2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
Banff Center, Banff, Canada, October 5-8, 2017

CSnet: Constructing Symptom Network based on Disease-Symptom Relationships

Sohee Hwang
Department of Computer Science,
Yonsei University
Seoul, Korea
soheeee2@yonsei.ac.kr

Jungrim Kim
Department of Computer Science,
Yonsei University
Seoul, Korea
kimgogo02@yonsei.ac.kr

Jeongwoo Kim
Department of Computer Science,
Yonsei University
Seoul, Korea
jwkim2013@yonsei.ac.kr

Sanghyun Park*
Department of Computer Science,
Yonsei University
Seoul, Korea
sanghyun@yonsei.ac.kr

Abstract— **A symptom is the physical indication of an unstable state or the beginning of diseases. Symptom analysis is an essential factor in the medical area, where it is used for disease diagnosis, drug prescription, and the development of new pharmaceuticals. Commensurate with its importance, symptom analysis has been the subject of various studies in recent years. However, prior literature on this topic has been largely limited to studying symptoms for a specific disease. Our paper attempts to expand and build on previous studies by introducing a network-based symptom analysis. Symptom analysis that can provide a basis for analyzing symptoms related to various diseases.**

For a universal symptom analysis system, we proposed a network-based symptom analysis. In order to construct a symptom network, we utilized Medical Subject Heading (MeSH) terms and the PubMed search engine which are maintained and developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM). We identified symptom-disease relationships with two measurements, the term frequency-inverse document frequency (TF-IDF) and frequent occurrence of two terms (co-occurrence) from PubMed articles. Symptom-symptom pairs, which is the outline for symptom network, were built up based on symptom-disease relationships. As a result, we constructed a symptom network with 223 nodes and 5313 edges. Evaluations were performed in two ways, compared with two symptom clusters and demonstrated with previous researches. Additionally, proposed method has shown possibility for a guideline of clinical demonstration and a discovery of potential symptoms pair.

Keywords—symptom analysis, text mining, bioinformatics, network-based analysis

I. INTRODUCTION

The rapid growth of data mining has brought about fundamental changes in modern society. We now realize that the problem is not the lack of information, but rather the excess of it. A natural consequence is the issue of dealing with such abundant data, one such example being biological data. Various studies have produced valuable results from analyzing biological data, which contributed to human life. Bioinformatics is the one major part of biotechnology revolution which extracted valuable information from analyzing biological data.

One important subject of bioinformatics is analyzing medical data. As humans, diseases can result in various symptoms, such as pain, physical changes, or abnormal conditions. Symptoms are subjective factors, noticed by the patient, that demonstrate physical cues of an unstable state or disease. Symptom analysis comprises a large proportion of physical examination such as disease diagnosis, drug prescription etc. Accurate defining symptoms and correct understanding symptoms are essential to reduce problems related to physical examination. Erroneous diagnosis will occur frequently without symptom analysis. As shown in Figure 1, the number of publications about symptom analysis has increased over the past 8 years, illustrating the increased effort to analyze and address the importance of symptom analysis.

Substantial research has focused on symptom analysis and symptom based diagnosis and treatment. Clustering methods, agglomerative hierarchical method and exploratory factor analysis have been performed to analyze symptoms for advanced cancer. [18] In addition, three identified clusters were extracted by the aforementioned statistical technique. [7] Although symptom clustering is in its early stages, it is a very promising field of analysis.

Network-based approaches are one of the most rapidly increasing methods of biological analysis. Human disease network are constructed by a network based on the outcome of the symptom-based similarity of two diseases [25], [17]. Furthermore, PPI network is the primary example of a network based analysis. Several network-based approaches have focused on predicting novel targets and new uses for existing drugs. [2],[12]

*Corresponding author. Tel.: +82 2 2123 5714; fax +82 2 365 2579; E-mail address: sanghyun@yonsei.ac.kr).

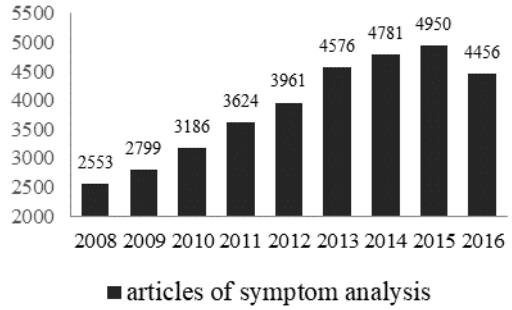

Fig. 1. The number of articles on symptom analysis over 8 years

The objective of this study is to build the basis for a systematic computational approach that presents the available symptom relationships. Additionally, this study is available to apply guideline for clinical demonstration and detected new symptom pairs. There are some prior literature works on symptom clustering which cover only one specific disease, rather than diverse diseases. Network based symptom analysis is valuable since symptom analysis is a promising field with growing interest, and network based analysis is an effective way of conducting such analysis. Furthermore, there is no formal research utilizing network based examination, which makes this paper unique.

The paper is organized as follows. In section 2, we present the datasets collected from MeSH and PubMed—a reliable dataset from NLM—and show identifying symptom-symptom relationships. In section 3, we evaluate the performance of our suggested method. In the last section, we conclude with the contribution of our research and propose potential directions for future work.

II. METHODS

In this part, we present our methodology for constructing symptom networks and examining the results. Our method proceeds as follows: 1) Collect required datasets, 2) Obtain symptom and disease relationships, 3) Extract symptom and symptom correlation, 4) Construct a symptom network. An overview of our method is illustrated in Figure 2. For step 1, we investigated relations between symptoms and diseases utilizing MeSH(Medical Subject Heading) terms and PubMed. In step 2, we described the process of extracting symptom-symptom pairs. In the last step, we constructed the symptom network which is the ultimate goal of our research.

A. Collecting required datasets

The construction of a symptom network requires two basic datasets, 1) A basic vocabulary of symptoms and diseases and 2) A corpus of data from which their relationships are extracted. We utilized the PubMed literature database for extracting relations and Medical Subject Heading(MeSH) for vocabulary data. PubMed provides references and abstracts about biomedical topics and life sciences. The United States National Library of Medicine(NLM) at the National Institutes of Health manages the database. MeSH is a contains a comprehensive set of vocabulary about life sciences, and serves to index articles and books which are organized in a hierarchical tree. As the MeSH is managed by NLM and the academic papers are consisted of literature, so we can safely conclude that MeSH and PubMed are reliable sources of information. MeSH data is one major resource for researchers who study biomedical text mining. 322 instances of "Symptoms and Sign terms" and 4442 instances of "disease terms" were collected from MeSH. Combinations of symptom and disease terms were utilized primarily.

B. Obtaining symptom and disease relationships

The key point of constructing a symptom-network is building relationships between diseases and symptoms. To figure out symptom-symptom relationships, we identified disease and symptom relationships first. There were limited research articles for directly detecting symptom-symptom relationship. Most articles that included symptom term pairs showed only enumerated symptom terms, instead of

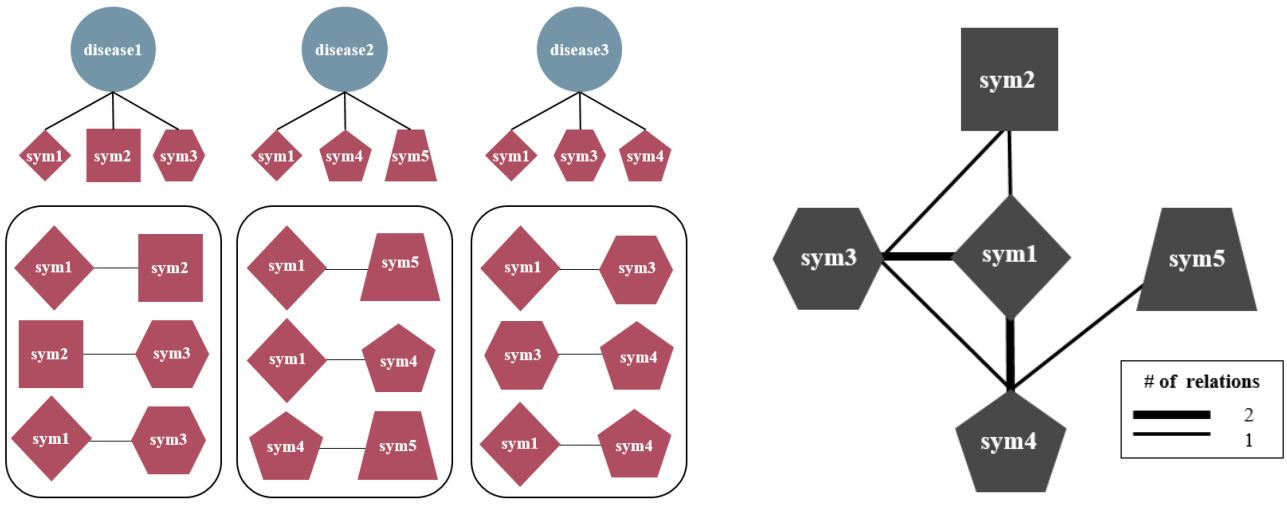

Fig. 2. Overview of construction symptom network

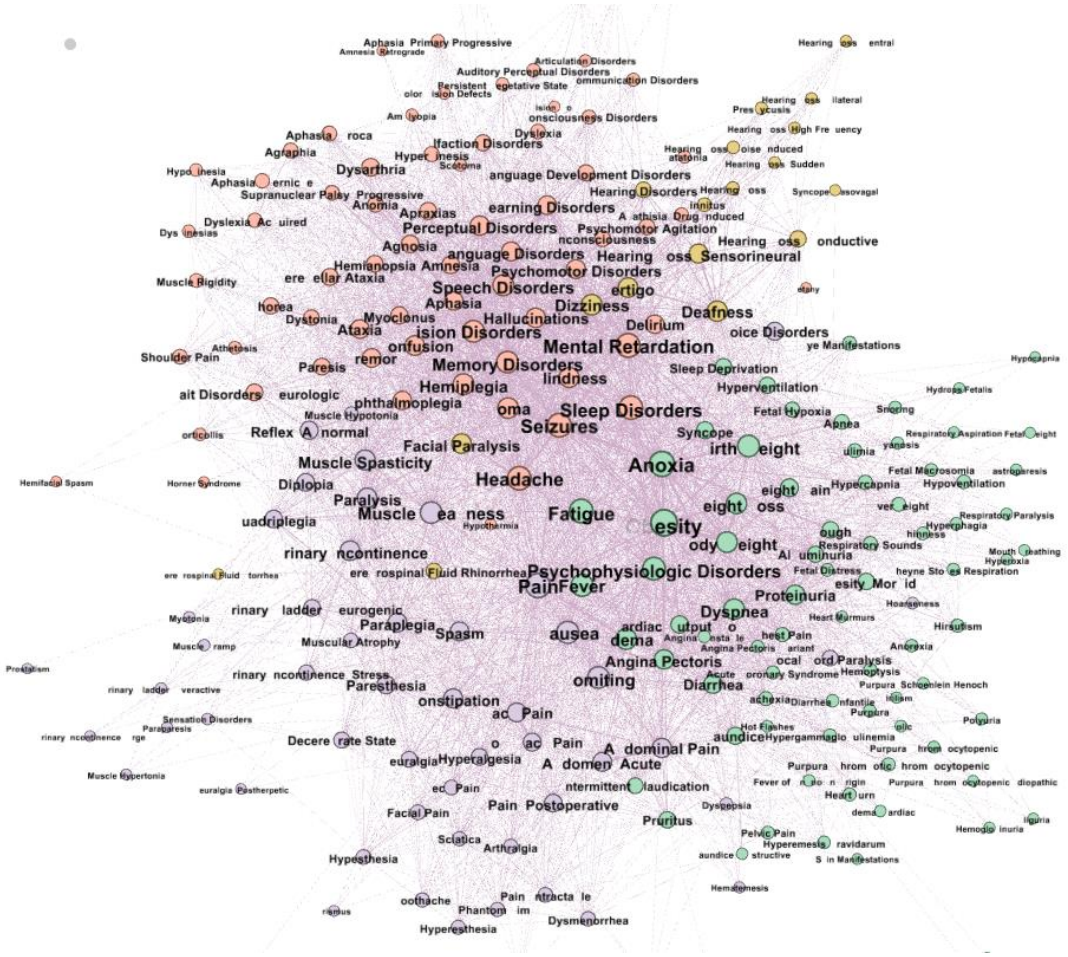

Fig. 3. Symptom network (the color of nodes indicated the clusters)

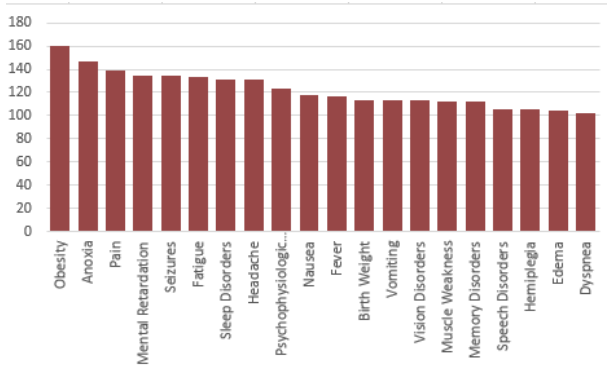

Fig. 4. The top twenty term frequency of symptom terms

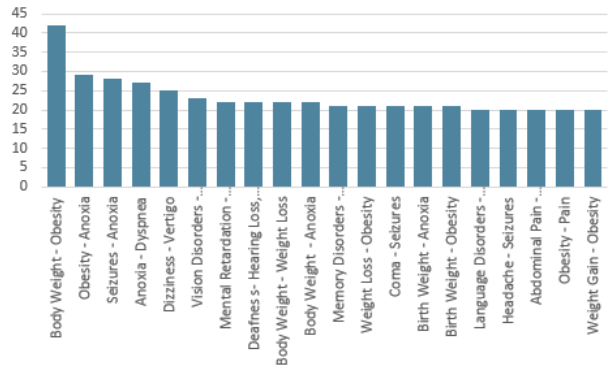

Fig. 5. The top twenty relations of symptom terms

mentioning them in a relevant context. Therefore, in order to extract precise symptom correlation, we utilized disease and symptom terms. The diseases and symptoms have a reciprocal effect. For instance, a patient who has gastro-intestinal trouble, will likely suffer from severe symptoms such as vomiting, diarrhea and dyspepsia which are related to the stomach. On the other hand, if the patient suffered from such symptoms related with stomach, it is very likely that gastro-intestinal disorder would have caused the symptoms. Therefore, we can assume that symptoms provide meaningful information about the disease causing it, and vice versa, to some extent.

In PubMed, each article is associated to metadata that includes a list of keywords describing core topics of a certain article. To acquire the association between diseases and symptoms, we applied metadata containing keyword lists PubMed and adopted two measurements as term co-occurrences and TF-IDF to describe relevance degree. Term co-occurrence is a measurement of text-mining, which presents the number of times specific terms appear together.

Additionally TF-IDF is the numerical statistic that reflects importance of a term is in a document. For detail, TF represents the importance of a word in a document and IDF represents the importance of a word in a set of documents. The equation of TF-IDF is shown in equation(1).

$$Tf\text{-}idf(t,d,D)=tf(t,d)idf(t,D) \quad (1)$$

Where $tf(t,d)$ represents the number of times term t appeared in document d, and $idf(t,D)$ is the value that we get by dividing the total number of documents by the number of documents containing the term.

Using these two measurements, we developed a Java program to acquire PubMed identifiers whose keywords contained any of the symptom or disease terms according to MeSH.

C. Extracting symptom and symptom correlation

To extract symptom and symptom relationship, we identified symptom-disease co-occurrence first. It means particular diseases are associated with specific symptoms and high co-occurrence value implies a strong association. Furthermore, symptoms, which are connected with identical diseases, were inferred from a connection between symptoms. In the case of flu, frequent symptoms were cough, fever, sore throat and a runny nose. We can therefore extract four disease-symptom pairs (e.g. flu-cough, flu-fever, flu-sore throat, flu-runny nose). Using such proposed inference, we can get symptom-symptom pairs(e.g. cough-fever, fever-sore throat, sore throat-runny nose etc.). The weight of edges between symptoms is calculated by the number of shared diseases. In the case of fever and sore throat, two symptoms share the flu as a disease. Therefore, the weight is counted as 1. This method is illustrated in Figure 2.

D. Fitering significant symptom-symptom relations

Gephi is an open-source software written in java which is widely used for network analysis and visualization. It was employed for constructing symptom networks. The full symptom network was too dense with all possible pair links being shown, therefore we filtered some pairs using thresholds to avoid the dense problem. We set 3 thresholds, which limited the values of co-occurrence, TF-IDF and degree of network.

- $Co-occurence \leq \sum_{i=0}^{n} \frac{ith\ co-occurrence}{n}$ (i= order of symptom pairs)
- $TF-IDF \leq \sum_{i=0}^{n} \frac{ith\ TF-IDF}{n}$
- $edge-degree < \sum_{i=0}^{n} \frac{ith\ edge-dgree}{n}$

III. Results

A. Symptom-Symptom networks

147,978 disease-symptom relationships were extracted with 4,210 disease terms and 322 symptom terms. It covers 98.5% of the symptoms and 95% of the diseases extracted from the total terms from MeSH. Using the method previously illustrated, we calculated co-occurrence and the TF-IDF value of relationships. To construct networks, we set co-occurrence threshold as 11 based on an average of total co-occurrences (11.887675). Likewise, the threshold of TF-IDF was set as 16 on account of average of total TF-IDF values (16.673). Total symptom relations were 21,788. Consequently, 223 nodes and 5313 edges were constructed by the symptom network. The most frequency terms in the symptom network were Body Weight and Anoxia with 160 and 147 instances respectively. Furthermore, the strongest connection of symptom-symptom relations is Body Weight-Obesity, the number of presented diseases is 42. The details of these results are shown in Figure 4, 5.

TABLE I. DATASETS

Table Column Head		
data	***Total***	
Disease-symptom relation	*147,978*	
Symptom-symptom relation	*21,788*	
Threshold		*16,475*
Symtpom network	**Nodes**	**edges**
	223	5313

B. Performance evaluations of symptom networks

Two experiments and one demonstration were performed to validate our approach. There was no previous research done on symptom networks, therefore the best performance evaluation of the research was compared with the same subject. Therefore, we adopted clustering method on behalf of comparing same subject. To further test the reliability of the obtained symptom connections, we created 4 clusters from the obtained symptom networks based on edge weight. From the proposed network, clusters of symptoms were constructed based on the edge weight. Symptoms are divided into 4 clusters (C_0, C_1, C_2, C_3), which contain 36.68% , 29.07%, 20.42%, 13.84% of symptoms respectively. We computed the probability of symptoms in each cluster being included in the same cluster extracted by the symptom network. However, since the terminology of both studies was different, we pre-processed the terminology. Therefore, 24% of terms were filtered out while 76% were covered.

1) Symptom clustering

To evaluate the proposed symptom network, we studied various literature related to symptom analysis, and the most relevant studies include: symptom clustering in advanced cancer. [5] In this paper, they identified symptoms that occur simultaneously using data from patients with advanced cancer. However, this research is only limited to one type of disease.

The 7 symptom clusters of advanced cancer are:

- The fatigue cluster: easy fatigue, weakness, anorexia, lack of energy, dry mouth, early satiety, weight loss, taste changes
- The neuropsychological cluster: sleep problems, depression, anxiety

- The upper gastrointestinal cluster: dizzy spells, dyspepsia, belching, bloating
- The nausea and vomiting cluster: nausea, vomiting
- The aerodigestive cluster: dysphagia, dyspnea, cough, hoarseness
- The debility cluster: edema, confusion
- The pain cluster: pain, constipation

For computing the probability of symptoms in each cluster being included in the same cluster extracted by the symptom network, , the pb_value is calculated by equation (1).

$$\text{pb\_value} = \sum_{i=1}^{7} \sum_{j=1}^{n} \frac{\max_{0 \le x \le 3} P(x_{ij})}{7} \quad (2)$$

x_{ij} is the cluster number of symptom network that includes the j^{th} symptom from symptom network whose cluster number is i of previous research. Therefore $P(x_{ij})$ shows the probability of the cluster number. The pb_value is 0.7054, which presents that over 70% of symptoms are divided significantly based on previous research. A result indicated high portion of symptom pairs were precise division based on edge weight.

2) *MeSH hierarchical subjects*

Descriptions of MeSH terms are arranged in a hierarchy. These terms are allocated in the same level if the terms meet the conditions (related descriptors, list of synonyms or very similar terms). The tree locations carry systematic labels known as tree numbers, and consequently one descriptor can carry several tree number. For instance, the descriptor "Sign and Symptoms" has a tree number of C23.888. Likewise, every descriptor carries a unique alphanumerical ID such as C23.888.144.699, which is the ID of "Over Weight" as shown in Figure 6.

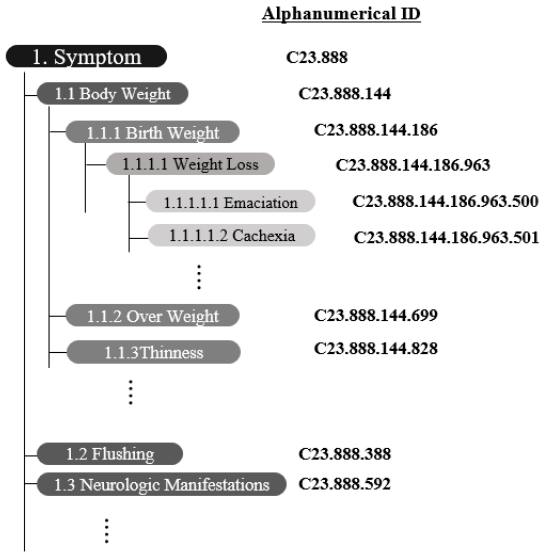

Fig. 6. Hierarchical structure of MeSH term (Sign and Symptom)

To further evaluate the reliability of the obtained symptom connection, we utilized MeSH hierarchical structure with hierarchical clustering. Hierarchical clustering is a method of cluster analysis performed two ways, agglomerative(bottom up) and divisive(top down). The performance evaluations were performed in agglomerative hierarchical clustering. In the case of Sign and Symptoms tree, the first level contained 384 nodes of symptoms, each symptom terms identified by ID. For example, ID of first level has formed as C23.888.xxx, and second level has C23.888.xxx.xxx. To handle more symptom terms, we picked the first level and the second level clustering which contain all symptom terms and 92.96% of the symptom terms respectively. For examining the accuracy of network clustering, we used equal verification method in previous evaluations.

$$\text{pb\_value} = \sum_{i=1}^{c\_num} \sum_{j=1}^{n} \frac{\max_{0 \le x \le 3} P(x_{ij})}{c\_num} \quad (3)$$

c_num is cluster number from previous research. The pb_value of the first level was 0.93 which covered almost all the symptoms terms. This indicates that clusters from the symptom network had 93% accuracy based on the MeSH terms, while the pb_value of the second level, which covered 90.36% of symptom terms, was 0.79. According to observed value, it proved that symptom clusters extracted from symptom network has high accuracy of divisions according to MeSH data which is reliable dictionary of terms.

3) *performed clinical demonstration*

To demonstrate effectiveness of symptom network as guideline for clinical demonstration, we investigated previous researches of symptom pairs. In practice, 55% of top 20 pairs were studied from using clinical test. We handled one symptom pair as representative, Seizure-Anoxia is the third top relationships of symptom network. There are 15,600 articles associated with Seizure-Anoxia relationship, one of articles studied that seizures were predominately seen in patients who with acute anoxic episodes receiving resuscitation[16], and another article presented anoxia could produce dysfunction of the developing brain that includes increased sensitivity to seizure with rat experiment.[18] Likewise these researches, there are various clinical test to determined Seizure-Anoxia relationships. And From previous researches, it is clear that symptom pairs from symptom network were have been studied. Symptom network is available to utilize as not only a guideline for clinical test, but also detected new symptom relationships.

IV. Conclusion

Symptom-analysis delivers a more practical and realistic guideline to the clinical problem, symptom analysis will remain an intricate topic because most symptom experiences occur with confounding factors. While most research works

dealt with symptoms of only one disease category, our research covered all disease categories.

The main goal of this research is building a basis for a systematic computational approach that which presents the available symptom relationship. In this work, a symptom-network was constructed based on disease-symptom relationship for a reliable outcome. We obtained useful information from MeSH and PubMed, which are frequently used by researchers who are interested in biomedical text mining. Even though most articles emphasized the importance of symptom analysis, there has been no study focusing on a network-based approach. This work provides a useful tool to design a more precise network or to evaluate symptom analysis. Additionally, this work is available to apply guideline for clinical demonstration and detected new symptom pairs.

ACKNOWLEDGMENT

This research was supported by the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT & Future Planning (NRF-2015M3C4A7065522)

REFERENCES

[1] Barabási, Albert-László, Natali Gulbahce, and Joseph Loscalzo. "Network medicine: a network-based approach to human disease." Nature Reviews Genetics 12.1 2011, pp.56-68.

[2] Berger, Seth I., and Ravi Iyengar. "Network analyses in systems pharmacology." Bioinformatics 25.19 2009, pp.2466-2472.

[3] Bredenoord, A. J., B. L. A. M. Weusten, and A. J. P. M. Smout. "Symptom association analysis in ambulatory gastro-oesophageal reflux monitoring." Gut 54.12 2005, pp.1810-1817.

[4] Cha, Junbum, Jeongwoo Kim, and Sanghyun Park. "GRiD: Gathering rich data from PubMed using one-class SVM." Systems, Man, and Cybernetics SMC, 2016 IEEE International Conference on. IEEE, 2016.

[5] Chen, Mei-Ling, and Ho-Ching Tseng. "Symptom clusters in cancer patients." Supportive Care in Cancer 14.8 2006, pp.825-830.

[6] Dodd MJ, Miaskowski C, Paul SM. Symptom clusters and their effect on the functional status of patients with cancer. Oncol Nurs Forum 2001;28:465–70.

[7] Fan, Grace, L. Filipczak, and E. Chow. "Symptom clusters in cancer patients: a review of the literature." Curr Oncol 14.5 2007, pp.173-179.

[8] Madison, D., and E. Niedermeyer. "Epileptic seizures resulting from acute cerebral anoxia." Journal of Neurology, Neurosurgery & Psychiatry 33.3 1970, pp.381-386.

[9] National library of medicine. [Online]. Availablepp.https://www.ncbi.nlm.nih.gov/mesh

[10] Shimomura, Chieko, and Hisashi Ohta. "Behavioral abnormalities and seizure susceptibility in rat after neonatal anoxia." Brain and Development 10.3 1988, pp.160-163.

[11] Sun, Peng Gang. "The human drug–disease–gene network." Information Sciences 306 2015, pp.70-80.

[12] Wu, Zikai, Yong Wang, and Luonan Chen. "Network-based drug repositioning." Molecular BioSystems 9.6 2013, pp.1268-1281.

[13] Bredenoord, A. J., B. L. A. M. Weusten, and A. J. P. M. Smout. "Symptom association analysis in ambulatory gastro-oesophageal reflux monitoring." Gut 54.12 2005, pp.1810-1817.

[14] De Jong, Hidde. "Modeling and simulation of genetic regulatory systems: a literature review." Journal of computational biology 9.1 (2002): 67-103.

[15] Hopkins, Andrew L. "Network pharmacology: the next paradigm in drug discovery." Nature chemical biology 4.11 (2008): 682-690.

[16] National library of medicine. [Online]. Availablepp.https://www.nlm.nih.gov/

[17] Sharma, Amitabh, Amir Bashan, and Alber-Laszlo Barabasi. "Network Approach to Disease Diagnosis." APS Meeting Abstracts. Vol. 1. 2014.

[18] Tsai, Jaw-Shiun, et al. "Significance of symptom clustering in palliative care of advanced cancer patients." Journal of pain and symptom management 39.4 2010, pp.655-662.

[19] Walsh, Declan, and Lisa Rybicki. "Symptom clustering in advanced cancer." Supportive Care in Cancer 14.8 2006, pp.831-836.

[20] Wellman, Barry, and Stephen D. Berkowitz. Social structures: A network approach. Vol. 2. CUP Archive, 1988.

[21] Weusten, Bas LAM, et al. "The symptom-association probability: an improved method for symptom analysis of 24-hour esophageal pH data." Gastroenterology 107.6 1994, pp.1741-1745.

[22] Xu, Xue, et al. "Drug-symptom networkingpp.Linking drug-likeness screening to drug discovery." Pharmacological research 103 2016, pp.105-113.

[23] Yates, Andrew, Nazli Goharian, and Ophir Frieder. "Learning the Relationships between Drug, Symptom, and Medical Condition Mentions in Social Media." ICWSM. 2016.

[24] Yildirim, Muhammed A., et al. "Drug--target network." Nature biotechnology 25.10 2007, pp.1119.

[25] Zhou, XueZhong, et al. "Human symptoms–disease network." Nature communications 5 2014.

[26]