A Biomedical Search Support System for Improving the Data Search Accuracy

Vikas Ransore, Jagdish Raikwal
Institute of Engineering & Technology,
Devi Ahilya University, Indore, India

ABSTRACT
Medical domain is a huge scientific domain where a large amount of data is available for analysis and discovery. In this proposed work a search improvement technique is suggested which provides guidelines for finding appropriate data during medicine search. That is because the basic search systems contains three major modules first query interface, second for search methodology and finally search ranking. User provide input to the search interface than the search system find the data from database according to the user query, finally results are ranked according to the relevancy of the user query. The proposed improvement is implemented on the query input phase for improving the user query for finding the more accurate and nearer medicines from the database.

Keywords
Text mining, support vector machine (svm), principal component analysis (pca), biomedical research.

1. INTRODUCTION
The new era reflect the technology is growing continuously and their advantages are reflecting in our daily life. Now in these days the computer users are directly connected with internet and frequently usage search engines to find the information from web. Web search engines search the user specified data from entire web and produces the results. The web search engines are includes three key elements first the user query, which is provided by the user to perform search. Second the Search algorithm that accept the user query and find the desired data from web and finally the ranking function that provide the relevancy on search results according to the user query.

In the presented study a new kind of search engine is indented to design which provide the search results for the medicines [Gang Luo, "Design and Evaluation of the iMed Intelligent Medical Search Engine"] according to the user.

Input symptoms and the medicine name [RaduDragusina, Paula Petcua, Christina Lioma,BirgerLarsend,Henrik L. Jørgensene, Ingemar J. Coxa, Lars Kai Hansena, PeterIngwersend, Ole Winthera, "FindZebra: A search engine for rare diseases"]. Using various text processing techniques the proposed system is implemented. The proposed search engine is designed to enhance the user query because the wrong search query produces the incorrect results. In addition of that a machine learning based search engine is designed which includes the learning about the user query and their results to optimize the search results next time similar user query search. In order to learn the search and their search patterns the system utilizes the SVM (support vector classifier) [AartiKaushik, Gurdev Singh & Anupam Bhatia, "SVM Classification in Multiclass Letter Recognition System"] that helps to learn the search content and their results. This section demonstrates the simple overview of the proposed medical search engine and their initial components.

The key objective of the proposed work is to enhance the medical record search process by utilizing the concept of query optimization and data mining classification techniques. Therefore the following works are included in presented study.

Study of medical search engines: in this phase various medical databases search techniques are studied which provides the understanding about the contents and their issues during search.

Study of user query relevance data search techniques: in order to perform the search how user query is processed and data are extracted from the database is reported in this phase of the work.

Study of query optimization techniques: in this phase the query optimization techniques are studied.

Design and implementation of a new medical search enhancement technique: in this phase the new medical search engine is designed and implemented using JAVA technology.

Performance study of the proposed technique: In this phase the performance of the proposed search system is evaluated and results are populated.

2. PROPOSED WORK
Text processing and text search [Ian H. Witten, "Text mining"] is a domain of data mining and knowledge processing. Text is used to represent knowledge and also used for communication. Due to this a document contains different kinds of knowledge and data. Therefore text processing and recognition faces various issues and challenges such as lingual issues, semantics, and classification and categorization issues. In this presented work a text search engine is developed for medical record search. Medical science is a where a number of similar kinds of records are available for analysis and discovery. In this proposed work a search improvement technique is suggested which provides guidelines for finding appropriate data during medicine search. That is because the basic search systems contains three major modules first query interface, second for search methodology and finally search ranking. User provide input to the search interface than the search system find the similar data from database according to the user query, finally results are ranked according to the relevancy of the user query. In order to rank the outcomes according to the relevancy of data sorting is performed.

According to the discussion the search outcomes are enhanced in all the three phases. In the proposed system improvement is taken place on the query input phase. When user search a

medicine different medicine are available which having similar spellings but having different chemical compositions. Thus proper record search is necessary and essential. Therefore the user query is manipulated and enhanced for finding the more accurate results. In addition of that a predictive system is also required which measures the current query input and produces the next spelling which is possible according to availability of records.

The main aim of the proposed system is to improve the medical data base search technique. Therefore the query optimization techniques are data classification techniques are utilized for finding the user relevant content search from the medical database. The proposed system architecture is based on the concept of improving the search query using the previous medical search queries and optimizing the results discovery by improving the relevancy of the search technique. The above given figure demonstrate the proposed system architecture of the query improvement technique. The proposed technique involves a search interface which accepts the user query for search. There are two aspects of the user query input first during input and after input of user query. During input of user query system identify the key strokes and according to the changing inputs the new query suggestion is provided through the most nearest attributes which is performed previously most ***relevant previous user queries*** are extracted from data base. This extracted user queries are sorted and listed using the ***auto complete text box*** for guiding the user for writing the correct query. Then after the similar query words are collected as queries are provided as input to the ***PCA algorithm***. Where the essential features are extracted from data and using selected features and the ***medicinal data base*** a ***SVM classifier*** is trained. The SVM classifier can accept the user current input values and predict the actual the content which is actually user want to find in medicinal data base. On the other hand the data final user query which is used for finding the outcomes from the database is then preserved into the query database.

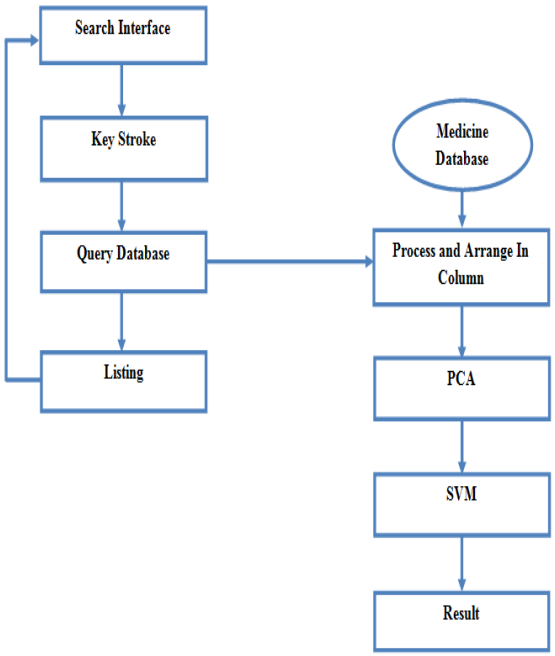

Fig. 1 proposed architecture

3. ALGORITHM

3.1 Principle Component Analysis

The weighted average is more formally called a linear combination: it is a way of combining the original variables in a linear way. It is also sometimes called a latent variable where, in contrast, the original variables are manifest. The data are collected in a matrix X with I rows (i = 1, …, I; samples/objects) and J columns (j = 1, …, J; variables), hence of size I × J. The individual variables (columns) of X are denoted by x_j (j = 1, …, J) and are all vectors in the I-dimensional space. A linear combination of those x variables can be written as $t = w_1 X x_1 + \cdots + w_j X x_j$, where t is now a new vector in the same space as the x variables (because it is a linear combination of these). In matrix notation, this becomes$t = Xw$, with w being the vector with elements W_j (j = 1 …z, J). Since the matrix X contains variation relevant to the problem, it seems reasonable to have as much as possible of that variation also in t. If this amount of variation in t is appreciable, then it can serve as a good summary of the x variables [9].

Hence, the fourteen variables of X can then be replaced by only one variable t retaining most of the relevant information. The variation in t can be measured by its variance, var(t), defined in the usual way in statistics. Then the problem translates to maximizing this variance choosing optimal weights $w_1 \ldots w_j$. There is one caveat, however, since multiplying an optimal w with an arbitrary large number will make the variance of t also arbitrary large. Hence, to have a proper problem, the weights have to be normalized. This is done by requiring that their norm, i.e. the sum-of-squared values, is one. Throughout we will use the symbol $\|.\|^2$ to indicate the squared Fresenius norm (sum-of-squares). Thus, the formal problem becomes

$$\underset{|w| = 1}{argmax\ var(t)}$$

which should be read as the problem of finding the w of length one that maximizes the variance of t (note that ||w|| = 1 is the same as requiring $\|w\|^2 = 1$). The function argmax is the mathematical notation for returning the argument w of the maximization function. This can be made more explicit by using the fact that t = Xw:

$$\underset{|w| = 1}{argmax(t^T t)} = \underset{|w| = 1}{argmax(w^T X^T X w)}$$

Where it is assumed that the matrix X is mean-centered (then all linear combinations are also mean-centered). The latter problem is a standard problem in linear algebra and the optimal w is the (standardized) first eigenvector (i.e. the eigenvector with the largest value) of the covariance matrix $X^T X / (n-1)$ or the corresponding cross-product matrix$X^T X$.

3.2 Support Vector Machine

In device learning, SVM is administered learning replicas with connected learning algorithms. SVMs belong to relations of global linear categorizers and can be understudied as an expansion of the perceptions. SVMs are a group of supervised learning methods that can be applied to classification or regression. It is primarily a two class classifier. SVMs can efficiently perform non-linear categorization using what is called the essence function; indirectly map their inputs into high-dimensional feature spaces. It can also solve multiclass problem with the help of kernel methods and kernel function. It aims to maximize the width of the margin between classes, that is, the vacant area between the decision boundary and the nearest training pattern. The basic idea of SVM classifier is to choose the hyper plane that has maximum margin. The dash

appearance strained similar to the unraveling line spot the distance between the separating line and the neighboring vectors to the line. The space among the dashed appearance is called the edge. The vectors (points) that limit the width of the edge are the holdup vector. Suppose the two classes can be presented by two hyper planes parallel to the optimal hyper plane [8].

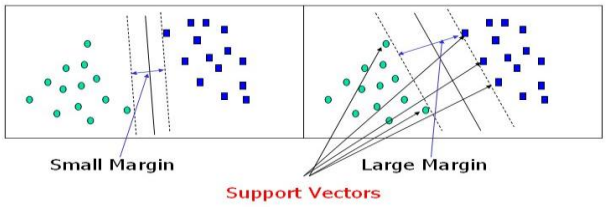

Figure 3.1: Support Vectors

$w_n x_t + b \geq 1$ for $y_t = 1$, t = 1,2,3.....,k

$w_n x_t + b \leq -1$ for $y_t = -1$

Where w = {w1, w2, w3,.....,w_n} is a vector of n constituent

Figure 2.1 signify the little edge, big edge and support vectors through categorization of a two class dataset.

3.3 Kernel Method Functions

The essence method consists of two components: First one is the option of essence and the second one is the technique which receives essence as input. The basic thought of essence technique is to drawing the information from input space to feature space F using Ø [7], Ø: X → F where X = "inputs", F = "feature space", Ø = "feature map". The liberty of the unique data is called input space we say that k(x, y) is a essence purpose if there is a characteristic map Ø such that for all x, y K(x, y) = Ø(x).Ø(y). In example credit a characteristic space is a theoretical space where every prototype model is represent as a point in n-dimensional space. Its measurement is firm by the amount of characteristics used to explain the prototypes. The notion of an essence mapping purpose is very influential. It permits SVM prototypes to present partings even with very complex boundaries. Figure 3.2 shows that how to map the data from low dimensional space to higher dimensional space.

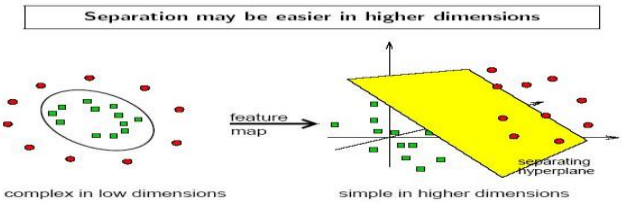

Figure 3.2: Non-linear Separable Data [7]

The mapping function requires to be calculated because of a instrument called essence deception. The essence deception is a arithmetical instrument which can be applied to any technique which exclusively depends on the dot product among two vectors. Every place a dot product is used; it is substituted by a kernel function. When appropriately functional, those applicant linear techniques are converted into non-linear techniques. Those non-linear techniques are reporter to their linear exclusive in service in the variety space of a characteristic space Ø. However, because kernels are used, the Ø purpose doesn't require to be still clearly calculated. Essence purpose must be incessant, symmetric, and most quite should have a optimistic (semi-) exact Gram matrix. Essences which are said to please the Mercer's theorem are optimistic semi-definite, sense their essence matrices has no non-negative Eigen values. A optimistic specific essence assure that the optimization difficulty will be curved and resolution will be exclusive.

Types of kernel functions:

1. Linear Kernel

The Linear essence is the easiest essence purpose. It is known by the inner product <x,y> plus an optional constant c.

$$k(x, y) = (x^T y + c)$$

2. Polynomial Kernel

The Polynomial essence is a non-motionless essence. Polynomial essences are fined appropriate for difficulties where all the teaching data is regularized.

$$k(x, y) = (\alpha x^T y + c)^d$$

Adaptable limits are the incline alpha, the steady term c and the polynomial amount d.

3. Gaussian kernel

The Gaussian essence is an instance of radial basis purpose essence.

$$k(x, y) = e^{\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)}$$

4. IMPLEMENTATION

The implemented system and their user interface are described in this section, which reflect the planning of the implementation.

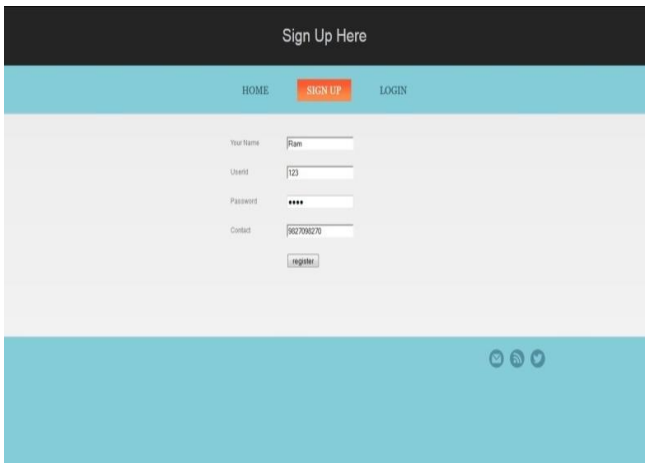

Fig 4.1 new user signup

The figure 4.1 shows the initial project screen that contains a registration for new user of the system after successfully registration of the user can use the proposed search system. The next screen shows the login window that accepts the user id and password as the parameters and validates the user input using the database registration as given in figure 4.2. If the user provides the correct login credentials then user get access to the search screen.

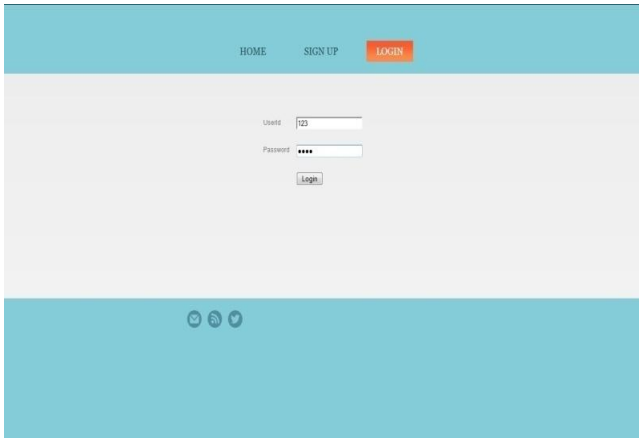

Fig 4.2 login screen

The initial search window of the project is given using the figure 4.3.

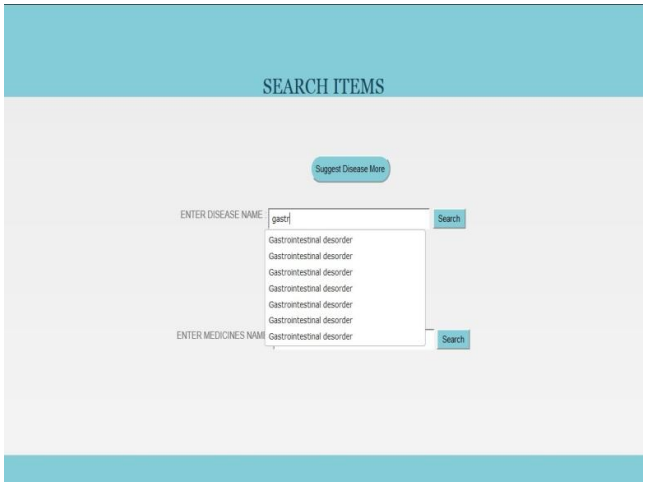

Fig 4.3 search using disease

In this screen user provides the disease name as input and their medicine names are produced as results. The result screen of the search is given using figure 4.4.

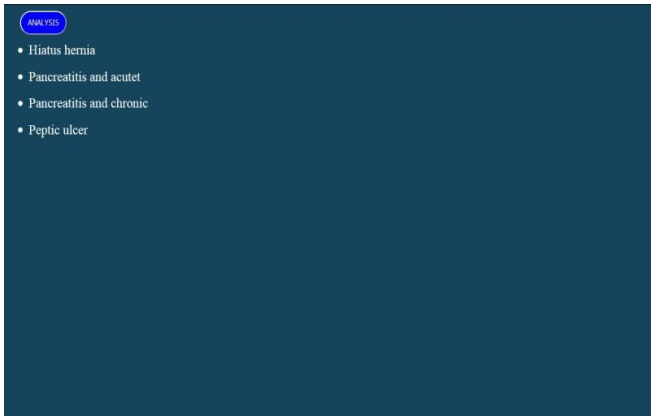

Fig 4.4 results screen

In this screen we can see the result for disease search

5. RESULT ANALYSIS

5.1 Memory consumption

The amount of memory required to execute the algorithm for finding the results is termed as the memory consumption.

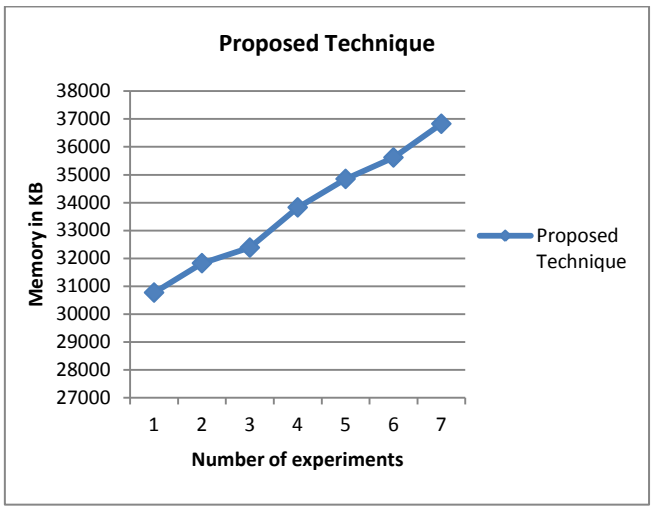

Fig 5.1 memory consumption

The memory consumption of the system is simulated using figure 5.1 in this diagram the X axis shows the number of experiment performed and the Y contains the amount of main memory consumed. According to the obtained results as the number of instances in data base for classification is increases the amount of memory requirement is increases.

5.2 Time consumption

The amount of time required to execute the algorithm for evaluation of data is known as the time consumption. The figure 5.2 shows the time consumed for finding and learning of the data to retrieve the accurate text from database.

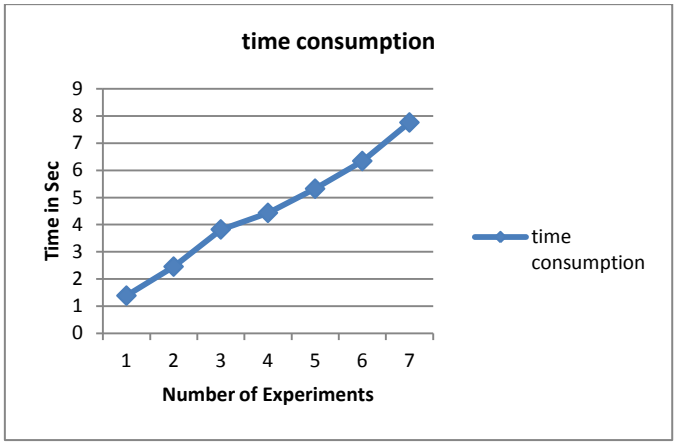

Fig 5.2 time consumption

In this diagram the X axis shows the number of experiments performed with the system and Y axis shows the time consumption in terms of second. According to the given diagram the performance of the proposed algorithm is depends upon the amount of data in database thus when the experimental data is increases the amount of time is also increases with the experiments.

5.3 Precision rate

In search systems the precision is a fraction of search results which is most relevant to the input query. The provided precision of the proposed system and their filtering options are given using figure 5.3. This can be evaluated using the user feedback basis and can be evaluated using the following formula.

$$precision = \frac{releventdocument \cap retrieveddocumets}{retrieveddocuments}$$

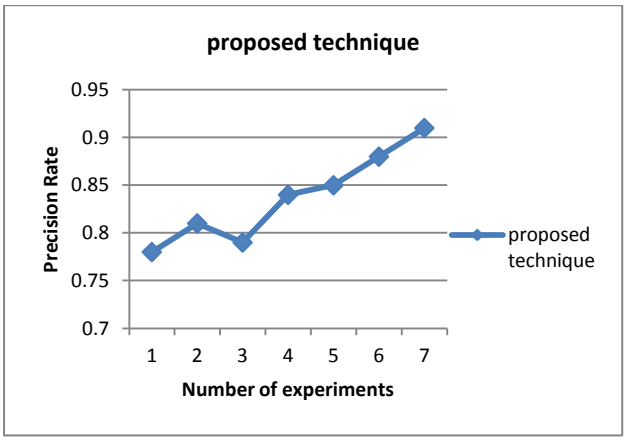

Fig 5.3 precision

The given figure 4.3 shows the precision of the proposed and traditional system, in this diagram the X axis shows the different experiments performed with the system and the Y axis shows the precision rate of the system. According to the obtained results the performance of the proposed system is adoptable.

5.4 Recall

The search recall values are measured in this section, that is an accuracy measurement in terms of relevant document retrieved according to the input search query. This can be evaluated using the following formula.

$$recall = \frac{relevantdoucment \cap retrieveddocuments}{relevantdocuments}$$

The recall rate of the system is given using figure 5.4 in this diagram the X axis shows the number of experiments performed and the Y axis shows the recall of the proposed and traditional system. According to the obtained results the performance of the proposed technique is much adoptable.

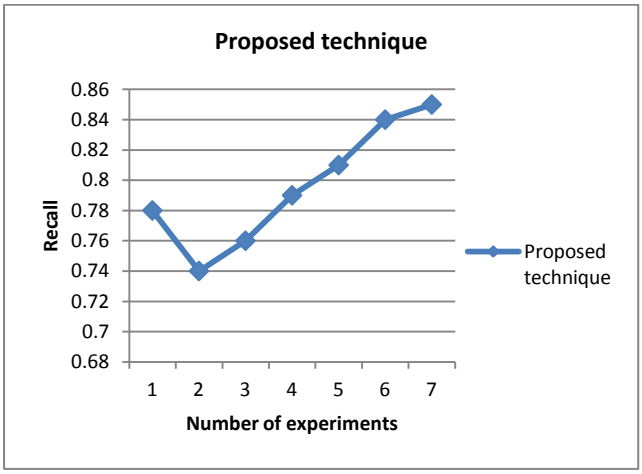

Fig 5.4 recall

That is estimated using the precision and recall values estimated using the search or document retrieval technique. That represents the harmonic mean of the system, and can be evaluated using the following formula.

$$F - measures = 2 . \frac{precision . recall}{precision + recall}$$

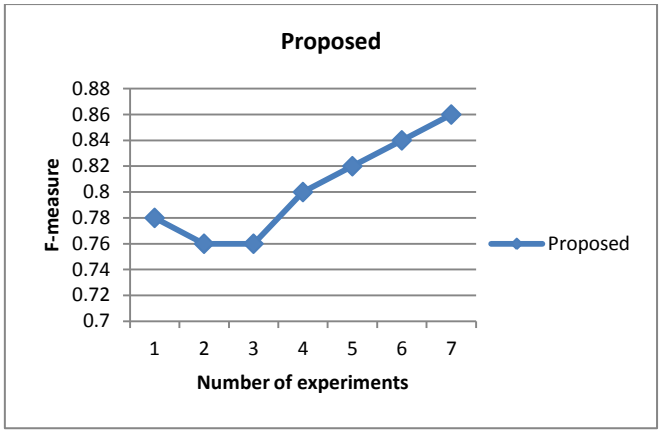

Fig 5.5 f-measure

The f-measure values of the proposed search engine is given using figure 4.5 the given values are computed on the basis of the precision and recall using the above listed formula. In this diagram the X axis contains the number of different experiments performed with the different query and the Y axis contains the estimated f-measure values. According to the obtained results the f-measures shows the effective results by improving the quality of search outcomes.

6. CONCLUSION AND FUTURE WORK

The internet is a huge support for new generation applications and the data search. A number of users frequently access the internet to find the data of interest. In this presented work the medical data search technique for finding the medicine and disease is implemented and their performance is demonstrated in this paper. After implementation of the proposed medical record search engine the accuracy of the proposed data model is found promising additionally produces more accurate results among the available set of data.

In near future that is required to enhance the data and collect a significant amount of data for the medical dataset development. In addition of that it is also required to enhance the model's performance in terms of their complexity in terms of memory consumption with the huge amount of data repository.

7. REFERENCES

[1] Radu Dragusina, Paula Petcua, Christina Lioma,Birger Larsend, Henrik L. Jørgensene, Ingemar J. Coxa, Lars Kai Hansena, PeterIngwersend, Ole Winthera, "FindZebra: A search engine for rare diseases", 23 February 2013, DOI:10.1016/j.bbr.2011.03.031

[2] Aarti Kaushik, Gurdev Singh & Anupam Bhatia, "SVM Classification in Multiclass Letter Recognition System", Global Journal of Computer Science and Technology, Software & Data Engineering, Volume 13 Issue 9 Version 1.0 Year 2013

[3] NATHAN HALKO, PER-GUNNAR MARTINSSON, YOEL SHKOLNISKY, AND MARK TYGERT, "AN ALGORITHM FOR THE PRINCIPAL COMPONENT

ANALYSIS OF LARGE DATA SETS", http://amath.colorado.edu/faculty/martinss/Pubs/2010_07_05_outofcore.pdf

[4] Gang Luo, "Design and Evaluation of the iMed Intelligent Medical Search Engine",IEEE 25th International Conference on Data Engineering, 2009. ICDE '09

[5] Ian H. Witten, "Text mining", Computer Science, University of Waikato, Hamilton, New Zealand

[6] Rasmus Bro and Age K. Smilde, "Principal component analysis", DOI: 10.1039/C3AY41907J (Tutorial Review) Anal. Methods, 2014, 6, 2812-2831

[7] James Kwok, "Kernel Methods in Machine Learning", Department of Computer Science and Engineering Hong Kong University of Science and Technology, 2006.

[8] Asa Ben-Hur, Jason Weston, "A User's Guide to Support Vector Machines," Department of computer Science Colorado State University.

[9] Ramus Bro and Age K. Smilde, "Principal component analysis", DOI: 10.1039/C3AY41907J (Tutorial Review Anal. Methods, 2014, 6, 2812-2831)