A Review On Recent Trends and Developments In Web And Document Based Information Retrieval

Dr. Shyamal Tanna , Shardul M Upadhyaya
Department of Computer Engineering
L.J. Institute of Engineering and Technology, Ahmedabad, India

***Abstract*— Information retrieval is method of extracting useful information form big data available over internet. Data available over internet has semi-structured in nature and available to user as collection of large set of documents.to retrieve useful information from this semi structured data base user query based indexing approach is used which include extraction of relevant keyword and indexing them as per terms of relevance to the user query. In this paper we survey various extraction and indexing methods available in information retrieval system in web or in document management system. The goal of this review paper is provide insight on some latest and efficient approaches in field of information retrieval system.**

***Index Terms*— Keyword Extraction, Information Retrieval, Automatic Indexation, Vector Space Model (VSM), singular value decomposition (SVD), Recommendation Systems.**

I. INTRODUCTION

Text databases consist of huge collection of documents. They collect these information from several sources such as news articles, books, digital libraries, e-mail messages, web pages, etc. Every Day Increase in the amount of information, the text databases are growing rapidly. The data is semi structured Document contain structured fields title, author, publishing date, etc. unstructured text components abstract and contents. To know what could be in the documents:- formulate effective queries for analyzing and extracting useful information from the data. Require tools to Compare the documents and rank their importance and relevance. Information retrieval deals with the retrieval of information from a large number of text-based documents. Examples of information retrieval system include Online Library catalogue system, Online Document Management Systems, Web Search Systems etc. Information Filtering. Recommender Systems. Note: Some of the database systems are not usually present in information retrieval systems because both handle different kinds of data. Main Parts Of information Retrieval System:- Information Filtering. Recommender Systems. The main problem in an information retrieval system: - locate relevant documents in a document collection based on a user's query. This kind of user's query consists of some keywords describing an information need. [12].

Before moving to survey knowledge of frequently used classic textual similarity measures is important. Various textual similarity measures used in information retrieval system to calculate the similarity between two documents.

A. TEXTUAL SIMILARTY MEASURES

Textual similarity measures are used to calculate the similarity between two documents. For this purpose, documents are represented as vectors in the space of terms, i.e. a term document matrix. In Information Retrieval (IR) this way of representing documents is also known as Vector Space Model (VSM). For focused crawling, the input topic and each web page, or some parts of it, are considered as documents to be shown as vectors. During next subsections, we explain TF-IDF and LSI measures used as the intelligent and deciding parts of focused crawlers.

1) Term Frequency-Inverse Document Frequency: TF-IDF is the most common weighting measure to assign an importance value to each of terms that occurs in a document. It is the product of Term Frequency (TF) and Inverse Document Frequency (IDF). TF measures the frequency of a term inside a document and IDF measures inverse frequency of documents containing that term. The tf-idf weight is composed by two terms: the first computes the normalized Term Frequency (TF), aka. The number

of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

TF: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

TF (t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as ”is”, ”of”, and ”that”, may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following. [10]

$$IDF(t) \models \log_e(\frac{Totalnumberofdocuments}{Numberofdocumentswithtermtinit})$$

2) Latent semantic indexing:

Latent semantic indexing (LSI) is an indexing and retrieval method that uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts. LSI is also an application of correspondence analysis, a multivariate statistical technique developed by Jean-Paul Benzcri in the early 1970s, to a contingency table built from word counts in documents. This method is called Latent Semantic Indexing because of its ability to correlate semantically related terms that are latent in a collection of text, it was first applied to text at Bellcore in the late 1980s. The method, also called latent semantic analysis (LSA), uncovers the underlying latent semantic structure in the usage of words in a body of text and how it can be used to extract the meaning of the text in response to user queries, commonly referred to as concept searches. Queries, or concept searches, against a set of documents that have undergone LSI will return results that are conceptually similar in meaning to the search criteria even if the results dont share a specific word or words with the search criteria. [11]

B. Commonly Used Performance Measures in Information Retrieval:

There are three fundamental measures for assessing the quality of text retrieval Precision Recall F-score Precision is the percentage of retrieved documents that are in fact relevant to the query.

Precision is the percentage of retrieved documents that are in fact relevant to the query. Precision can be defined as

$$Precision = |Relevant \cap Retrieved|/|Retrieved|$$

Recall is the percentage of documents that are relevant to the query and were in fact retrieved. Recall is defined as

$$Recall = |Relevant \cap Retrieved|/|Relevant|$$

F-score is the commonly used trade-off. The information retrieval system often needs to trade-off for precision or vice versa.

F-score is defined as harmonic mean of recall or precision as follows

$$F - score = recallxprecision/(recall + precision)/2.$$

[12]

II. RELATED WORK

In A Novel Statistical and Linguistic Features Based Technique for Keyword Extraction Gupta, A. et al. proposed the statistical and language model based keyword extraction which overcomes the shortcomings existing solutions, require either training models or domain specific information for automatic keyword extraction. This Approach works on an individual document without any previous parameter adjustment and takes full advantage of all the features of the document to extract the keyword. The extracted keywords can than assist in domain specific indexing. This system consists of two major subcomponents: Keyword Extractor Module and Domain Extractor Module. Keyword Extraction module takes input from document collection. These *input HTML pages are analyzed for extracting significant keywords from the HTML pages. The candidate keywords are supplied to Domain Extraction module for matching with ontological constructs to determine the domain of the corresponding webpage. The web page along with specified domain and significant keywords is then stored in domain specific webpage repository. The simplicity and efficiency of the proposed algorithm makes it applicable for a wide range of documents and collections and may be useful in any application where there is a need of representing the content of a web page within small set of keywords. [1]*

In Low Frequency Keyword Extraction with Sentiment Classification and Cyberbully Detection Using Fuzzy Logic Technique J.I.Sheeba et al. proposed framework to detect the online social cruelty attack from meeting transcripts like twitter, blog and face book also identify the low frequency keywords. Cyberbullying is a social aggressive and it has powerful negative effects on individuals, specifically adolescents. Proposed framework will detect both Implicit and explicit expressions and it will classify the Positive, Negative and Neutral words and also to identify the topic of the particular meeting transcripts like twitter, blog and face book also identify the low frequency keywords. In this proposed framework it additionally to detect the online social cruelty attack from meeting transcripts like twitter, blog and face book also identify the low frequency keywords And Some additional features improved by using fuzzy logic technique On results. In this task it includes 3 Steps A. Extracting keywords from the Transcripts using fuzzy logic technique B. Detecting Implicit and Explicit expressions of words from the Transcripts using fuzzy logic technique C. Detecting Cyberbully words from the Transcripts using fuzzy logic technique. Cyberbullying is a social aggressive and it has powerful negative effects on individuals, specifically adolescents so it used to detect cyber bullying attack. [2]

In Keywords Extraction for Automatic Indexing of e- Learning Resources Hendez, M. et al. propose an approach to help the indexing operation. This approach consists in automatically extracting a set of relevant terms describing the educational content of a resource it is based on the TFIDF algorithm, the usage of a domain lexicon and exploits the structure of educational documents In two principal steps: (1) learning resources pre-processing; (2) Keywords extraction and ranking It was applied to facilitate the searching and retrieving from online educational resources By Performing automatic indexing. [3]

In Keyword Extraction for Mining Meaningful Learning- Contents on the Web Using Wikipedia Toyota, T. Res. et al. present a method for extracting appropriate keywords to identify meaningful learning contents on the Web using Wikipedia. They use PF-IBF method to calculate degrees of association between the articles and the keywords and improve the accuracy of learning-related keyword extraction. Firstly, they calculated the relevance of articles to learning unit names based on PF-IBF from Wikipedia category structure and internal link data, thus obtaining related articles. Then, using course of study data, they ranked names of the related articles, from which they defined high ranked names as learning-related keywords. They proposed a method that would allow dynamic adjustment to weighted relevance to learning in response to the school year or level of the learner, but accuracy was low for some learning items, indicating that there is still room for improvement. So it useful in application which suggest article based learning area of the student's school year. [4]

In Sentiment Enhanced Hybrid TF-IDF for Microblogs Atakan Simsek et al. propose a new approach, sentiment supported hybrid TF-IDF, in order to extract keywords to represent a user's profile more effectively from microblogs so it can be used to build recommendation systems to satisfy various needs of different types of user on social network. Past techniques are limited to domain dependent solution this new approach support domain independent solution. [5]

In Keywords Extraction Method Based on Deleting Actor Index ZHANG Lisheng. Et al. Proposed algorithm to extract high frequent terms as well as important terms with low frequency. The approach is based on term network and deleting actor index, for effective keywords extraction the algorithm. Firstly algorithm lists of terms are got by document preparation. Secondly, constructing the term network, and calculate the value of Total loss of each term and then top m terms are selected as keywords. [6]

In Improving Multi-term Topics Focused Crawling by Introducing Term Frequency-Information Content

(TF-IC) Measure Ali Pesaranghader et al. proposed a new measure called TF-IC, to prioritize keywords in multi-term topics for predicting relevant links on the web pages.in multi-term topics focused crawling Term Frequency-Information Content (TFIC) outperforms other Measures like Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Semantic Indexing (LSI) for multi-term topics focused crawling as TF-IDF suffers from IDF obtained from small number of windows, and Latent relations discovered by LSI may mislead the crawler. [7]

In Micro-blog Commercial Word Extraction Based on Improved TF-IDF Algorithm Xing Huang et al. proposed application of conventional TF-IDF algorithm in term weight calculation. They combine the relative knowledge of information theory and analyzing the distribution of keywords within the class, improved TF-IDF algorithm to be used in term weight calculation. Finally the algorithm is ported to the Hadoop Distributed framework to implement fast and high accuracy extraction terms of commercial value in the mass data of micro-blog. [8]

In Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information Yutaka Matsuo et al. proposed an algorithm to extract keywords from a single document. In this approach frequent terms are extracted first, then a set of co-occurrence between each term and the frequent terms, i.e., occurrences in the same sentences, is generated. Co-occurrence distribution shows importance of a term in the document as follows. If probability distribution of co-occurrence between term and the frequent terms is biased to a particular subset of frequent terms, then term a is likely to be a keyword. The degree of biases of distribution is measured by the 2-measure. .Main advantages of this method are its simplicity without requiring use of a corpus and its high performance comparable to tf-idf. Method will be useful application such as domain-independent keyword extraction. [9]

III. CONCLUSION

Keywords Extraction plays a very important role in the information retrieval domain, since the keywords can represent the asserted main point in a document. Accurate and effective method to find more relevant keywords based on user query is need of current information era. We saw various approach which enhance the capacity of keyword extraction and indexing and facilitate the searching from very large unstructured web data. We saw that extracted keyword is used for information filtering and recommendation system purpose.in future to make fast and effective web search and web based recommendation more research should be done on improving accuracy and speed of the various keyword extraction techniques. Effective methods to improve indexation and dimensionality reduction from extracted keywords to find more relevant result is needed in future.

REFERENCES

(Periodical style)

[1]Gupta, Arpan, Abhishek Dixit, and Arvind Kumar Sharma. ”A novel statistical and linguistic features based technique for keyword extraction.” Information Systems and Computer Networks (ISCON), 2014 International Conference on. IEEE, 2014.

[2]Sheeba, J. I., and K. Vivekanandan. ”Low frequency keyword extraction with sentiment classification and cyberbully detection using fuzzy logic technique.” Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on. IEEE, 2013.

[3]Hendez, Marwa, and Hadhemi Achour. ”Keywords extraction for automatic indexing of e-learning resources.” Computer Applications And Research (WSCAR), 2014 World Symposium on. IEEE, 2014.

[4]Toyota, Tetsuya, and Yuan Sun. ”Keyword extraction for mining meaningful learning-contents on the Web using Wikipedia.” Frontiers in Education Conference (FIE), 2014 IEEE. IEEE, 2014.

[5]Simsek, Atakan, and Pinar Karagoz. ”Sentiment Enhanced Hybrid TFIDF for Microblogs.” Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on. IEEE, 2014.

[6]Mtafya, Ambele Robert, Dongjun Huang, and Gaudence Uwamahoro. ”On Objective Keywords Extraction: Tf-Idf based Forward Words Pruning Algorithm for Keywords Extraction on YouTube.” International Journal of Multimedia and Ubiquitous Engineering 9.12 (2014): 97-106.

[7]Pesaranghader, Ahmad, Norwati Mustapha, and Nurfadhlina Mohd Sharef. ”Improving multi-term topics focused crawling by introducing term Frequency-Information Content (TF-IC) measure.” Research and Innovation in Information Systems (ICRIIS), 2013 International Conference on. IEEE, 2013.

[8]Huang, Xing, and Qing Wu. ”Micro-blog commercial word extraction based on improved TF-IDF algorithm.” TENCON 2013-2013 IEEE Region 10 Conference (31194). IEEE, 2013.

[9]Matsuo, Yutaka, and Mitsuru Ishizuka. ”Keyword extraction from a single document using word co-occurrence statistical information.” International Journal on Artificial Intelligence Tools 13.01 (2004): 157-169.

[10]InformationRetrievalandTextMining’,2015.[Online].Available:http://www.tfidf.com/.[Accessed: 27-Nov-2015].

[11]Latentsemanticindexing’,2015.[Online].Available:https://en.wikipedia.org/wiki/Latent semantic indexing.[Accessed:27-Nov-2015].

[12]DataMiningMiningTextData’,2015.[Online].Available:http://www.tutorialspoint.com/data mining/dm mining text data.htm.[Accessed: 26-Nov-2015].