhilai@yahoo.com

Abstract—*Text Mining is the analysis of data contained in natural language text. Text Mining works by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analysed using traditional data mining techniques. This field of information retrieval was born in 1950's. Information retrieval is generally concerned with the searching and retrieving of knowledge- based information from database. Information retrieval techniques works basically on unstructured documents as the data stored in the text database are semi structured. In this paper we represent various techniques for information retrieval, by describing different indexing methods for reducing search space and searching techniques. We have briefed a bit on the latest information retrieval system of relational keyword search technique. Traditional keyword search scheme was having a drawback of huge memory consumption. Here we describe how IR techniques works together in the process of transferring information from the generator to the user, summarize the issues and report the results.*

Keywords*: Text Mining, Information retrieval, IR models, Searching techniques, Keyword Search, Applications*

1. INTRODUCTION

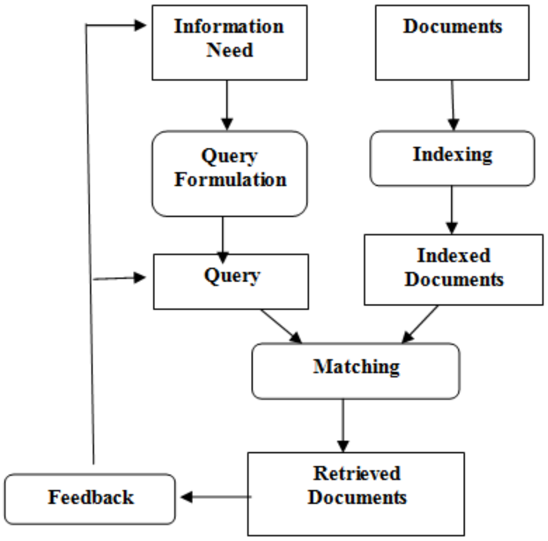
Text Mining is a variation on a field called data mining that tries to find interesting patterns from large databases. Text Mining or Information Retrieval is the process to retrieve data from dataset and transform it to the user in understandable form, so user easily gets that information.

Data stored in most text databases are semi structured data in that they are neither completely unstructured nor completely structured. Information retrieval is considered as a subfield of computer science that deals with the representation, storage and access of information. To achieve the main purpose of information retrieval, following processes has to be implemented:-

1. In Indexing process the documents are represented in summarized content form.
2. In filtering process all the stop words and common words are removed.
3. Searching is the main process of information retrieval

Following are the three basic processes an information retrieval system has to support:-

1. The representation of the content of the documents.
2. The representation of the user's information need.
3. The comparison of the two representations.



In the above figure, square boxes indicate data and rounded boxes indicate processes.

Representing the documents is called indexing process, which is done off-line, that is the end user is not directly involved. The process of representing the user's information need is referred to as query formulation process. The resulting representation is the query. Comparing the two representations is known as the matching process. The final result of this process is the retrieval of the required documents.

2. TEXT MINING

Text mining deals with the machine supported analysis of text. It uses techniques from information retrieval, information extraction, as well as natural language processing (NLP) and connects them with the algorithms and methods of KDD, data mining, machine learning and statistics to texts with the goal of finding useful patterns. For this purpose it is necessary to pre-process texts accordingly. Many authors, natural language processing or some simple pre-processing steps in order to extract data from texts. To the extracted data then data mining algorithms can be applied [2,3].

3. TECHNIQUES OF TEXT MINING

a. **Information Retrieval:** Information Retrieval is a field that has been developing in parallel with database systems for many years. Retrieval of information is basically focussed on large number of text-based documents. Since information retrieval and database systems each handle different kinds of data, some database system problems are usually not present in information retrieval systems, such as concurrency control, recovery, transaction management and updating. Some common information retrieval problems are unstructured documents, appropriate search based on keywords, and the notion of relevance. Due to the abundance of text information, information retrieval has found many applications. There exists many information retrieval systems, such as on-line library catalogue systems, on-line document management systems, and many web search engines. A information retrieval problem is to locate relevant documents in a document collection based on a user's query, which is often some keywords describing an information need. In such a search problem, a user takes the initiative to "pull" the relevant information out from the collection; this is most appropriate when a user has some adhoc information need, such as finding information to buy a used car. When a user has a ling-term information need, a retrieval system may also take the initiative to "push" any newly arrived information item to a user if the item is judged as being relevant to the user's information need such an information access process is called information filtering.

i. **Measures of Text Retrieval:** The set of documents relevant to a query be denoted as {Relevant}, and the set of documents retrieved be denoted as {Retrieved}. The set of documents that are both relevant and retrieved is denoted as {Relevant} ∩ {Retrieved}

There are two measures for accessing the quality of text retrieval.

Precision: It is the percentage of retrieved documents that are in fact relevant to the query.

$$\text{Precision} = \frac{\{\text{Relevant}\} \cap \{\text{Retrieved}\}}{\{\text{Retrieved}\}}$$

Recall: It is the percentage of documents that are relevant to the query and were, in fact, retrieved.

$$\text{Recall} = \frac{\{\text{Relevant}\} \cap \{\text{Retrieved}\}}{\{\text{Relevant}\}}$$

Recall measures the quantity of relevant results returned by a search, meanwhile precision is the measure of the quantity of the results returned. Recall is the ratio of the relevant results returned divided by all relevant results. Precision is the number of relevant results returned divided by the total no of results returned.

ii. Information Retrieval Models: An Information Retrieval model specifies the details of the document representations, the query representation and the retrieval functionality.

The fundamental IR models can be classified into:

- Boolean Model
- Vector Space Model
- Probabilistic Model
- Inference Network Model

Boolean Model: The Boolean Model is the first model of Information Retrieval. The model can be explained by thinking of a query term as unambiguous definition of a set of documents. Using the operators of George Boole's mathematical logic, query terms and their corresponding sets of documents can be combined to form new sets of documents. The Boolean Model allows for the use of operators of Boolean algebra, AND, OR and NOT, for query formulation but has one major disadvantage; a Boolean system is not able to rank the returned list of documents. In the Boolean Model, a document is associated with a set of keywords. Queries are also expressions of keywords separated by AND, OR or NOT. The retrieval function treats a document as either relevant or irrelevant.

Vector Space Model: VSM is an algebraic model for representing text documents, as vectors of identifiers, such as, index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings. Its first use was in the SMART information retrieval system.

Documents and queries are represented as vectors:

$d_j = (w_{1,j}, w_{2,j}, w_{3,j},\text{-------------}w_{t,j})$

$q = (w_{1,q}, w_{2,q}, w_{3,q},\text{-------------}w_{t,q})$

Each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero. Vector operations can be used to compare documents with queries.

VSM has the following advantage over the standard Boolean model:

1. Term weights not binary.
2. Allows computing a continuous degree of similarities between queries and documents.
3. Allows ranking documents according to their possible relevance.

Following are the limitations of VSM:

1. Long documents are poorly represented.
2. Search keywords must precisely with similar context but different term vocabulary won't be associated, resulting in a "false negative model".

Probabilistic Model: Probabilistic Relevance Model was devised by Robertson and Jones as a framework probabilistic model to come. It is a formalism of information retrieval useful to derive ranking functions used search engines and web search engines in order to rank matching documents according to their relevance in a given search query.

It makes an estimation of the probability of finding if a document d_j is relevant to a query q. This model assumes that this probability of relevance depends on the query and document representations. It assumes that there is a portion of all documents that is preferred by the user as the answer set for query q. Such an ideal answer set is called R and should maximize the overall probability of relevance to that user. The prediction is that the documents in this set R are relevant to the query, while which are not present in the set are non-relevant.

Inference Network Model: In this model, document retrieval is modelled as an inference process in an inference network. In the simplest implementation of this model, a document instantiates a term with certain strength, and the credit from multiple terms is accumulated given a query to compute the equivalent of a numeric score for the document. The strength of the instantiation of a term for the document can be considered as the weight of the term in the document, and documents ranking in the simplest form of this model**.**

b. Text Indexing: When dealing with a small number of documents, it is possible for the full-text-search engine to directly scan the contents of the documents with each query, a strategy called "Serial Scanning".

When the number of documents to search is large, then the problem of full-text search is divided into two tasks: Indexing and Searching.

The Indexing stage will scan the text of all documents and build a list of search terms. In the Searching stage, when performing a specific query, only the index is referenced, rather than the text of original documents.

Indexing Techniques:

a) Signature File: In Signature file method each documents yields a bit strong using hashing on its words and superimposed coding. The resulting document signatures are used sequentially in a separate file called signature file, which is much smaller than the original file, and can be searched much faster.

b) Inversion Indices: Each document can be represented by a list of keywords which describe the contents of the document for retrieval purposes. Fast retrieval can be achieved if we invert on those keywords. The keywords are stored in the index file for each keyword we maintain list of pointers to the qualifying documents in the postings file.

c. Information Extraction: Information extraction directly with text mining process by extracting useful information from texts. Information extraction deals with the extraction of specified entities, events and relationships from unrestricted text sources.

The goal is to find specific data in natural language texts. Therefore the IE task is defined by its input and its extraction target. The input can be unstructured documents like free texts that are written in natural language or the semi-structured documents that are pervasive on the web, such as tables and enumerated lists. Using IE approach events, facts and entities are extracted and stored into a structured database.

IE are developed using following 3 steps:

i. Text Pre-Processing- Level ranges from text segmentation into sentences and sentences into tokens and tokens into full syntactic analysis.

ii. Rule Selection – the extraction rules are associated with triggers, the text is scanned to identify the triggering items and the corresponding rules are selected.

iii. Rule Application – It checks the conditions of the selected rules and fill in the form according to the conclusion of the matching rules.

a. Natural Language Processing: NLP is a technology that covers with Natural Language Generation(NLG) and Natural Language Understanding(NLU). NLG uses some level of underlying linguistic representation of text, to make sure that the generated text is grammatically correct and fluent. NLG systems include a syntactic realise to ensure that grammatical rules are obeyed, and text planner to decide how to arrange sentences, paragraphs and other parts coherently. NLU is a system that computes the meaning representation. NLU consists of at least one of the following components: tokenization, morphological or lexical analysis, syntactic analysis and semantic analysis.

4. SEARCHING TECHNIQUES

There are various searching algorithms including Linear Search, Binary Search, Brute Force Search etc.

1) The Linear Search algorithm is a method of finding a particular element or keyword from a list or array that checks every element in the list, one at a time and in sequence. Linear search is a simplest search algorithm also called sequential search, one of the most important drawbacks of linear search is slow searching speed in ordered list.
2) Binary Search algorithm, finds specified position of the element by using the key value within a sorted array. In each step, the algorithm compares the search key value with the key value of the middle element of the array. If the keys match, then a matching element has been found and its index, or position is returned. Otherwise, if the search key is less than the middle element key, then the algorithm repeats its action on the sub-array to the left of the middle element or, if the search key is greater, on the sub-array to the right.
3) Brute Force search is general purpose solving technique that consists of systematically enumerating all possible candidates for the solution and checking whether each candidate satisfies the problem's statement. It is simple to implement and always finds a solution if it exists.

Keyword Search Technique

Keyword Search is the most famous information discovery technique because the user does not need to know either a query language or the underlying structure of the data. It can be implemented on both structured and semi-structured databases; also it is possible on graph structure which combines relational, HTML and XML data.

1) **Schema Based Approaches**: It supports keyword search over relational databases using execution of SQL commands [4]. These techniques are combination of vertices and edges including tuples and keys. These are some techniques are existed for schema based approaches:

   a. DISCOVER- Techniques that multiple information retrieval approaches. It allows its user to issue keyword queries without any knowledge of the database schema. DISCOVER returns qualified joining network of tuples, which is the set of tuples that are associated because they join on their primary and foreign keys. DISCOVER uses static optimization.

   b. SPARK- With the increasing of the text data stored in relational databases, there are increase a demand for RDBMS to support keyword query search on text data. This techniques focus on effectiveness and efficiency of keyword query search.

2) **Graph Based Approaches**: It assumes that the database is modelled as a weighted graph where the weights of the edges indicate the importance of relationships. s

   a. BANKS- It enables user to extract information in a simple manner without any knowledge of schema. BANKS algorithm is an efficient heuristics algorithm for finding and ranking query results. BANKS focus on browsing and keyword searching.

   b. BLINKS- In query processing over graph-structured is a top-k keyword search query on a graph finds the top answered according to some ranking criteria.

5. APPLICATIONS

Information retrieval systems were firstly developed to help to manage the huge amount of information. Many universities, corporate and public libraries now use information retrieval systems to provide access to books, journals and other documents.

1) Digital Library – A Digital library in which collections are stored in digital formats and accessible by computers. The digital content may be stored locally, or accessed remotely via computer networks. A digital library is a type of information retrieval system.
2) Search Engines – A search engine is one of the most practical applications of information retrieval techniques to large scale text collections. Web search engines are best-known examples, but many other searches exist like Desktop search, Enterprise search, mobile search and social search.
3) Media Search – An image retrieval system is a computer system for browsing, searching and retrieving images from a large database of digital images.

REFERENCES

[1] R.Sagayam, S,Srinivasan, S.Roshni, "ASurvey of Text Mining: Retrieval and Indexing Techniques", International Journal of Computational Engineering Research, Volume:2, Issue:5, ISSN:2250-3005, September 2012

[2] V.Nahm and R.Mooney, "Text Mining with Information Extraction", In proceedings of AAAI 2002 Spring Symposium on Mining answers from texts and knowledge bases, 2002

[3] R.Gaizauskas, "An information extraction perspective on text mining: Tasks, technologies and prototype applications", http://www.itri.bton.ac.uk/projects/euromap/TextMiningevent/Rob_G.aizauskas.pdf, 2003

[4] A.Baid, I.Rae, J.Li, A.Doan and J.Naughton, “Toward scalable keyword search over Relational Data”, Proceedings of the VLDB endowment, Volume:3, Issue:1, 2010

[5] S.Cohen, J.Mamou, Y.Kanza and Y.Sagiv, Xsearch: A Semantic search engine for XML, in VLDB, 2011

[6] W.Webber, “Evaluating the effectiveness of keyword search”, IEEE data engineering Bulletin, Volume:3, Issue:1, 2010

[7] G.Salton and M.J. McGill, “editors, Introduction to Modern Information Retrieval”, McGraw-Hill, 1983