Exploring Clustering Based Knowledge Discovery towards Improved Medical Diagnosis

Rajendra Prasath
Dept. of Business Information Systems, University College Cork
Ireland
R.Prasath@ucc.ie

Philip O'Reilly
Dept. of Business Information Systems, University College Cork, Ireland
Philip.OReilly@ucc.ie

ABSTRACT

We propose to develop a framework for an intelligent reasoner with capabilities that support complex decision making processes in medical diagnosis. *Identifying* the causes, *reasoning* the effects to explore information geometry and *learning* the associated factors, from medical forum information extracted, are the core aspects of this work. As part of the proposed framework, we present an approach that identifies semantically similar causes and effects for any specific disease from medical diagnosis literature using implicit semantic interconnections among the medical terms. First we crawled MedHelp[1] forum data and considered two types of information: *forums* data and *posts* data. Each forum link points to a specific disease and consists of several topics pertaining to that disease. Each topic consists of multiple posts that carry either users' queries/difficulties or doctor's feedback pertaining to the issue(s) of the users. We use graph based exploration on the terms (diseases) and their relations (in terms of causes/effects) and explore the information geometry pertaining to similar diseases. We performed a systematic evaluation to identify the relevance of the contextual information retrieved for a specific disease and similar factors across different diseases. The proposed approach looks promising in capturing similar causes and/or effects that pertain to multiple diseases. This would enable medical practitioners to have a multi-faceted view of a specific disease/condition.

Keywords

Causes and Effects, Medical Diagnosis, Semantically Similar diseases, Information Geometry, Graph Analysis

1. INTRODUCTION

Understanding the causes and effects pertaining to a specific disease is key to better prediction in medical diagnosis and improved patient management. Diseases/Conditions may have similar or semantically related causes and effects. Furthermore, gaining insight on how diseases are diagnosed and managed would enable medical practitioners to make more informed decisions on disease management. Illustrating the role of machine learning as an enabler of this for more informed decision support is a key aspect of this paper. In this paper, we present an approach to identify causes and effects that exist across different diseases using a graph clustering based knowledge discovery approach.

Understanding causation and correlation between individual factors is key to decision making in multiple domains including medicine and business. Traditionally, such association between elements has been identified through human interpretation of text. However, this is a very time consuming, manual, labour intense process and is limited by human capacity. The ability to mine textual content, identify association and the nature of that association between elements using machine learning techniques provides significant opportunities. Specifically in the medical domain, where a significant amount of content is textual in nature (e.g. medical notes), having the ability to identify causation and correlation between elements in large medical datasets provides significant opportunity for advancing medical research and enabling better decision making pertaining to condition diagnosis and patient management.

In this paper, we attempt to identify and create term clusters using graph clustering approach and then perform topic classification. This approach improves the document classification task by putting the terms, that are semantically related, in the same cluster. Users could search for a specific disease and explore information pertaining to its causes and effects by means of semantically related texts.

2. CLUSTERING BASED KNOWLEDGE DISCOVERY

To incorporate natural language understanding, commonsense and domain specific knowledge could be used to improve the text representation by including more generated informative features to perform deep understanding of the document text than the mere *Bag-of-Words* approach [4, 2, 8]. Mitra *et* al. [6] proposed an unsupervised feature selection algorithm suitable for data sets, based on measuring similarity between features whereby redundancy therein is removed. Pedersen and Kulkarni [7] presented a system called SenseClusters to cluster similar contexts in natural language text and assigns identifying labels to these clusters based on their content. In addition to clustering similar contexts, it can be used to identify

[1] http://www.medhelp.org/forums/list

MedIR 2014, July 11, 2014, Gold Coast, Australia.
ACM SIGIR.

synonyms and sets of related words[2]. To incorporate this kind of additional common-sense/domain knowledge, Gabrilovich and Markovitch [3] used world knowledge from open source knowledge repository like Wikipedia to generate additional features. Similar intuition is adopted to form word clusters that generate features enriching the document content in a better way. Then the documents are represented in the knowledge-rich space of generated features. This leads to better organization of semantically related text representation. In this proposed scheme, given a knowledge repository, the text documents are examined and their representation is enriched in a completely mechanical way. Motivated by the above considerations, our aim is to empower machine learning techniques for text representation with a substantially wider body of knowledge like the one obtained from the superior inference capabilities of humans.

2.1 Mathematical Formulation

In this section, we characterize the medical forum data in a formal way. Each post pertaining to a specific topic is informally written and we focus on terms and their co-occurrences. We have n textual descriptions, viz-a-viz, posts: $P = \{p_1, p_2, \cdots, p_n\}$ and each post can be formally represented as a sequence of terms, illustrating the scenario of the underlying disease, as follows: $p_i = \{t_1, t_2, \cdots, t_m\}$ where $1 \leq i \leq m$ and m varies differently for different post. We convert the entire text data of all posts into a graph, say $G = (V, E)$ where V represents the set of nodes (each term in P is considered as a node) and the co-occurrence of any pair of nodes across the posts is considered to be the edge representing the strength of association between the pair of nodes.

2.2 Graph Clustering

Clustering deals with identifying a pattern/structure in the bunch of unlabeled data. In general, clustering organizes data into groups whose members are related in some way and two or more data can be grouped into the same cluster if they are, in some way, falling close to each others' context. Clustering has many useful applications like finding a group of people with similar behavior, processing orders, grouping plants and animals, grouping web blog data to access similar patterns.

While exploring a variety of possibilities to identify the context of causes and effects of diseases from text fragments, feature space grows and it is hardly possible to limit the expansion of the new feature space containing the local contexts extracted from the informal writing of medical data. This could possibly be solved by using dimensionality reduction techniques to limit the size of the document to be classified. This attempt first makes word cluster vectors using unsupervised feature generation by identifying the related contexts. Then using the identified contexts, supervised learning is performed for categorizing the given text collection consisting of user posts.

First, we use the entire collection of medical diagnosis related posts and filter out the list of the distinguishable unique terms. Using these terms, we first build the weighted graph in which nodes represent terms and edges represent the weight - the number of documents in which the given pair of terms co-occurs across the collection of posts. For each unique term, the list of documents in which it occurs is retrieved. Using this data, we build the weighted graph in which the edge between two terms would represent their semantic association implicitly. This process is repeated for all features and a weighted graph for the overall data is constructed. Thus the problem is modeled into a graph clustering problem. This results in a graph $G = (V, E, A)$ where $|V| = n$ represents the number of unique terms; $|E|$ represents the number of edges and the adjacency matrix; and A is $|V| \times |V|$ whose nonzero entries correspond to the edge weight between a pair of terms (adjacency list is assumed in case of sparse matrix - in this case, number of rows in the graph represents the total number of terms in the graph).

We use the kernel-based multilevel clustering algorithm proposed by Dhillon *et* al. [1] on the weighted input graph with the number of desired partitions. This algorithm uses three steps: coarsening, base-clustering, and refinement. In coarsening, the given graph is repeatedly transformed into smaller subgraphs. This process is repeated until a few nodes remain in the graph. Then during base clustering, regional growing approach [5] could be used with these few nodes. The quality of the resulting clusters depends on the choice of the initial nodes. The refinement process is applied as follows: If a node in G_i is in cluster c, then all nodes in G_{i-1} formed from that node are in cluster c. For more details, please refer to [1, 5].

In this work, we generate term clusters from extracted word graphs, using co-occurrence information of terms. The task is to partition the graph into clusters so that terms could be grouped into a few subsets and dimensionality of the new term space is reduced. Now based on the generated term clusters, we perform classification to identify similar causes and effects across various diseases.

2.3 Proposed Approach

The proposed approach works as follows: From the posts of each topic, textual descriptions are extracted. Unique terms (after removing stop words) are considered as nodes in the graph and the number of times a pair of terms co-occurs in the entire corpus is considered as the weight of the edge connecting the pair of terms. At first, we build word cluster vectors using graph clustering algorithm.

Algorithm 1 Building Word Clusters

Input: A set of n textual descriptions (posts)
$P = \{p_1, p_2, \cdots, p_n\}$
A set of predefined category labels $C = \{c_1, c_2, \cdots, c_l\}$

Build Word Cluster Vectors:

1: Extract text from posts and build the unique word list
2: **for** each unique term t_i in the word list **do**
3: Identify the existence of edges from t_i to all other terms with nonzero positive weight.
4: Store the co-occurring term with its corresponding edge weight in the adjacency list
5: **end for**
6: Use kernel-based multilevel graph clustering algorithm on the adjacency list and perform clustering to generate cluster IDs
7: For every cluster ID, construct word clusters
8: Store these word cluster vectors

[2] *word* and *term* are used interchangibly

Secondly, we use these word cluster vectors to re-represent text documents that discuss the causes and effects of a specific disease. Then we perform classification of the re-represented text documents. The proposed algorithm inherently applies clustering semantically related terms and performs classification of them. Based on the word clusters found in the first step, we perform the classification on the clustered space of label pertaining to the terms in the original documents.

Algorithm 2 Classification using Word Clusters

1: Preprocess the text documents by removing numbers, punctuations and stop-words (using SMART[3] word list)

2: **for** each processed text (post) data p_i in P **do**
3: **for** each unique term in p_i **do**
4: Identify its cluster id
5: Map the given feature in terms of its cluster ID
6: Augment text fragments with cluster ID mappings
7: **end for**
8: **end for**
9: Build classifier on these mapped/expanded text data (containing only cluster IDs) and use it to predict the class of the text having similar causes and effects
10: Compute the classification accuracy

Output: The category label(s) for texts (post) that have similar causes and effects across diseases.

3. EXPERIMENTAL RESULTS

3.1 Corpus

We crawled a subset of MedHelp[4] forum data from the world wide web. The MedHelp forum is organized as a set of topics, each representing a specific disease and under each topic, there are several subtopics. Each subtopic consists of several posts from both users (may be patients or their dependents) and doctors as well.

For our experiments, we have selected 15 categories covering the most widely discussed topics in the MedHelp forum. The details of this experimental corpus is given in Table. 1

We have used 3 different types of classifiers, namely *Naive Bayes*, *k−Nearest Neighbours (k−NN)*, and *Support Vector Machines (SVM)*, to test the effect of the proposed approach. We have used *rainbow*[5] to build the classification models and during classification, we have used 60% of the data for training and 40% for testing.

3.1.1 Evaluation Methodology

We have used Precision, Recall, F-Measure and the classification Accuracy to evaluate the quality of the identified diseases having similar causes and effects. We use the two-way confusion matrix given in Table. 2 to derive the evaluation measures:

Precision is defined as follows:

$$Precision(P) = \frac{TP}{TP + FP}$$

[4] http://www.medhelp.org/forums/list/

[5] The 'Bow' Toolkit - http://www.cs.cmu.edu/~mccallum/bow/

S.No	Class	#Diseases (Posts)
1	Asthma	35 (43)
2	Breast-Cancer	35 (37)
3	COPD	41 (47)
4	Cosmetic	57 (64)
5	Dental	97 (100)
6	Embarazo	41 (52)
7	Genetic-Disorder	59 (68)
8	Hepatitis	110 (147)
9	Kidney	69 (89)
10	Liver-Transplant	64 (62)
11	Oral	35 (47)
12	Pathology	31 (36)
13	Respiratory	48 (54)
14	Thyroid-Cancer	58 (63)
15	Varicose-Veins	36 (52)

Table 1: Corpus Statistics

	Correctly Classified	Wrongly Classified
Actual=Yes	True Positive (TP)	False Negative (FN)
Actual=No	False Positive (FP)	True Negative (TN)

Table 2: 2-way confusion matrix

Recall is measured by the following equation:

$$Recall(R) = \frac{TP}{TP + FN}$$

F-Measure is computed by considering the ratio between Precision and Recall and hence calculated as follows:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

The balanced F_1-measure with $\beta = 1$ (when P and R are weighted equally), is given by

$$F_1 = \frac{2pr}{P + R}$$

Classification accuracy is measured from confusion table as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

3.1.2 Discussion

We have observed that the proposed approach performs well in classifying the medical text having causes and issues expressed by either the patients or their friends or relatives. Figure. 1 shows the classification accuracy of proposed approach with top 15 classes using three different types of classifiers. While using the Naive Bayes method, the accuracy for the class "Asthma" goes down due to the fact that certain specific causes are pertaining to the "Respiratory" related disease. Similar misclassification takes place across "Dental" and "Oral" classes and these misclassified instances share common causes. Even though, the diseases like "Breast-cancer" and "Thyroid-cancer" share a common cause for cancer, misclassification is significantly less. Additionally the misclassification takes place across the diseases: "Hepatitis" and "Liver-Transplant".

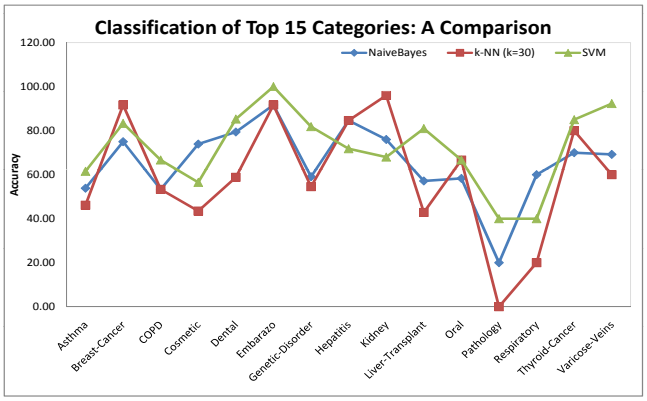

Figure 1: Classification of top 15 classes: A comparison of Naive Bayes, $k-$NN and SVM methods

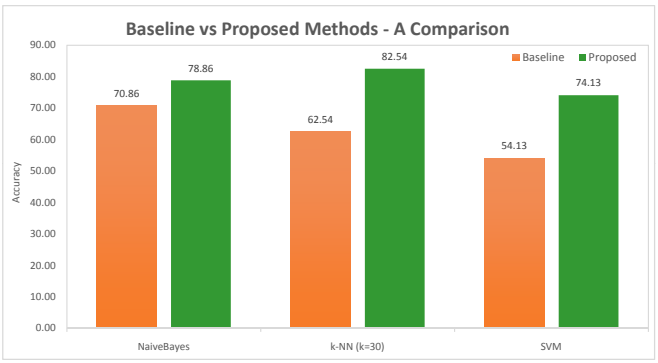

Figure 2: Comparison of the overall classification accuracy: Baseline vs Proposed approaches

We have observed the classification accuracy of $k-$NN classifier for various values of k (=10, 20, 30, 50) and found that for k =30, the system performs very well. In this classification task, we have observed that the maximum number of instances from the class "Liver-Transplant" is misclassified under "Hetatitis" class as the causes of these common diseases coincide by and large. At the same time, none of the instances is classified correctly for the class "Pathology" as these causes and effects are very similarly described as that of "Hepatitis" and "Thyroid-cancer". While using the SVM classifier, we noticed that some instances are misclassified across the classes: "Breast-cancer" and "Cosmetic". In this case, user raised cross reference related queries with post surgical treatment of the Breast-cancer disease. Similar misclassification is found across the classes: "Chronic Obstructive Pulmonary Disease" (COPD) and "Resporatory". Subsequently, we would like to apply this approach for effective retrieval of causes and effects pertaining to a specific disease. Also we will draw the information geometry of prominent diseases that share the common causes/effects in our subsequent experiments.

4. CONCLUSION

We proposed a method to enable greater understanding of various conditions, their symptoms, treatment and management by *identifying* similar scenarios, *reasoning* the effects to explore information geometry and *learning* the associated contextual factors, from medical forum information extracted from health services data. This approach identifies semantically similar causes and effects for any specific disease/condition, using implicit semantic interconnections among the medical terms. We use graph based exploration on the terms and their relations (causes/effects) across the collection of posts and explore the information geometry pertaining to the similar diseases. We evaluated the relevance of the contextual information retrieved for a specific disease and/or similar factors across different diseases. The proposed approach looks promising in capturing similar scenarios pertaining to multiple diseases. This would enable medical practitioners to have a multi-faceted view about any specific disease, towards better decision making.

Acknowledgments: This research is co-funded by the Irish Government (Enterprise Ireland) and European Union (European Regional Development Fund). Dr. Philip O'Reilly is the Principal Investigator responsible for this research and can be contacted at `Philip.Oreilly@ucc.ie`.

5. REFERENCES

[1] Dhillon, I. S., Guan, Y., and Kulis, B. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE Trans. Pattern Anal. Mach. Intell. 29*, 11 (2007), 1944–1957.

[2] Gabrilovich, E. *Feature Generation for Textual Information Retrieval Using World Knowledge*. PhD thesis, Technion - Israel Institute of Technology, Haifa, Israel, 2006.

[3] Gabrilovich, E., and Markovitch, S. Feature generation for text categorization using world knowledge. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (San Francisco, CA, USA, 2005), IJCAI'05, Morgan Kaufmann Publishers Inc., pp. 1048–1053.

[4] Giles, J. Internet encyclopaedias go head to head. *Nature 438*, 1 (2005), 900–901.

[5] Karypis, G., and Kumar, V. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput. 20*, 1 (1998), 359–392.

[6] Mitra, P., Murthy, C. A., and Pal, S. K. Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 3 (2002), 301–312.

[7] Pedersen, T., and Kulkarni, A. Identifying similar words and contexts in natural language with senseclusters. In *Proc. of the 20th national conf. on Artificial intelligence* (2005), AAAI'05, AAAI Press, pp. 1694–1695.

[8] Prasath, R., and Sarkar, S. Unsupervised feature generation using knowledge repositories for effective text categorization. In *Proceedings of the 2010 Conference on ECAI 2010: 19th European Conference on Artificial Intelligence* (Amsterdam, The Netherlands, The Netherlands, 2010), IOS Press, pp. 1101–1102.