iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** XI **Month of publication:** November 2014

DOI:

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

An Efficient Way of Classifying and Clustering Documents Based on SMTP

U.Umamaheswari[1], Mr. G. Shivaji Rao M.E.[2]
[1]*M.E scholar,* [2] *Assistant Professor,*
Department of Computer Science and Engineering,
Sree Sowdambika College of Engineering, Aruppukottai-626 134, Tamil Nadu, India

Abstract: ***In text processing, the similarity measurement is the important process. It measures the similarities between the two documents. In this project we proposed the new similarity measurement. The computation of similarity measurement is based on the feature of two documents. Our proposed system contains three case to compute the similarity. The three cases are, both two documents contains features, only one document contains feature, there is no feature into the two documents. In first case, the similarity is increased when the differences of feature value is decreased between the two documents. Then the given differences are scaled. In second case, fixed value is given to the similarity. In third cases there is no contribution to the similarity. Finally our proposed measure method achieves the better performance compared than other measurement methods.***

Index Terms—Document classification, document clustering, entropy, accuracy, classifiers, clustering algorithms

I. INTRODUCTION

Text processing plays an important role in information retrieval, data mining, and web search. A document is any content drawn up or received by the Foundation concerning a matter relating to the policies, activities and decisions falling within its competence and in the framework of its official tasks, in whatever medium (written on paper or stored in electronic form , including e-mail, or as a sound, visual or audio-visual recording). The term classification means the allocation of an appropriate level of security (as confidential or restricted) to a document the unauthorised disclosure of which might prejudice the interests of the Foundation, the EU or third parties. Documents are confidential when their unauthorised disclosure could harm the essential interests of an individual, the Foundation or the EU.

Documents are restricted when their unauthorised disclosure could be Disadvantageous to the Foundation, the EU or a third party. Documents are restricted when their unauthorised disclosure could be disadvantageous to the Foundation, the EU or a third party. The term originator means the duly authorised author of a classified document. The term downgrading means a reduction in the level of classification. The term declassification means the removal of any classification.

A. Rules For Classification

Foundation documents that are not public shall be classified in one of the following categories: confidential or restricted. Criteria and guidance for classification are set out in Annex 1 to this decision. The classification of a document shall be decided by the originator based on these rules. All classified documents shall be recorded in a register of classified documents. Applications for access to classified documents shall be examined by the Director. If a classified document is to be made available in response to a request from a member of the public, it shall be first declassified by a decision of the Director. Documents shall be classified only when necessary. The classification shall be clearly indicated and shall be maintained only as long as the document requires protection. The classification of a document shall be determined by the level of sensitivity of its contents.

Classification of documents shall be periodically reviewed. By request of the Document Management Officer (DMO), the originator of a document shall indicate if that document or information may be downgraded and declassified. Where a document or information is declassified, details shall be recorded in the register and the document shall be archived appropriately. Where classification is retained, details of the review shall be entered in the register. All classified documents shall be retained in a manner that ensures they are not disclosed to unauthorised individuals. Security and access levels to classified documents shall be recorded in the register by the DMO. Either the originator or the DMO may retain classified documents. Where the originator retains documents, they shall be physically safeguarded as specified by the DMO.

Individual pages, paragraphs, sections, annexes, appendices, attachments and enclosures of a given document may require

different classifications and shall be classified accordingly. The classification of the document as a whole shall be that of its most highly classified part. The originator shall indicate clearly at which level a document should be classified when detached from its enclosures. The classification shall appear at the top and bottom centre of each page, and each page shall be numbered. Each classified document shall bear a reference number and a date. All annexes and enclosures shall be listed on the first page of a document classified as confidential. The classification shall be shown on restricted documents by mechanical or electronic means. Classification shall be shown on confidential documents by mechanical means or by hand or by printing on pre-stamped, registered paper.

B. Document Classification

Document classification or document categorization is a problem in library science, information science and computer science. The task is to assign a document to one or more classes or categories. This may be done "manually" (or "intellectually") or algorithmically. The intellectual classification of documents has mostly been the province of library science, while the algorithmic classification of documents is used mainly in information science and computer science. The problems are overlapping, however, and there is therefore also interdisciplinary research on document classification.

The documents to be classified may be texts, images, music, etc. Each kind of document possesses its special classification problems. When not otherwise specified, text classification is implied. Documents may be classified according to their subjects or according to other attributes (such as document type, author, printing year etc.). In the rest of this article only subject classification is considered. There are two main philosophies of subject classification of documents: The content based approach and the request based approach.

C. Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

D. Hierarchical Clustering

In clustering algorithms can be categorized based on Text or documents. We are used in the connectivity based clustering and K-mean clustering. Connectivity based clustering also Know as Hierarchical clustering is based on the core idea of object being more related to nearby objects than to object farther away .This algorithms connect “Object “ to form “clusters” based upon their distance. This K mean clustering algorithm is represented by a central vector, which may not necessarily be a member of the data set. This cluster aims to partition n observations into K clusters in which each observation belong to the cluster with the nearest mean, serving as a prototype of the cluster.

E. Content Based Classification

Content based classification is classification in which the weight given to particular subjects in a document determines the class to which the document is assigned. It is, for example, a rule in much library classification that at least 20% of the content of a book should be about the class to which the book is assigned. In automatic classification it could be the number of times given words appears in a document.

F. Request Oriented Classification

Request oriented classification (or -indexing) is classification in which the anticipated request from users is influencing how documents are being classified.

Request oriented classification may be classification that is targeted towards a particular audience or user group. For example, a library or a database for feminist studies may classify/index documents differently when compared to a historical library. It is probably better, however, to understand request oriented classification as *policy based classification*: The classification is done according to some ideals and reflects the purpose of the library or database doing the classification. In this way it is not necessarily a kind of classification or indexing based on user studies. Only if empirical data about use or users are applied should request oriented classification be regarded as a user-based approach.

Automatic document classification tasks can be divided into three sorts: supervised document classification where some external mechanism (such as human feedback) provides information on the correct classification for documents, unsupervised

document classification (also known as document clustering), where the classification must be done entirely without reference to external information, and semi-supervised document classification, where parts of the documents are labeled by the external mechanism.

Similarity is the measure of how much alike two data objects are. Similarity in a data mining context is usually described as a distance with dimensions representing features of the objects. A small distance is indicating a high degree of similarity and a large distance indicating a low degree of similarity. Similarity is subjective and is highly dependent on the domain and application.

II. RELATED WORK

The documents are represented as vectors, each elements of vectors indicates the corresponding feature values of the documents. In existing system the feature values are termed as frequency, relative term frequency and tf-idf (term frequency and inverse document frequency). The document size is large and most of the vector value is zero in this system. In existing system non symmetric measure is used to measure the similarities. Canberra distance metric is one of the existing methods which are used to measure the similarity. This method is applicable when the vector elements are always non-zero. In existing system cosine similarity measure is used to measure the similarities between the two documents. It takes cosine of angle between the two vectors. The phase based similarity measure is used in one of the existing system. So, we are our some problem in following as the document is large in existing system so it takes more memory to store the documents. It contains zero vector values so it makes severe challenges to measure the similarity. It is not efficient and does not provide accurate results. The performance is low. It takes more time to produce the result.

In our proposed system new measure is used to measure the similarities between the two documents. This new measure is symmetric measure. The similarities between the two documents are measured with respect to the features. Three feature cases are used to measure the similarities. In first case the two documents contains features value, in second case, only one document contains the feature value and in third case there is no feature value in both documents. This measure is applied in many applications such as single label and multi label classification, clustering and so on. In our proposed system we k-NN based single label classification (SL-KNN) and k-NN base multi label classification (ML-KNN) for classification purpose. The documents are formed as cluster in our concept for this purpose we used Hierarchical Agglomerative Clustering (HAC). It is the one type of k means clustering algorithm. We three type of data sets such as WebKB, Reuters 8 and RCV1. These data sets are stored in the form of XML. So, In our project improve following as It achieves better performance compared than other measure, It provides efficient results, It achieves accuracy in similarity measurement, It take less time for similarity measure.

III. PROPOSE SIMILARITY MEASURE

In our proposed system new measure is used to measure the similarities between the two documents. This new measure is symmetric measure. The similarities between the two documents are measured with respect to the features. Three feature cases are used to measure the similarities. In first case the two documents contains features value, in second case, only one document contains the feature value and in third case there is no feature value in both documents. This measure is applied in many applications such as single label and multi label classification, clustering and so on. In our proposed system we k-NN based single label classification (SL-KNN) and k-NN base multi label classification (ML-KNN) for classification purpose. The documents are formed as cluster in our concept for this purpose we used Hierarchical Agglomerative Clustering (HAC). It is the one type of k means clustering algorithm. We three type of data sets such as WebKB, Reuters 8 and RCV1. These data sets are stored in the form of XML.

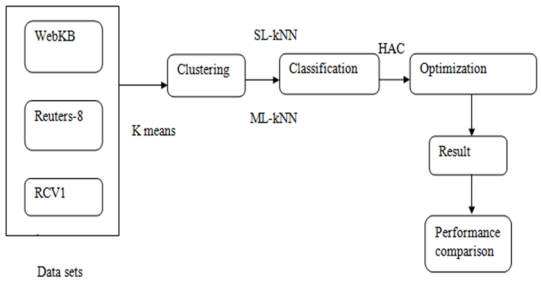


Fig 1: System architecture

In our project, the three dataset is collecting the data. That data should be k-mean algorithm use for clustering purpose and classification used for SL-KNN, ML-KNN cluster. The classify data should be use HAC cluster because HAC cluster is used for similarity measure based upon weight and distance. Finally, to produce the original result for similarity.

The following properties, among other ones, are preferable for a similarity measure between two documents:

1. The presence or absence of a feature is more essential than the difference between the two values associated with a present feature.

Let as Consider two features *wi* and *wj* and two documents **d**1 and **d**2.

wi = **d**2 (some relationship)

wi ≠ **d**1 (no relationship)

In this case, **d**1 and **d**2 are dissimilar in terms of *wi*. If *wj* appears in both **d**1 and **d**2. Then *wj* has some relationship with **d**1 and **d**2 simultaneously. In this case, **d**1 and **d**2 are similar to some degree in terms of *wj*. For the above two cases, it is reasonable to say that *wi* carries more weight than *wj* in determining the similarity degree between **d**1 and **d**2.

2. The similarity degree should increase when the difference between two non-zero values of a specific feature decreases.
3. The similarity degree should decrease when the number of presence-absence features increases.
4. Two documents are least similar to each other if none of the features have non-zero values in both documents. Let

d1 = < *d*11, *d*12, . . . , *d*1*m* > and

d2 = < *d*21, *d*22, . . . , *d*2*m* >.

If *d*1*id*2*i* = 0,

*d*1*i* + *d*2*i* > 0

for 1 ≤ *i* ≤ *m*, then **d**1 and **d**2 are least similar to each other. As mentioned earlier, **d**1 and **d**2 are dissimilar in terms of a presence-absence feature.

5. The similarity measure should be symmetric. That is, the similarity degree between **d**1 and **d**2 should be the same as that between **d**2 and **d**1.
6. The value distribution of a feature is considered, i.e., the standard deviation of the feature is taken into account, for its contribution to the similarity between two documents. A feature with a larger spread offers more contribution to the similarity between **d**1 and **d**2.

IV. METHODOLOGY

A. Data Set Selection

Here we used three types of data sets. They are WebKB, Reuters 8 and RCV1. These data sets are stored in the form of HTML and text or SGM. Each data set is divided into two types. They are training and testing data set. WebKB contains web pages as the document which is collected by the world wide knowledge base. It does not contain predefined training set and testing set. So we randomly divide these data sets as training and testing set. Reuter's data set contains predesigned training set and testing sets. If the resultant data set contains 71% of training set and 29% of testing set then it is called as Reuters 8. RCV1 also contains predesigned training and testing data sets.

B. Clustering

After the data set collection then we perform the cluster formation. Cluster contains the more than one data set. For cluster formation we used k-means algorithm. In our proposed system, we used HAC (hierarchical agglomerative clustering). It is the one type k means clustering algorithm and it used bottom up approach for clustering. The HAC is mostly used to measure the performance.

In our proposed K-mean clustering algorithm is used for clustering information in randomly choice the data. This clustering is used for classification data. The HAC is used for the performance measurement in data .Hierarchical Agglomerative cluster (HAC) is a bottom up approach. This cluster is each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

C. Perform Classification

After form the cluster then we performs the classifications. We used SL-kNN and ML-kNN method is used to perform the classifications. SL-KNN is used for only one category and ML-KNN is used for more than two categories. These are used to measure the similarity between the documents. It gives best result for similarity. Classification is used to retrieve process. It

increases the fast of retrieve the data. It easily identifies the similarity between the documents. It gives related result efficiently. SL-KNN is used for single document categories.SL-KNN and ML-KNN are used for text classification and clustering to measure the similarity efficiently.

Here the similarity degree is increased and no of feature is decreased. Multi-label classification should not be confused with multiclass classification, which is the problem of categorizing instances into more than two classes.

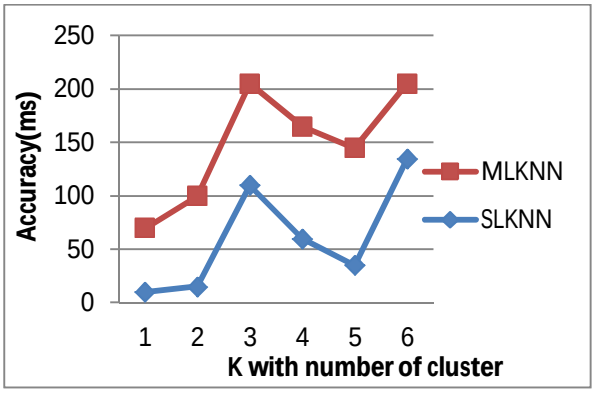


Fig 2: performance classification

D. Find Similarities

After select the classification methods then we finds the similarities between the data sets. The similarity measurement is based on the features of documents. It finds the word weight and distance to measure the similarities. Similarity measure has three properties they are, the presence and absence feature is more important than the difference of two values that are presented in the feature, the similarity degree is increased when the difference between the non zero values of feature is decreased, the similarity degree is decreased when the presence and absence of feature is increased.

The similarity measurement is used for two or different data set. The similarities find we are used in HAC cluster, this cluster is used for weighted and unweighted average in measurement.

E. Performance Comparison

After perform all the process then we compared the performance of process. In our proposed system, the performance is high compared than all existing system.

V. CONCLUSION

The main concept of our proposed system is to find the similarity between the documents. For similarity measure we used KNN based single label classification and KNN based multi label classification. Here we collect three data sets such as WEBKB, reuters-8 and RCV1. Classification is used for retrieve process and it increases the fast of retrieve the data. It easily identifies the similarity between the documents. It gives related result efficiently. Here we used HAC for clustering and it is the one type of k means clustering algorithm. It also measures the performance. In our proposed system, the data sets are stored in the form the XML. Here we used tf-idf (term frequency and inverse document frequency) to retrieve the document terms and It measures the similarity by calculating the weight. Here we used optimization and it identifies the no of cluster that we formed in the document. It increases the accuracy of similarity measurement.

REFERENCES

[1] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, *Member, IEEE,"* A Similarity Measure for Text Classification and Clustering", in IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 7, July 2014 1575

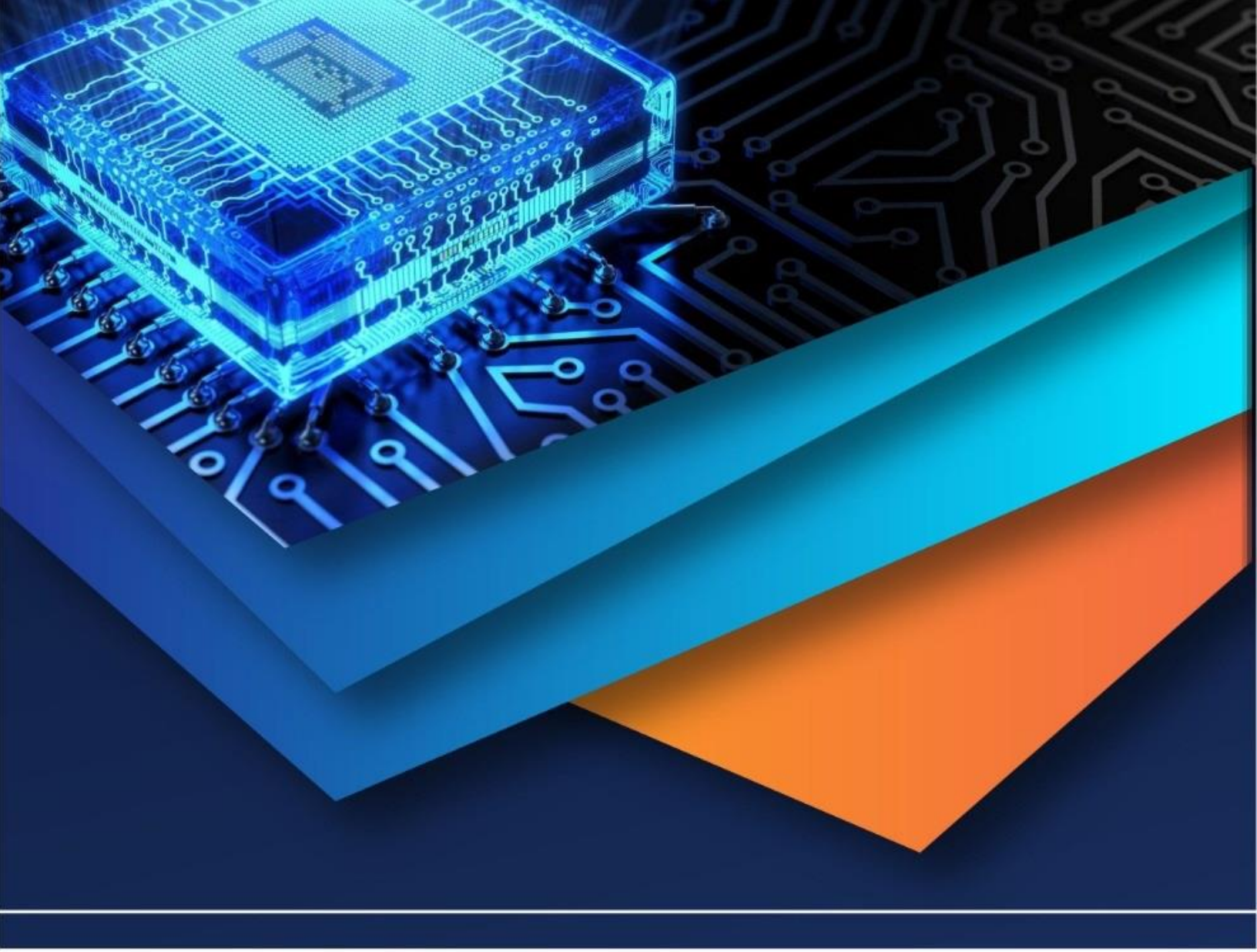
[2] D.Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," IEEE Trans. Knowl. Data Eng., vol. 17, no. 12, pp. 1624–1637, Dec. 2005.

[3] H. Chim and X. Deng, "Efficient phrase-based document similarity for clustering," IEEE Trans. Knowl. Data Eng., vol. 20, no. 9, pp. 1217–1229, Sept. 2008.

[4] S. Clinchant and E. Gaussier, "Information-based models for ad hoc IR," in Proc. 33rd SIGIR, Geneva, Switzerland, 2010, pp. 234–241.

[5] K. M. Hammouda and M. S. Kamel, "Hierarchically distributed peer-to-peer document clustering and cluster summarization," IEEE Trans. Knowl. Data

Eng., vol. 21, no. 5, pp. 681–698, May 2009.
[6] Y. Zhao and G. Karypis, "Comparison of agglomerative and partitional document clustering algorithms," in *Proc. Workshop Clustering High Dimensional Data Its Appl. 2nd SIAM ICDM*, 2002, pp. 83–93.
[7] T. Zhang, Y. Y. Tang, B. Fang, and Y. Xiang, "Document clustering in correlation similarity measure space," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1002–1013, Jun. 2012.
[8] M. L. Zhang and Z. H. Zhou, "ML-*k*NN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, 2007.
[9] Strehl and J. Ghosh, "Value-based customer grouping from large retail data-sets," in *Proc. SPIE*, vol. 4057. Orlando, FL, USA, Apr. 2000, pp. 33–42.
[10] F. Sebastiani, "Machine learning in automated text categorization," *ACM CSUR*, vol. 34, no. 1, pp. 1–47, 2002.
[11] T. W. Schoenharl and G. Madey, "Evaluation of measurement techniques for the validation of agent-based simulations against streaming data," in *Proc. ICCS*, Kraków, Poland, 2008.
[12] D. D. Lewis, Y. Yang, T. Rose, and F. Li, "RCV1: A new benchmark collection for text categorization research," J. Mach. Learn. Res., vol. 5, pp. 361–397, Apr. 2004.
[13] V. Lertnattee and T. Theeramunkong, "Multidimensional text classification for drug information," *IEEE Trans. Inform. Technol. Biomed.*, vol. 8, no. 3 pp. 306–312, Sept. 2004.
[14] S.-J. Lee and C.-S. Ouyang, "A neuro-fuzzy system modeling with self-constructing rule generation and hybrid SVD-based learning," *IEEE Trans. Fuzzy Syst.*, vol. 11, no. 3, pp. 341–353, Jun. 2003.
[15] G. Amati and C. J. V. Rijsbergen, "Probabilistic models of information retrieval based on measuring the divergence from randomness," *ACM Trans. Inform. Syst.*, vol. 20, no. 4, pp. 357–389, 2002.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY