iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** VII **Month of publication:** July 2019

DOI: http://doi.org/10.22214/ijraset.2019.7154

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VII, July 2019- Available at www.ijraset.com

A Survey of Designing Multilingual Document Retrieval and Ranking in Cloud

Manju More E[1], Dr. Sunil Kumar G[2]
[1]*School of C&IT, REVA University, Bengaluru*
[2]*Department of CSE, Vijaya Vittala Institute of Technology, Bengaluru*

Abstract: ***The aim of this paper is to extract and fetch simple and efficient features to enhance multilingual document ranking (MLDR).Our approach is to extract monolingual and multilingual similarity features using a bilingual dictionary. In order to make this approach extensible for all other languages, no language-specific tools are preferred to be used.The process of ranking the documents of various languages based on their relevancy to the query irrespective of query's language is ranking for multilingual information retrieval (MLIR). There are some approaches which focuses on merging the relevant scores of different retrieval settings but do not learn the concept of ranking.The concept of web MLIR ranking in learning-to-rank(L2R) framework is preferred to be used. We create a ranking model to findout the relations among the documents and also to findout the joint relevance probability for the documents. We can improve the relevant estimation of documents in all the languagesusing this method.***

I. INTRODUCTION

Multilingual Information Retrieval (MLIR) is essential and desirable because of the increase of information in various languages. MLIR plays and involves the task of Cross Lingual Information Retrieval for each different desired languages. As the development of globalization and digital online information in Internet is growing, there is a great demand for MLIR. In order to produce a single result, which is obtained from different languages, we come across a merging step, which results in ranking of documents of multilingual results obtained based on the relevancy of the results. The problem of CLIR has been well studied in the past decade especially with the help of CLEF, NTCIR, TREC and FIRE forums. In the realm of CLIR theproblem of ranking multilingual result lists is a very challenging task. The task of identifying whether two different language documents talks about the same topic is itself very challenging. There are few early attempts on ranking multilingual documents (Round robin merging [1], raw-score merging [1]). These merging processes have to make some simplifying assumptions. For example, one may assume that the similarities calculated for different language result lists are comparable; so the result lists can be merged according to their raw similarity values [1]. One can also normalize the similarities first; but this approach implicitly assumes that the highly ranked documents in different languages are similar to the query at a comparable level. These assumptions are not true. Until recent past [2], [3], [4], [5], [6], there was little focus on merging multilingual result lists. The recent work concentrated more on extracting semantic information such as multilingual topics from documents. These methods are highly dependent upon language specific tools like named-entity recognizer, part-of-speech tagger etc., hence they cannot be extended for languages with fewer resources, i.e., they do not achieve high-multilinguality. If we have a requirement to approach a ranking in order to apply in various languages, There is a major challenge in achieving language specific development pose. When we try to merge multilingual list of results achieved ,techniques that is suitable for one language may not be suitable for the other language .There are some applications which deals with limited number of languages while others require lot of different languages. We try to implement language-independent approaches which will benefit multilingual retrieval which encourages MLIR community. In paper[7], Using multilingual documents and topics we extract efficient features which is useful for the enhancement of the performance in Multilingual Document Ranking using the similarities in candidate documents. After the result lists of different languages is obtained along with their queries, we calculate similarity measures along with the various metrics. The tool given for translation will produce large number of translations which is acceptable for a given set of queries [8]. There is a limited set of availability of tools for a specific number of languages.In this regard some language-specific tools are eliminated while measuring the document based on similarity metrics. Therefore in order to calculate the similarity of multilingual document, we can use the bilingual dictionaries, Wikipedia for gaining the knowledge. The approach can also be used to other language to provide the availability of basic language resources.

Some experiments are carried out on FIRE2010 corpus which was conducted by using several ranking algorithms on various features and the results were extracted and combined using the NDCG as the metric for evaluation and the extracted results were verified and compared against the BM25 baseline ranking system[30].

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VII, July 2019- Available at www.ijraset.com

II. SYSTEM OVERVIEW

Usually in MLIR, When a query is given in any language, the CLIR will be performed on monolingual collections. After obtaining the result list from collection, the lists are combined into multilingual result list. The work considers a query in different languages with their result list as the starting point. There is no information how the result lists are combined and produced.There are some judgements made for all query related documents visualizing whether the document is relevant or not.The document which is relevant and similar are extracted and examined. The term vector is constructed for all the relevant documents and they are modeled using different algorithms for ranking. We use probabilities to the documents that are assigned by the different ranking algorithms. The estimated relevance probabilities assigned to the documents by the ranking algorithms are used in ordering the documents.

A. Feature Engineering

In case of Information Retrieval and Natural Language Processing, it is a process of answering a question in a natural language. A pre-structured database or a collection of natural language documents are used to find the answers and address the different challenges[2].

1) Justifying the relevance of a answer is a process of finding whether the answer is relevant to the question and the technique to identify relevant answers in a group of irrelevant ones. Remaining answers can be ranked efficiently if irrelevant answers are eliminated by using some knowledge base algorithm.

2) Examining the redundancy in answers. If there are repeated number of answers in a given list, then that particular answer will be ranked higher.

There is a need for identifying the relevant documents and to be ranked higher, in case they are found to be similar when compared with other documents. Hence we use features engineering to address these two challenges.

The features can be extracted using 3 levels 1) Similarity in query-document 2) Similarity in Monolingual Document and 3) Similarity in multilingual Document

We can discuss the features in detail as follows:

a) *Similarityin Query-Document:* tf and idf measures are used for measuring the relevancy in a document for a query. Tf-idf value is calculated for document 'j' for every term k in a query Q for the collection of document D for the same language. The tf-idffeature value for a document j[2].The scores can be calculated and added up in order to get tf-idf feature value for document j for all the query terms. we can normalize tf-idf value by collecting the document based on the formula

$$TF - IDF = \frac{\Sigma \forall K \in QTF - IDFKJ}{\Sigma \forall \in D \Sigma \forall K \in QTF - IDFKJ}$$

b) *Similarity in Monolingual Document:* If there are 2 documents of the same language, we can measure the similarity in different ways. The terms in a document can be assigned their weights in tf-idf weights which can be calculated within the range of that document collection D. We can also include the Wikipedia redirection terms corresponding to every other term present in the vector [2]. The concept of Wikipedia redirections can be explained by calculating the similarity measure using the formula.

$$simk(di) = \frac{\Sigma j = 1 to\ d(i \neq j) simk\ to\ 1(di, dj)}{\Sigma \forall i\ to\ d \sum j = 1 tod(i + j) sim\ k\ to\ i(di, dj)}$$

c) *Similarity in multilingual Document:* We can compare 2 multilingual documents, in order to map them for a common representation.Given a document in English and kannada, we can map English document terms to kannada representation using bilingual dictionary and Wikipedia redirections[24]. If we find any terms in dictionary, it can be replaced using its synonyms in Kannada. Every word will have more than one possible synonyms. We can take into account, the Wikipedia redirections for every term in the synonyms. Finally, we can represent the English document in a vector of kannada terms and vice-versa[2]

III. CLASSIFICATION AND RANKING

There are different types of classification and Ranking methods for a document:

A. Bipartite Ranking

This technique was introduced by Cohen[10] by taking the help of information retrieval systems where the results are taken in the form of ordered list of objects. The framework which was designed introduced an algorithm in order to learn a new form of supervision such as preference relations over the examples. This algorithm was also used to optimize the criteria which is related to the performance for predictor for ranking. The original ranking algorithm was used for preference relation, The task for learning a scoring function was completely reduced by the subsequent proposals.[11] [12].

We can create the ranking by sorting the examples in decreasing scores. The supervision is a bipartite graph for a special case of ranking algorithm.[12].. It is similar to the information routing problems where, we fix the query and examples can be relevant or irrelevant to the query[13]. Bipartite Ranking is a process of learning a scoring function by the process of optimizing the area under the ROC curve [14].The conclusion is that when the data is imbalanced the ranking methods should be superior and more Effective[21][26].

B. Single view Semi-Supervised Ranking

The approach to multiview for supervised and semi-supervised learning to rank the functions in the bipartite setting can be approached and implemented using algorithms for binary classification. The approaches of single view semi-supervised learning of classifiers is not easily adoptable to ranking[25]. We have made assumptions which was used in single-view semi-supervised classification as decision boundary which is very easy to detect the set of unlabeled data[26]. The task of ranking is used to detect the scoring function which was induced as the best possible ordering completely for the observations. The observations for the ranking is calculated using the probability which is relevant. [15].We do not consider some criteria for classification:

Based on given observation, some algorithms need the most repeatitive class label.Basic works were carried out on a single view semi-supervised bipartite ranking with some results obtained.[16],in an pseudo-labeling step uses information from neighbourhood in order to optimize the objective function of ranking on some labeled training sets.In [17].,we use the unlabeled data in order to change the representation spaces using Gaussian approach. By this we consider the fact that bipartite ranking data has the form of data classified in the form of binaries[28].

C. Supervised Ranking

Semi-supervised process uses an algorithm for bipartite ranking functions in a fully supervised setting. We focused on a linear SVM for ranking, as linear functions with a bag-of-words representation are known to perform well on textual data.

For each view V, the training set available at some given iteration of the algorithm, we can learn linear scoring function is the dot product of Euclidean space which denotes the representation of document X in the Vth language and wv is the parameter vector to learnt for view v.

IV. MULTILINGUAL INFORMATION RETRIEVAL

The Multilingual Information Retrieval is the process of retrieving required and appropriate information in which the user requests for information and the collected document against which we try to match the results in different human understandable language[23]. So therefore there is no language barriers for the users to get and access the information.There are different approaches which is used by translation module in CLIR[29]

A. Query Translation

In this approach we are mapping the query representation into the document representation. The user sends the request and then it is translated into different language and the information that was queried is searched in a set of different documents written in that specific language. The user then retrieves the result. The major problem with this approach is that query lacks the context information so therefore the ambiguity is amplified.

We use different translation approaches such as:

1) Dictionary-based query translation- In this approach, the given query is processed and is translated using machine readable dictionaries.
2) Query translation using Corpora-We use the technique of parallel corpora or comparable corpora in order to translate the user request.
3) Query translation using machine translation – We use machine translation systems to produce the sufficient quality translations using machine translation systems[33].

B. Document Translation

In this approach, the document representation is mapped into the query representation. As there is more amount of information is available. So therefore the ambiguity problem is solved. There is more translation work involved as we translate the documents in multiple languages. More amount of storage is also required to store the translated documents. Machine translation systems is used to translate the document[32].

C. Inter-lingua Translation

In this approach we map both the documents and user requests to a third language. We use this approach when there is no document for direct translation. The performance is low when compared to other methods[31].

D. Challenges facing CLIR

The major challenges to CLIR systems include [18].

1) There is a lack of contextual information which occurs because of the presence of homonymy and polysemy.
2) Word inflection is the major problem used in the queries which can be solved using stemming.
3) Dictionary based translation is a major problem when the query contains the grammar.

V. REPRESENTATION OF DOCUMENTS

The documents in different languages like can be represented in a model named Classical Vector space model[19][22]. The documents is represented in the of 'Bag of Words' notation which has no information ordering. There is a stopword list which is maintained for every language which is nothing but the word which appears in more than 50% of the document can be called as stopword. The stopwords can be removed from the documents even then several times the documents still contains noise. So therefore we can consider only top-k keywords for each document. The Experiment was carried out considering k values from 40% to 100% in which we can increment atleast 10%. We can achieve the best results when K=50%[20]

VI. ENRICHING THE DOCUMENT REPRESENTATION

In order to group the contents of Wikipedia, its details are categorized. The Wikipedia is sinked with the references and hyperlinks to different articles and denote the Outlinks for that article. Those Outlinks are used to create the interlinks among other articles which is a detailed description about a topic or a concept. Some equivalent topics can be represented with different phrases which are grouped together by the directed links. It also contains a hierarchical categorization system, where each article belongs to atleast one category. Therefore Wikipedia has become a major resource which is used and exploited by most of the users to enhance the clustering in the text document.

VII. CONCLUSION

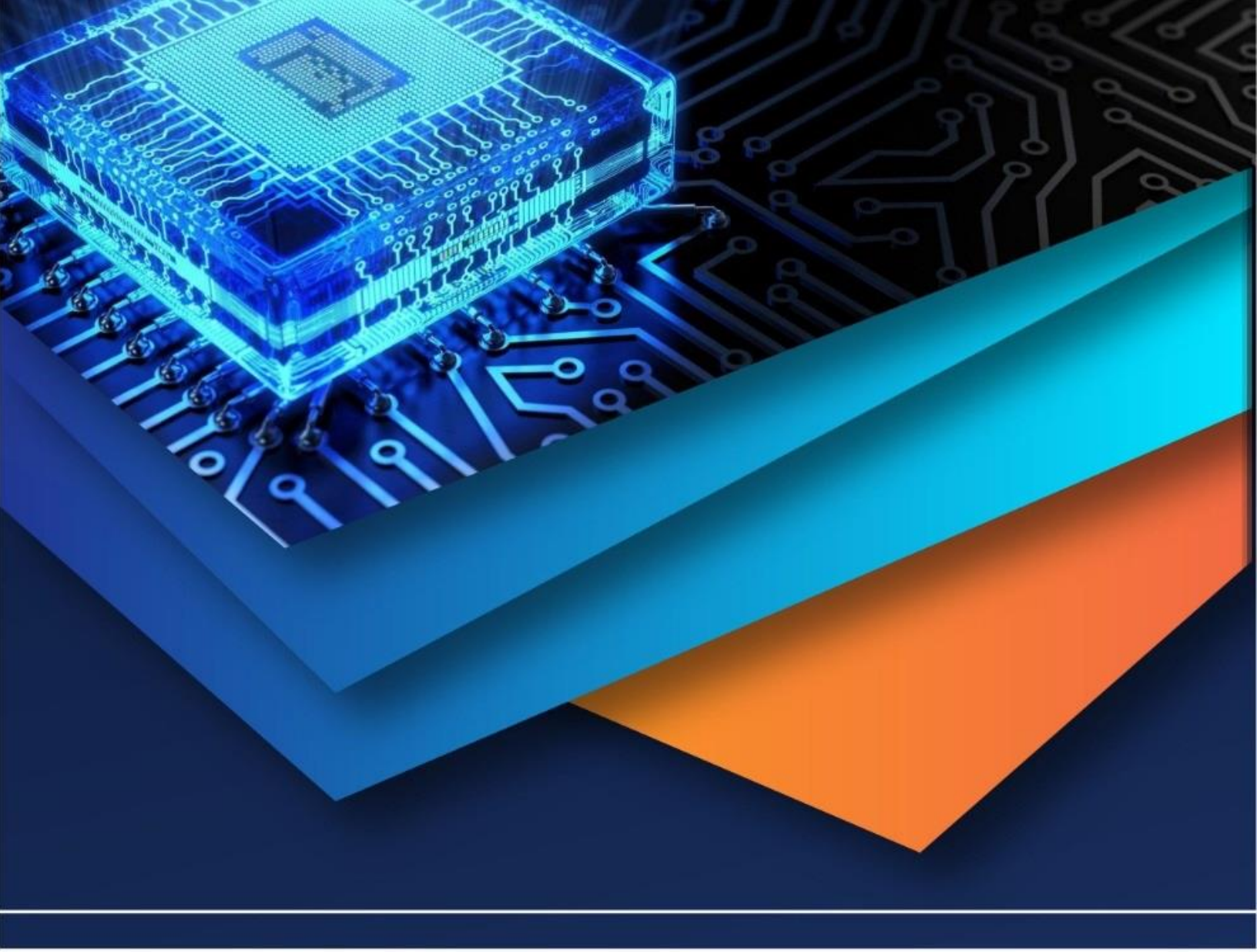
In this paper we have presented about the System overview of MLIR. We have discussed the concepts that is already existing in Feature Engineering. We have approached different classification and ranking methods, in order to rank the multilingual documents in different perspective. We have also learnt how to retrieve the multilingual documents and translate them to different representations.

We have come up with an approach to develop an Ranking algorithm algorithm for searching web data in multiple languages in cloud, and Multilingual framework which can be used by developers for developing multilingual search engine for any languages.

REFERENCES

[1] Ackley, D.H., Hinton, G.E., Sejnowski, T.J.A.: Learning Algorithm for Boltzmann Machines. Cognitive Science 9, 147–169 (1985).
[2] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender,G.: Learning to Rank using Gradient Descent. In: Proc. of ICML, pp. 89–96 (2005)
[3] Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. In: Proc. TREC-2 (1994).
[4] Freund, Y., Iyer, R., Schapire, R., Singer, Y.: An Efficient Boosting Algorithm forCombining Preferences. Journal of Machine Learning Research 4, 933–969 (2004)
[5] Herbrich, R., Graepel, T., Obermayer, K.: Large Margin Rank Boundaries for Ordinal Regression. In: Advances in Large Margin Classifiers. MIT Press, Cambridge(2000)
[6] Jaakkola, T.S.: Variational Methods for Inference and Estimation in Graphical Models. Ph.D. Thesis, MIT (1997)
[7] WeiGao, Cheng Niu, MingZhoun and KamFaiWong: Joint Ranking for Multilingual Web Search, Springer-Verlag Berlin Heidelberg 2009,114-125.
[8] J¨arvelin, K., Kek¨al¨ainen, J.: IR Evaluation Methods for Retrieving Highly Relevant Documents. In: Proc. of ACM SIGIR, pp. 41–48 (2000)
[9] Santosh GSK, Kiran Kumar N, VasudevaVarma:Ranking Multilingual Documents using Minimal Language Dependent Resources.
[10] W. W. Cohen, R. E. Schapire, and Y. Singer.Learning to order things. Journal of Artificial Intelligence Research, 10:243–270, May 1999.
[11] T. Joachims. Optimizing search engines using click through data. In KDD, 2002.
[12] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer.An efficient boosting algorithmfor combining preferences. JMLR, 4:933–969, 2003
[13] S. Robertson and J. Callan. Routing and filtering, chapter 5, TREC: Experimentand Evaluation in Information Retrieval, pages 99–121. MIT Press, 2005.]
[14] S. Agarwal, T. Graepel, R. Herbrich, S. HarPeled, and D. Roth. Generalization bounds for the area under the roc curve. JMLR, 6:393–425, 2005.
[15] S. Cl´emen¸con, G. Lugosi, and N. Vayatis. Ranking and scoring using empirical risk minimization. In COLT-05, pages 1–15, 2005.
[16] M. R. Amini, T. V. Truong, and C. Goutte. A boosting algorithm for learning bipartite ranking functions with partially labeled data. In SIGIR'08, 2008.

[17] L. Ralaivola. Semi-supervised bipartite ranking with the normalized Rayleigh co-efficient. In ESANN-09, 2009
[18] N. A. Nasharuddin and M. T. Abdullah.Cross-lingual Information Retrieval.Electronic Journal of Computer Science & InformationTechnology, 2(1), 2010.
[19] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun.ACM 18 (1975) 613{620].
[20] Kiran Kumar N, Santosh GSK, VasudevaVarma Multilingual Document Clustering usingWikipedia as External Knowledge
[21] Liu, T.-Y., Xu, J., Qin, T., Xiong, W.Y., Li, H.:LETOR: Benchmark Dataset for Research on Learning to Rank for Information Retrieval. In: Proc. of ACM Workshop on Learning to Rank for Information Retrieval, Amsterdam, The Netherland(2007).
[22] Yue, Y., Finley, T., Radlinski, F., Joachims, T.: A Support Vector Method for Optimizing Average Precision. In: Proc. of ACM SIGIR, pp. 271–278 (2007)
[23] Lin, W., Chen, H.: Merging Mechanisms in Multilingual Information Retrieval.In Peters, C., Braschler, M., Gonzalo, J., Kluck, M., eds.: Third Workshop of theCLEF. Volume 2785 of LNCS., Springer (2002) pages 175-186.
[24] Huang, A.: Similarity measures for Text Document Clustering. In: Proceedings of New Zealand Computer Science Research Student Conference. (2008) pages 49-56.
[25] M. Belkin and P. Niyogi. Semi-supervised learning on riemannian manifolds. MachineLearning, 56:209–239, 2004.
[26] O. Chapelle, B. Sch¨olkopf, and A. Zien, editors.Semi-Supervised Learning.MITPress, Cambridge, MA, 2006.
[27] L. Ralaivola. Semi-supervised bipartite ranking with the normalized Rayleigh co-efficient.In ESANN-09, 2009.
[28] X. Zhu. Semi-supervised learning literature survey. Technical report, CarnegieMellon University Department of Computer Sciences, 2006.
[29] F. Ture and E. Boschee.Learning to translate: A query-specific combination approach for Cross-lingual Information Retrieval.
[30] M. R. Warrier and M. S. S. Govilkar. A survey on various CLIR techniques..
[31] Hu, J., Fang, L., Cao, Y., Zeng, H.J., Li, H., Yang, Q., Chen, Z.: Enhancing textclustering by leveraging wikipedia semantics. In: SIGIR '08: Proceedings of the 31st annual International ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2008) 179-186.
[32] Jones, G. J. F., Burke, M., Judge, J., Khasin, A., Lam-Adesina, A., Wagner, J. 2004. Dublin City University at CLEF 2004: Experiments in Monolingual, Bilingual and Multilingual Retrieval. InC. Peters(Ed.),Results of the CLEF2004 cross-language evaluation forum.
[33] BRILL,E.,DUMAIS,S.,AND BANKO, M. 2002. An analysis of the AskMSR question answeringsystem.In Proceedings of the Conference on Empirical Methods in Natural Language Processing.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY