International Journal of Engineering & Technology, 7 (2) (2018) 832-836

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i2.11664
Research paper

Term weight measures influence in information retrieval

K. Pradeep Reddy [1] *, T. Raghunadha Reddy [2],G. Apparao Naidu [3], B. Vishnu Vardhan [4]

[1] Associate Professor, Dept. of CSE, Tirumala Engineering College, Hyderabad
[2] Associate Professor, Dept of IT, Vardhaman College of Engineering, Hyderabad
[3] Professor, Dept of CSE, JBIET, Hyderabad
[4] Professor, Dept of CSE, JNTUHCEJ, Jagtiyal, Karimnagar
**Corresponding author E-mail: pradeep529@gmail.com*

Abstract

Indexing was majorly used in different applications like information retrieval (IR), Document categorization. In the field of IR, indexer is used by search engines to represent the content of a document with short and content-bearing terms so that the retrieval process obtained great performance. The text index systems produce better results based on the assignment of suitable weights to the terms. These results crucially depend on the selection of the efficient term weighting measures. In this work, the experimentation carried out with different types of term weight measures to assign weights to the terms in the query and document representation. Cosine similarity measure is used to find the similarity between the query vector and document vector. The experimentation is performed on four standard datasets and recall as a performance evaluation measure. The results obtained in this work are promising than most of the approaches in IR field.

Keywords: *Information Retrieval; Term Weight Measures; Tfidf; Recal; Cosine Similarity Measure*

1. Introduction

Since the digital technology has emerged, massive amount of data has been produced in electronic format. As the available electronic data in every sector got increased, accessing valuable information was become a critical issue. In order to dig up information from huge repositories with less time and energy, we need to use an effective indexing mechanism. Indexing is a way of locating documents using representative terms or concepts to make information searching or document categorizing easy and fast. There are two main categories of indexing such as manual and automatic indexing. Manual (Human) indexing is the process of representing documents by domain experts without computerized systems whereas automatic indexing is done with automated systems without any human intervention.

Indexing was used in various applications like information retrieval (IR), Document categorization. In the case of document categorization, index terms are used to identify the predefined category of a document. Various researches [1], [2] proposed different types of approaches in the area of document indexing for different languages with an intention of bringing a means for better document processing and retrieval. In general, there are two main categories for these indexing approaches such as keyword-based and concept-based indexing (semantic indexing). Keyword-based methods are not capable of capturing the implicit relation among terms or the semantics of the words in the document. To remove this weakness, concept-based indexing comes into existence.

The important phases in the IR process are user interface, indexing, query evaluation and ranking. The user interface provides an environment for specifying the needs of the user. Simple and relaxed interfaces improve the performance and user friendliness. The most commonly used interfaces in the IR literature were keyword based, form based, natural language based and graph based interfaces. Indexing plays a vital role in IR system to retrieve relevant results to the user query. Indexing is the process of organizing the terms, phrases and concepts which differentiate the types of documents. Documents are indexed so that searching for the relevant documents for user query become faster. Most of the researchers proposed various Indexing techniques to improve the query performance and the response time. Indexing is one of the components in IR which plays a great role in making searching easy, quick and effective. The classical indexing mechanism, exact term matching, was not capable of dealing with these two vital properties of the language (Synonymy and Polysemy).

In query evaluation phase, the user search statements were converted into the internal form that the system can understand. In this process, various processing tasks were performed such as stop word removal and stemming. The terms in the search statement were compared with the indexed terms, which are maintained by the IR system. The matched document results were forwarded to the ranking phase. Ranking phase is used to assign ranks to the documents given in the query evaluation phase. The ranks are assigned based on the similarity score between the query and document. The documents were displayed according to the similarity score, high similarity score documents were displayed in top.

IR systems was majorly classified into three types of models such as Boolean model, vector space model and probabilistic model based on the representation used for query and document. In Boolean Model, the queries are just Boolean expressions and the documents were represented as set of terms. A document is considered as relevant if it contains all the terms specified in the Boolean expression. The vector space model represents both the queries and documents as vectors of term weights. The documents are retrieved based on the degree of similarity between the document and query vectors. The vector space model also retrieves partially matched documents, when compared with the Boolean model. In probabilistic model, the relevant documents were retrieved to the query according to the probabilities derived using the bayes theorem.

This paper is structured in 7 sections. Section 2 analyzes the different approaches proposed to indexing techniques in information retrieval area. The dataset characteristics and evaluation measures were explained in section 3. The term weight measures used to represent the document vectors were described in section 4. Section 5 explains our approach and experimental results of our approach were presented in section 6. Section 7 concludes this work with future directions.

2. Related work

Indexing is the process of representing documents so as to make searching information out of documents and locating them easily, quickly and effectively. There are two broad categories of indexing, manual and automatic indexing. Manual indexing, also called as human indexing, is one kind of indexing methods where indexing experts, human indexers, use their domain knowledge to understand the contents of the documents and select candidate terms which they think are capable of representing the meaning or concept of the documents [1]. Automatic indexing refers to an indexing mechanism where indexing is done by computer algorithms without the involvement of humans. As it is indicated in [1], manual indexing is slow and expensive whereas automatic indexing is cheaper, fast, and easy to modify. Automatic indexing was used in different computer applications like document categorization, information retrieval, information extraction and so on.

Indexing is one component in document categorization which has a substantial impact on the performance of the categorizer [2]. Classifying documents into the correct category depends on the performance of the indexer. As the quality of the indexer increases the performance of the categorizer also increases. Additionally, indexer is a vital component of information retrieval (IE) systems and its performance has a considerable impact on the overall system performance.

According to the authors in [3], the different indexing approaches are grouped into three major categories such as statistical, probabilistic and linguistic methods depending on the extraction techniques used to extract index terms from documents. The linguistic approaches have two main subdivisions such as syntactic and semantic methods. In statistical approach, the terms which are believed to be capable of reflecting the content of the documents are extracted by applying statistical methods on the words that appear in the entire document collections. The basic notion of this approach is to explore the occurrences of concept bearing words in one document and in the collection as a whole. Inverse Document Frequency (IDF) is one of the well-known statistical methods. It takes those words which occur in a few documents from the entire collection but appear frequently in a single document as index terms [4]. Term frequency and document frequency needs to be calculated before computing IDF. Term frequency is the number of occurrences for the term in a document, whereas document frequency is the number of occurrences for the term in the whole collection [5]. There are other techniques which belong to this approach like N-gram based method to extract concept-bearing terms from documents and Statistical Corpus-Based Term Extractor [6].

The probabilistic techniques are based on the interdependency among terms and the probability of these closely interconnected terms exist in relevant documents to extract more document-bearing complex index terms [7]. As it is stated in [3], even though the dependencies among terms were explicitly defined by users, this approach and the statistical approach are not giving quality index terms because it is not possible to say that terms that occur together are necessarily related semantically.

The syntactic approach makes use of the sentence structure of the document in order to extract the relationship among words so that it can identify the appropriate index terms to represent the document [8]. Using syntactic information makes this approach able to overcome the problem of generating incorrect phrases in non syntactic approaches. However, this approach is incapable of capturing the semantics of the content of documents. There are some research works conducted using this indexing approach such as The Effect of Syntactic Phrase Indexing on Retrieval Performance for Dutch Texts [9] and Syntactic Approaches to Automatic Book Indexing [4].

The semantic approach is more dedicated to capture the meaning of the contents of the documents using concepts rather than words as index terms. The concepts were expressed using different terms in different documents. This approach looks at the semantic relationship among index terms (concepts) to overcome the problems of the keyword based approaches [10]. Semantic methods are expected to handle the synonymy, polysemy and other challenges of natural languages. Latent Semantic Indexing (LSI) is one of the known approaches that identify implicit concepts from documents using mathematical computations [11].

In [12], the researcher applied latent semantic indexing (LSI) with Singular Value Decomposition (SVD) method in an attempt to solve the problem of VSM scheme. The researcher had tried to construct a semantic indexer which considers documents which do not share common words with the query by exploring the LSI method. The LSI extracts concepts from a given corpus by looking at the words that occur together frequently without giving emphasis to the relationship between concepts. However, this approach is incapable of handling indirect queries because it does not consider the relationship among concepts. The other problem of the LSI methodology is providing irrelevant query results out of the user's demand. The other problem of this methodology is as the size of the documents gets enlarged, the performance of the indexer degrades.

3. Dataset characteristics and evaluation measures

The table 1 shows the characteristics of standard corpuses used in this work. In this table, LISA stands for Library and Information Science Abstracts. The files in this directory contain the LISA collection as provided by PeterWillett of Sheffield University. The LISA documents contain just the title and abstract fields. The NPL collection is a collection document titles. Cranfield dataset is a collection of abstracts and a set of queries with relevance judgments. Medline dataset contain a collection of articles from a medical journal.

Table 1: Dataset Characteristics

Dataset Name	Number of Documents	Number of Queries	Size of Dataset (MB's)
LISA	5872	35	3.4
NPL	11429	93	3.1
Cranfield	1400	225	1.6
Medline	1033	30	1.1

The researchers used various evaluation measures such as Recall, precision, and F-measure to evaluate the performance of the information retrieval systems. Recall is the ratio of the number of documents retrieved correctly to the total number of relevant documents in the document collection whereas precision is the ratio of the number of documents retrieved correctly to the total number of documents retrieved. F-measure is the standard measure for evaluating IR by combining recall and precision techniques. F-measure is the harmonic mean of precision and recall. Precision and recall are used to show how many of the relevant documents are captured and missed by the IR system for each query whereas the F-measure shows the overall performance of the system for each query by combining the recall and precision values. The harmonic F-measure gives equal weight for recall and precision. In this work, recall measure was used to evaluate our system.

4. Term weight measures

The main utilization of term weighting system is improvement of effectiveness of relevant information retrieval. This system identifies the relevant documents for the given query as well as rejection of irrelevant documents in the results of a query. The content bearing words were used to represent the document and query vectors. Term weight measures allocate suitable weight to the terms based on the information of terms distribution in the dataset [13]. Traditional term weighting measures are Term Frequency (TF), binary and Term Frequency Inverse Document Frequency (TFIDF). Binary weight measure assigns 1 or 0 to the term based on the term presence or absence in a document [14]. TF measure computes the frequency of a term in a document. TF may assign large weights to the common terms (a, an, the, of, etc.,) which are weak in text discrimination [15]. To overcome this shortcoming, TFIDF measure is proposed by the researchers to reduce the weight of common terms [16]. In TFIDF measure, the IDF allocate more weight to the terms that were appeared in less number of documents [17]. The next sections explain the different term weight measures proposed in the text categorization domain.

4.1. Term frequency inverse document frequency (TFIDF)

The TFIDF measure [18] computed using (1).

$$TFIDF(t_i,d_k) = tf\ (t_i,d_k) \times \log\left(\frac{|N|}{DF_i}\right) \quad (1)$$

Where, tf(ti, dk) is the number of times ti occurred in document dk, N is the number of documents in the dataset, DFi is the number of documents in the dataset which contain the term ti.

4.2. Non uniform distributed term weight measure (NDTW)

The NDTW measure assigns more weight to the terms which are distributed non uniformly across the documents [19]. Equation (2) shows the NDTW measure.

$$W_{t_{ij}} = \log(TOTF_{ti}) - \sum_{k=1}^{m}\left(\frac{tf\ (t_i,d_k)}{TOTF_{ti}}\ \log\left[\frac{1+tf\ (t_i,d_k)}{1+TOTF_{ti}}\right]\right) \quad (2)$$

Where, $TOTF_{ti}$ is the total occurrence of term t_i in all the documents of a dataset, tf (t_i, d_k) is the frequency of term ti in document d_k.

4.3. Normalized document length term weight measure (NDLTW)

A NDLTW Measure is used to avoid the differentiation of small sized and large sized documents [20]. Equation (3) represents the NDLTW measure.

$$W_{t_{ij}} = \frac{(1+\log(TF_i))/(1+\log(AVGTF_i))}{\sum_{k=1}^{m}(1-slope)\times AVGUT_k + slope \times UT_k} \quad (3)$$

Where, TF_i is the number of times the term t_i is occurred in document D_j, $AVGTF_i$ is the ratio of TF_i to total number of terms in all the documents of a dataset, slope =0.2, UT_k is number of unique terms in document d_k, $AVGUT_k$ is the ratio of UT_k to total number of terms in document d_k.

5. Approach

Figure 1 shows the model of our approach. First, two preprocessing techniques such as stop word removal and stemming were performed on the dataset. Stop word removal removes non informative words and stemming techniques convert words into its root words to avoid the different forms of same word. After cleaning the text, identify the distinct terms and represented as a bag of terms. Represent the documents and query as a vector using this bag of terms. Each term value in a vector is computed using term weight measures. The similarity between query and document is calculated using cosine similarity measure. The results of similarity measure are given to performance evaluation (recall) measure. Recall measure gives the number of relevant documents was identified for the given query. In this model, (D_1, D_2, D_m) is a set of documents in the dataset, (T_1, T_2, T_n) is a set of distinct terms identified in the dataset after performing stop word removal and stemming, WDT_n, WQT_n are the weight of term t_n in document and query vector respectively. The next section explains the cosine similarity measure.

5.1. Cosine similarity measure (CSM)

In VSM, the sets of documents and queries are viewed as vectors. Cosine similarity measure is a popular method for calculating the similarity value between the vectors [21]. With document and queries being represented as vectors, similarity signifies the proximity between the two vectors. Cosine similarity measure computes similarity as a function of the angle made by the vectors. If two vectors are close, the angle formed between them would be small and if the two vectors are distant, the angle formed between them would be large. The cosine value varies from +1 to -1 for angles ranging from 0 to 180 degrees respectively, making it the ideal choice for these requirements. A score of 1 evaluates to the angle being 0o, which means the document are similar. While a score of 0 evaluates to the angle being 90o, which means the documents are entirely dissimilar. The cosine weighting measure is implemented on length normalized vectors for making their weights comparable. Equation (4) gives the formula for Cosine Similarity.

$$CSIM(q,d_j) = \frac{\sum_{i=1}^{m} w(t_i,q)\times w(t_i,d_j)}{\sqrt{\sum_{i=1}^{m} w(ti,q)^2} \times \sqrt{\sum_{i=1}^{m} w(ti,dj)^2}} \quad (4)$$

Where, w (t_i,q), w(t_i,d_j) are the weights of the term t_i in query q and document d_j respectively.

6. Experimental results

Table 2: The Accuracies of Recall Measure

Term Weight Measure /Dataset	Medline	Cranfield	LISA	NPL
TF	0.7725	0.7595	0.8031	0.8315
TFIDF	0.7960	0.7815	0.8380	0.8416
NDTW	0.8270	0.8235	0.8515	0.8771
NDLTW	0.8405	0.8320	0.8605	0.8839

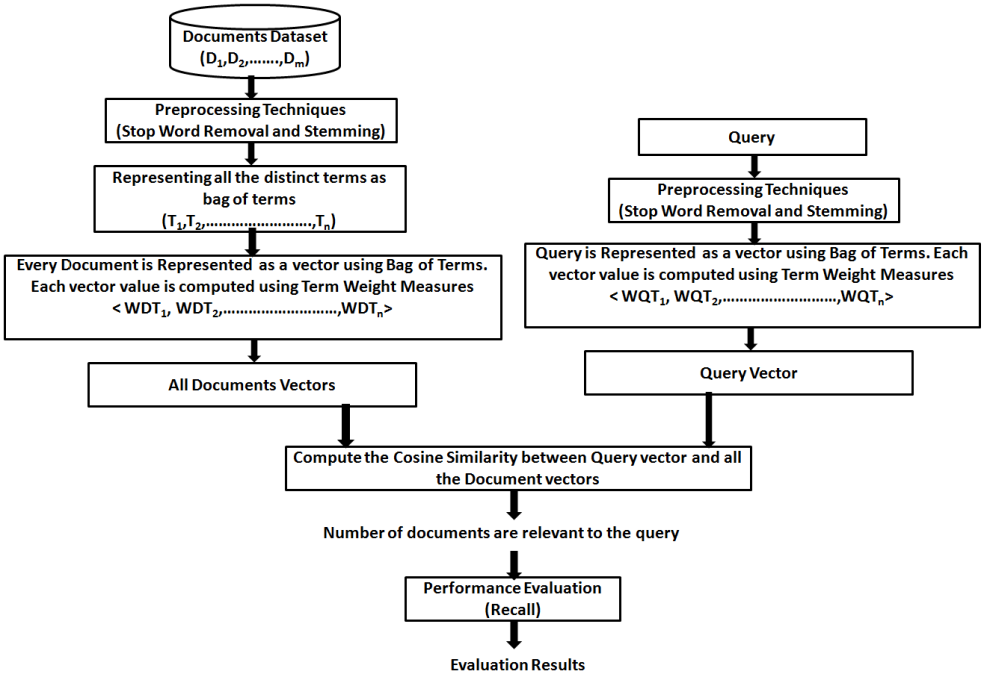

Fig.1: The Model of Our Approach.

Table 2 shows the results of our approach on different datasets when term weight measures were used to represent the document vectors. The NDLTW measure obtained good recall value for all datasets.

It was observed that the datasets used in our experimentation contains documents of varied sizes. In general if the document size is large, there is a possibility of the document is matching to more number of queries and the term frequency also more in that documents. The terms that are occurred in more number of documents have more weight and the terms that occurred in less number of documents have less weight.

The NDLTW is one measure it considers these factors and normalizes the weights of the terms. It normalizes the document lengths whether it contains more number of terms or less number of terms. This is the reason the NDLTW measure obtains good performance to match relevance documents. NPL dataset obtained good recall value, because this dataset contains less number of queries and more number of documents under each query. In general, more number of documents in the dataset is useful for computing effective weights of the terms.

7. Conclusion

The proposed approach in this work obtained good recall value of 0.8839 for NPL dataset. The NDLTW obtained best recall value when compared with other term weight measures in all the datasets experimented in this work. In this approach the terms are independently participated in the document and query representation. The relationship between terms in a document and query was not specified in the representation. This is one reason for not getting best recall value.

In our future work, it was planned to propose an approach which capture the semantic similarity between the terms by proposing a new semantic similarity measure. It was also planned to apply deep learning techniques to improve the performance of information retrieval systems.

References

[1] Tony I. Obaseki, "Automated Indexing: The Key to Information Retrieval in the 21st Century," Library Philosophy and Practice (e-journal) Libraries at University of Nebraska-*Lincoln*, 2010.

[2] MeronSahlemariam, MulugetaLibsie, and Daniel Yacob, "Concept-Based Automatic Amharic Document Categorization," AMCIS 2009 Proceedings, 2009.

[3] Dow Jones Markets, Vijay V. Raghavan, William I. Grosky, Rajesh Kasanagottu, and Venkat N. G Udivada, "Information retrieval on the World Wide Web.," in IEEE Internet Computing, 1997.

[4] Gerard Salton, "Syntactic Approaches to Automatic Book Indexing," Proceedings of the 26th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp. 204-210, 1988. https://doi.org/10.3115/982023.982048.

[5] S. E. Robertson and K. S. Jones, Simple, proven approaches to text retrieval. Cambridge: Computer Laboratory, University of Cambridge, 1997.

[6] Patrick Pantel and Dekang Lin, "A Statistical Corpus-Based Term Extractor," Advances in Artificial Intelligence, pp. 36-46, 2001. https://doi.org/10.1007/3-540-45153-6_4.

[7] Melvin Earl and John L. Kuhns. Maron, "On Relevance, Probabilistic Indexing and Information Retrieval," Journal of the ACM (JACM) 7.3, Vol. 7, No. 3, pp. 216-244, 1960.

[8] L Fagan Joel, "Experiments in Automatic Phrase Indexing for Document Retrieval: A comparision of Syntactic and non-Syntactic Methods," 1987.

[9] Renee Pohlmann and Wessel Kraaij, "The Effect of Syntactic Phrase Indexing on Retrieval Performance for Dutch Texts," In Proceedings of RIAO'97, pp. 176-187, 1997.

[10] Barón Marco Suárez and Valencia Kathleen Salinas, "An approach to semantic indexing and information retrieval," Revista Facultad de Ingeniería Universidad de Antioquia, pp. 174-187, 2009.

[11] Barbara Rosario, "Latent Semantic Indexing: An overview," Techn. Rep. INFOSYS 240, 2000.

[12] Tewodros Hailemeskel Gebermariam, "Amharic Text Retrieval: An Experiment Using Latent Semantic Indexing (LSI) with Singular Value Decomposition (SVD)," Master's Thesis, School of Information Studies for Africa, Addis Ababa University, Addis Ababa, Ethiopia, Unpublished 2003.

[13] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "Author profile prediction using pivoted unique term normalization", Indian Journal of Science and Technology, Vol 9, Issue 46, Dec 2016.

[14] Raghunadha Reddy T, Vishnu Vardhan B, GopiChand M, Karunakar K, "Gender prediction in Author Profiling using ReliefF Feature Selection Algorithm", Proceedings in Advances in Intelligent Systems and Computing, Volume 695, PP. 169-176, 2018.

[15] Raghunadha Reddy T, Gopichand M, Hemanath K, "Location Prediction Of Anonymous Text Using Author Profiling Technique", International Journal of Civil Engineering and Technology (IJCIET), Volume 8, Issue 12, December 2017, pp. 339-345.

[16] Swathi Ch, Karunakar K, Archana G, T. Raghunadha Reddy, "A New Term Weight Measure for Gender Prediction in Author Profiling", Proceedings in Advances in Intelligent Systems and Computing, Volume 695, PP. 11-18, 2018.

[17] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "A Survey on Author Profiling Techniques", International Journal of Applied Engineering Research, March 2016, Volume-11, Issue-5, pp. 3092-3102.

[18] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "Profile specific Document Weighted approach using a New Term Weighting Measure for Author Profiling ", International Journal of Intelligent Engineering and Systems, Nov 2016, 9 (4), pp. 136 - 146. DOI: 10.22266/ijies2016.1231.15 https://doi.org/10.22266/ijies2016.1231.15.

[19] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "N-gram approach for gender prediction", IEEE 7th International Advance Computing Conference, jan 5-7, 2017, pp.860-865

[20] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "A Document weighted Approach for Gender and Age Prediction", International Journal of Engineering, Volume 30, No. 5, 2017, PP. 647-653.

[21] Moheb Ramzy Girgis, Abdelmgeid Amin Aly & Fatima Mohy Eldin Azzam, "The Effect Of Similarity Measures On Genetic Algorithm-Based Information Retrieval", International Journal of Computer Science Engineering and Information Technology Research, Vol. 4, Issue 5, pp. 91-100, Oct 2014.