

# Disease Detection and Symptom Tracking by Retrieving Information from the Web

Lun-Wei Ku\*, Wan-Lun Li†, Ting-Chih Chang†

Institute of Information Science

Academia Sinica\*

lwku@iis.sinica.edu.tw

Department of Computer Science and Information Engineering

National Yunlin University of Science and Technology†

{u9817022, u9817024}@yuntech.edu.tw

## Abstract

This paper proposes techniques for preliminary disease detection and personal symptom tracking adopting concepts and methods of web information retrieval. The proposed approaches are inspired by web users' behavior. People look for information of symptoms from Internet. Therefore, considering information in Web pages, the developed system proposes possible diseases related to one or more queried symptoms. Moreover, these queried symptoms would be recorded in the query log so that the user could utilize these records to trace the history of symptoms, further to manage their own health or provide them to doctors as reference. As ranking detected diseases needs professional knowledge, we instead evaluate relevancy of retrieved sentences containing detected diseases in both strict and lenient metrics. Experimental results support the proposed ranking approach. The techniques described in this paper are also implemented to develop an Android application called "Health Generation". In this application, the detected disease is further linked to its Wikipedia introduction and the nearby clinics are listed. Users can utilize the GPS function provided by cell phones to plan the route for them. Through the proposed approaches and the application to provide medical information and solutions according to users' need and further to help users manage their health is the aim of this research.

## Introduction

As the Internet has been a basic infrastructure, and web surfing devices like computers or mobiles are getting popular, searching information from the Web before acting has become a common sense. At the same time, health care draws attention in the society; self-care, the basic level of health care, involves self-evaluation of symptoms and the

later self-decision regarding reactions to disease, now also commonly with the help of information from the Internet (Hesse et al. 2005). When people find some symptoms appearing in their bodies, they may not be able to go to see a doctor immediately due to the lack of time, medical information (not knowing where or who they should go for help), or simply not willing to go in some culture. Therefore, there are getting more and more medical sites and discussion forums on the Web, and governments in most developed countries also provide web sites to release the newest medical knowledge and policies for education purposes. From them rich medical information is available and could be utilized by all kinds of medical systems (Dickerson et al. 2004).

In the past, the researchers in the bio-informatics society tried to extract name of diseases (Bernhard 2006) and genes (Chun et al. 2006). They mined relations between diseases and genes, drugs, or symptoms but from literature or some specific database such as PubMed<sup>1</sup>. The mined information was usually for medical staff or professionals instead of general users. Comparing to searching from the Internet, its quality may be higher but less information is provided or even lack of information. In addition, there is no experience sharing, which may not fit users' need. Most important of all, the mining process does not conform to the general users' behavior pattern.

Several hospitals do provide web pages for their patients to look up symptoms and the corresponding diseases<sup>2</sup>. However, their systems are usually not in a freely input

<sup>1</sup> <http://www.ncbi.nlm.nih.gov/pubmed/>

<sup>2</sup> [http://www.vghtc.gov.tw/portal/m2/portalhome/home?tag=2\\_9\\_1](http://www.vghtc.gov.tw/portal/m2/portalhome/home?tag=2_9_1)

<http://health.sohu.com/medisearch.html>

<http://www.tmn.idv.tw/chiaungo/symptom/index.html>

[http://hlm.tzuchi.com.tw/index.php?option=com\\_content&view=categor&layout=blog&id=172&Itemid=513&lang=zh](http://hlm.tzuchi.com.tw/index.php?option=com_content&view=categor&layout=blog&id=172&Itemid=513&lang=zh)

style so that the users cannot search for arbitrary symptoms or multiple symptoms to find the most related disease. The update of their information also relies on the system administrator.

In this research, we proposed an approach to find the most possible diseases according to the symptoms queried by the users. Two parts of information retrieval techniques were utilized: one is to detect diseases from retrieved information and the other is to track symptoms from the search log (Salton and McGill 1986). Web articles containing both symptoms and diseases were retrieved and processed, and then possible relevant diseases were listed to save the time needed for searching and reading these articles. Moreover, the introduction of the detected diseases extracted from Wikipedia was provided to the users and the nearby related clinics were listed. The search log was recorded so that the users could provide the symptoms they have had/searched recently to the doctors. By developing the techniques and application, we aim to provide a convenient way for users to manage their health and find the proper medical help. Experimental results for proposing diseases were evaluated and possible ways to enhance the performance were discussed in later sections.

Concealing the malady for fear of taking medicine is common concept in the traditional Chinese culture. Therefore, experiments were performed on Chinese articles and the application interface was in Chinese, too. However, the proposed approach can also be applied to materials of other languages as there is no language specific process in it.

## Materials

### Resources

In the very beginning, we tried to utilize a Chinese medical dictionary to collect the names of diseases. However, the dictionary which we found included not only names of diseases but also other medical terms, which confused the program in the search process. Therefore, we decided to obtain names of diseases from Wikipedia. A total of 295 names of diseases were collected at the time of doing experiments. An advantage of getting information from Wikipedia is that we are able to know the alias of diseases which could help increase the recall rate when searching for diseases related to a certain symptom.

### Experimental Materials

The experimental materials are web articles retrieved by Yahoo! search engine using symptoms as queries. Though it is possible to consider all retrieved articles of each query, for efficiency, only top 20 links are further utilized to find the related diseases. A total of 25 symptoms were selected

for evaluation, that is, 500 links were utilized. However, some of these links were malformed, lack of contents, or no longer available. At last we had 299 articles and within them names of diseases were found in 1,473 sentences. These 25 symptoms for experiments include (1) itchy throat, (2) backache, (3) coughing pus, (4) fear of cold, (5) thirsty, (6) anorexia, (7) no appetite, (8) severe cough, (9) fever, (10) dizziness, (11) ear pain, (12) frequent micturition, (13) hemoptysis (coughing blood), (14) stomachache, (15) cold sweat, (16) shivering/trembling, (17) dyspnea, (18) chest tightness, (19) rhinocnesmus, (20) nausea, (21) solar ardor/heart burn, (22) cold hand and feet, (23) terry stool, (24) halitosis/bad breath, and (25) arthralgia.

## Disease Search

As mentioned, the proposed approach imitated users' behavior of looking for medical information from the Web. The basic concept was illustrated in Figure 1. When there are multiple queried symptoms, i.e., symptom  $i$ ,  $j$ ,  $k$ , we want to find the disease  $d$  which will cause most of them. We also need a mechanism to decide which diseases among  $d$ ,  $d_1$ ,  $d_2$ , and  $d_3$  we should propose. It now becomes a ranking problem (Geng et al. 2012), and two major issues should be considered:

- (1) A disease covers more symptoms should be ranked higher.
- (2) A disease which is more relevant to the symptoms should be ranked higher.

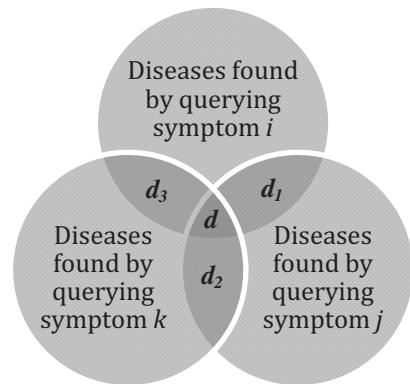


Figure 1. Syndromes and Corresponding Detected Diseases

Now we define the problem and propose our approach. Suppose there is a set of queries  $Q$ , which include  $n$  symptoms the user probably have now. First, for each symptom  $q_i$  in  $Q$ , it will be submitted to the search engine. Second, top 20 retrieved articles will be downloaded, and the number of sentences containing any disease  $d_j$  found in

Wikipedia is counted and denoted as  $N_{q_i, d_j}$ . These diseases found by querying  $q_i$  are sorted by  $N_{q_i, d_j}$  in a descending order. That is, in the disease list  $D_{q_i}$ ,  $d_j$  was ranked higher if  $N_{q_i, d_j}$  was larger. To find the diseases to report, *DCG* (Discounted Cumulative Gain, Jarvelin and Kekalainen 2002), the scoring function for ranking, widely used to evaluate the performance of information retrieval results, was adopted to calculate the score of each detected disease. The formula of *DCG* for information retrieval is as follows:

$$DCG_p = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2 i} \quad (1)$$

where  $p$  is a particular ranked position and  $rel_i$  is the relevant score of a document  $i$ . In our approach, relevant diseases were proposed instead of relevant documents. Therefore, the formula of medical *DCG* (*medDCG*) for each disease  $d_j$  in disease list  $D$  is defined in formula (2).

$$medDCG_{d_j} = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2 i} \\ rel_i = \begin{cases} 1 & \text{for } disease_i = d_j \\ 0 & \text{for } disease_i \neq d_j \end{cases} \quad (2)$$

where  $p$  is the number of detected diseases, and  $disease_i$  is the detected disease reported in the ranked position  $i$  in the disease list  $D$ . As mentioned, we had one disease list  $D_{q_i}$  for each symptom  $q_i$ , so we can calculate the  $medDCG_{d_j}$  for each  $q_i$ , denoted as  $medDCG_{d_j, q_i}$ . The final score  $S_{d_j}$  for disease  $d_j$  was calculated by formula (3):

$$S_{d_j} = \sum_{q_i} medDCG_{d_j, q_i} \quad (3)$$

The maximum score of a candidate disease is equal to the number of queried symptoms, and the reported diseases were the top three with maximum scores. Ranks of diseases with the same scores were determined by their ranks in the first disease list containing them. If they are still the same, the query symptom index of these first disease lists will determine their ranks. This is because we postulate that users usually input the major symptoms before the minor ones. For example, assume we have  $d_i$  and  $d_j$  where  $S_{d_j}$  equals  $S_{d_i}$ , and for query index 1 to  $n$ , we first detect  $d_i$  in  $D_{q_k}$  with rank  $p$  and  $d_j$  in  $D_{q_m}$  with rank  $p'$ , we will compare  $p$  and  $p'$  first, if they are still the same, we compare  $k$  and  $m$  to determine the final ranks of  $d_i$  and  $d_j$ .

## Android Application Illustration

With the proposed approach to detect diseases, we further developed an application “Health Generation” in the android platform so that users can use it with cell phones. Figure 2 illustrates its functions and interfaces.



Figure 2. Illustration of Health Generation

This application was designed for the following scenario:

- (1) Users feel ill but don't know why. Users feel ill but cannot go to see a doctor immediately. Users feel ill but don't know where to see a doctor.
- (2) Users look for information of disease, ways to ease the pain, and the location of the clinics.
- (3) Users go to see the doctor when available and describe the symptoms in the past few days.

According to this scenario, Health Generation provided the following functions.

- (1) Search related diseases according to the queried symptoms
- (2) Provide the related articles and the Wiki introduction of the detected diseases.
- (3) List the information of clinics (address, phone number, and distance) which provide the medical help for the detected diseases in the neighborhood. Users can further utilize the GPS to plan route directly.

## Evaluation and Discussions

It is always difficult to evaluate medical systems because of the need for knowledgeable persons. In this research, if we want to know whether the proposed detected diseases are appropriate, their relatedness and ranks should be evaluated and only professional doctors are competent to do it. However, we had only non-medical background annotators, i.e., college students. Therefore, two evaluations were made in this research instead. In the first evaluation (strict metric hereafter), we asked annotators to label whether the retrieved sentence described the relation of detected disease and the queried symptom; in the second evaluation (lenient metric), we asked annotators to label whether the detected disease in the retrieved sentence was related to the queried symptom according to their prior knowledge. From these two evaluations, we hope to know the relations between symptoms and the detected diseases, and the characteristics of different symptoms. For deeper analyses, we also asked annotators to label the degree of irrelevance: some irrelevant (scoring 1), irrelevant (scoring 2), mostly irrelevant (scoring 3) and totally irrelevant (scoring 4). The results are shown in Table 1 and Table 2. Table 1 and Table 2 show the number of sentences containing names of diseases (Num S), number of sentences containing relevant diseases (Rel S), average degree of irrelevance of the irrelevant diseases (Deg of Irr), and the precision of detecting relevant diseases (Prec %).

The average precision in Table 2 (75.51%) is much higher than that in Table 1 (33.19%). In other words, when diseases co-occur with the queried symptom in one article, most of them are relevant to the symptom no matter whether the retrieved sentence describes their relations. The precision 75.51% is expected to be an underestimate, as not knowing the symptom is related to the detected disease is more common than mis-judging it as related according to the prior knowledge. This observation also supports the proposed ranking mechanism, which considers sentences where diseases are found. Notice that, in order to analyze correct contexts (in this research, sentences), noise filtering techniques for web pages which clean hyperlinks and advertisements should be applied before detecting possible diseases.

Symptom#	Num S	Rel S	Deg of Irr	Prec (%)
1	101	3	3.51	2.90
2	43	4	3.08	9.30
3	37	3	2.94	8.10
4	74	8	3.38	10.80
5	23	9	3.36	39.10
6	23	11	2.50	47.80
7	16	6	2.50	37.50
8	104	103	4.00	99.00
9	87	77	2.70	88.50
10	86	86	0.00	100.00
11	52	3	3.67	5.70
12	61	39	3.64	63.90
13	86	27	3.42	31.30
14	38	20	3.39	52.60
15	54	9	3.69	16.70
16	16	5	3.45	31.25
17	32	3	3.41	9.38
18	60	5	3.56	8.30
19	133	2	3.88	1.50
20	25	0	3.76	0.00
21	43	1	2.79	2.30
22	41	14	3.26	34.10
23	75	7	3.56	9.30
24	114	93	2.76	81.60
25	49	19	2.00	38.80
<b>Average</b>	<b>58.92</b>	<b>22.28</b>	<b>3.13</b>	<b>33.19</b>

Table 1. Evaluation Results under the Strict Metric

Symptom#	Num S	Rel S	Deg of Irr	Prec (%)
1	101	78	3.61	77.23
2	43	33	1.80	75.00
3	37	37	0	100
4	74	44	3.17	59.46
5	23	17	2.33	73.91
6	23	11	2.50	47.80
7	16	6	2.50	37.50
8	104	103	4.00	99.00
9	87	77	2.70	88.50
10	86	86	0	100
11	52	38	3.79	73.07
12	61	61	0	100
13	86	86	0	100
14	38	38	0	100
15	54	47	3.71	87.03
16	16	4	3.25	25.00
17	32	32	0	100
18	60	60	0	100
19	133	108	4.00	81.20
20	25	15	4.00	60.00
21	43	33	1.00	76.74
22	41	31	1.00	75.60
23	75	23	1.29	30.67
24	114	53	2.00	46.50
25	49	36	1.38	73.47
<b>Average</b>	<b>58.92</b>	<b>46.28</b>	<b>1.92</b>	<b>75.51</b>

Table 2. Evaluation Results under the Lenient Metric

Degree of irrelevance also draws a similar conclusion. The average degree of irrelevance is 3.13 with the strict metric and 1.92 with the lenient metric respectively, which shows that if we consider the degree of irrelevance according to the prior knowledge, the occurrences of these diseases are more relevant. Those sentences judged as irrelevant could include information of complications or similar diseases for the disease causing the queried symptom, and they can still provide information for reference. In other words, we found most relevant sentences containing detected diseases, which made it possible to adopt the sentence frequency as a guide to rank the relevancy; the other irrelevant ones are minority and they can also help us determine the relevancy to some degree.

We also find an interesting phenomenon from the symptom whose difference of precisions between Table 1 and Table 2 is large. These symptoms are mentioned more in an informal way in articles like blogs. In these articles, one sentence seldom describes the relation of disease and symptom. Instead, people talked about the experience of having these symptoms or diseases. This phenomenon may also tell us that applying different approaches like we did here when detecting diseases from the Web instead of the medical literature might be a correct direction.

## Conclusion and Future Work

Health care is getting important in societies of nowadays, and utilizing the information from the Internet may benefit it. In this paper, we adopted the concept and techniques of information retrieval to imitate users' behaviors of searching solutions when they find symptoms of diseases. We proposed an approach to detect and rank possible diseases related to the multiple queried symptoms. We further developed an Android application, Health Generation, which implemented the proposed approach to detect diseases and provided further functions. After reporting detected diseases, this application provided Wikipedia information of the diseases and listed nearby clinics. We hope the proposed approach can provide information according to users' need, and through this application, self-caring and finding proper medical assistance or solutions become more feasible.

Though we think that the proposed approach and application are satisfactory, they could be improved from many aspects. First, Wikipedia helps to know the alias of diseases, but we also need to know the different ways of expressing symptoms. For example, "itchy foot" is the same as "itchy skin on foot"; "itchy foot" includes "red itchy foot". Therefore, a good method to expand the queried symptom is necessary to detect diseases properly. Second, we could search the general web logs (Jansen,

Spink and Taksa 2009) to find those queries which are symptoms, and then utilized their query logs to improve the performance. Third, the retrieved web pages and found diseases were of equal weights in the current approach. We may give those web pages from authoritative sources and those web pages which include commonly seen diseases more weights. As to the evaluation, because ranking diseases is difficult, we evaluate the relevancy of the retrieved sentences instead. To have a more precise evaluation, diseases and their known symptoms can be collected and utilized as the testing data in the future. However in this way, we can only check whether the testing diseases are detected by querying the testing symptoms. Knowing the ranks of the other detected diseases is still a tough work.

## Acknowledgements

Research of this paper was partially supported by National Science Council, Taiwan, under the contract NSC101-2628-E-224-001-MY3.

## References

- Bernhard, D. 2006. Multilingual Term Extraction from Domain-specific Corpora Using Morphological Structure. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), Trento, Italy.
- Chun, H. W., Tsuruoka, Y., Kim, J. D., Shiba, R., Nagata, N., Hishiki, T. and Tsujii, J. 2006. Extraction of Gene-Disease Relations from Medline Using Domain Dictionaries and Machine Learning. In Proceedings of Pacific Symposium on Biocomputing 2006, 4-15.
- Dickerson, S., Reinhart, A. M., Feeley, T. H., Bidani, R., Rich, E., Garg, V. K., Hershey, C. O. 2004. Patient Internet Use for Health Information at Three Urban Primary Care Clinics. *Journal of the American Medical Informatics Association* 11(6):499-504.
- Geng, B. , Yang, L., Xu, C. and Hua, X.-S. 2012. Ranking Model Adaptation for Domain-Specific Search. *IEEE Transactions on Knowledge and Data Engineering* 24(4):745-758.
- Hesse, B. W., Nelson, D. E., Kreps, G. L., Croyle, R. T., Arora, N. K., Rimer, B. K. and Viswanath, K. 2005. Trust and Sources of Health Information. The Impact of the Internet and Its Implications for Health Care Providers: Findings From the First Health Information National Trends Survey. *Arch Intern Med* 165(22):2618-2624. doi:10.1001/archinte.165.22.2618
- Jarvelin K. and Kekalainen, J. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems* 20(4):422-446.
- Salton, G. and McGill, M. J. 1986. *Introduction to Modern Information Retrieval*: McGraw-Hill, Inc.
- Jansen, B. J., Spink, A. and Taksa, I. eds. 2009. *Handbook of Research on Web Log Analysis*. Information Science Reference: Hershey.