A Content-based Recommendation System for Medical Concepts: Disease and Symptom

Anupam Mondal **Dipankar Das** **Sivaji Bandyopadhyay**
Department of Computer Science and Engineering
Jadavpur University, Kolkata, India
link.anupam@gmail.com, dipankar.dipnil2005@gmail.com,
sivaji_cse_ju@yahoo.com

Abstract

Dealing with healthcare data is becoming difficult because decision-making becomes crucial to extract information from a huge volume of medical concepts being evolved on daily basis. Moreover, unstructured and semi-structured medical corpora and lack of domain-experts fueled more challenges in this research arena. In order to face one of such challenges, we have developed a baseline model of Medical Recommendation System (MRS). Primarily, MRS helps the experts (e.g. medical practitioners and doctors) by suggesting relevant diseases and symptoms as well as their in-between similarities. Here, we have used a content-based approach to identify similar types of diseases and symptoms by employing two well-known distance metrics, Manhattan and Euclidean. Evaluation based on perplexity score reveals that the performance of MRS is equally well for identifying relevant diseases and symptoms.

1 Introduction

During last few decades, medical information retrieval and extraction behavior are largely observed in the web. A recent survey says that 59% and 49% of U.S. and Indian internet users [1] are looking for online health information e.g., diseases, diagnosis, and treatments (Fischer et al., 2014). Such information helps the doctors as well as patients in their decision-making process for treatment.

Besides, medical experts face difficulties in identifying relevant information from the web due to information overloading (Sommerhalder et al., 2009). In order to overcome such challenges, various domain-specific information extraction systems are essential to help personalized delivery by identifying relevant information (Roitman et al., 2010).

In the present task, we have developed a Medical Recommendation System (MRS), an information extraction system that assists in recommending similar type of diseases as well as symptoms with respect to a particular symptom and disease, respectively. Therefore, we have employed two similarity matrices, disease and symptom. The disease similarity matrix contains similar diseases which have common symptoms, whereas symptom similarity matrix presents similar symptoms with respect to common diseases.

In order to develop the similarity matrices and prepare a disease-symptom relational matrix, we have employed WordNet of Medical Events (WME 3.0) (Mondal et al., 2016), a domain-specific lexicon. Thereafter, the similarity matrices and two well-known distance measurement techniques namely Manhattan and Euclidean have been used to build the proposed MRS. Additionally, we have observed the following challenges to design this recommendation system.

A. *How to identify the categories of disease and symptom for medical concepts?*

B. *How to detect the relation between diseases and symptoms?*

C. *How to frame matrices of similar diseases and symptoms?*

D. *How to recommend the disease and symptom based on the number of user-provided symptoms and diseases, individually?*

E. *How to evaluate the proposed MRS?*

In order to address these challenges, we have employed WME 3.0 lexicon and two well-known similarity measurement techniques such as Man-

[1] http://www.prmoment.in/category/pr-news/survey-shows-that-49-of-indians-use-the-internet-for-health-information

Proc. of ICON-2018, Patiala, India. December 2018 ©*2018 NLPAI*, pages 115–121

hattan and Euclidean distance. Additionally, we have prepared a disease-symptom matrix in the presence of WME 3.0 lexicon and a Healthline resource [2]. Finally, we have applied Latent Semantic Indexing (LSI) method on the disease-symptom matrix to identify the hidden relations between them.

2 Background of the Work

2.1 Medical Concepts and their Categories Assignment

The research on biomedical information extraction is demanding to extract medical concepts and their relations from the daily produced large amount of unstructured and semi-structured medical corpora. In order to present a structured corpus and extract subjective information from corpora, we have observed that the domain-specific ontologies and lexicons are essential (Borthwick et al., 1998). To this end, the standard vocabularies and ontologies, namely UMLS (Unified Medical Language System), GATE (General Architecture for Text Engineering), and SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms), and lexicons namely MEN (Medical WordNet) and WME (WordNet of Medical Event) were used by the researchers (Smith and Fellbaum, 2004; Kilgarriff and Fellbaum, 2000; Chaturvedi et al., 2017; Mondal et al., 2015, 2016).

These ontologies and lexicons help to extract the relevant information from the corpus such as medical concepts, their categories, and relations between them. Besides, the medical terms or concepts extraction from a clinical corpus is treated as an ambiguous task (Styler IV et al., 2014). A group of researchers introduced a sense selection and pruning strategy to expand the ontology in the medical domain (Widdows et al., 2006).

Eklund (Eklund, 2011) developed an annotation system to extract the relations as *diseases* for *treatments* from the scientific medical corpus. Yao, et al. (Yao et al., 2010) extracted relations such as *cures, prevents, and side effects*, which describe the distinctive nature of the biomedical text (medical papers) (Abacha and Zweigenbaum, 2011; Frunza and Inkpen, 2010). Franzen et al. (Franzén et al., 2002) have annotated Yapex corpus with 200 medical abstracts to extract the category as *proteins*. These ontologies are fundamentally looking for extracting *protein-protein* interaction and *disease-treatment* relations from corpora under a BioText project (Rosario and Hearst, 2005).

2.2 Recommendation System

Since last decades, recommendation systems are attracting in healthcare services along with the online shopping systems (Ricci et al., 2015; Adomavicius and Tuzhilin, 2005). The recommendation system provides support to extract the relevant and novel information from the corpus and increases the diversity of recommendation.

Adomavicius and Tuzhilin (Adomavicius and Tuzhilin, 2005) generalized the recommendation problem as a utility function u: C x S $->$ R, where C is the set of users and S is the set of recommendable items. u(C, S) returns a real value $>= 0$, where larger values presume a higher interest of C in S and u(C, S) = 0 presumes no interest of C in S. Initially, u is only a partially defined function, where known item ratings are given via users' profiles.

In order to build a recommendation system and compute u(C, S), primarily three approaches are usually followed such as content-based, collaborating-filtering, and hybrid. Content-based approach presents a sparse matrix according to the liking and disliking of the items of the user (Balabanović and Shoham, 1997). On the other side, the collaborating-filtering works with item-item and user-user similarity matrix based techniques with respect to the rating of the users (Cheung and Tian, 2004). The hybrid approach helps to combine the above-mentioned two approaches in an effective way for improving the accuracy (Zhang et al., 2017).

2.3 Medical Recommendation System

Recommendation is a useful technique that helps to find the relevant item for the users. Primarily, we have observed that the recommendation system is used to overcome the information overloading challenges with various type of items such as books, movies, and medical conditions namely diseases and treatments. In the present work, we have developed a Medical Recommendation System (MRS) to recommend subjective information from the textual content in healthcare services (Paruchuri, 2016).

In connection to MRS, Eysenbach and Jadad (Eysenbach and Jadad, 2001) developed a healthcare recommendation system to link the personal online accessible health records

[2] http://www.healthline.com/

with general health information from evidence-based resources. On the other hand, Roitman et al. (Roitman et al., 2010) observed the personalized recommendation, a valid approach to increase patient safety by avoiding so-called adverse drug reactions (ADR) (Wiesner and Pfeifer, 2014). We have also noticed that the content recommendation merely supplies medical information such as diseases and symptoms from the web (Agarwal et al., 2013; Wendel et al., 2013).

The earlier mentioned study motivates to build a medical recommendation system for diseases and symptoms using a content-based approach in this research.

3 Dataset Preparation

This sub-section presents, how we have prepared an experimental dataset which helps to build the proposed Medical Recommendation System (MRS). In order to start with, we collected the medical corpus from two different resources namely SemEval-2015 Task-6 [3] and MedicineNet [4]. Initially, we have converted all the acquired texts from the resources into context, which refers each sentence in a corpus. We collected 3647 number of medical contexts from both of the resources and prepared an experimental dataset with 2624 number of unique medical contexts.

Thereafter, we have applied a well-defined medical concept identification system developed by Mondal et. al. (Mondal et al., 2016) to identify medical concepts from contexts. Thereafter, to assign the categories of medical concepts, we have employed an auto-categorization technique developed by Mondal et. al. (Mondal et al., 2017). They have annotated the medical concepts into five different categories (*diseases, symptoms, drugs, human_anatomy*, and *Miscellaneous Medical Terms (MMT)*, an unspecified and undetectable category). Among all these categories, we have selected only two primary frequent categories of medical concepts such as diseases and symptoms for the current research.

On the other hand, we have used healthLine [5] resource to recognize the relationship between the assigned diseases and symptoms in a context for our experimental dataset. The relations help to prepare a disease-symptom matrix, which contains 5069 and 1124 number of unique diseases and symptoms individually. The disease-symptom matrix assists in designing disease-disease and symptom-symptom matrices to recommend similar diseases and symptoms for a particular disease and symptom, respectively. These matrices are processed through a content-based approach for building the proposed MRS system.

4 MRS Implementation

In order to implement the system, the primary required recommended information are similar diseases and symptoms according to the user-provided diseases and symptoms, individually (Mondal et al., 2018). Hence, we have prepared one disease-disease and another symptom-symptom similarity matrix from our experimental dataset. Thereafter, we have employed Manhattan and Euclidean distance techniques as a part of the content-based approach to design the MRS. MRS has been presented by two different types of recommendation systems namely RSDS (recommendation for similar diseases and symptoms) and RDS (Recommendation based on diseases and symptoms). RSDS provides the similar type of diseases and symptoms with respect to a particular disease and symptom consequently. On the other hand, RDS presents common diseases as well as symptoms for a number of symptoms and diseases supplied by users, individually. Both of the recommendation systems under MRS have been illustrated in the following subsections.

4.1 Recommendation for Similar Diseases and Symptoms (RSDS)

In order to recognize similar diseases as well as symptoms for a particular disease and symptom individually, we have developed a content-based approach with the help of Euclidean and Manhattan distance technique. Both of the techniques have been applied to the disease-symptom (Di-Sy) matrix for obtaining disease-disease (Di-Di) and symptom-symptom (Sy-Sy) matrices, respectively. Initially, the Di-Sy matrix presents relevant (score 1) and non-relevant (score 0) between a disease and a symptom. The scores have been assigned through the knowledge-based relation between them as mentioned in our experimental dataset. Unfortunately, we have observed

[3] http://alt.qcri.org/semeval2015/task6/
[4] http://www.medicinenet.com/script/main/hp.asp
[5] http://www.healthline.com/

that the scores are not assigning any partial relations between them due to versatile nature of medical concepts. Hence, we have used two different types of distance measurement techniques namely Euclidean and Manhattan to assign the fractional relations between diseases and symptoms. In the following paragraphs, we have illustrated, how Euclidean and Manhattan distances have been used to calculate the score.

Euclidean Distance: Euclidean distance refers to the straight-line distance between two points in Euclidean space (Greenacre and Primicerio, 2008). In this research, we have represented Di-Sy matrix as a Euclidean space where diseases and symptoms appear as points. Besides, we have identified similar diseases as well as symptoms based on similar symptoms and diseases respectively, which presented as a content-based recommendation system.

We have observed that the Euclidean distance does not provide an adequate accuracy due to the high dimension of disease and symptom vectors (Charulatha et al., 2013). Therefore, we have employed Manhattan distance to overcome the mentioned challenge and improve the accuracy.

Manhattan Distance: Manhattan distance function computes the distance between two items by summing up the differences of their corresponding components (Madhulatha, 2012). The Manhattan distance helps to prepare another set of Di-Di and Sy-Sy matrix to develop the proposed RSDS system.

Thereafter, we have combined Euclidean distance (ED) and Manhattan distance (MD) for both diseases as well as symptoms using equation 1 to identify the similar diseases and symptoms. We have selected a threshold value as $>$ 3.00 to recognize the similar diseases and symptoms for a provided disease and symptom under RSD.

$$Similarity_S = (w_1 * ED) + (w_2 * MD) \quad (1)$$

where w_1 = 0.8 and w_2 = 0.2 present the weight for both of the techniques, individually.

On the other hand, the following subsection describes the development steps of another type of recommendation system namely disease recommendation using various symptoms and symptom recommendation using various diseases.

4.2 Recommendation based on Diseases and Symptoms (RDS)

In case of designing recommendation systems in healthcare, we have observed that the identification of a particular symptom or disease is very difficult with respect to a specific disease or symptom, individually. Hence, we have developed a recommendation system that identifies common symptoms based diseases as well as common diseases based symptoms as suggested by a group of medical practitioners. These assumptions offer an adequate accuracy for the proposed MRS.

Thereafter, the following algorithm assists in recommending the common diseases for a particular set of symptoms and vice-versa.

Step-1: Initially, we have presented symptom vectors as SV_{Di-Sy} respect to all diseases from Di-Sy matrix.
Step-2: Take n number of input symptoms (S_i) and generate their corresponding symptom vectors (SV_i).
Step-2.1: If $SV_i \in SV_{Di-Sy}$:

$$SV_i =< a_1, ..., a_{5069} >,$$

where a refers 0 or 1.
Step-2.2: Else:

$$SV_i =< b_1, ..., b_{5069} >$$

where b presents only 0.
Step-3: Common diseases for all n number of symptoms present by CD vector.

$$CD = \bigcup_{k=1}^{n} SV_k \quad (2)$$

Step-4: CD vector helps to recommend common diseases based on the value 1.

5 Evaluation

In order to validate both the recommendation systems under MRS, we have used perplexity distribution approach on Di-Sy matrix, which has been treated as a baseline. Perplexity presents a measurement of how well a probability distribution or probability model predicts a sample in information theory. Equation 3 defines the perplexity of a discrete probability distribution (PD).

$$PD = 2^{\tilde{H}_r} \quad \text{where} \quad \tilde{H}_r = -\frac{1}{T}\log_2 p(w_1, \ldots, w_T) \quad (3)$$

where $\{w_1, \ldots, w_T\}$ is held out test data that provides the empirical distribution $q(\cdot)$ in the cross-entropy using equation 4.

$$\tilde{H} = -\sum_{x} q(x) \log p(x) \tag{4}$$

and $p(\cdot)$ is the recommended system estimated on a training set.

$$H(X) = E[-\log(p(x))] \tag{5}$$

$$\tilde{H} = -\sum_{x} q(x) \log p(x) \tag{6}$$

Perplexity provides a score of difficulty label of the prediction problem, where information entropy [6] measures the unpredictability. Equation 5 and equation 6 refer the entropy (H(X)) of random variable X for linear and discrete domain individually. These equations help to calculate the perplexity score for the different set of diseases as well as symptoms of Di-Sy matrix. Table 1 shows the distribution of perplexity scores for all sets of diseases over symptoms that are initially indicated as Di-Sy matrix and vice-versa.

# Diseases	**# Symptoms**	**Perplexity Score**
5069 (Overall)	1124	**283.50**
1-1267 (First Quarter)	1124	*107.69*
1268-2535 (Second Quarter)	1124	109.74
2536-3802 (Third Quarter)	1124	110.94
3803-5069 (Fourth Quarter)	1124	110.00
# Symptoms	**# Diseases**	**Perplexity Score**
1124 (Overall)	5069	**86.00**
1-281 (First Quarter)	5069	36.21
282-562 (Second Quarter)	5069	34.61
563-843 (Third Quarter)	5069	*33.86*
844-1124 (Fourth Quarter)	5069	36.89

Table 1: A detailed statistics of perplexity scores for various combination of diseases over symptoms and vice-versa.

		Symptoms		
		Ear-Discharge	Breast-Pain	Clubfoot
Diseases	asthma	0	0	0
	HIV	0	0	0
	Lung Cancer	1	1	1
	Pneumonia	1	1	1
	Narcolepsy	0	0	0
	SVD for LSI Score	3.039	2.183	1.414

Table 2: A sample LSI output of the disease-symptom matrix under MRS.

In addition to validate the output of the proposed MRS, we have applied Latent Semantic Indexing (LSI) method. It helps to discover the hidden relation between diseases and symptoms from the baseline Di-Sy matrix. Each disease and symptom is presented as a vector with elements corresponding to these symptoms and diseases, individually. Each element in a vector refers to the weighted association between the concepts as the category of diseases and symptoms. This method assists in describing the efficiency of the prepared Di-Sy matrix in the process of developing MRS along with Singular Value Decomposition (SVD) technique [7]. Table 2 shows a sample output of the disease - symptom matrix of MRS using the BlueBit calculator [8].

The result indicates the developed RDS provides a better prediction over RSDS due to the structure of our experimental dataset.

6 Conclusion and Future Work

In this article, we have attempted to build a medical recommendation system (MRS) using a content-based approach to better services in healthcare. Our primary motivation behind this research is to help the medical experts and non-experts to understand the domain-specific knowledge and their in-between relations. So, we have distributed the overall task into four sub-tasks such as 1) experimental dataset preparation, 2) a relational matrix building namely disease-symptom (Di-Sy), 3) development of similar diseases and symptoms recommendation system (RSDS), and 4) symptoms based diseases and diseases based symptoms recommendation system (RDS).

In order to prepare the experimental dataset and baseline matrix, we have employed WME 3.0, a domain-specific lexicon, Healthline [9] resource. Thereafter, Euclidean and Manhattan distance techniques have been applied to various disease and symptom vectors as content-based recommendation system to build both RSDS and RDS systems.

In future, we will try to improve the accuracy of the proposed MRS by enriching the experimental dataset. We will also focus on design a ranking based technique viz. collaborative filtering instated of the applied content-based to recommend the adequate output for the MRS.

[6] http://pespmc1.vub.ac.be/ENTRINFO.html

[7] http://webhome.cs.uvic.ca/ thomo/svd.pdf

[8] http://www.bluebit.gr/matrix-calculator

[9] http://www.healthline.com/

References

Asma Ben Abacha and Pierre Zweigenbaum. 2011. A hybrid approach for the extraction of semantic relations from medline abstracts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 139–150. Springer.

Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749.

Deepak Agarwal, Bee-Chung Chen, Pradheep Elango, and Raghu Ramakrishnan. 2013. Content recommendation on web portals. *Communications of the ACM*, 56(6):92–101.

Marko Balabanović and Yoav Shoham. 1997. Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Proc. of the Sixth Workshop on Very Large Corpora*, volume 182.

BS Charulatha, Paul Rodrigues, T Chitralekha, and Arun Rajaraman. 2013. A comparative study of different distance metrics that can be used in fuzzy clustering algorithms. *International Journal of Emerging Trends and Technology in Computer Science (IJETTICS)*.

Iti Chaturvedi, Edoardo Ragusa, Paolo Gastaldo, Rodolfo Zunino, and Erik Cambria. 2017. Bayesian network based extreme learning machine for subjectivity detection. *Journal of The Franklin Institute*.

Kwok-Wai Cheung and Lily F Tian. 2004. Learning user similarity and rating style for collaborative recommendation. *Information Retrieval*, 7(3-4):395–410.

Ann-Marie Eklund. 2011. Relational annotation of scientific medical corpora. In *LOUHI 2011 Third International Workshop on Health Document Text Mining and Information Analysis*, page 27.

Gunther Eysenbach and Alejandro R Jadad. 2001. Evidence-based patient choice and consumer health informatics in the internet age. *Journal of medical Internet research*, 3(2).

Shira H Fischer, Daniel David, Bradley H Crotty, Meghan Dierks, and Charles Safran. 2014. Acceptance and use of health information technology by community-dwelling elders. *International journal of medical informatics*, 83(9):624–635.

Kristofer Franzén, Gunnar Eriksson, Fredrik Olsson, Lars Asker, Per Lidén, and Joakim Cöster. 2002. Protein names and how to find them. *International journal of medical informatics*, 67(1):49–61.

Oana Frunza and Diana Inkpen. 2010. Extraction of disease-treatment semantic relations from biomedical sentences. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 91–98. Association for Computational Linguistics.

Michael Greenacre and R Primicerio. 2008. Measures of distance between samples: Euclidean. *Fundacion BBVA Publication (December 2013). ISBN*, pages 978–84.

Adam Kilgarriff and Christiane Fellbaum. 2000. Wordnet: An electronic lexical database.

T Soni Madhulatha. 2012. An overview on clustering methods. *arXiv preprint arXiv:1205.1117*.

Anupam Mondal, Erik Cambria, Dipankar Das, Amir Hussain, and Sivaji Bandyopadhyay. 2018. Relation extraction of medical concepts using categorization and sentiment analysis. *Cognitive Computation*, pages 1–16.

Anupam Mondal, Erik Cambria, Antonio Feraco, Dipankar Das, and Sivaji Bandyopadhyay. 2017. Auto-categorization of medical concepts and contexts. In *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on*, pages 1–7. IEEE.

Anupam Mondal, Iti Chaturvedi, Dipankar Das, Rajiv Bajpai, and Sivaji Bandyopadhyay. 2015. Lexical resource for medical events: A polarity based approach. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1302–1309. IEEE.

Anupam Mondal, Dipankar Das, Erik Cambria, and Sivaji Bandyopadhyay. 2016. Wme: Sense, polarity and affinity based concept resource for medical events. *Proceedings of the Eighth Global WordNet Conference*, pages 242–246.

Venkata A Paruchuri. 2016. Med-hyrec: A recommendation system for medical domain. In *Information Systems Design and Intelligent Applications*, pages 605–611. Springer.

Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B Kantor. 2015. *Recommender systems handbook*. Springer.

Haggai Roitman, Yossi Messika, Yevgenia Tsimerman, and Yonatan Maman. 2010. Increasing patient safety using explanation-driven personalized content recommendation. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 430–434. ACM.

Barbara Rosario and Marti A Hearst. 2005. Multi-way relation classification: application to protein-protein interactions. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 732–739. Association for Computational Linguistics.

Barry Smith and Christiane Fellbaum. 2004. Medical wordnet: a new methodology for the construction and validation of information resources for consumer health. In *Proceedings of the 20th international conference on Computational Linguistics*, page 371. Association for Computational Linguistics.

Kathrin Sommerhalder, Andrea Abraham, Maria Caiata Zufferey, Jürgen Barth, and Thomas Abel. 2009. Internet information and medical consultations: experiences from patients and physicians perspectives. *Patient education and counseling*, 77(2):266–271.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Sonja Wendel, Benedict GC Dellaert, Amber Ronteltap, and Hans CM van Trijp. 2013. Consumers intention to use health recommendation systems to receive personalized nutrition advice. *BMC health services research*, 13(1):126.

Dominic Widdows, Adil Toumouh, Beate Dorow, Ahmed Lehireche, et al. 2006. Ongoing developments in automatically adapting lexical resources to the biomedical domain. In *Fifth International Conference on Language Resources and Evaluation, LREC, Genoa, Italy*.

Martin Wiesner and Daniel Pfeifer. 2014. Health recommender systems: concepts, requirements, technical basics and challenges. *International journal of environmental research and public health*, 11(3):2580–2607.

Lin Yao, Cheng-Jie Sun, Xiao-Long Wang, and Xuan Wang. 2010. Relationship extraction from biomedical literature using maximum entropy based on rich features. In *2010 International Conference on Machine Learning and Cybernetics*, volume 6, pages 3358–3361. IEEE.

Yin Zhang, Min Chen, Dijiang Huang, Di Wu, and Yong Li. 2017. idoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization. *Future Generation Computer Systems*, 66:30–35.