An efficient approach for medical text categorization based on clustering and similarity measures

Assistant Lecturer :Amal Hameed Khaleel

College of Science, University of Basrah

Abstract

The huge amount of medical information available in the medical document, makes the use of automated text categorization methods essential in clinical diagnosis and treatment. Automatic categorization of a text can provide information about classes which a text belongs to. This paper can serve as a medical diagnosis tool for categorization patient records by propose text categorization algorithm based on the similarity cluster centers for the categorization of patients with eye diseases records. We propose VEMST algorithm as update to EMST algorithm by using variance to find cluster centers. A text categorization algorithm is developed using two similarity measures (cosine , common words) to classify the categorical data. The results showed that when the number and size of medical documents used great for training the classification accuracy increases, as we noticed when we use comparing medical terms method in the preprocessing phase, the accuracy is better than the use of frequency of all terms in medical document, as well as the execution time at least. Finally, we found the performance of our system when we use the cosine similarity measure is better than his performance with the use of the similarity of common words scale.

Keywords

Data mining, Text mining, Text categorization(TC), Midline, Euclidean minimum spanning tree (EMST), Cosine similarity, Common word similarity.

1. Introduction

There are a huge volumes of data growing on the internet in the form of research papers and web documents. The amount of medical literature continues to grow and specialize, most of the data is contained by the journal of medicines and biology which makes this type of textual mining a central and core problem [1]. The access to a large quantity of textual documents turns out to be effectual because of the growth of the technical documentation, medical data and more. These textual data comprise of resources which can be utilized in a better way. Thus, text mining (TM) is a prominent and tough challenge due to the value and ambiguity of natural language which is employed in majority of the existing documents. Data mining techniques can be adapted in a better way to mine text. Thus, Text mining refers to one of the application of data mining techniques to automated discovery of valuable or interesting information from unstructured text [2].

Text Categorization (TC) is one of the important tasks in information retrieval and data mining. Automated TC involves assigning text documents in a test data collection to one or more of the pre-defined classes/categories based on their content. Unlike manual classification, which consumes time and requires high accuracy, so automated TC makes the classification process fast and more efficient since it automatically categorizes documents. The goal of TC task is to assign class labels to un labelled text documents from a fixed number of known categories [3]. Clustering is an example of unsupervised classification and algorithms in clustering methods are mainly divided into two categories: A partition algorithms and hierarchical algorithms. A partition clustering algorithm partition the data set into desired number of sets in a single step [4]. A hierarchical clustering algorithm divides the given data set into smaller subsets in hierarchical fashion and this hierarchical clustering approaches are related to graph theoretic clustering [5]. The minimal spanning tree (MST) is one of the hierarchical clustering which is known to be capable of detecting clusters with various shapes and size, so we used in this paper.

This paper is organized as follows: related works and detailed view are described in Section 2,3. In Section 4,5 system architecture and experiment results are discussed respectively , and finally conclusions and future works are given in Section 6.

2. Related works

A review of some recent researches related to clustering techniques especially spanning tree algorithm are represented in this section.

Zhao and Karypis in 2001[6] they build an agglomerative tree for the documents belonging to each one of the clusters, and then we combine these trees by building an agglomerative tree, whose leaves are the partitionally discovered clusters. This approach ensures that the k-way clustering solution induced by the overall tree is identical to the k-way clustering solution computed by the partitional algorithm. Laszlo and Mukherjee in 2005 [7] present an MST-based clustering algorithm that puts a constraint on the minimum cluster size rather than on the number of clusters. This algorithm is developed for micro aggregation problem, where the number of clusteres in the data set can be figured out by the constraints of the problem itself. Grygorash et al. in 2006 [8] proposed two MST-based clustering algorithms called the Hierarchical Euclidean Distance based MST clustering algorithm (HEMST) and the Maximum Standard Deviation Reduction clustering algorithm (MSDR) respectively. Pakhomov et al. in 2008 [9] used a bag-of-words approach to process the text of physical examination sections of in-patient and out-patient clinical notes in order to identify whether the findings of structural, neurological, and vascular components of a foot examination revealed normal or abnormal findings or if they were not assessed. Wang et al. in 2009 [10] proposed a new approach called Divide and Conquer Approach to facilitate efficient MST-based clustering by using the idea of the "Reverse Delete" algorithm. which by using an efficient implementation of the cut and the cycle property of the minimum spanning trees, can have much better performance than O(N^{2}).

Peter and Victor in 2010 [5] proposed two minimum spanning trees based clustering algorithm. The first algorithm produces k clusters with center and guaranteed intra-cluster similarity. The radius and diameter of k clusters are computed to find the tightness of k clusters. The variance of the k clusters are also computed to find the compactness of the clusters. The second algorithm is proposed to create a dendrogram using the k clusters as objects with guaranteed inter-cluster similarity. Peter et al. in 2010 [11] present algorithm uses a new cluster validation criterion based on the geometric property of data partition of the data set in order to find the proper number of segments. The algorithm works in two phases, the first phase of the algorithm creates optimal number of clusters/segments, whereas the second phase of the algorithm further segments the optimal number of clusters/segments and detect local region outliers. Chakrabarty and Roy in 2014 [12] proposed work evolves a text clustering algorithm where clusters are generated dynamically based on minimum spanning tree incorporating semantic features. The proposed model can efficiently find the significant matching concepts between documents and can perform multi category classification. The formal analysis is supported by applications to email and cancer data sets. The cluster quality and accuracy values were compared with some of the widely used text clustering techniques which showed the efficiency of the proposed approach.

3. Basic Principles

3.1 Minimum Spanning Tree Clustering Algorithm (MST)

A spanning tree is an acyclic sub graph of a graph S, which contains all vertices from S and is also a tree. The minimum spanning tree (MST) of a weighted graph is the minimum weight spanning tree of that graph [13]. The Euclidean minimum spanning tree (EMST) method starts by constructing an MST from the points in S. The weight of an edge in the tree is the Euclidean distance between the two end points. Next, the average weight w_{avg} of the edges in the entire EMST and its standard deviation σ are computed; any edge with a weight $w > w_{avg} + \sigma$ is removed from the tree. This leads to a set of disjoint subtrees ST ={T1, T2, . . .}. Each of the subtrees Ti is treated as a cluster, which has a centroid ci.

There are several possibilities in the construction of the clusters in the subtrees as follows:

If the number of the subtrees $|ST| < k$, $k - |ST|$ additional longest edges are removed from the entire edge set of ST to produce k disjoint subtrees.

If $|ST| > k$, a representative point(centroid) is identified for each subtree. Once all the representative points are found, each point in a particular subtree is replaced with the representative point(centroid) of the subtree, thus reducing the number of points in S to $|ST|$.

If $|ST| = k$, the clustering process is considered complete, having produced the required k clusters [8].

3.2 Clustering Validation

There are a number of validity measures that have been proposed clustering validation, which can be divided into three main types (External, Internal and Relative). Where external validation measures employ criteria that are not inherent to the dataset. Internal validation measures employ criteria that are derived from the data itself. Relative validation measures are used to compare different clustering obtained by varying different parameters for the same algorithm, and to choose the number of clusters k.

The silhouette coefficient is one of internal validation measures, where it's a measure of both cohesion and separation of clusters, and is based on the difference between the average distance to points in the closest cluster and to points in the same cluster [14].

The silhouette coefficient (S) for each document i works as follow:

let a(i) be the average dissimilarity of i with all other data within the same cluster. This paper used Euclidean distance as the similarity measure.

Let b(i) be the lowest average dissimilarity of to any other cluster which is not a member. The cluster with this lowest average dissimilarity is said to be the "neighboring cluster" of because it is the next best fit cluster for point.

The silhouette coefficient s(i) defined as:

$$S(i) = \begin{cases} 1 - a(i)/\, b(i) & \text{If } a(i) < b(i) \\ 0 & \text{If } a(i) = \ b(i) \\ b(i)/a(i) - 1 & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1) \end{cases}$$

From the above definition it is clear that $-1 \leq s(i) \leq 1$, where

If s(i) to be close to 1, we require a(i)<b(i). As a(i) is a measure of how dissimilar i is to its own cluster, a small value means it is well matched. Furthermore, a large b(i) implies that i is badly matched to its neighboring cluster. Thus an s(i) close to one means that the data is appropriately clustered.

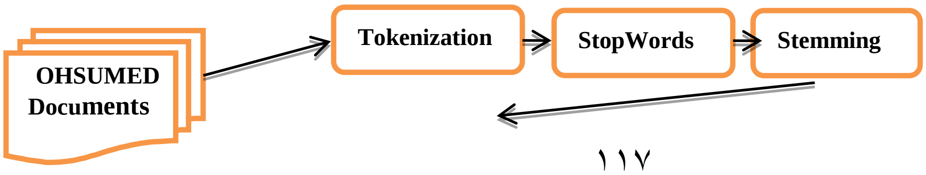
If s(i) is close to negative one, then by the same logic we see that i would be more appropriate if it was clustered in its neighboring cluster.

If s(i) near zero means that the data is on the border of two natural clusters.

The average s(i) of a cluster is a measure of how tightly grouped all the data in the cluster, thus the average s(i) of the entire dataset is a measure of how appropriately the data has been clustered [13].

4. System Architecture

The proposed system consists of three phases : Pre-processing, processing and post-processing . The block diagram of the proposed system is shown in the figure (1) :



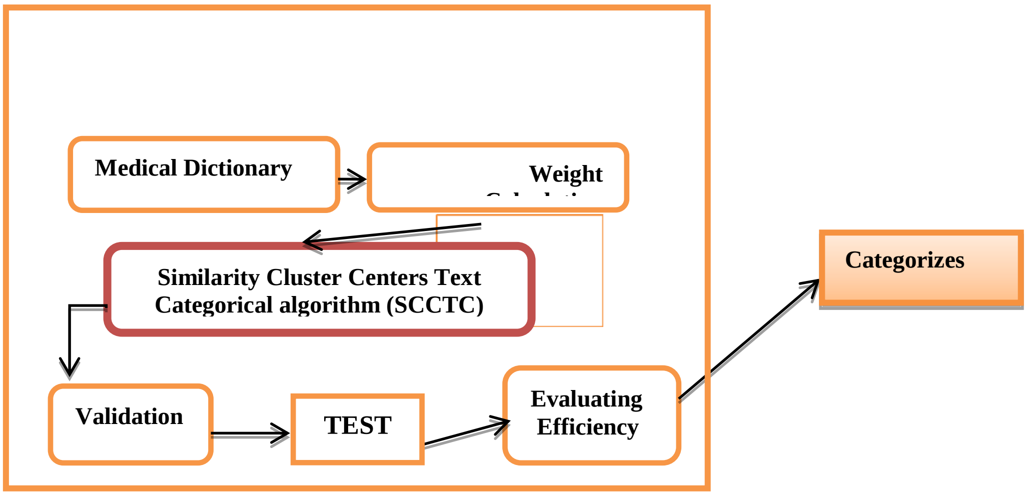


Figure (1): The Block diagram of the proposed system

4.1 Dataset

Ohsumed Data Corpus: The OHSUMED collection is a clinically-oriented MEDLINE subset from year 1987 to year 1991, consisting of 348, 566 references covering all references from 270 medical journals. The OHSUMED in year 1991 includes 74337 documents but only 50216 of which having abstracts [15].

Because the eye is very sensitive and important member in humans, so any error in determining the type of eye disease lead to give the wrong medication and thus damage to the eye, and hence the identify the type of disease is very important. In this paper, we have chosen medical texts that contain the most common diseases and interlocks in their symptoms that can infect the human eye. The data set which used consist of 660 documents of different lengths, each document was manually labeled based on its contents and the domain that it was found within and each document is stored in a separate file; these documents are categorized into seven categories (Cataract, Presbyopia, Glaucoma, Allergy, Macular degeneration, Blurred vision, Floater in vision) as show in table (1).

Table (1): Number of documents per Category in Dataset

Category name	No of documents
Cataract	120
Presbyopia	70
Glaucoma	160
Allergy	90
Macular degeneration	4o
Blurred vision	150
Floater in vision	30

4.2 pre-processing phase

It is the initial phase in this proposed system, which include the following steps:

4.2.1 Tokenization

The first step of preprocessing which aim to exploration of the word by splitting the input text into smaller units such as sentence and word. However a common delimiters are dot between sentences and space between words.

4.2.2 Stop Words Removal

Every language has its own list of stop words, stop words mean high frequency and low discrimination such as pronouns, prepositions, conjunctions and others. These words should be filtered from the text to improve the performance in text mining system.

4.2.3 Stemming

Stemming is one of the most important factors of preprocessing tasks. It is defined as the process of reducing words to their base form (stem) by removes the prefixes (letters, which are added in the beginning of the word root) and suffixes (letters, which are added in the end of the word root).

4.2.4 Medical Dictionary

After having pre-processing phase on the text, we want to get only required special medical terms of eye. For this purpose, we are going to match all the terms after stemming from step above with database of special medical terms of eye disease which has been previously established as shown in table (2) which display samples of special medical terms. Thus, we obtain the terms that contain only the medical terms in sentences.

Table (2): Samples of special medical terms of eye disease

Medical Dictionary for eye disease		
Medical Terms	Medical Terms	Medical Terms
Cornea	Aniridia	Vascular
Retina	Astigmatism	Conjunctivitis
optic nerve	Migraine	Endophthalmitis
Eye membrane	Mucormycosis	Episcleritis
Lens	Retinitis	Uveitis
Albinism	Stroke	Blepharoconjunctivitis
Amblyopia	Toxocariasis	Dacryoadenitis

4.2.5 Weight Calculation

The aims of this step is minimizing storage requirements by eliminating redundant terms, which can done in two steps:

Step1: normalized term frequency occurred in document by formula (2)

tfij = fij/max (fij) ……………………………………………………………………(2)

Where Fij is occurrences of term tj in document di.

Then calculate term weighting of each term (wij) is computed by multiplying the term frequency with the inverse document frequency (tf *idf) to improve the performance of text categorization.

idfj=1+ log (d/dfj) ………………………….…………………………………(3)

wij = tfij * idfj) ………………………..…………………………………………(4)

Where, d is the total number of documents and dfj is number of documents that contains term tj.

Step2: Select Wij weights above a certain threshold Φ, and also remove zero values. This step removes terms which have less/no significance, where threshold is determine by high weight(wij) ≥ [maximum weight (wij) /2].

4.3 Processing Phase

In work, proposed VEMST algorithm, which is EMST algorithm based on Variance to generate the desired k clusters (for k disease sets) and record the number of samples in each cluster. The VEMST algorithm is update on EMST algorithm which partitions the point set into a number of more compact clusters. Subsequently, a new partitioning process is repeated on the EMST constructed from a much smaller set of representative points. Each representative point with smaller variance value is close to the centroid of the subset created, where variance for each subtree (cluster) is computed to find the compactness of clusters. A smaller the variance value indicates, a higher homogeneity of the objects in the data set. The detailed pseudo code (VEMST) is given as follow:

Algorithm: VEMST

```
Input : S the point set
Output : k number of clusters with C (set of center points)
Let e be an edge in the EMST constructed from S
Let We be the weight of e
Let σ be the standard deviation of the edge weights
Let ST be the set of disjoint subtrees of the EMST
Let nc be the number of clusters
 Repeat
Construct an EMST from S
Compute the average weight of $W_{avg}$ of all the edges
Compute standard deviation σ of the edges
Compute variance of the set S
ST = ø; nc = 1; C= ø;
For each e EMST
    If ($We > W_{avg} + \sigma$) or (current longest edge e)
        Remove e from EMST
        nc ← nc + 1
        ST = ST∪ {T }   //T  is the new disjoint subtree
If nc < k         //  If the number of clusters nc is less than k
   While  nc ≠ k
        Remove the current longest edge
        nc ← nc + 1
        ST = ST∪ {T }   //T  is the new disjoint subtree
If  nc > k       //  If the number of clusters nc is greater than k
    Compute the centroid ci of each Ti ∈ ST
    Find the representative  ri  with smaller variance value ∈ Ti closest to  ci
    ST = $\cup_{Tii}$∈ ST {ri}
       Until  nc = k
     Return k clusters with C
```

After that, we calculate the centers weight value of each category (or cluster) using the formula mentioned below.

$$y(Wij,Cj) = \left(\sum_{w=1}^{di} Wij\right) / \, Nci \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

where Nci indicates the number of samples in category Ci

4.4 Post-processing Phase

4.4.1 Validation of Clusters

After formation of k-clusters, validated using Silhouette coefficient is used as one of efficient validation technique to determine how well each document lies within the cluster.

The steps involved in merging of clusters are given below:

Step 1: Select the cluster with least number of patient documents.

Step 2: The documents in the selected cluster are relocated based on the Cluster Criteria.

Step 3: The above steps are repeated until no more merging is possible.

4.4.2 Testing of Clusters

Use cluster centers as new training sets to classify the test document X, after that Judge document X to be the category by calculating the probability of X belong to category Cj which has largest P(X,Cj).

$$P(X,Cj) = \sum SIM(X,Wij).Y(Wij.Cj) \quad \text{…………………………………………………….….(6)}$$

Here, two similarity measures are used to calculate the similarity SIM(X,Wij) from equation (6). We try to run proposed algorithm twice to calculate the similarity to classify a new document into its category.

The first, we run algorithm using Common words Similarity Measure (CWSM) to calculate SIM(X,Wij) from the following equation [16] :-

$$CWSM(xi,yi) = \frac{nc\ (xi,yj)}{nmax\ (xi,yj)} \quad \text{………...…………………………………….………..(7)}$$

Where nc(xi,yj) is the number of common words between documents xi and yj

Nmax (xi,yj) the maximal number of words between documents xi and yj.

The second, we run algorithm using Cosine Similarity Measure (CSM) to calculate SIM(X,Wij) from the following equation [17]:

$$CSM(x,y) = \frac{\sum_{i=1}^{n} xiyi}{\sqrt{\sum_{i=1}^{n}(xi)2\sqrt{\sum_{i=1}^{n}(yi)2}}} \quad \text{……………………………………………...………..(8)}$$

Where xi is the frequency of words in document xi, yi is the frequency of words in document yi

In our proposed algorithm, we used two similarity measures to work a comparison between them to find the best as show in the result section.

4.5 Proposed Algorithm (Similarity Cluster Centers Text Categorical Algorithm)

Following is the used (SCCTC) algorithm:

Input: Documents in category Ci ,{ di1 ,di2 , …,din }

Output: Document X's category is Ci .

1. For all Documents.

Tokenize all sentences .

Remove all stopwords.

Perform stemming.

Compare with Medical Dictionary .

Calculate weight of terms. ((Wij = TFij x IDFj))

2. Constructer spanning tree to training sets

3. Using VEMST to find cluster centers. ((y(Wij,Cj) = ($\sum_{w=1}^{di} Wij) / Nc$))

4. Validation of clusters

5. Test of clusters using two of similarity measures (CSM , CWSM) to compute the similarity between test documents X and training documents, hence judge document X to be the category which has largest similarity.

5. Experiments Results

The experimental evaluation of the SCCTC algorithm applied on Ohsumed data corpus by used Delphi software.

5.1 Performance measures

The experimental data used in this paper is related to patient case study of eye disease diagnosis which is collected from Ohsumed dataset. We used split method(holdout), where 70% of data used as training and the remaining 30% used as testing. The training dataset involves different type of eye diseases like (Cataract, Presbyopia, Glaucoma, Allergy, Macular degeneration, Blurred vision, Floater in vision), and the testing dataset is used to test the efficiency of the method proposed (SCCTC) in this paper. Table-3 show specific quantity of samples in each category.

Table (3) : Experimental data used

Category of Patient data	Quantity of training documents ((70%))	Quantity of testing documents ((30%))
Cataract	84	36
Presbyopia	49	21

Glaucoma	112	48
Allergy	63	27
Macular degeneration	28	12
Blurred vision	105	45
Floater in vision	21	9

We used three evaluation measures (Recall, Precision, and F-measure) as the bases of our comparison [2], where

Precision(P) $=\frac{categories\ found\ and\ correct}{total\ categories\ found} = \frac{x}{(x+y)}$
……………………………….……….(9)

Recall(R) $=\frac{categories\ found\ and\ correct}{total\ categories\ correct} = \frac{x}{(x+Z)}$
……………………………..………..(10)

F-measure$=\frac{2*Precision*Recall}{Precision+Recall}$
…………………………………………………….………..(11)

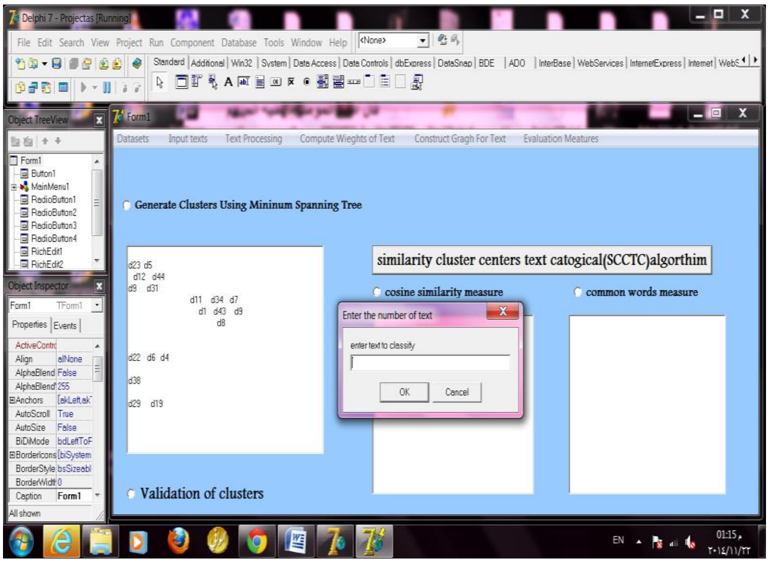
Where

x is number of the documents which are Cj category in fact and also the classifier judge them to Cj category is x.

y is the number of the documents which are not Cj category in fact but the classifier judge them to Cj category.

z is number of the documents which are Cj category in fact but the classifier don’t judge them to Cj category.

We can run the our system used Delphi program as show in the following figure (2) :


Delphi 7 - Projectas [Running]
File Edit Search View Project Run Component Database Tools Window Help
<None>
Standard Additional Win32 System Data Access Data Controls dbExpress DataSnap BDE ADO InterBase WebServices InternetExpress Internet WebS
Object TreeView
Form1
Button1
MainMenu1
RadioButton1
RadioButton2
RadioButton3
RadioButton4
RichEdit1
RichEdit2
Object Inspector
Form1 TForm1
Properties Events
ActiveContr
Align alNone
AlphaBlend False
AlphaBlend 255
Anchors [akLeft,ak
AutoScroll True
AutoSize False
BiDiMode bdLeftToF
BorderIcons [biSystem
BorderStyle bsSizeabl
BorderWidth 0
Caption Form1
All shown
Form1
Datasets Input texts Text Processing Compute Wieghts of Text Construct Gragh For Text Evaluation Meatures
○ Generate Clusters Using Mininum Spanning Tree
similarity cluster centers text catogical(SCCTC)algorthim
○ cosine similarity measure
○ common words measure
OK
Cancel
○ Validation of clusters
EN
01:15 م
٢٠١٤/١١/٢٢

```
d23 d5
 d12  d44
d9   d31
                d11   d34  d7
                  d1   d43   d9
                        d8

d22  d6  d4

d38

d29   d19
```

Enter the number of text

enter text to classify

Figure (2): Run the system using SCCTC algorithm in Delphi

5.2 Discussion and Evaluation

We have studied on the results of the performance tests in order to found out the most important factors that effects the performance of the system. The factors that have been distinguished, can be listed as follows;

1. The number of training documents: As the number of the training documents increase, the accuracy of the program increases also. This is due to the fact that, more number of terms related to a category at hand, results in a better classification of sample documents.

Assign appropriate weights to terms only related with special medical terms of eye to improve the performance of text categorization: The performance of the program increases with the increase in the relatedness of the terms included in the training documents to the category that they have.

2. Notice, when we used two similarity measures (cosine , common words), the accuracy of program using cosine similarity is better than the accuracy of algorithm with common words similarity because whenever increase the value of F-measure this means increased accuracy of our proposed algorithm, as shown in the following results:

Table (4) and Figure (3), show the sample of results for (SCCTC) algorithm using precision , recall and F-measure by using common words similarity to classify the test document for its category.

Table (4) : Results of precision, recall and F-measure for (SCCTC) using common words similarity measure

SCCTC algorithm using			
Category	Common Words Similarity		
	Precision	Recall	F-Measure
Cataract	0.75	0.631	0.695
Presbyopia	0.669	0.484	0.561
Glaucoma	0.85	0.732	0.787
Allergy	0.753	0.631	0.686
Macular degeneration	0.631	0.58	0.521
Blurred vision	0.767	0.681	0.721
Floater in vision	0.573	0.326	0.415

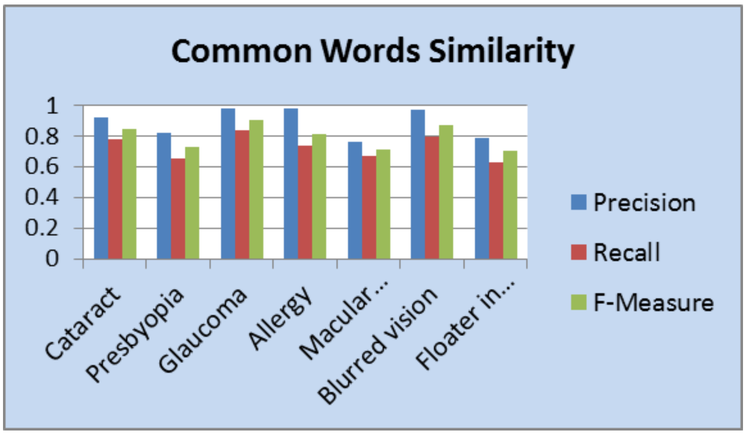


Figure (3): Result of common words similarity measure

In Table (5) and Figure (4), show the sample of results for (SCCTC) algorithm using precision , recall and f-measure by using cosine similarity to classify the test document for its category.

Table (5): Result of precision, recall and f-measure for (SCCTC) using cosine similarity measure

SCCTC algorithm using			
Category	Cosine Similarity		
	Precision	Recall	F-Measure
Cataract	0.921	0.776	0.842

Presbyopia	0.821	0.652	0.726
Glaucoma	0.977	0.84	0.903
Allergy	0.98	0.735	0.814
Macular degeneration	0.763	0.666	0.711
Blurred vision	0.968	0.792	0.871
Floater in vision	0.784	0.631	0.699

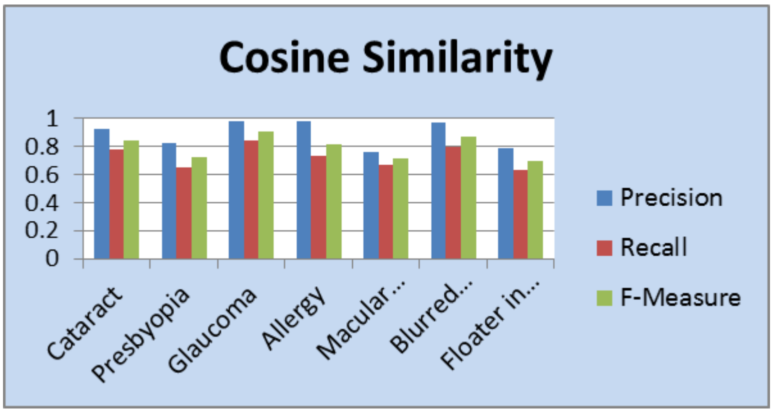


Figure (4): Result of cosine similarity measure

We can be seen from tables (4,5) that show the results as a comparison between the values of three evaluation measures (precision, recall and F-measure) which obtained it from (SCCTC) algorithm and that twice, the first using common words measure(equation-7) to calculate the similarity to classify the document in the test data to its category and secondly, another measure is used to calculate the similarity called cosine similarity measure (equation-8). Through the evaluation of the results of our proposed algorithm, noticed the values of three measures when using cosine similarity is higher than the values of these measures using common words similarity.

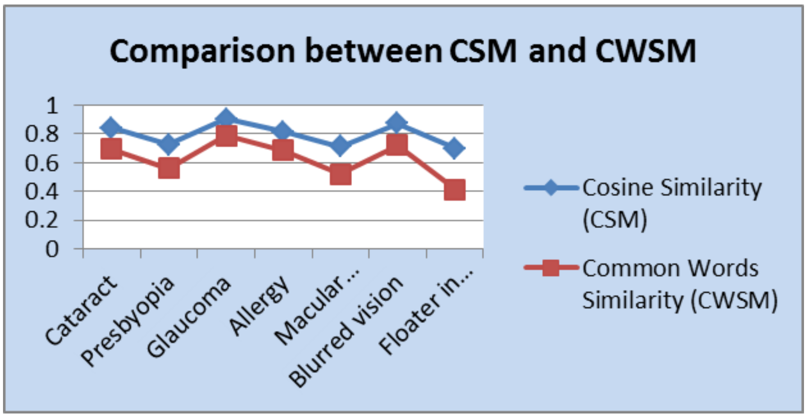


Figure (5): Show the results for both cosine similarity and common words similarity

6. Conclusion and Future works

This paper describes the development of the system for automated categorization of medical documents. The proposed system is similarity cluster centers text categorical (SCCTC) algorithm based on clustering technique, which can classify a new medical document automatically to a desired cluster based on similar weight value of cluster centers.

This improved algorithm dealt with all training categories by VEMST algorithm for developing k clusters of eye diseases and finding the cluster center weights which were subsequently used as the new training samples. The clusters were validated using silhouette coefficient, and we used two similarity measures (cosine , common words), so when evolution the accuracy of system, we notice the results with using cosine similarity is better than common words similarity measure. The source of documents used OHSUMED is a subset of the MEDLINE database, where the our corpus consists of 660 medical documents for eye diseases that belong to seven categories.

As future work, improvement system by using Naïve Bayes algorithm rather than VEMST clustering algorithm to categorization of medical documents , or used VEMST clustering algorithm with consider cases when the no. of clusters is not known. Finally the same algorithm (SCCTC) can also be applied with different datasets.

7. References

[1] M. W. Hussain et al., " Finding Semantic Relationship among Associated Medical Terms", Proceedings of Int. J. Emerg. Sci., Vol.2, No.2,pp. 300-309, ISSN: 2222-4254, June 2012 .

[2] S. M. Krishna and S. D. Bhavani, " An Efficient Approach for Text Clustering Based on Frequent Itemsets", Proceedings of European Journal of Scientific Research, ISSN 1450-216X, pp.385-396, Vol.42, No.3, pp.385-396 ,2010.

[3] S. Alsaleem et al., "Automated Arabic Text Categorization Using SVM and NB", Proceedings of International Arab Journal of e-Technology, Vol. 2, No. 2,pp. 124-128, June 2011 .

[4] G Manimekalai et al, " A Survey on Various Approaches in Document Clustering ", Proceedings of Int. J. Comp. Tech. Appl., Vol 2 , No.5, pp.1534-1539, IJCTA , sept-act 2011.

[5] S. J. Peter and S. P. Victor, " A Novel Algorithm for Informative Meta Similarity Clusters Using Minimum Spanning Tree" , Proceedings of International Journal of Computer Science and Information Security, Vol. 8, No. 1,pp 112-120, April 2010.

[6] Y. Zhao and G. Karypis, " Criterion Functions for Document Clustering: Experiments and Analysis", Proceedings of Technical Report #01-34, University of Minnesota, MN 2001.

[7] M. Laszlo and S. Mukherjee, " Minimum spanning tree partitioning algorithm for microaggregation", Proceedings of IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 7, pp. 902-911, July 2005.

[8] O. Grygorash al., " Minimum Spanning Tree Based Clustering Algorithms", Proceedings of IEEE Inn Conf. Tools with Artificial Intelligence, pp. 73-81, 2006.

[9] Pakhomov et al., "Automatic Classification of Foot Examination Findings Using Clinical Notes and Machine Learning", Proceedings of Journal of American Medical Informatics Association, Vol.15, No.2, 2008.

[10] X. Wang et al., " A Divide-and-Conquer Approach for Minimum Spanning Tree-Based Clustering", Proceedings of IEEE Transactions on Knowledge & Data Engineering, Vol.21, No. 7, pp. 945-958, July 2009.

[11] Peter et al., "Minimum Spanning Tree-based Structural Similarity Clustering for Image Mining with Local Region Outliers", Proceedings of International Journal of Computer Applications, Vol. 8, No.6, pp. 0975 – 8887, 2010.

[12] A. Chakrabarty and S. Roy , "Dynamic Clustering Based on Minimum Spanning Tree and Context Similarity for Enhancing Document Classification", Proceedings of International Journal of Information Retrieval Research, Vol.4, No.1, pp. 46-60, January-March 2014.

[13] A. Chakrabarty, " A Framework for Medical Text Mining using a Novel Categorical Clustering Algorithm ", Proceedings of International Journal of Computer Applications , Vol. 70, No.20, pp.0975 – 8887, May 2013.

[14] G. Brock et al., " clValid , an R package for cluster validation", Proceedings of Journal of Statistical Software, Vol. 25, pp. 1-22, Issue 4, March 2008.

[15] M. LAN et al., " Text Representations for Text Categorization: A Case Study in Biomedical Domain", Proceedings of IEEE International Joint Conference, ISSN :1098-7576, pp. 2557 – 2562, 12-17 Aug. 2007.

[16] S. B. Dbabis and L. H. Belguith, " Automatic summary revision based on a multicriteria analysis: sentence similarity detection", Proceedings of the Third International Conference on Modeling, Simulation and Applied Optimization Sharjah,U.A.E January. 20-22 2009.

[17] S. Xie and Y. Liu, " Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization", Proceedings of International Conference on Acoustics, Speech, and Signal Processing - ICASSP , pp. 4985-4988, 2008.

طريقة فعالة لتصنيف النصوص الطبية بالاعتماد على العنقدة ومقاييس التشابه

المستخلص

ان وجود كميات هائلة من المعلومات الطبية في المستندات الطبية، جعل استخدام أساليب التصنيف الآلي للنصوص ضروري في التشخيص والعلاج السريري. التصنيف الألي للنص يستطيع أن يوفر معلومات حول توقع الصنف الذي ينتمي إليه النص. هذا البحث يمكن أن يكون بمثابة أداة تشخيص طبي لتصنيف سجلات المرضى وذلك باقتراح خوارزمية تصنيف النص بالاعتماد على تشابه المراكز العنقودية لتصنيف سجلات المرضى المصابين بأمراض العين. اقترحنا خوارزمية (VEMST) كتحديث لخوارزمية (EMST) وذلك باستخدام التباين لإيجاد المراكز العنقودية وتم تطوير خوارزمية تصنيف النص باستخدام مقياسي التشابه (جيب التمام، الكلمات المشتركة) لتصنيف البيانات المعنقدة. حيث أظهرت النتائج أنه عندما يكون عدد وحجم الوثائق الطبية المستخدمة للتدريب كبير فأن دقة التصنيف تزداد، كذلك لاحظنا عند استخدامنا طريقة مقارنة المصطلحات الطبية في مرحلة المعالجة الأولية، ان الدقة تكون افضل من استخدام التكرار لكل الكلمات في النص الطبي بالإضافة ان وقت التنفيذ أقل. أخيراً، وجدنا أداء نظامنا عندما نستخدم مقياس التشابه جيب التمام هو أفضل من ادائه مع استخدام مقياس التشابه للكلمات المشتركة.