LANGUAGE INDEPENDENT DOCUMENT RETRIEVAL USING UNICODE STANDARD

Vidhya M[1] and Aji S[2]

Department of Computer Science, University of Kerala
Thiruvananthapuram, Kerala, India

ABSTRACT

In this paper, we presented a method to retrieve documents with unstructured text data written in different languages. Apart from the ordinary document retrieval systems, the proposed system can also process queries with terms in more than one language. Unicode, the universally accepted encoding standard is used to present the data in a common platform while converting the text data into Vector Space Model. We got notable F measure values in the experiments irrespective of languages used in documents and queries.

KEYWORDS

Language independent searching, Information Retrieval, Multilingual searching, Unicode, QR Factorization, Vector Space Model.

1. INTRODUCTION

The digital world is becoming a universal storehouse of knowledge and culture, which has allowed an unswerving sharing of ideas and information in an unpredictable rate. The efficiency of a depository is always measured in terms of the accessibility of information, which depends on the searching process that essentially acts as filters for the richness of information available in a data container. The source of information or data in the depository can be any type of digital data that have been produced or transformed into the digital format such as electronic books, articles, music, movies, images, etc. The language is also not becoming a barrier for expressing the ideas and thoughts in the digital form; as a result, the diversity of data is also increasing along with the volume.

Research and findings in Information Retrieval methods that support Multilanguage has a vital role in the upcoming years of information era. Even though there are lots of findings in the information retrieval mechanisms [1, 11], there is not that much works in the language independent information retrieval methods. The output of the CLEF (Cross-Language Evaluation Forum) workshop [18] also points the need for increasing research in multi-lingual processing and retrieval.

Text is the natural way recording the thoughts and feelings of human being and as a result; the text data became the major component in the entire digital data. In traditional Information Retrieval (IR) methods [1], the text is tokenized into words and a corresponding Vector Space Model (VSM) [11] is created in the initial phase. The IR algorithms such as LSI [19], QR Factorization [10], etc. are applied to retrieve the documents which are most relevant to the query given by the user. Since the steps stemming and stop word elimination [20] are purely language dependent, the VSM formed in most of the works related with IR are language dependent, Naturally this type of VSM cannot accommodate the documents which are written in a couple of

DOI:10.5121/ijcsit.2014.6413

different languages. This paper gives a new mechanism that will convert a document in VSM irrespective of the language used. A special module in our work called alltoOne will convert the tokens in a document, irrespective of language used, into Unicode representation. The query can also be specified in different languages. The result obtained in the experiment shows that the proposed method can do some better improvements in the existing document retrieval models.

2. Methods and Practices

The Information retrieval (IR) systems are indented to identify the documents which are more relevant to a particular query given by the user. Most of the information retrieval systems such as search engines use the generalized or knowledge imparted version of the document listing problem, a collection of D text documents $d_1, d_2 .....d_k$, each d_i being a string of alphabets and the document listing problem tries to find the set of all documents that contain one or more copies of a pattern string p of length m. The said version of the document listing problem is called document mining [21] problem that will find the set of all documents that contain p text at least K, a predefined threshold, times.

Formally, the output is $\{i \mid there\ exist\ at\ least\ K\ \ j's\ such\ that\ d_i\ [j......j+m-1]= p\}$.

The works and findings put forward by various database groups [7] supplement the modifications of the document mining problem. The later developments in this area have evolutionary significance in molecular data, which are extensively used in the computational Biology as well [2, 15].

2.1. Methods of Information Retrieval

The works that have been evolved in the IR can be classified according to the methods that used to measure the relevance of the document to be retrieved for a specific query. In mathematical modeling, method to explain features and characteristics of a problem with mathematical equations and techniques, the works can be classified into three- Probabilistic Relevance Models, Probabilistic Inference Models and Similarity-based Models [22].

It is hard to measure the true relevance of a document for a given query, and the probabilistic relevance model is the mechanism to estimate it. Consider the random variables D, Q and R that represent the document d, query q and the relevance of d for q. The probabilistic relevance model will be used to estimate the values of R such as, $p(R = r \mid D, Q)$. There are two possible values for R, r(relevant) and $\bar{r}$ (not relevant), which can be calculated using either directly by the discriminative (regression) model or indirectly by a generative model.

The earlier works in the regression model [3] deals with features that characterize the matching of D and Q. Later the polynomial regression [4] came to the picture to approximate relevance. The generative model finds the value of R as

$$p(R = r \mid D, Q) = \frac{p(D, Q \mid R = r)\, p(R = r)}{p(D, Q)} \qquad (1)$$

The probabilistic inference model is trying to prove that the query supplied to the system is from a document in the collection. The measure of uncertainty associated with this inference is treated as the relevance of a document with respect to the query. The logic-based probabilistic inference model [5], Boolean retrieval model and the general probabilistic inference model [6] are some of the published works in the probabilistic inference model.

In Similarity-based Models, the correlation between the query and document is treated as the measure of relevance. In order to find the correlation, the text data need to be converted into some another common representation. The vector space model [11], an algebraic framework for processing text documents, is a common platform for finding the correlation. In the vector space model, each document is treated as a vector of frequencies of elements (words or terms) in it. That is, each document can be represented as a point in a space of all documents in the collection. The degree of closeness of points in this space shows the semantic similarity.

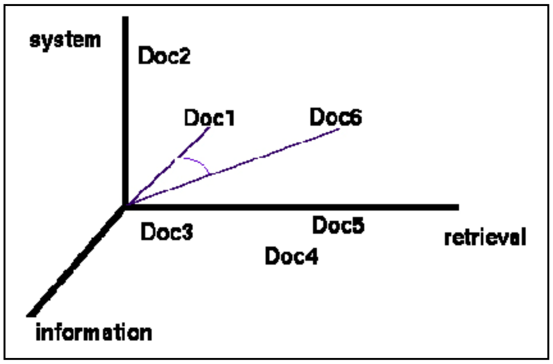


Figure 1.Representation of documents in Vector Space Model

The flexibility of the vector space models in IR is that it can easily incorporate different indexing models [8, 9].

2.2. QR Factorization

The size of Vector Space Model will increase along with the number of unique documents in the collection. Once the high-dimensional data is mapped into a low dimensional space, IR models can be effectively applied to retrieve the information [10]. QR Factorization is one of the well known techniques in dimension reduction in Information Retrieval [13, 11]. The basis of the column space of term-document-matrix W with t unique terms and d documents can be computed through QR factorization [11]:

$$W = QR \tag{2}$$

where R is a t x d upper triangular matrix and Q is a t x t orthogonal matrix.

In order to find the dependence among the columns of W by examining the matrix Q, we can rewrite the relation as

$$[w_1\ w_2\ \ldots\ w_n] = [q_1 r_1^T\ q_2 r_2^T\ \ldots\ q_n r_n^T] \tag{3}$$

The following block multiplication is used to show that the basis of W is contained in the independent columns of Q.

$$\begin{aligned} W &= \begin{pmatrix} Q_W & Q_{\frac{1}{W}} \end{pmatrix} \begin{pmatrix} R_W \\ 0 \end{pmatrix} \\ &= Q_W R_W + Q_{\frac{1}{W}} * 0 = Q_W R_W \end{aligned} \qquad (4)$$

where R_W is the non-zero part of R with W rows, Q_W is the first W columns of Q and $Q_{\frac{1}{W}}$ will be other part. This partitioning shows that the columns of $Q_{\frac{1}{W}}$ do not contribute to the value of W and that the ranks of W, R and R_W are equal. Thus the columns of Q_W constitute a basis for the column space of W.

The similarity between two vectors can be measured using the equation,

$$\cos\theta = \frac{\langle D_1 . D_2 \rangle}{\|D_1\| . \|D_2\|} \qquad (5)$$

So that the similarity measure between the jth document w_j^T and q will be

$$\cos\theta_j = \frac{w_j^T q}{\|w_j\| . \|q\|} \qquad (6)$$

by substituting the property of equation 4.4, we will get

$$\cos\theta_j \qquad = \frac{(Q_W r_j)^T q}{\|Q_W r_j\| . \|q\|} \qquad (7)$$

by applying the properties of an orthogonal we can write the cosine similarity as

$$\cos\theta_j = \frac{r_j^T \left(Q_W^T q\right)}{\|r_j\| . \|q\|} \qquad (8)$$

The query vector q as the sum of its components in the column space of W and in the orthogonal compliment of the column space as

$$\begin{aligned} q &= Iq = QQ^T \\ &= \left[Q_W Q_W^T + Q_{\frac{1}{W}} \left(Q_{\frac{1}{W}} \right)^T \right] q \\ &= Q_W Q_W^T q + Q_{\frac{1}{W}} \left(Q_{\frac{1}{W}} \right)^T q \\ &= q_W + q_{\frac{1}{W}} \end{aligned} \qquad (9)$$

It is noted that QA is a basis for the column space of A and QQT is the projection matrix onto the column space of Q. Then

$$Q_W Q_W^T q + Q_{\frac{1}{W}}\left(Q_{\frac{1}{W}}\right)^T q = q_A + q_{\frac{1}{W}} \qquad (10)$$

where q_W and $q_{\frac{1}{W}}$ are the projections of q onto the spaces of Q_W and $Q_{\frac{1}{W}}$ respectively. Therefore, the properties of the projection allow us to say that q_W is the best approximation of the query vector in the column space of W.

Now recall the similarity measure specified above that is, similarity between the jth document w_j^T and q will be

$$\begin{aligned}\cos\theta_j &= \frac{w_j^T q}{\|w_j\|.\|q\|}\\ &= \frac{w_j^T q_W + w_j^T q_{\frac{1}{W}}}{\|w_j\|.\|q\|}\\ &= \frac{w_j^T q_W + w_j^T Q_{\frac{1}{W}}\left(Q_{\frac{1}{W}}\right)^T q}{\|w_j\|.\|q\|} \qquad (11)\end{aligned}$$

Note that w_j is in the column space of W, which is an orthogonal complement to the column space of Q_W^1. Then $w_j^T Q_W^1 = 0$ Then the formula becomes

$$\begin{aligned}\cos\theta_j &= \frac{w_j^T q_W + 0 * \left(Q_{\frac{1}{W}}\right)^T q}{\|w_j\|.\|q\|}\\ &= \frac{w_j^T q_W}{\|w_j\|.\|q\|} \qquad (12)\end{aligned}$$

3. Proposed Method

Even though the meaning of the term information retrieval is very broad, the people around our digital world visualising or simplifying it with the searching process in the information repositories, especially in the web contents. Web searching has undergone remarkable improvements, but researches in multi language supported searching are still in the childhood stage. The method explained in this paper is an attempt for boost up the works in language supported information retrieval. The working of our proposed method is abstracted in the following block diagram.

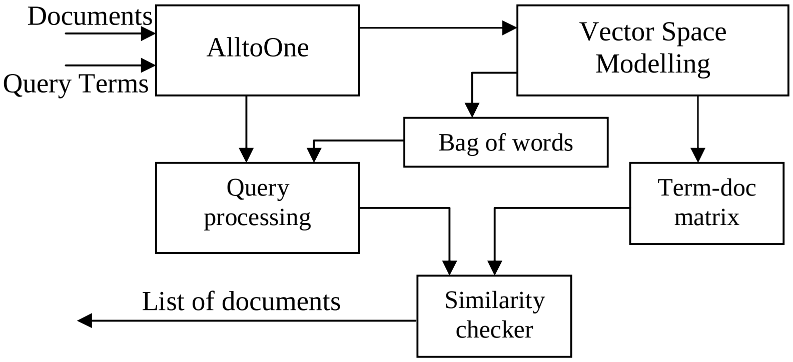


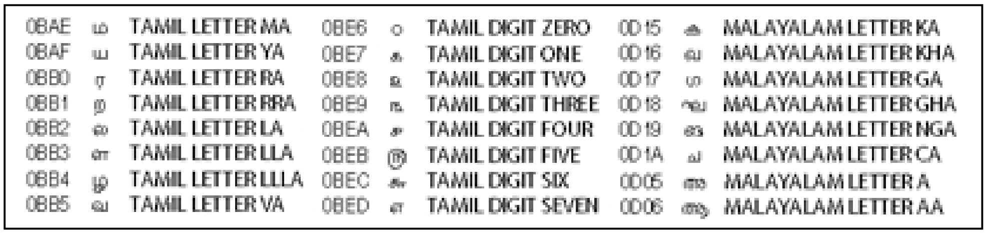
Figure 2: Block diagram of proposed system

As mentioned in the introduction, the AlltoOne module will convert all the unique words in the documents into an Unicode representation, the universally accepted encoding schema for symbols. The vector space model gives two outputs-the term document matrix (TDM) and bag of unique words in the entire collection. The TDM is a weighted matrix which has properly applied TF-IDF weight function. The query processing module will convert the query terms into a query vector with the help of the bag of words. The similarity checker will be used to measure the relevance of each document with regard to the query vector using the QR-factorization. Finally, the list of documents, if any, will be given out as output.

3.1. Unicode Standard

There are several coding schemes, but no two schemes were compatible. The Unicode Standard not only solved these problems, but made multilingual computing challenge a less daunting task to tackle [23]. The Unicode specification includes a huge number of different letters, symbols, characters, mathematical and musical symbols, dingbats, etc (known collectively as 'glyphs'). Unicode provides a unique number for every glyph, no matter what the platform, no matter what the program, no matter what the language. Any software that has Unicode enabled fonts containing the relevant glyphs can, in theory, display any of the glyphs.

The latest version of Unicode (7.0) adds 2,834 new characters. This latest version adds the new currency symbols for the Russian ruble and Azerbaijani manat, approximately 250 emoji (pictographic symbols), many other symbols, and 23 new lesser-used and historic scripts, as well as character additions to many existing scripts.



0BAE	ம	TAMIL LETTER MA	0BE6	௦	TAMIL DIGIT ZERO	0D15	ക	MALAYALAM LETTER KA
0BAF	ய	TAMIL LETTER YA	0BE7	௧	TAMIL DIGIT ONE	0D16	ഖ	MALAYALAM LETTER KHA
0BB0	ர	TAMIL LETTER RA	0BE8	௨	TAMIL DIGIT TWO	0D17	ഗ	MALAYALAM LETTER GA
0BB1	ற	TAMIL LETTER RRA	0BE9	௩	TAMIL DIGIT THREE	0D18	ഘ	MALAYALAM LETTER GHA
0BB2	ல	TAMIL LETTER LA	0BEA	௪	TAMIL DIGIT FOUR	0D19	ങ	MALAYALAM LETTER NGA
0BB3	ள	TAMIL LETTER LLA	0BEB	௫	TAMIL DIGIT FIVE	0D1A	ച	MALAYALAM LETTER CA
0BB4	ழ	TAMIL LETTER LLLA	0BEC	௬	TAMIL DIGIT SIX	0D05	അ	MALAYALAM LETTER A
0BB5	வ	TAMIL LETTER VA	0BED	௭	TAMIL DIGIT SEVEN	0D06	ആ	MALAYALAM LETTER AA

Figure 3: Unicode of some letters in Tamil and Malayalam languages

The Unicode Standard has been adopted by industry leaders such as Apple, HP, IBM, Microsoft, Oracle, SAP, Sun, etc. The emergence of the Unicode Standard and the availability of tools supporting it, are among the most significant recent global software technology trends.

3.2. Term frequency-inverse document frequency (tf-idf)

Term weighting models project the term's ability to represent a document's content, otherwise to distinguish it from other documents. The weighting models give more weight to surprising events and less weight to expected events [14]. The term 'inverse document frequency' [16], a simple and classical approach in term weighting, was proposed in 1972.

The TF factor has been used for Term Weighting for years in text analysis, especially in classification. The IDF is inversely proportional to the number of documents (n) to which a term is assigned in a set of documents N. A typical IDF factor is log (N / n) [17]. So the best index terms to identify the contents of a document are those able to distinguish certain individual documents from the rest of the set. This implies that the best terms should have high term frequencies, but low overall collection frequencies. A reasonable measure of the importance of a term can be obtained, therefore, by the product of term frequency and inverse document frequency (TF x IDF). Hence the weight can be derived as

$$w_{ij} = tf_{ij}\ X\ \log\left(\frac{n}{dfi}\right) \tag{13}$$

where tf_{ij} is the number of occurrence of term i in j^{th} document, df_i is the number of documents that contain the i^{th} term and n is the number documents in the entire collection.

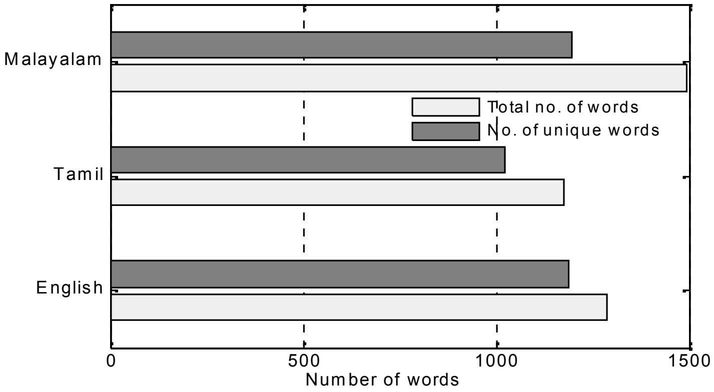
4. EXPERIMENT AND RESULTS

We conducted the experiments using a bunch of text documents in English and two South Indian languages-Tamil and Malayalam. A collection of text documents is prepared from these documents by randomly mixing different language contents. We prepared the documents under the condition that there should be at least 300 words and 12-15 sentences in a single document. A minimum of 30% words and three sentences should be from a single language. Same sentences can be placed in more than one documents, and so that we can directly check the retrieval process. A sample, randomly selected five documents, statistics of words and sentences after the initial processing is tabulated below.

Table 1: Sentence and word statistics of five random documents

Doc No.	Sentence			Words			Unique Words		
	English	Tamil	Mal.	English	Tamil	Mal.	English	Tamil	Mal.
27	4	6	5	86	138	142	69	121	118
58	7	5	6	163	115	139	147	102	112
96	6	4	5	137	93	108	119	79	88
115	8	5	4	174	111	72	153	94	61
176	4	4	5	95	90	83	87	69	72

It is also noted that 13-17 percentage of words in the total number of words are duplicate irrespective of language and the following figure reveals the same.



There are 3278 unique words in the entire collection of documents prepared for the experiments. This bag of words is used to create VSM of the text collection and prepare the query vector.

4.1. Results and Analysis

As noted in the initial sections of this paper, the language dependent searching is familiar mechanism to all information seekers, while our experiments concentrate on the language independent searching. In the first phase of the experiments, the bag of words will be converted into Unicode, the unique and standards representation of alphabets. The TF-IDF weight will be applied to the term document matrices of VSM to keep the semantic relations in the document.

The query set for testing is also generated by randomly picking the words, irrespective of languages, from the collection of documents. There can be words from more than one language in the query. Three different sets of queries are generated for testing-query with 3 words, query with 7 words and query with 12 words. The query will be converted into query vector using the bag of words identified in the VSM, and the query vector will be used to identify the documents which are more relevant to query.

The first experiment was carried out using the query with 3 words, and the result is shown below.

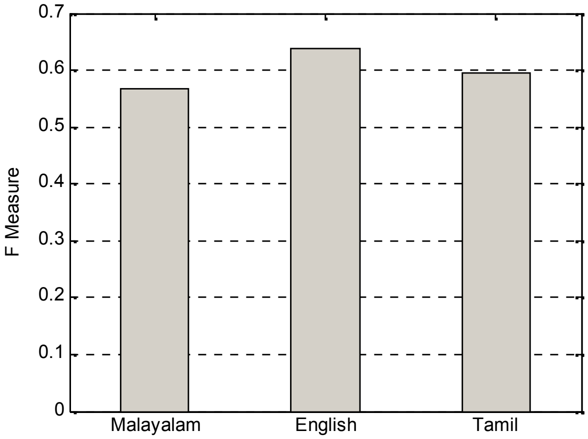


Figure 4: F Measures of different languages with Type 1 Query

It is noticed that the experiments with all languages obtained almost same range of F measure [24]. We have conducted the rest of the experiments with other sets of queries, and the result obtained is abstracted in the following figure.

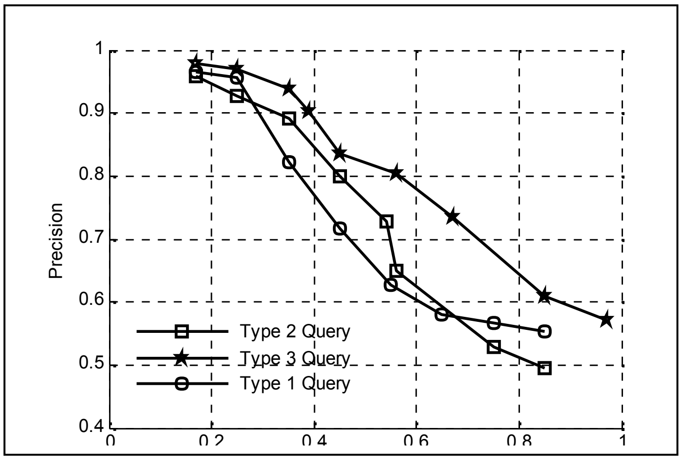


Figure 5: Precision-Recall curve of entire experiments

It has seen that the value of F measure increases with the number of terms in the query and there was no influence of words with different languages in the query. We got an F measure value of 0.664 which is a remarkable result and proves that there are enough possibilities in the area of language independent searching using a standard encoding mechanism such as Unicode.

5. CONCLUSIONS

As the human can express his feelings without any barrier of languages, he should have the facility to access the information with different languages and style. In this work, the searching process was carried out in the VSM which has generated from the common Unicode representation. We obtained an overall F measure of 0.642, in the different experiments with words from different languages. Even though the morphological operations could not address in the work, the result shows that the proposed method will generate a new way of thinking among the researchers in IR, especially language independent searching.

REFERENCES

[1] Salton, G. (1971). The SMART Retrieval System—Experiments in Automatic Document Retrieval. Prentice Hall Inc., Englewood Cliffs, NJ.

[2] D. Gusfield. (1997) Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge Univ Pr; ISBN: 0521585198 ; Dimensions (in inches): 1.44 x10.30 x 7.41

[3] Fox, E. (1983). Expending the Boolean and Vector Space Models of Information Retrieval with P-Norm Queries and Multiple Concept Types. PhD thesis, Cornell University.

[4] Fuhr, N. and Buckley, C. (1991). A probabilistic learning approach for document indexing. ACM Transactions on Information Systems, 9(3):223.248.

[5] Van Rijsbergen, C. J. (1986). A non-classical logic for information retrieval. The Computer Journal, 29(6).

[6] Wong, S. K. M. and Yao, Y. Y. (1995). On modeling information retrieval with probabilistic inference. ACM Transactions on Information Systems, 13(1):69.99.

[7] J. Han, L. Lakshmanan and J. Pei. (2001). Scalable Frequent-Pattern Mining Methods: An Overview. 7th ACMConf on Knowledge Discovery and Data Mining (KDD), Tutorial.

[8] Bookstein, A. and Swanson, D. (1975). A decision theoretic foundation for indexing. Journal for the American Society for Information Science, 26:45.50.

[9] Harter, S. P. (1975). A probabilistic approach to automatic keyword indexing (part I & II). Journal of the American Society for Information Science, 26:197.206 (Part I), 280.289 (Part II)

[10] Ravi Kanth K.V, Divyakant Agrawal, Amr El Abbadi, Ambuj Singh. (1999). Dimensionality reduction for similarity searching in dynamic databases. Computer Vision and Image Understanding: CVIU, 75(1–2):59–72.

[11] Michael W. Berry; Zlatko Drmac; Elizabeth R. Jessup. (1999). Matrices, Vector Spaces and Information Retrieval. SIAM Review, Vol. 41, No. 2. pp. 335-362.

[12] R. Baeza-Yates and B. Ribeiro-Neto. (1999). Modern Information Retrieval. Addison Wesley Longman Publ.

[13] Barry Schiffman, Kathleen R. McKeown. (2005). Context and Learning in Novelty Detection, HLT/EMNLP.

[14] Peter D. Turney, Patrick Pantel. (2010). From Frequency to Meaning: Vector Space Models of Semantics. J. Artif. Intell. Res. (JAIR) 37: 141-188.

[15] G. Benson and M. Waterman. (1994). A Method for Fast Database Search for All k-nucleotide Repeats. Nucleic Acids Research, Vol 22, No. 22.

[16] Robertson, S. (2004). Understanding Inverse Document Frequency: on Theoretical Arguments for IDF. In: Journal of Documentation, Vol. 60, pp. 503-520.

[17] Salton G. Buckley C. (1987). Term Weighting Approaches in Automatic Text Retrieval. Technical Report TR87-881, Department of Computer Science, Cornell University. Information Processing and Management Vol.32 (4), pp. 431-443.

[18] Information Access Evaluation. Multilinguality, Multimodality, and Visualization 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings Series: Lecture Notes in Computer Science Vol. 8138.

[19] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., Harshman, R. A. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science (JASIS), 41 (6), pp:391-407.

[20] Hankyu Lim., Ungmo Kim, (1995). Word recognition by morphological analysis, Intelligent Information Systems. ANZIIS-95. pp: 236 – 241

[21] S. Muthukrishnan. (2002). E_cient algorithms for document retrieval problems. In SODA, pp: 657-666.

[22] ChengXiang Zhai. (2008). Statistical Language Models for Information Retrieval A Critical Review, Journal Foundations and Trends in Information Retrieval archive Volume 2 Issue 3, Pages 137-213.

[23] Valentin Tablan, Cristian Ursu, Kalina Bontcheva, Hamish Cunningham, Diana Maynard, Oana Hamza, Tony McEnery, Paul Baker, Mark Leisher. (2002). A Unicode-based Environment for Creation and Use of Language Resources. LREC.

[24] Raghavan, V., Bollmann, P., & Jung, G. S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans. Inf. Syst., 7, pp: 205-229.