International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

# Information Retrieval for Bridging Vocabulary Gap between Health Seekers and Providers

**Venkatesh .M, Hemavathi .D**

Department of Information Technology, SRM University, Kattankulathur, India

**Abstract:** ***In this paper we describe how to bridge vocabulary gap between health seekers and providers using novel scheme. To code medical records by jointly using local mining and global mining. Local mining uses individual medical records to drive a conclusion about individual health map into the authenticated terminology. Global mining combines medical records of similar types and analysis it to drive conclusion. We use a terminology to space a gap between local mining and global mining.***

**Keywords:** local mining, global mining, corpus aware terminology

## 1. Introduction

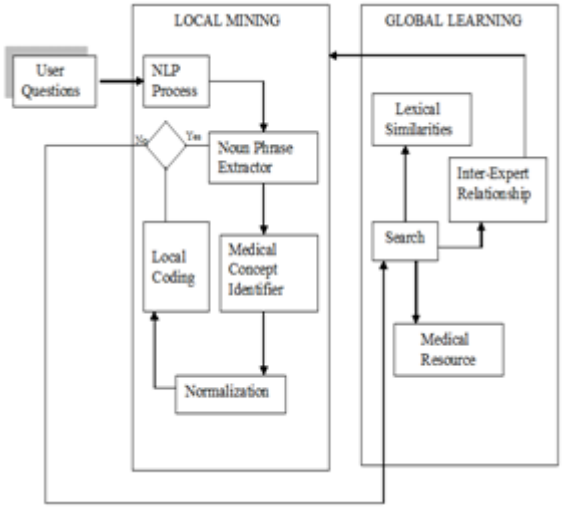
Patients seeking online information about their health, connecting patients with doctors worldwide to know about their health via question and answering. Doctors able to interact with many patients about particular issue and provides instant trusted answers for complex and sophisticated problems. Previously we used to relate medical data with external dictionary which was not that much sufficient enough. Here we incorporate corpus aware terminology which is used to relate the natural language medical data with medical terminology this narrow down the path between health seekers and health providers. For example: heart attack can also be said as myocardial disorder.

We use tire stage frame work to accomplish the task
i. Noun phase extraction
ii. Medical concept identifier
iii. Medical concept normalization

Due to loss of information missing of key components we compliment the global mining in a graph based approach. By graph based approach we are able to map the missing information by combining all other related records.
A. Inter expert relationship → historical data
B. Inter terminology relationship → external ontology relationship

The main contribution of our project is A) the first work on automatically coding the community generated health data which is more complex, inconsistent and ambiguous compared to hospital generated health data. B) Generate the corpus-aware terminology vocabulary with the help of external knowledge. C)Builds a global learning model collaboratively enhance local coding results. Rule based technique is used to discover and construct effective rules by making strong uses of morphological, syntactic, semantic and pragmatic aspects of natural language. Machine learning approach is to build inference model from medical data with known annotations then apply the trained models to unseen data from terminology prediction.

Medical sires are among the most popular internet sites today through which people can get more knowledge about their health conditions. The practice of medicine is experiencing a shift from patients who passively accept their doctors orders to patients who actively took online information to know briefly about their health because doctor are very busy with many patients and hence they cannot give brief description about their health issue to each and every patient. This is the reasons why health seekers normally use online medical sites.




Most of the medical sites such as mayo clinic, Medscape are consumer oriented and provide their sound advice about general medical topics. The vocabulary used is readily comprehensive when health seekers search for more detailed information about a very specific topic. Due to tremendous number of records have been accumulated in their repositories and in most circumstances user may directly locate good answers by searching rather than waiting for experts to answer. However users with diverse background do not necessary share same vocabulary, the same question may be written in different native languages by their health seekers which is difficult for other health seekers to understand to bridge vocabulary gap we use corpus aware terminology.

## 2. Existing System

We incorporate local mining and global mining to find the answers asked by the health seekers. Here health seekers has to wait for the experts come online and answer for the question, this system lacks in reading the pdf file format.

## 3. Local Mining

Three stage framework is being implemented. First a medical record given → noun phrases are extracted from the record → medical concepts are being identified → normalize the detected medical record.

### 3.1 Noun phase extraction

Initially assign part of speech to each word given in the medical record, pos tagger assigns parts of speech to each word.
(Adjective|noun)*(noun
preposition)?(Adjective|noun)*noun.

The noun phase contains zero or more adjective or more adjective or noun followed optional noun or proposition followed by an adjective or more than a single noun.

### 3.2 Medical concept detection.

To differentiate the medical concepts from other general noun phrase
$CE_I(c) = - \sum_{i=1}^{2} P(d_{i|c})\log P(d_j|c)$
C → concept
$D_1$, $D_2$ → medical corpus
$P(d_j|c)$ → probability that C is related to $D_I$
Hence
$P(D_I|c)=\frac{count\ (c,Di)}{count\ (c)}$
To remove different medical corpus we normalize it as
$Pn(Di|c) =(P(Di|c)/Li)/\ (\sum_{j=1}^{2} P(Dj\ |c)/Lj)$

$L_i$ → is the sum of document length in $D_i$
$L_j$ → is the sum of document length in $D_j$
0.693 is max value reached with medical corpus computer processable

$$\text{Specificity(c)} = \begin{cases} 1-\propto c\ E_1(c)\ if\ P_n\left(\frac{D_1}{C}\right) > P_n(\frac{D_2}{C}) \\ \propto CE_1(c)\ otherwise \end{cases}.$$

Where $\propto = \frac{0.5}{0.693}$threshold to detect medical concepts

### 3.3 Medical concept normalization

Medical domain specific noun phrase cannot ensure that they are standardized terminology example health seekers used the word “birth control” but this should be mapped as “contraception” in medical terms ICD,UML, SNOMEDCT map words to external dictionary.

## 4. Proposed System

We try to incorporate natural language processing the main aim of using NLP process is in documentation is retrieval is that it allows users to frame their questions in a natural way. We also use a method tracker and chucker for any given question, it is likely that someone has written the answer down somewhere to tracker these answer we use chucker certain medical reportsof the patient may be in pdf format usually all system does not have a built in adobe reader so there arise a difficulty in reading the pdf file hence in our project we try to incorporate a pdf boxer which is used to convert pdf file to normal text file, similar medical records traced using concept based mining.

When a particular question is asked by the health seekers the answer is being searched in the local mining and global mining if the answer is not found in both then it returns that answer is not found, where in our proposed system if the answer is not found the question that goes to the pending state, when doctors come online they reply for the answers and the answer is being stored in the database for future reference.

## 5. Graph Based Global Learning

Let $Q=\{q_1,q_2,\ldots q_n\}$, $T=\{t_1,t_2,\ldots t_m\}$ denotes the repository of medical records and their associated locally mined terminology, t is global terminology is annoyed to the medical record q.

### 5.1 Relationship identification.

The inter terminology and inter expert relationship are not initially seen or implied from medical records so it is known as implicit relationship.

### 5.2 Inter terminology relationship.

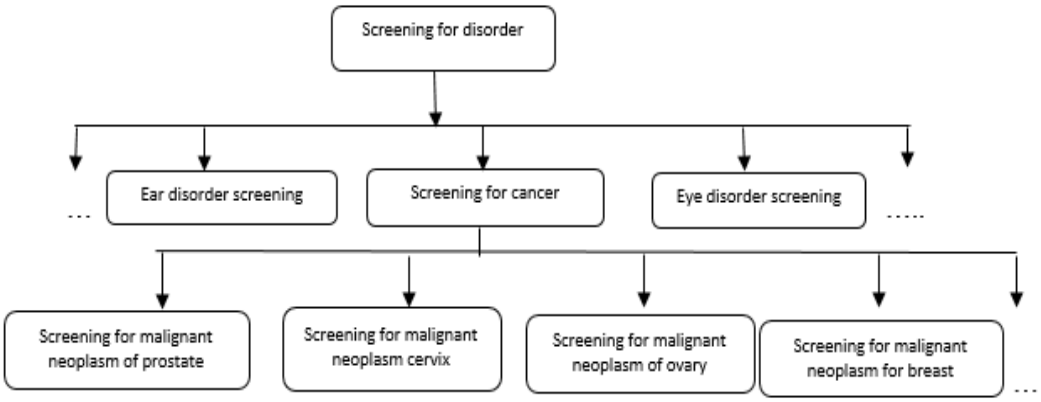
A well-defined ontology is able to semantically capture the inter terminology hierarchical relationship. Two terminologies $t_i$ and $t_j$

$$R_{ij} = \begin{cases} \frac{1}{29}\ if\ ancestor - child\ relationship \\ 0\ otherwise \end{cases}$$

Where p is the length of ancestor child path between both code $t_i$ and $t_j$. R is the weighted matrix for representing inter terminology relationship. The medical terminology will enhance our scheme in 2ways 1.trackles the granularity mismatch problem. 2. The hierarchical relationship boost the coding accuracy via filtering out the sibling terminology.

### 5.3 Inter expert relationship.

Related to specific medical areas mainly related to the historical data (ie..) the number of questions they have co-answered.
$J(U_i, U_j) = \frac{|u_i \cap u_j|}{|u_i \cup u_j|}$
$u_i$ is a set of medical records that expert $u_i$ has answered.




### 5.4 Probabilistic Hypergraph Construction

The graph based learning can be categorized into simple graph and hyper graph based approach. Simple graph conveys the pairwise relationship of vertices and then overlooks the relations in higher orders, sensitive radius is used to calculate the similarities. Hyper graph contains summarized local grouping information by allowing each hyper graph to connect with more than two vertices simultaneously

A hyper graph is composed of $G(\gamma, \varepsilon, \omega)$ $\gamma$ denotes vertex,$\varepsilon$ is hyper edge,
$\omega$ is diagonal matrix
$\varepsilon$ is a family of arbitrary subsets $\gamma, \varepsilon$
Such that $v_{e\varepsilon\varepsilon} = V$ e is assigned with W(e), a probabilistic hyper graph can be represented by $|\gamma| \times |\varepsilon|$ incidence matrix

$h(v_i,e_j) = \begin{cases} p(v_i, e_j)\ if\ v_i \epsilon e_j \\ o\ otherwise \end{cases}$

$p(v_i,e_j) \rightarrow$ describes the probability that vertex $v_i$ falls into hyper edge $e_j$, based on vertex of $v_i \epsilon \gamma$ is estimated as
$d(v_i) = \sum_{e_j \epsilon \varepsilon} W(e_j) h(v_j, e_j)$ for hypergraph $e_j \in E$, its degree is defined as
$\delta(e_j) = \sum_{vi\,2ej} h(vi;\ ej)$ We denote the vertex degrees and hyperedge degrees
by **D***v* and **D***e*.

If suppose there are N medical records each forms Q are connected by three types of hyper edge. 1. Each vertex as centroid and forms a hyper edge by dividing around its K-nearest neighbour based on medical record content similarities. 2. Terminology sharing network a group of medical records sharing the same terminology together. Example there are 2 medical records "what are the signs of pregnancy in first week?", "is it safe to colour hair during pregnancy?". In terminology sharing network these two medical records are connected to each other because both the records has the word pregnancy, but these records are not grouped together because they belong to different topics. Hence this terminology sharing network is capable capturing semantic relationship in sub topic level. 3. Takes the users social behaviour to consideration by rounding up all the questions answered by closely associated aspects.N +M+U hyperedges are constructed in our hypergraph, where U is the number of involved experts.

$$P(vi; ej) = \begin{cases} 1\ \text{Inter} - \text{expert Relationships} \\ K(qi;\ qj)\ \text{Content Similarity} \\ 1\ \text{Terminology} - \text{Sharing} \end{cases}$$

$K(qi; qj) = \exp(-\frac{||q_i - q_j||^2}{\sigma^2})$
The weight of each hyperedge is,

$W(e_j) = \sum_{v_i \in e_j} h(v_i, e_j)$

### 5.5 Global learning optimization.

The optimization of global learning technique is being done it has mainly three objectives 1. It should guarantee the relevance probability function in continuous and smooth semantic space. 2. Related to empirical loss function which forces the relevance probability. 3. Encourage the values of medical record, which are connected by hierarchical structured terminologies should be similar to each other.

### 5.6 Pseudo label estimation.

The initial relevance scores are being seen detail.$Y_{N \times M}$is a label biases matrix, where N and M respectively denotes the number of medical records and the number of terminologies. $Y_{ij}$stands for the initially estimated relevance between medical recordsi and terminology j.
$Y_{ij} = \frac{1}{|X_j|} \sum_{q_c \epsilon X_j} K(\boldsymbol{q_i}, \boldsymbol{q_c})$ $X_j$ is a set of medical records, the closeness of unknown medical record is being determined.

### 5.7 Complexity analysis.

$O(E^3 + 2NE^2 + 2EN^2 + N3 + dN^2)$

Where d denotes the dimension of extracted features, N and E respectively represents the number of involvedmedical records and hyper edges.Pre-clustering the medical records during the data collection stage into several subgroups to avoid complexity.

## 6. Experiments

### 6.1 Experimental Settings

More than 109 thousands records has been taken from health which contains both questions and answers, the answers are being answered by many experts

| Unique Question # | Answer # | Duplicate Answer # | Unique Expert # |
|---|---|---|---|
| 109,843 | 160,736 | 6,444 | 5,958 |

Nearly 54% of experts have answered 4 questions, more than 33.2% of experts have answered at least 2 questions. The questions with only one answer or multiple answers but from different doctors are being removed because they are unable to contribute relationship they also provided noise so they are being eliminated. The experts who has answered less than 4 questions are also being removed, the non-active doctors who reply less frequently are also being removed because they have very less care in answering questions.

### 6.2 Local mining analysis.

Nearly 8910 records obey the noun phrase as described in section 3. These records also suits the power law of distribution, according to the law data is nothing but data with similar names or words. Example family names may common between many families. When analysing the medical concept we found 1. Not all the detected medical concept can be mapped to one entry is SNOMED CT. Example some experts has misspelt menses as mense while mense is not reachable. 2. Many medical records can be converted into similar medical records 3. Less than 15% of records are same with their normalized terminologies.

The representive medical concepts and their corresponding terminologies after normalization.

| S.no | Medical concepts | Normalized technologies |
|---|---|---|
| 1. | Birth control | Contraception |
| 2. | Blood loss | Haemorrhage |
| 3. | Breast cancer | Malignant tumour of breast |
| 4. | Condom | Uses of contractive health |
| 5. | Home pregnancy test | Pregnancy test |
| 6. | Late menses | menstrual period late |
| 7. | Sex | Finding sexual intercourse |
| 8. | Period pain | Dysmenorrhea |

### 6.3 Graph based global learning analysis.

PR feedback, RW re-ranking, CHL learning, CG learning are some of the approaches or methods used for doing graph analysis between records.

### 6.4 Medical terminology

We adopt 2 metrics that able to characterizing precisions. 1. S@K finding relevant medical records S@K is 1, if relevant medical records are found S@K is 0, if no relevant records found. 2. P@K average similar records

P@K = $\frac{|C \cap R|}{|C|}$ where c is similar record, R is manually labelled positive one.

## 7. Conclusion and Future Work

This paper presents a medical terminology assignment which bridges the gap between health seekers and health care knowledge. The scheme compress of local compass of local mining and global mining however local mining suffers from missing key words and presence of relevant medical concepts, so we use global learning which provides a detailed description and outcomes the problem of local mining. We also incorporate NLP process for native users to understand the answers in their own language. In future we can incorporate the spell check of medical words and also try to improve the accuracy, speed receiving of answers.

## References

[1] A. e-HIM Work Group on Computer-Assisted Coding, "Delving into computer-assisted coding," *Journal of AHIMA*,2004.
[2] G. Leroy and H. Chen, "Meeting medical terminology needs-the ontology-enhanced medical concept mapper," *IEEETransactions on Information Technology in Biomedicine*, 2001.
[3] G. Zuccon, B. Koopman, A. Nguyen, D. Vickers, and L. Butt, "Exploiting medical hierarchies for concept-basedinformation retrieval," in *Proceedings of the Australasian Document Computing Symposium*, 2012.
[4] E. J. M. Laur´ıa and A. D. March, "Combining bayesian text classification and shrinkage to automate healthcare coding: A data quality analysis," *Journal of Data and Information Quality*,2011.
[5] L. Yves A., S. Lyudmila, and F. Carol, "Automating icd-9-cm encoding using medical language processing: A feasibilitystudy," in *Proceedings of the AMIA Annual Symposium*, 2000.
[6] C. Dozier, R. Kondadadi, K. Al-Kofahi, M. Chaudhary, andX. Guo, "Fast tagging of medical terms in legal text," in*Proceedings of the International Conference on Artificial Intelligenceand Law*, 2007.
[7] L. V. Lita, S. Yu, S. Niculescu, and J. Bi, "Large scale diagnosticcode classification for medical patient records," in *Proceedingsof the Conference on Artificial Intelligence in Medicine*, 1995.
[8] L. S. Larkey and W. B. Croft, "Automatic assignment of icd9 codes to discharge summaries," *PhD Thesis, University ofMassachusetts at Amherst*, 1995.
[9] H. Suominen, F. Ginter, S. Pyysalo, A. Airola, T. Pahikkala, S. Salanter, and T. Salakoski, "Machine learning to automatethe assignment of diagnosis codes to free-text radiology reports: a method description," in *Proceedings of the ICMLWorkshop on Machine Learning for Health-Care Applications*, 2008.
[10] J. Patrick, Y. Wang, and P. Budd, "An automated system for conversion of clinical notes into snomed clinical terminology," in *Proceedings of the fifth Australasian symposium on ACSW frontiers*, 2007.