iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4272

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: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

An Automated System for Bridging Health Care Seekers and Providers

Balika . J. Chelliah[1], J .Johnson *Rajasingh*[2]
[1, 2] *SRM Institute of Science & Technology*

Abstract*: *Data accessing in community based health services have encumbered the vocabulary gap between the medical experts and users seeking help. This paper puts forward a proposal to code the medical directory by combining local and global learning techniques which will help in eradicating the vocabulary gap. In local mining, each of the record in the medical directory is coded, this process allows each record to be extracted and corroborated with the terminologies. An off-shoot of this approach would be a corpus-aware vocabulary which will be used as terminology space for global learning. However, the local mining techniques may possess the disadvantages of having lower precision, ambiguity and a database with irrelevant medical concepts. Thus the global learning approach enhances the local mining technique with the help of various information cues. In practice this unsupervised method could have the potential to handle large datasets.
Index Terms: Healthcare knowledge, Local mining, Global learning, Machine learning, Question and Answer

I. INTRODUCTION

Emerging technologies have brought in drastic changes in medicine, in diagnosis as well as medical assistance. With the booming online trends, many forums have been developed with provides the seeker with instant medical advice or recommendations. Example of one such forum would be WebMD. Such forums favor both the experts/professional and the user seeking information. In case of the professionals, it provides a platform to interact with renowned medical experts, strengthen their knowledge, attract new patients. Also such forums provide an opportunity to increase their reputation among their colleagues and patients. For users, these yield instant answers to there queries and from trusted sources. These forums rely on community related data rather than waiting for the expert's response or browsing to see various related information from the web and picking the accurate information from relevant information, thus relying on community generated data.

However it might not be advisable to directly use the community generated data as it may result in a vocabulary gap, where users from different backgrounds might not be able to understand the various medical terms that has been used as that might not use the same vocabulary. In case of medical advisory forums, the questions asked may be narrated by different individuals in different manner. Also the expert may provide answers which many contain multiple possible meanings and the medical terminologies might not be standardized. Due to such complications few of these forums have encouraged experts to provide footnotes for the medical terms. For example, bronchiolar disorder, sibilant rhonchi all can be rooted to wheezing. Community generated data are often discordant and contravene the data analysis and data management process. Also there has been problems reported regarding the reusing the database as it does not contain any corpus aware system. Thus the proposed system is valuable for patients whose knowledge of medical vocabularies is inadequate to find the desired information, and for medical experts who search for information outside their field of expertise.

For the current health providing systems the data is organized and maintained manually, and there are two common approaches. One, Rule-based learning and Second, Machine learning. Rule based methods have a significant impact on the real-world applications and are proved to be faster. But rule construction can work well only with a stable database or corpus. Hence it is considered to be a challenging task which it comes to working with different corpuses. The second approach, Machine based learning rectifies this problem and enhances the learning process by having an inference model to get unseen data by applying the trained data.

We propose a machine learning system that uses the corpus aware terminology and maps the medical terminologies with them. Local mining for this unsupervised system can be explained in three steps. At first, the data is automatically coded with the footnotes. Then these medical concepts are detected. Finally these tracks will be normalized. Concept Entropy Impurity (CEI) is used for detection and normalization process. By-product of Concept Entropy Impurity method would be a corpus aware terminology. However the local mining technique may have certain drawbacks such as ambiguity, lower precision etc. To

©IJRASET (UGC Approved Journal): All Rights are Reserved

compensate such issues we propose global learning techniques. It complements the local mining approach by incorporating various information cues like inter-expert relationship, inter-terminology relationship. Thus the system will be able to handle real word complex large data sets.

II. LOCAL MINING

Local mining approach of the system can be explained as a three stage frame work. Given a medical record, the local mining technique starts with analyzing and extorting the noun phrase embedded in it. Once we extract the noun phrase, the medical concepts are detected. Medical concept detection is done by measuring the specificity. Final step in local mining is normalization.

A. Noun Phrase Extraction

Noun phrase extraction in local mining helps in extracting all the noun phrases from the given context. It is done by assigning a part-of-speech tag to each word in the medical record with Stanford POS tagger. For the extraction process there must be a pattern that is to be followed. The pattern is computed as:

(Adjective | Noun)* (Noun Preposition)?

(Adjective | Noun)* Noun

Any sequence of words which follows such a pattern is assured to be a noun phrase.

B. Medical Concept Detection

Second stage of local mining is detecting the medical concepts from the extracted noun phrase. This process helps in individualizing the medical concepts. Concept entropy impurity measures the domain relevance of a concept and is computed as:

$CEI(c) = \sum^{2}_{i=1\ P} (D_i|c) \log P (D_i|c)$

Where D1 and D2 represent medical corpus and general corpus respectively; and P (Di/c) denotes the probability that a concept c is related to a specified domain Di. Larger the CEI of a concept more relevant will be the concepts.

C. Medical Concept Normalization

Once the noun phrase has been extracted and the domain has been specified, the final step is to normalize the terms. Even with the domain relevant terms we cannot be confident about the standardized medical terminologies. Normalization acts as the key to bridge the vocabulary gap that exists between the health seeker and the medical experts.

Keeping in mind all the existing dictionaries, we are going to use SNOMED CT. SNOMED CT is usually used for electronic purposes as it provides core terminologies and the hierarchical representation of logic.

In normalization, the medical terms and their descriptions are mapped. Then a term is searched against the SNOMED CT library. The result for each is thus returned and for those with multiple results, the google distance is calculated to find the accurate answer out of the generated multiple results by exploring their closeness and occurrence in google.

$d(ti,c) = \underline{\max(\log r(ti), \log r(c)), \log r(ti,c)}$

$\log G - \min(\log r(ti), \log r(c))$

Where, G is the total number of documents retrieved from Google; ti and c respectively represent the terminology and the medical concept.

However, these techniques of local mining may not be sufficient enough to act as an effective and efficient system. It suffers from incompleteness, lower precision. This might be due to the irrelevant concepts embedded in the data set.

III. DISCUSSIONS

There are three drawbacks of local mining techniques like information loss, lower precision and over-widened space. The main reason for using global learning approach is to get over these drawbacks.

The information loss occurs due to the missing key concept in some medical record. The pseudo label estimation determines the high relevance score. By doing so the global learning approach is able to discover the missing key and link them to the given medical record.

The lower precision depends upon the presence of irrelevant concepts that have grammatical meaning for interaction between humans and understanding their intent. These concepts sparsely distribute in semantically similar data space. By using the empirical loss function the regularized term brings the relevance term down, this is how Global Learning is able to overcome irrelevant concepts.

©IJRASET (UGC Approved Journal): All Rights are Reserved

The over-widened space is depleted by merging all the locally defined terminologies in the data collection by naturally forming a corpus-aware terminology. It also utilizes the global learning approach by learning and propagating medical terms within scope.

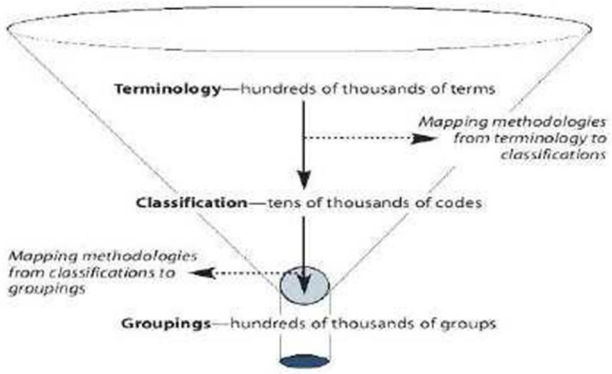


Fig .1 SNOMED CT standardizes medical terms, enabling accurate communication among diverse systems

IV. GLOBAL LEARNING

The aim of global learning technique is to learn suitable concepts and terminologies from the global terminology space and provide footnote to each medical record in the repositories of medical records. Out of all the global learning techniques, graph based techniques promises better performance. For this, we have to consider a repository of medical records and their associated locally mined terminologies which are denoted by Q = {q1; q2; . . . ; qn} and T = {t1; t2; . . . ; tm} respectively. This model will also be able to consider various information cues such as inter-expert relationship, inter-terminology relationship etc.

A. Inter-Terminology Relationship

In SNOMED CT, medical terms are organized into acyclic taxonomic (is-a) hierarchies. In such hierarchies each terminology may have multiple parents. To capture the inter-terminology hierarchical relationships we need to have a well-defined. The hierarchical relationship is quantitatively estimated as:

Rij = 1/2p; if ancestor-child relationships;
0; otherwise;

Allows the connection of more than two vertices thus summarizing the local grouping information. We extend the hyergraph model for the application. A hypergraph G (V, E, W) is composed of the vertex set, the hyperedge set E, and the diagonal matrix of hyper edge weight W. A probabilistic hypergraph G can be represented By a |V| x|E| incidence matrix H.

For our work, the medical records are considered as the vertices and they are connected by three types of hyper edges The first type takes each vertex as a centroid and forms a hyperedge by circling around its k-nearest neighbours based on medical record content similarities. The second type is based on terminology-sharing network. For each terminology, it groups all the medical records sharing the same terminology together. The third kind actually takes the users' social behaviours into consideration by rounding up all the questions answered by closely associated experts. As a consequence, up to N + M + U hyperedges are constructed in our hypergraph, where U is the number of involved experts. For each hyperedge, the likelihood of each constituent medical record belonging to its local group is defined according to its hyperedge type as follows:

$P(v_i , e_{ij}) = 1$ Inter-expert Relationships;
$K(q_i , q_j)$ Content Similarity;
1 Terminology-Sharing;

Where we have two terminologies ti and tj , p is the length of ancestor-child path between code ti and tj. And R is a matrix representing the weighted inter-terminology relationships.

The medical terminology hierarchy is used as it enhances our scheme in two ways. First, granularity mismatch problem is corrected and is done by applying appropriate weights to the ancestral nodes. And secondly, it boosts the accuracy of the coded data by discarding the sibling terminologies.

©IJRASET (UGC Approved Journal): All Rights are Reserved

B. Inter-Expert Relationship

The inter-expert relationships will be considered for professional experts and it will be viewed stronger if the experts are working in the same or related specific medical fields. This can be analyzed from their historical data that is the number of questions they have answered together. The relationship between two experts' u i and uj is calculated as:

$$J(Ui, Uj) = \frac{|Ui \quad Uj|}{|Ui \ U \ Uj|}$$

Where, Ui is the set of medical records that expert ui have involved. This is inspired by The Jaccard coefficient and it measures the similarity between two objects, which are represented by two unordered sets.

C. Probablistic Hypergraph

Global learning can be enhanced using graphical learning methods. Graph-based learning methods are classified into 2 major categories, graph based and hyper graph based. In both the cases vertices are taken as the samples. Simple graph technique takes into account only the pair-wise relationship between the samples or the vertices of the graph, overlooking any possible higher order relations. Whereas the hyper graph

Where K(. , .) is the Gaussian similarity function [32], which is a measure of similarity between two feature vectors. To explicitly incorporate the well-structured inter terminology Relationships, one more regularize term need to be incorporated.

$$\arg\min \sum_{i=1}^{M} \{ \Omega(f_i) + \lambda L(f_i) + \mu \sum_{i=1}^{M} R_{ij} \|f_i - f_j\|^2$$

Where Rij is the inter-class relationship between class i and class j and m is a regularization constant to regulate the effect of the third term.

V. EXPERIMENT

After gathering 1500 datasets and doing analysis, the record that contained question, answers, and all the experts who answered the question showed that the questions were lexically different had shared the same answers. By this the inference drawn was that the vocabulary gap among users was very large. The questions that were answered multiple times by the same expert were isolated and experts that had answered less were removed. This was done so that the efficiency does not decrease. The question samples collected consisted of short sentences and this does not provide co-occurrences for effective similarities. This limits the correctness that is achieved by general learning techniques. In our system, the answers are incorporated that supplements the short questions. This resolves the extra data problem. The ground truth was constructed with three master degrees and the labelers were trained with demonstrating examples and short tutorials.

VI. LOCAL MINING ANALYSIS

The noun phrases were extracted after the preprocessing of the data. The prediction guaranteed 100 percent results based on noun-phrase extraction.

Each noun phrase was estimated by comparing the frequencies of two different corpora. The medical concepts that have higher frequencies were more generic and less informative. Whereas the medical concepts that had lower frequencies were specific and descriptive. There were some processes that were seen on performing medical concept normalization:

The detected concepts were not mapped onto one entry in SNOMED CT. Some experts miss pelt menstruation as menses which is not considered as searchable.

Several medical concepts may be changed into a similar terminology and also the vocabulary varied among information providers or generators.

More than 85 percent of medical concepts are different in terms of their normalized terminologies. Thus the utility of a secured terminology is very sparse and these are desired to be normalized.

The proposed normalization approach was validated by the voting done by the annotators and the accuracy was about 82 percent.

VII. GRAPH-BASED GLOBAL LEARNING ANALYSIS

A. The effectiveness of the Global Learning Technique Used Was Explained

1) PR Feedback- Pseudo-Relevance Feedback had a SVM classifier that established the relevance score between a term and the medical record. The training data was created based on the idea that the samples that were at the highest rank were more similar that the lower ones.

©IJRASET (UGC Approved Journal): All Rights are Reserved

2) RW Re-ranking- Random Walk based Re-ranking was a simple graph that was constructed based on the relevance score between the terminologies and the medical records.

3) CH Learning- Conventional Hyper graph Learning was an approach where an inter-terminology hierarchical relationship was not taken into account. The results were achieved from multi-label transductive learning problem instead of regularizing a terminology.

4) GG Learning- Global Learning Approach: All the above mentioned approaches involved parameters that were carefully tuned and their performances were compared. The methods contained the local mining results that had been coded locally. The ground truth was obtained by manual labeling process. Random selection of medical records was done and for each record four different ranking list was generated. Voting was done by the annotators that resulted in final relevance of each terminology. The kappa method was used to analyze the inter-rater reliability. The kappa metric is a quantity measure that can be interpreted as expressing the amount of agreement that was seen among raters whose results ranged from 0 to 1. The CH Learning and the GG Learning evaluated the depth of NDCG that resulted in better performance that the other two approaches. This was due to two reasons. One being that the hyper-graph-based learning captured the high-ordered relationship that summarized the local group among medical records. The PR Feedback and RW Re-ranking approaches characterized simple pair wise relationship. Integration of heterogeneous information cues was done by the hyper graph-based learning.

VIII. MEDICAL TERMINOLOGY ASSIGNMENT

The annotation task includes that precision being more important than recall. Two matrices were used that characterized precision from different aspects. First, measures the probability of finding similar terminology among top recommended ones by the means of average that is, S@K over the testing records. If there is even a single relevant terminology then it is desired to be in the top recommendations or else its 0. Second is for proportion of recommended terminologies that are similar, that is, P@K average. The difference is that the annotators were asked to label the top terminologies as positive or negative. Methods that compared our proposed system with other coding approaches:

A. *Tag Collective*- was a tag recommendation based on collective knowledge which is a statistical and data-driven approach. Terminologies were derived based on their repetitive occurrences of their local mining terminologies. Vote-based strategy was used for aggregations

B. *Tag Assist*- generates the search query by annotating a medical record from the given medical record. Searching from the whole of the consisting medical record and selecting suitable terminologies from the retrieved records was a part of this approach.

The other two were Local Mining and Local + Global scheme. The compared results were in the form of averages that is S@K and P@K. The local mining approach gave the worst results that mapped irrelevant terms to the medical concept. Our proposed system assures one among the four terminologies to give 100 percent desired result. Thus we can estimate the effectiveness of the global learning approach. TagCollective ignores the lexical property which is more reliable than terminologies co-occurrences based on similarity of medical records and holds good only for the repetitive terminologies. TagAssist is better than TagCollective since it considers pair wise similarities but not the hierarchical relationship among social connection between the experts. ANOVA-Analysis Of Variance test was performed with a single factor which resulted in a variety of values that had varied based on the techniques that were used for comparison between the proposed system and competitive coding approaches. It also indicated the statistical analysis that indicated significant statistical improvements. We also inferred that the local mining terminologies were irrelevant to the medical records and some key terms were missing. This resulted in missing key concepts of medical terminologies of the medical record. The global learning approach enhanced and comprehended the terminologies and made it more reliable.

IX. CONCLUSION

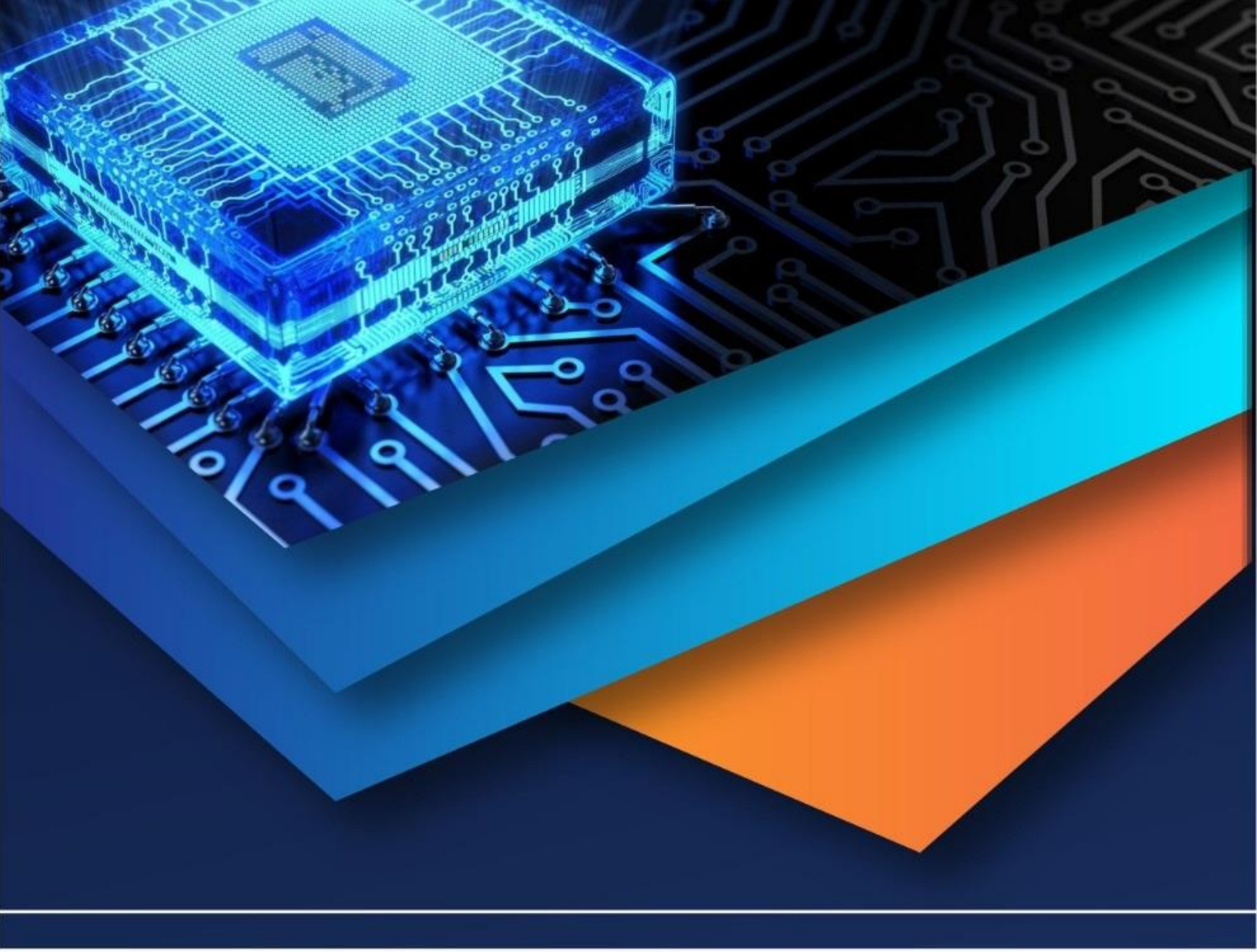
This paper presents a scheme which helps in eliminating an vocabulary gap that could exist between the users and the health care experts. It follows unsupervised learning, which comprises of two main components , local mining and global learning. The former has a three stage framework where the data is automatically coded, followed by medical concept
detection and normalization. However, the local mining approach itself does not prove to be efficient enough, due to lower precision and presence of irrelevant data. Hence we inculcate the global learning technique which enhances the local code with the help of various information cues. Thus our scheme has the potential to hold a large data set as it comes under unsupervised learning approach.

©IJRASET (UGC Approved Journal): All Rights are Reserved

REFERENCES

[1] L. Nie, M. Akbari, T. Li, and T.-S. Chua, "A jo int localglobal approach for medical terminology assignment,"in Proc. Int. ACM SIGIR Conf., 2014.
[2] L. Nie, T. Li, M. Akbari, and T.-S. Chua, "Wen zher: Comprehensive vertical search for healthcare domain," in Proc. In t. ACM SIGIR Conf., 2014, pp. 1245–1246.
[3] M.-Y. Kim and R. Goebel. Detection and normalization of medical terms using domain-specific term frequency and adaptive ranking. In Information Technology and Applications in Biomedicine, IEEE International Conference on, 2010.
[4] G. Leroy and H. Chen, "Meeting medical terminol ogy needs-the ontology-enhanced medical concept mapper," IEEE Tra ns. Inf. Technol. Biomed., vol. 5, no. 4, pp. 261–270, Dec. 2001.
[5] G. Zuccon, B. Koopman, A. Nguyen, D. Vickers, and L. Butt, "Exploiting medical hierarchies for concept-based i nformation retrieval," in Proc. Australasian Document Comput. Symp. 2012, pp. 111–114.
[6] E. J. M. Laurıa and A. D. March, "Combining Bay esian text classi-fication and shrinkage to automate healthcare coding: A data quality analysis," J. Data Inf. Qu
[7] R. L. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. IEEE Transactions on Knowledge and Data Engineering, 2007.
[8] C. Dozier, R. Kondadadi, K. Al-Kofahi, M. Chaudhary, and X. Guo. Fast tagging of medical terms in legal text. In Proceedings of the International Conference on Artificial Intelligence and Law, 2007.
[9] S. Fox and M. Duggan. Health online 2013. Survey, Pew Research Center, 2013.

©IJRASET (UGC Approved Journal): All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)