Integrated Disease Analysis Using Similarity Measure for Preventing the Disease

R. Pradheepa
HOD, Department of Information Technology, Sankara College of Science and Commerce, Coimbatore, Tamil Nadu, India

R. Menaka
M.Phil Scholar, Department of Computer Science, Sankara College of Science and Commerce, Coimbatore, Tamil Nadu, India

K.Mohanapriya
M.Phil (Computer Science), Kovai Kalaimagal College of Arts and Science, Coimbatore, Tamil Nadu, India.

Abstract – The growth of medical field has led to an inevitable growth in curing the diseases. In the evolving disease-network, each disease would expose some point of connection with one or more disease. With the lack of patient's awareness the disease disorder of the existing disease may connected with one another "hidden disease" contains some set of same number of disorders which is unknown to the patient. This can be achieved by some set of predefined parameters related to disease documents. The classification scheme can be obtained by similarity clustering with different constraints through which the identical symptoms of various diseases will fall under a respective cluster-n. This paper investigates a classification scheme to categorize the integrated disease and helps to create the awareness between the patients who is suffering from one disease can able to prevent them from the hidden disease.

Index Terms – Inevitable- Disease Network – Cluster.

1. INTRODUCTION

A. Introduction to Cluster

The Clustering techniques play an important role in this paper. Mainly in the field of categorizing the diseases related to the estimated disease symptoms. Cluster Analysis separates data into meaningful clusters. It is one of the useful starting-point for the Data-summarization. The study mainly focuses on the similarity cluster, which uses the similarity criteria as "Distance". Hence the distance based clustering uses the calculation, that is based on if two or more objects are "close to each other" according to the given geometrical distance then those two objects will belong to the same cluster.

B. Uses of Cluster

There has been n-number of real-time applications that use the cluster analysis to solve the practical problems. Some of the popular examples that use the "Clustering for Understanding" (Meaningful groups of objects that share the common characteristics which help the people to analyze and define the world) are given below.

- Biological Research
- Information Retrieval System
- Weather Condition Estimation System
- Research in Psychology and Medicine
- Business Analyst

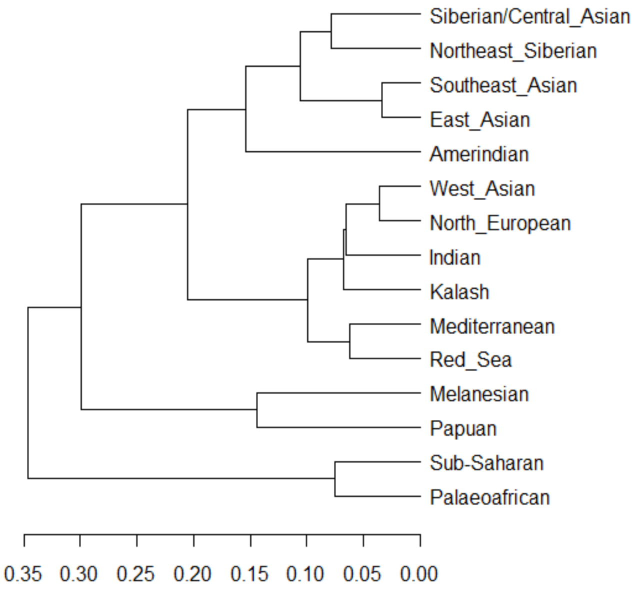

Fig. 1 Example of Hierarchical Clustering.

The above figure shows the tree-like structure representation of human variation that experienced the lateral Gene flow. Through which the four major divisions of mankind that are separated quite distinctly from each other. They are, i) East Eurasians, ii) West Eurasians, iii) Australo-Melanesians, and iv) Sub-Saharan Africans. (See from top to bottom of fig 1.1)

C. Overview of Medical Research Council

The Medical Research Council (MRC) has evolved from being a massive evidence storehouse for the economic development and social welfare of developing countries. In that evolution, many of the research findings for "Health" were based on "Gene" spectrum. Human being's health is not only based on the genetic factor but also based on their food lifestyle, climatic conditions, and some other factors too. By analyzing the classification of diseases and their cause with the existing one, the preventing strategies can be aided.

This paper addresses the following issues:

- Literature Review
- Problem Definition
- About Classification scheme
- Proposed detection approach

2. LITERATURE REVIEW

A. The rise of International Classification of Disease.

Initially, classification for medical field was made by John Graunt in the year 1700 and it gets enclosed by 13 kinds of the diseases of the early population. The work Nosologia Methodica wrote by the French doctor Francois Boissier de Sauvages from Monpell (year 1706-1767) in which he divides the diseases into 10 major groups, it contains 295 various species and also with 2,400 kinds. The essential parts of scientific methodology in the healthcare are nomenclature and classifications.

The nomenclature of the current state in the healthcare activity permits the optimal application of the computer technology in the processing and the recovery of the medical data or information. One of the ancient and most important classifications in medicine is ICD classification (International Classification of Disease). It is mainly used in the field of statistics and as a coding system in medical databases. Most of the physicians use this classification.

B. Usage of International Classification of Disease (ICD)

Healthcare provider institutions, such as hospitals are focuses that should facilitate implementation of medical applications that track the patient medical condition and facts connected with him. The list of measures with their prices can be found in all hospitals and is used by economist. The database should provide the entire picture of an existing situation. It is very important that the researchers have tools for analysis of clinical data. Analysis of data is the only mode to improve the prevention of future errors and persuade reduction of costs of hospitalization. Using the database it is possible to reveal all advantages and disadvantages of some technique.

C. The result of Gene expression.

In an Integrated analysis of numerous gene expressions, the researchers had collected and curated a large collection of gene expression profiles from diverse diseases, developed and tested several approaches for classifying patient samples coining from each disease. For those diseases whose classification was authenticated successfully and they developed specific biomarker genes and concise them in the context of protein interaction, mutation and drug target data. Whereas in Genome, wide-ranging expression profiling has transformed biomedical research; vast amounts of expression data from frequent studies of many diseases are now available. Building the best use of this resource in order to better understand disease practices and treatment remains an open challenge. In particular, disease biomarkers detected in case-control studies suffer from less reliability and are only feebly reproducible.

D. Limitations

The difficulties of the nomenclature of the healthcare services, the classifications, the identification and coder for the needs of the development and functioning of the informational systems in healthcare are the weakest link in our conditions. This classification is not appropriate in cases where few or no information about the patient is presented. In such case, merely symptoms of the disease can be coded that can be initiated by numerous different medical conditions that can be regularly coded if we have enough information to confirm the diagnosis.

E. Example for Disease Group Code.

Below mentioned Table1 shows the code for the set of diseases that fall under a specific group, based on ICD research findings.

TABLE I
INTERNATIONAL STATISTIC CLASSIFICATION OF THE DISEASES AND PROBLEMS CONNECTED WITH HEALTHCARE

CHAPTER	GROUP OF THE DISEASES	CODE
I	Certain infectious and parasitic diseases	AOO - B99
II	Neoplasm	COO - D48
III	Diseases of the blood and blood-forming organs and certain	D50 -D89

	disorders involving the immune mechanism	
IV	Endocrine, nutritional and metabolic diseases	E00 -E90
V	Mental and behavioural disorders	F00 -F99
VI	Diseases of the nervous system	G00 -F99
VII	Diseases of the eye and adnexa	H00 -H59
VIII	Diseases of the ear and mastoid process	H60 -H99
IX	Diseases of the circulatory system	I00 -I99
X	Diseases of respiratory system	J00 -J99
XI	Diseases of the digestive system	K00 -K93
XII	Diseases of the skin and subcutaneous tissue	L00 -L99
XIII	Diseases of the musculoskeletal system and connective tissue	M00 -M99
XIV	Diseases of the genitourinary system	N00 -N99
XV	Pregnancy, childbirth and the puerperium	O00 -O99
XVI	Certain conditions originating in the perinatal period	P00 -P96
XVII	Congenital malformations deformations and chromosomal abnormalities	Q00 -Q99
XVIII	Symptoms signs and abnormal clinical and laboratory findings not elsewhere classified	R00 -R99
XIX	Injury poisoning and certain other consequences of external causes	S00 – T98
XX	External causes of morbidity and mortality	V01 –Y98
XXI	Factors influencing health status and contact with health services	Z00 –Z99
XXII	Codes for special purposes	U00 –U99

3. PROBLEM DEFINITION

The problem of finding frequent item set of disease is first initiated through which the use of frequent symptoms to find association rules in large disease related databases. In clustering a given set of text documents from neighbor set is proposed. In the classification of text files or documents is done by considering Gaussian membership function and making use of it to obtain clusters by finding word (symptoms) patterns. Each cluster is identified by its word (symptoms) pattern calculated using fuzzy based Gaussian membership functions once clusters are formed. In this paper the idea is to first obtain frequent set of symptoms for each document using existing association rule mining algorithms either by horizontal or vertical approach. Once we find frequent symptoms sets in each disease then we form a Boolean matrix with rows indicating disease and columns indicating unique frequent symptoms from each disease related document. This is followed by the calculation of a ternary feature vector for each document pair, represented as a 2D matrix by redefining the XNOR function as hybrid XNOR logic with slight modification in the function by introducing high impedance variable as Z. The idea of maximum capturing is taken as the base framework for clustering. Finally the cause for the symptoms in the cluster will be reported through which the prevention strategies can be aided.

4. PROPOSED WORK

A. Role of Similarity Measure

To perform clustering, we define a similarity measure which may be used to find the similarity between any pair of Disease as given in Table II. Disease description details could be disease analytic dataset or files. The proposed similarity measure Siml (A1, A2) is a function of any two attributes A1 and A2. We consider A1 and A2 as the attributes present in two different files F1 and F2 respectively. The input for component clustering algorithm is a set of disease descriptions with properties predefined and the output is a set of highly cohesive report with low coupling feature.

TABLE II

PROPOSED SIMILARITY MEASURE

A1	A2	Siml(A1,A2)
Absent	Absent	Neglect (Say Z)
Absent	Present	0
Present	Absent	0
Present	Present	1

B. Clustering Algorithm

The algorithm may be used to cluster disease description documents or program codes. If the dimensionality is very high, then to reduce the dimension of the documents we may eliminate stop words and stemming words by forming a list of stop words separately and storing in a file. When clustering English documents we may use porter stemmer algorithm. In case, if the dimensionality of the documents is too large to handle, then to further reduce the dimension we may apply Singular value decomposition as given in step1 of the algorithm.

Algorithm Name: Algorithm for Component Clustering.

Input: Document set, frequent items.

Output: set of clusters.

```
Begin of Algorithm
Step1:
For each document D do
Begin
Step1.1 Remove the stop words and stemming words from each disease related document.
Step1.2 Find distinctive words (symptoms) in each document and count of the same.
Step1.3 Find compact dimension set by applying Singular Valued Decomposition to all the document set.
End for
Step 2: Form a word (symptoms) set W consisting of each word (symptoms) in condensed item set of each document of step1.
Step 3: With each row and column corresponding to each Document and each word respectively, form Dependency Boolean Matrix
For each file in file set do
Begin
For each word (symptom) in word (symptom) set do
Begin
If (word wi in Word (symptom) set W is in file Di)
Begin
Set D[Di, wi] = 1
Else
Set D[Di, wi] = 0
End if
End for
End for
Step 4: Find the Feature vector similarity matrix by estimating similarity value for each component pair applying similarity measure defined in Table II to obtain the matrix with feature vectors for each component or document pair.
Step 5: Interchange the corresponding cells of matrix by count of number of zeroes in tri state feature vector.
Step 6: At each one of the step, find the cell with maximum value and document pairs containing this value in the matrix.
Join such document pairs to form clusters. Also if document pair (A, B) is an individual cluster and document pair (B, C) is in another cluster, form a new cluster containing (A, B, C) as its items.
Step 7: Repeat Step6 until no modules or documents exist or we reach the stage of first minimum value leaving zero entry.
Step 8: Finally, report the set of clusters obtained.
Step 9: Label the clusters by considering disease entries.
End of the algorithm
```

5. EXPERIMENTAL RESULT

Consider the sample parameters for the disease analysis, based on different symptom issues.

Let the parameter contain the domain as given in the below mentioned table (Table III)

TABLE III

SYMPTOM DOMAIN DETAILS AND ITS PARAMETER DEFINITION

DOMAIN	PARAMETER
Communicable Disease (C.D)	{Yes: 1, No: 0}
Germs (GE)	{Bacteria: 1, Fungi: 2, Virus: 3, Protozoans: 4, none: 0}
Cough (CO)	{Mild: 1, Heavy: 2, with blood: 3, none: 0}
Sleeping Disorder (S.D)	{Yes: 1, No: 0}
Bleeding Disorder (B.D)	{Anemic: 1, Internal Bleeding: 2, External Bleeding: 3, none: 0}
Respiratory Disorder (R.D)	{Irritation in nose/throat: 1, wheezing: 2, Shortness of breath: 3, Dehydration: 4, None: 0}
Weight Loss (W.L)	{Yes: 1, No: 0}
Vision Disorder (V.D)	{Yes: 1, No: 0}

Hearing Disorder (H.D)	{Yes: 1, No: 0}
Skin Infection (S.I)	{Yes: 1, No: 0}
Types of Pain (T.PN)	{Nociceptive: 1, Neuropathic: 2, Mixed: 3, none: 0}
Common Effects (C.E)	{Nausea: 1, Dizziness: 2, Nausea + Dizziness: 3, none: 0}
Movement Disorder (M.D)	{possible: 1, not possible: 2}
Urine Problem (U.P)	{Yes: 1, No: 0}
Treatment Period (T.P)	{Days: 1, Weeks: 2, Months: 3, Years: 4, Continue: 5}

By forming the 2D matrix, with rows indicating the symptom domain and the column indicating the respective elements set, considered for symptom evaluation using direct approach.

TABLE IV- A
MATRIX SHOWING THE COMBINATION OF SYMPTOM DOMAIN VS. PARAMETER

Symptom_Domain Vs Parameter	C.D	GE	CO	S.D
Dibetes	No	Virus	No	Yes
Jaundice	No	Virus	Mild	Yes
Malaria	Yes	Protozoans	No	No
Migraine	No	Bacteria	No	Yes
Heart Attack	No	Bacteria	With blood	Yes
Acute Bronchitis	Yes	Virus	Heavy	Yes
Asthma	No	Virus	Heavy	Yes
Lung Cancer	No	Virus	With blood	Yes

TABLE IV- B
MATRIX SHOWING THE COMBINATION OF SYMPTOM DOMAIN VS. PARAMETER

Symptom_Domain Vs Parameter	B.D	R.D	W.L	V.D
Dibetes	Anemia	Shortness of breath	Yes	Yes
Jaundice	Anemia	Shortness of breath	Yes	Yes
Malaria	Anemia	Shortness of breath	Yes	No
Migraine	External Bleeding	Dehydration	No	Yes
Heart Attack	External Bleeding	Shortness of breath	Yes	Yes
Acute Bronchitis	External Bleeding	Irritation in nose/ throat	Yes	No
Asthma	Anemia	Shortness of breath	No	Yes
Lung Cancer	External Bleeding	Shortness of breath	Yes	Yes

TABLE IV- C
MATRIX SHOWING THE COMBINATION OF SYMPTOM DOMAIN VS. PARAMETER

Symptom_Domain Vs Parameter	H.D	S.I	T.PN	C.E
Dibetes	Yes	Yes	Neuropathic	Nausea + Dizziness
Jaundice	Yes	Yes	Nociceptive	Nausea + Dizziness
Malaria	No	No	Nociceptive	Nausea
Migraine	Yes	Yes	Neuropathic	Dizziness
Heart Attack	Yes	Yes	Neuropathic	Nausea + Dizziness
Acute Bronchitis	Yes	No	Nociceptive	none
Asthma	Yes	Yes	Nociceptive	Nausea
Lung Cancer	Yes	Yes	Neuropathic	Nausea + Dizziness

TABLE IV- D
MATRIX SHOWING THE COMBINATION OF SYMPTOM DOMAIN VS. PARAMETER

Symptom_Domain Vs Parameter	M.D	U.P	T.P
Dibetes	Possible	Yes	Continues
Jaundice	Possible	Yes	Weeks
Malaria	Possible	Yes	Weeks
Migraine	not Possible	No	Continues

Heart Attack	not Possible	Yes	Months
Acute Bronchitis	Possible	No	Weeks
Asthma	Possible	Yes	Continues
Lung Cancer	Possible	Yes	Continues

TABLE V – A

Matrix Attained after Index Number Generation

Symptom_Domain Vs Attributes	C. D	G E	C O	S. D	B. D	R. D	W. L	V.D
Dibetes (D1)	0	3	0	1	1	3	1	1
Jaundice (D2)	0	3	1	1	1	3	1	1
Malaria (D3)	1	4	0	0	1	3	1	0
Migraine (D4)	0	1	0	1	3	4	0	1
Heart Attack (D5)	0	1	3	1	3	3	1	1
Acute Bronchitis (D6)	1	3	2	1	3	1	1	0
Asthma (D7)	0	3	2	1	1	3	0	1
Lung Cancer (D8)	0	3	3	1	3	3	1	1

TABLE V – B

Matrix Attained after Index Number Generation

Symptom_Domain Vs Attributes	H. D	S. I	T.P N	C. E	M. D	U. P	T.P
Dibetes (D1)	1	1	2	3	1	1	5
Jaundice (D2)	1	1	1	3	1	1	2
Malaria (D3)	0	0	1	1	1	1	2
Migraine (D4)	1	1	2	2	2	0	5
Heart Attack (D5)	1	1	2	3	2	1	3
Acute Bronchitis (D6)	1	0	1	0	1	0	2
Asthma (D7)	1	1	1	1	1	1	5
Lung Cancer (D8)	1	1	2	3	1	1	5

TABLE VI

Matrix Shows the similarity measure from table V

	D1	D2	D3	D4	D5	D6	D7	D8
D1	X	12	6	8	10	5	11	13
D2	X	X	7	5	9	7	11	11
D3	X	X	X	1	3	7	6	4
D4	X	X	X	X	9	3	7	8
D5	X	X	X	X	X	4	7	12
D6	X	X	X	X	X	X	6	6
D7	X	X	X	X	X	X	X	10
D8	X	X	X	X	X	X	X	X

Stage 1: Select the cell with maximum value from the above mentioned table (Table.VI). Here (D1, D8) have value as 13.Group these cells to form an initial cluster containing most similar symptoms domain (D1, D8).

We thus have CLUSTER-1: {D1, D8}.

TABLE VII

Matrix Shows the Reduced similarity measure from Stage-1

	D2	D3	D4	D5	D6	D7
D2	X	7	5	9	7	11
D3	X	X	1	3	7	6
D4	X	X	X	9	3	7
D5	X	X	X	X	4	7
D6	X	X	X	X	X	6
D7	X	X	X	X	X	X

Stage 2: Choose the cell with next maximum value from the above mentioned table (Table-VII). Here (D2, D7) are having value 11. Group these cells to form a cluster containing the next set of similar symptoms domain (D2, D7).

We thus have CLUSTER-2: {D2, D7}.

Stage 3: Choose the cell with next maximum value from the Table-VIII. Here (D4, D5) have the value 9. Group these cells to form a cluster {D4, D5} of next similar symptom domain.

Hence we have the CLUSTER-3: {D4, D5}

TABLE VIII

MATRIX SHOWS THE REDUCED SIMILARITY MEASURE FROM STAGE-2

	D3	D4	D5	D6
D3	X	1	3	7
D4	X	X	9	3
D5	X	X	X	4
D6	X	X	X	X

Stage 4: The formation of final cluster is CLUSTER-4: {D3, D7}

TABLE IX

MATRIX SHOWS THE REDUCED SIMILARITY MEASURE FROM STAGE-3

	D3	D6
D3	X	7
D6	X	X

If we have a new entry with the respective parameters, we can evaluate to which cluster the new disease is similar and finally it helps the patient to be self-aware.

Summary of the clusters are,

CLUSTER-1: {D1, D8}

CLUSTER-2: {D2, D7}

CLUSTER-3: {D4, D5}

CLUSTER-4: {D3, D6}

ACKNOWLEDGMENT

In this paper, the similarity functions with specified constraints to find the similarity between any two diseases. Based on that the algorithm has been outlined, through which the classification of similarity between diseases will be obtained. In future, the research findings can be extended to classify the in-taking drugs for one disease has the possibility to stimulate the growth of one another hidden disease or not. In technical point of view, the algorithm can be extended to classify the parameters by using Fuzzy Logic.

REFERENCES

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4652780/

[2] http://dienekes.blogspot.in/2010/12/human-genetic-variation-first.html

[3] https://www.mrc.ac.uk/about/what-we-do/medical-research-foundation/

[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3789162/pdf/aim-2008-16-3-9.pdf

[5] http://www.umm.edu/health/medical/altmed/condition-symptom-links/conditions-with-similar-symptoms-as-diabetes-mellitus

[6] http://www.nytimes.com/health/guides/disease/chronic-obstructive-pulmonary-disease/diseases-with-similar-symptoms.html

[7] https://en.wikipedia.org/wiki/List_of_infectious_diseases_causing_flu-like_syndrome

[8] http://www.peteducation.com/article.cfm?c=16+2160&aid=2956

[9] https://www.studyread.com/examples-of-protozoa

[10] http://www.action-on-pain.co.uk/you-and-chronic-pain/different-types-of-pain/

[11] https://www.livescience.com/36328-top-food-borne-illness-germs-sick.html

[12] http://www.nytimes.com/health/guides/disease/acute-bronchitis/overview.html

[13] V.Susheela Devi, M. Narasimha Murthy. Text Book on Pattern Recognition. An Introduction. University Press.Congnan Luo, Yanjun Li, M. Chung. Text document clustering based on neighbours, Data & Knowledge Engineering, 2009; 68:1271–1288.

[14] Tianming Hu, Sam Yuan Sung, Hui Xiong, Qian Fu. Discovery of maximum length frequent itemsets, Information Sciences, 2008; 178:69–87.

[15] Wen Zhanga,, Taketoshi Yoshida, Xijin Tang, Qing Wang. Text clustering using frequent itemsets, Knowledge-Based Systems, 2010;23: 379–388 Wen Zhanga, Taketoshi Yoshida, Xijin Tang. A comparative study of TF*IDF, LSI and multi-words for text classification. Expert Systems with Applications, 2011; 38: 2758–2765.