International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:11, 2011

Concepts Extraction from Discharge Notes using Association Rule Mining

Basak Oguz Yolcular

***Abstract*—**A large amount of valuable information is available in plain text clinical reports. New techniques and technologies are applied to extract information from these reports. In this study, we developed a domain based software system to transform 600 Otorhinolaryngology discharge notes to a structured form for extracting clinical data from the discharge notes. In order to decrease the system process time discharge notes were transformed into a data table after preprocessing. Several word lists were constituted to identify common section in the discharge notes, including patient history, age, problems, and diagnosis etc. N-gram method was used for discovering terms co-Occurrences within each section. Using this method a dataset of concept candidates has been generated for the validation step, and then Predictive Apriori algorithm for Association Rule Mining (ARM) was applied to validate candidate concepts.

***Keywords*—**association rule mining, otorhinolaryngology, predictive apriori, text mining

I. Introduction

COMPUTERS help us to store and access data, and make searches, have a significant effect on our daily life. Recently the amount of data which are stored in databases, have grown significantly, creating a great demand for new tools for turning data into useful knowledge. To satisfy this need researchers from various technological areas, such as machine learning, pattern recognition, statistical data analysis, data visualization, neural networks, econometrics, information retrieval, information extraction etc., have been exploring ideas and methods. Text mining is one of these tools that saw an increasing interest in the 2000s, for enabling people to find unknown information and facts from the free-text data [1].

Information extraction is one of the effective approaches to text mining that transforms the unstructured text to the structured form. It may be useful to first use IE to transform the unstructured data in the document corpus into a structured database, and then use traditional data mining tools to identify patterns in this extracted data. An Information Extraction system generally converts unstructured text into a form that can be loaded into a database table. Useful information such as names of people, places, or organizations mentioned in the text is extracted without a "deep understanding" of the text [1].

B. O. Akdeniz University, Department of Biostatistics and Medical Informatics, Antalya, 07059 Turkey (phone: +90 242-249-6927; fax: +90 242-227-6999; e-mail: basakoguz@akdeniz.edu.tr).

Several systems were designed to solve the problem of extracting information from unstructured textual data in medical field. RADA is the one of the developed systems, that was designed to extract and structure the information contained in free text radiology reports. This system uses natural language processing techniques to transform free text description into structured information units that can be used to index and provide access to image databases [2]. Another developed system extracts structured information about specimens and their related findings in free text surgical pathology reports [3]. In the biomedical field, extracting information has specific problems, so majority of existing works have been focused on the detection of genes and proteins [4]. One of the most successful rules-based approaches to gene and protein NER in biomedical texts has been the AbGene system. It applies manually generated post-processing rules based on lexical-statistical characteristics that help further identify the context in which gene names are used and eliminate false positives and negatives [5].

In this paper, we used electronically available discharge notes from Otorhinolaryngology department of an university hospital. Otolaryngology is the branch of medicine that specializes in the diagnosis and treatment of ear, nose, throat, and head and neck disorders [6]. Discharge notes contain much useful information about the patient, useful not only for subsequent visits but for a variety of other tasks. To turn a collection of discharge notes into an effective resource for research or quality assurance, it is necessary to structure information in it. To make it usable, we need to extract the diseases and procedures and code them into a standard vocabulary. This study has therefore two objectives. First, to develop a system that transforms the unstructured discharge notes to the structured form. Second, to extract structured information from the discharge notes using ARM.

II. Methods

A. Text Collection

In our experiments, we deal with a collection of electronically available discharge notes from Otorhinolaryngology department of a university hospital. The database was consisted 592 documents generated between 2002 and 2007. The number of words each document contained was between 400 and 700. In order to decrease the system process time the documents were converted into text format.

B. System Feature

The system was developed on Microsoft Visual Studio.Net platform using C# programming language. It only supports the texts that are written in Turkish language. In the system, discharge notes can be loaded from a folder which includes discharge notes in Unicode text formats. Firstly, the discharge notes, gathered from otolaryngology department, were applied preprocessing to decrease the dimension and improve performance of the system. There were four main components for preprocessing of the documents including a stemming module, a stopword filter, spelling error corrector and n-gram generator. The input texts were separated into tokens by using whitespaces and the punctuation marks (periods, commas etc.) and whitespaces which are more than one were removed from the texts. Then each word in the discharge notes were converted into their base form using Zemberek library which is an open source project that has been developed to provide NLP (Natural Language Processing) solutions for Turkish language to the system developers [7].

After this step, stopwords, also called function words (for instance there, more, yet etc.), are eliminated from the discharge notes. In order to identify some standard section names in the discharge notes, including patient history, age, problems, and diagnosis etc., several word lists was constituted. Firstly, the program tries to find the sections by looking how these are formatted (e.g., upper case followed by a colon) and then compare the section names with the word lists. Secondly, it transforms the discharge notes into a data table using sections as column. Once the text within each individual section was determined, the next step is to identify candidate concepts using the terms co-Occurrences within each section by calculating n-gram words. For each "n", where "n" is the number of words in the concept, the algorithm passes through the data collection once. A word or a group of consecutive words that occurs frequently enough in the entire text collection is considered as a candidate concept and a list of concept candidates is generated.

C. Association Rule Mining

ARM is a major data mining technique, and is a most commonly used pattern discovery method. It retrieves all frequent patterns in a data set and forms interesting rules among frequent patterns. The principal parts of an association rule are the rule body (also referred to as antecedent) and the rule head (also referred to as consequent). The rule body contains the item or items for which the ARM function has found an associated item. The rule head contains the item found. Formally, an association rule R is an implication $X \Rightarrow Y$, where X and Y are sets of items in a given dataset. The confidence of the rule conf(R) is the percentage of transactions that contains Y amongst the transactions containing X. The support of the rule supp(R) is the percentage of transactions containing X and Y with respect to the number of all transactions [8].

Association rule mining has been widely used in medical data analysis. Brossette et al. [9] uncovered association rules in hospital infection control and public surveillance data. Paetz and Brause [10] discovered association rules in septic shock patient data. Sequential patterns have been found in chronic hepatitis data by Ohsaki et al. [11] and in adverse drug reaction data by Chen et al. [12]. Ordonez et al. used association rules to predict heart disease [13].

In this study, ARM was used in order to validate candidate concepts. All the concepts within each section were evaluated by examining the results of this analysis. Among algorithms of ARM rule computing, the best-known and most popular one is Apriori algorithm which computes the frequent itemsets in the database through several iterations [14]. The Apriori algorithm finds association rules in two steps. First, all item sets x with support of more than the fixed threshold "minimum support" are found. Then, all item sets are split into left and right hand side x and y (in all possible ways) and the confidence of the rules $[x \Rightarrow y]$ is calculated. All rules with a confidence above the confidence threshold "minimum confidence" are returned [14]. Similarly to the Apriori algorithm, the Predictive Apriori algorithm generates frequent item sets, but it uses a dynamically increasing minimum support threshold. It searches with an increasing support threshold for the best rules concerning a support-based corrected confidence value. A rule is added if: the expected predictive accuracy of this rule is among the "n" best and it is not subsumed by a rule with at least the same expected predictive accuracy [15].

In this study, Predictive Apriori algorithm was applied to evaluate dataset of candidate concepts. Its parameters were calculated, analyzed and visualized by the freely available software WEKA (The Waikato Environment for Knowledge Analysis). WEKA is a tool for data analysis and includes implementations of data pre-processing, classification, regression, clustering, association rules, and visualization by different algorithms [16]. By providing a diverse set of methods that are available through a common interface, WEKA makes it easy to compare different solution strategies based on the same evaluation method and identify the one that is most appropriate for the problem at hand [17].

III. Results

592 discharge notes were analyzed to discover the common theme and concepts. ARM was applied to the dataset of candidate concepts. In this paper, results of problem and diagnosis section are demonstrated.

The results of ARM for specific otorhinolaryngology problem are given in Table I. There are 24 association rules that can be significant concepts. The best 2-grams concept with higher accuracy rate is "Hoarseness (Ses Kısıklığı)". In 3-grams concepts, "Drainage from right Ear (Sag Kulak Akıntı)" has higher accuracy rate. Generally, 3-grams and 4-grams concepts include the concepts similar to 2-grams with the position of problem. As seen in Table I, 4-grams concepts can consist of two or higher number of different concepts.

TABLE I
RESULTS OF PREDICTIVE APRIORI ALGORITHM FOR PROBLEM SECTION

N-grams	Concepts (n)	Accuracy*
2-grams		
	Ses (126) ==> kisiklik (124)	0.99456
	Geniz (8) ==> akinti (8)	0.9915
	Burun (45) ==> tikaniklik (37)	0.80851
	Bas (42) ==> donmesi (34)	0.79545
	Dudak (10) ==> yara (8)	0.75852
	Solunum (14) ==> sikinti (11)	0.75048
	Nefes (35) ==> darlik (24)	0.67568
	Isitme (66) ==> azlik (37)	0.55882
	Boyun (86) ==> sislik (46)	0.53409
3-grams		
	Sag akinti (10) ==> kulak (10)	0.91658
	Kulak alti (19) ==> sislik (18)	0.90491
	Sol akinti (8) ==> kulak (8)	0.89993
	Sik bogaz (8) ==> enfeksiyonu(8)	0.89993
	Önü sislik (7) ==> kulak (7)	0.88883
	Burun geniz (5) ==> tikaniklik (5)	0.8571
	Agzi (11) ==> acik uyuma (10)	0.84624
	Agri yutma (4) ==> bogaz (4)	0.8333
4-grams		
	Kisiklik darlik (10) ==> ses nefes (10)	0.91658
	Sag azlik (8) ==> kulak isitme (8)	0.89993
	Ses yutma (8) ==> kisiklik gucluk (8)	0.89993
	Kisiklik gucluk (8) ==> ses yutma (8)	0.89993
	Sag taraf sislik (8) ==> boyun (8)	0.89993
	Agzi horlama (7) ==> acik uyuma (7)	0.88882
	Bas donmesi bulanti (7) ==> kusma (7)	0.88882

*Accuracy>0.5

The results of ARM for specific otorhinolaryngology diagnosis listed in Table 2. There are 12 association rules that can be significant concepts. According to the results of analysis, any 4-grams concepts can be found. The best 2-grams concepts with higher accuracy rate are "Larynx (Larenks) CA", "Septum Deviation (Septum Deviasyon)", and "Otitis Media". Similar to the results of problem section, 3-grams concepts include the same concepts as 2-gram concepts with the position or status of diagnosis.

TABLE II
RESULTS OF PREDICTIVE APRIORI ALGORITHM FOR DIAGNOSIS SECTION

N-grams	Concepts (n)	Accuracy*
2-grams		
	Larenks (107) ==> CA (103)	0.95412
	Otitis (12) ==> media (12)	0.92781
	Septum (18) ==> deviasyonu (17)	0.89988
	Dudak (7) ==> CA (7)	0.87777
	Glottik (6) ==> larenks (6)	0.85674
	Dil (32) ==> CA (27)	0.82326
3-grams		
	Kronik media (8) ==> otitis (8)	0.95325
	Opere CA (8) ==> larenks (8)	0.95325
	Alt (6) ==> dudak CA (6)	0.93238
	Glottik (6) ==> larenks CA (6)	0.93238
	Sag kitle (4) ==> parotis (4)	0.89491
	Transglottik (2) ==> larenks CA (2)	0.81852

*Accuracy>0.5

The results of ARM for specific otorhinolaryngology diagnosis listed in Table 2. There are 12 association rules that can be significant concepts. According to the results of analysis any 4-grams concepts can be found. The best 2-grams concepts with higher accuracy rate are "Larynx (Larenks) CA", "Septum Deviation (Septum Deviasyon)", and "Otitis Media". Similar to the results of problem section, 3-grams concepts include the same concepts with the position or status of diagnosis.

IV. DISCUSSION AND CONCLUSION

In this paper we explain our system that is developed to transform the unstructured discharge notes to the structured form and extract the concepts from the discharge notes. One disadvantage of this study is its data collection which is dependent to the domain. Therefore, the results of study are highly correlated to the data collection. Also, a term appears less than our frequency threshold will not be considered as a concept. Since some specific terms will only appear in certain types of otorhinolaryngology discharge notes, these terms will be missed by our system. In order to solve this problem discharge notes can be classified by their type. Especially in Turkey, it is an emerging research area where the need for new applications, thus our intention is to develop this system in order to contribute to future studies.

REFERENCES

[1] M. Konchady , *Text Mining Application Programming*. Boston: Charles River Media, 2006, ch. 1.
[2] D.B. Johnson, R.K. Taira, A.F. Cardenas, and D.R. Aberle, "Extracting Information from Free Text Radiology Reports", *Int. J. Digit Libr*., vol. 1, no. 3, pp. 297-308, Dec. 1997.
[3] G. Schadow , C.J. Mcdonald,. "Extracting Structured Information from Free Text Pathology Reports," in *Conf. 2003 AMIA Annu. Symp. Proc*., pp. 584-8.
[4] R.A. Erhardt, R. Schneider , and C. Blaschke, "Status of Text Mining Techniques Applied to Biomedical Text," *Drug Dicovery Today,* vol. 11, no. 7-8, pp. 315-25, Apr. 2006.
[5] A.M. Cohen, W.R. Hersh, "A Survey of Current Work in Biomedical Text Mining," *Briefings in Bioinformatics*, vol. 6, no. 1, pp. 57-71, Mar. 2005.
[6] Wikipedia, "Otolaryngology (Unpublished work style)," unpublished.
[7] Google, "Zemberek (Unpublished work style)," unpublished.
[8] DB2 Universal Database, "Associations (Unpublished work style)," unpublished.
[9] S.E. Brossette, A.P. Sprague, J.M. Hardin, K.W.T. Jones, and S.A. Moser , "Association rules and data mining in hospital infection control and public health surveillance," *Journal of American Medical Informatics Association*, vol. 5, pp. 373-81, 1998.
[10] J. Paetz, R.W. Brause, "A frequent patterns tree approach for rule generation with categorical septic shock patient data," in *Proceedings of the second international symposium on medical data analysis*, London: Springer-Verlag, 2001, pp. 207-12.
[11] M. Ohsaki, Y. Sato, H. Yokoi, and T. Yamaguchi, "A rule discovery support system for sequential medical data in the case study of a chronic hepatitis dataset," in *Proceedings of the ECML/PKDD 2003 discovery challenge workshop*.
[12] J. Chen, H. He, G.J. Williams, and Jin H, "Temporal sequence associations for rare events," in *Advances in knowledge discovery and data mining*, Berlin/Heidelberg: Springer, 2004, pp. 235-9.
[13] C. Ordonez, N.F. Ezquerra, and C.A. Santana, "Constraining and summarizing association rules in medical data," *Knowledge and Information Systems,*vol. 3, pp. 1-2, 2006.
[14] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," *in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, DC: SIGMOD Conference, 1993, pp. 207-216.

[15] T. Scheffer, "Finding Association Rules That Trade Support Optimally against Confidence," in *Proc of the 5th European Conf. on principles and Practice of Knowledge Discovery in Databases (PKDD'01)*, Freiburg, Germany: Springer-Verlag, 2001, pp. 424-435.
[16] I.H. Witten, E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations," San Francisco, 2005.
[17] E. Frank, M. Hall, L. Trigg, G. Holmes, and I.H. Witten, "Data Mining in Bioinformatics using Weka," *Bioinformatics*, vol. 20, no. 15, pp. 2479-2481, 2004.