Medical Recommendation System using Co-Occurrence Graphs

Yanakorn Ruamsuk[1, *], Withawin Tirasopitlert[1], Anirach Mingkhwan[1] and Herwig Unger[2]

[1]Faculty of Industrial Technology and Management
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
[2]Faculty of Mathematics and Computer Science
FernUniversität in Hagen
Hagen, Germany

*Corresponding author. E-mail: yanakorn.ruamsuk@gmail.com

ABSTRACT

Medical diagnosis [1] is the process of determining which disease explain a person's symptoms and signs. Correct diagnosis depend on the knowledge and experience of doctors, huge amount of knowledge is distributed in textbooks and documents. it takes a huge amount of time to learn a lot of information for diagnosis, moreover, diagnosis is often challenging, because many signs and symptoms are nonspecific or some symptoms can be many of diseases. Therefor the knowledge processing must be automated. A co-occurrence graph is incorporated from sets of documents that represent knowledge, which is similar to human brain, when people read a book, the book content will become to knowledge. For co-occurrence graph, it created by information extraction from documents, the information that extract from document is read sentence by sentence and words of each sentence will create relation to each other, then the co-occurrence graph will appear and represent to brain knowledge. In this paper will presents an experiment to bring the co-occurrence graph theory to apply with medical information and create the system that contribute the diagnosis from symptoms.

Keyword: Centroid, Spreading Activation, Diagnosis, Co-occurrence Graphs

INTRODUCTION

When humans are sick, they will go to see the doctor and the doctor will always ask about the symptoms for diagnosis, symptoms is a unusual state that happen when human have certain diseases, more symptoms that the doctor knows from patients will make the diagnosis more accurate, in contrast to less symptoms will make more difficult for the diagnosis because some symptom may involves to many of diseases. For instance, "cough" involves to many of diseases such as "influenza", "measles" and "sinusitis", etc. But if add more symptoms such as

Article history:
Received 12 March 2020; Received in revised form 20 April 2020;
Accepted 16 May 2020; Available online 2 June 2020

“cough”, “headache” and “fever” therefor “influenza” will have high possibility to be the disease of those symptom diagnosis. Then, the correlation of various pieces of information must be considered [2].

A co-occurrence graph similarly to human brain, when people learn or read something from books or documents, that content will become to knowledge, which human will understand the relation of the content (or words) that read from documents. For the co-occurrence graph consist of nodes and edges, nodes is represent to words that extracted from document, edges represent to relation between words and relation value between words that represented by path distance which derived from frequency that two words occur together [4], path distance that can used for analysis with many thing such as centroid calculation [4], clustering, and hierarchical clustering [5].

Medical recommendation system is an experiment to create the system that contribute the diagnosis from symptoms by using the co-occurrence graph as a brain or database of the system. The co-occurrence graph of the system will create base on medical information (or diseases information) which the format of information is an article of various diseases that get from Wikipedia (https://en.wikipedia.org), each article of diseases will add into co-occurrence graph which consist relations (co-occurrence) between word and then the relations value (or path distance) will use to analyze for disease (or centroid) of the symptoms that input into the system. Medical recommendation system will demonstrate the benefits of using co-occurrence approach.

REVIEW OF LITERATURE

Detecting source topics by analyzing directed co-occurrence graphs, is demonstrate a new method to determine the sources of topics, that influence the main topics in texts, by analyzing directed co-occurrence graphs using an extended version of the HITS algorithm. Additionally, this method can be used to identify characteristic terms in texts. In order to obtain the needed directed term relations the notion of term association is introduced to cover asymmetric real-life relationships between concepts and it is described how they can be calculated by statistical means [7].

Centroid terms are comfortable instruments to represent texts, compare them semantically and to even (hierarchically) cluster sets of documents using them. Their determination depends on their topical and conceptual context, i.e. the dynamically changing knowledge of a user represented by the co-occurrence graph. Herein, important properties of centroids as well as their applicability for tasks in natural language processing and text mining shall be discussed and their use justified by a

set of experiments. Based on the obtained results, a new approach to detect fine-grained similarities between text documents is derived [2].

MEDICAL RECOMMENDATION SYSTEM PROCESSING

A. Concept

To create medical co-occurrence graph, words that extracted from documents, only significant words (nouns) will add into co-occurrence graph. Begin with read first sentence of first document, all words in that sentence add into the co-occurrence graph and represent to nodes, other that all nodes in the same sentence will create relation to each other which represent to edges, then next sentence were read and add all words into the co-occurrence graph, if some word in current sentence are duplicated in the first sentence, do not add that node into co-occurrence graph again but increase the occur value of that node instead, also as edges, if some pair word that appear together in the same sentence again, increase the occur value of that relation (edge), the occur value of nodes and edges represent to the frequency which will use to calculate to edge distance. Moreover, the list of diseases and list of symptoms must be prepared, then, while adding words to the co-occurrence graph, examine words with list of diseases and list of symptoms for determining node types. Follow all those concepts from the first document to the final document the co-occurrence graph will be completed.

After the medical co-occurrence graph created, now the system has its own brain, then the system will use the co-occurrence graph to calculate for disease of symptoms that user input into the system, calculate for disease represented by centroid calculation which is the calculation to find which disease node that closest to all symptoms in the co-occurrence graph.

B. Co-occurrence Graphs Building Algorithm

There are four main steps to build the co-occurrence graph.

1) All words (nouns) in documents add into co-occurrence graph and represent to nodes, each node has the following properties:
   “name” – The actual word in base form, names are left unchanged.
   “occur” – In how many sentences did this word occur?
   “type” – What type of node is this? disease (“DS”), symptom (“ST”) or normal (“NN”)
2) Words that appear together in the same sentence create relation to each other and represent to edges, each edge has the following properties:
   “count” – In how many sentences did the connected words co-occur?
   “dice” – DICE significant value for relation.
   “cost” – Distance of connected words.
3) A list of diseases and a list of symptoms must be prepared, while words that extracted from documents and add to co-occurrence graph, specify

word type by check if that word exist in list of diseases, then this word is a disease (“DS”), on the other hand if that word exist in list of symptoms, then this word is a symptom (“ST”), but if word that not exist in both diseases and symptoms list, then that word will classified as normal word (“NN”).

4) After all documents added to co-occurrence graph and the co-occurrence graph is constant, then define distance of each connected word or “cost” of each edges, but to do that must defined significant value for relation [3] (“dice”) first, dice can be defined by:

$$Dice_{wi,wj} = \frac{2\,countAB}{node(\text{”occur”})_{wi} + node(\text{“occur”})_{wj}}$$

- Dice, is the significant value for relation between two nodes (w_i, w_j).
- node(“occur”), is the frequencies value of that node appear in sentences.
- countAB, is the smallest value of $node(\text{“}occur\text{”})_{wi}$, $node(\text{“}occur\text{”})_{wj}$ and $edge(\text{“}count\text{”})_{wi,wj}$.
- If dice value > 1, then dice = 1.

After defined dice value, then calculate edge distance [2] by:

$$Distance_{wi,wj} = \frac{1}{dice_{wi,wj}}$$

The high value of a dice, it means the two connected words has highly related, and when using the dice to calculate the distance, it also has a short distance.

C. Centroid Calculation Algorithm

After the co-occurrence graph build process finished, then create centroid calculation task for calculate to find which node is the most probably disease which is closest to all symptoms in the co-occurrence graph [2]. Centroid calculation is the average distance calculation and find which node is has shortest average distance to all symptoms, but in huge co-occurrence graph it takes very long time to calculate all path, therefor this experiment will demonstrate fast centroid calculation approach which called spreading activation (hop) [6], the approach has following process:

When input symptoms into the system, then these symptoms will be calculated to find candidate centroid.

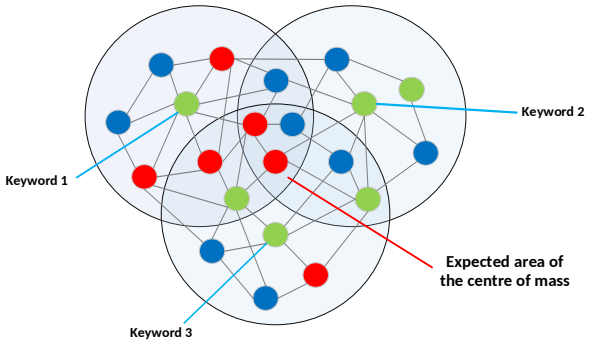

Figure 1 The spreading activation (hop)

1) Find node that connected direct (neighbors) to each symptom, then find which neighbors that connected to all symptoms.
2) If all neighbors of all symptoms that not had any node that connected to all symptoms in first hop, then keep finding in the next hop until set of nodes that connected to all symptoms were found, and that set of nodes are represent to centroid candidate.
3) After centroid candidate has found, then calculate the distance from each centroid candidate to all symptoms and find which one that has shortest average distance, the node that had shortest average distance will represent to centroid or the most probable disease of the symptoms.

Therefor, calculating a centroid using spreading activation (hop) approach will make centroid calculation faster even in huge co-occurrence graph because it no need to calculate every node in the graph to find which one is a centroid but calculate only nodes that represent to candidate of centroid.

EXPERIMENTAL

A. Goals

The experimental presented the result of the medical recommendation system using co-occurrence graph, the goals of the algorithms are:

- Create the co-occurrence graph base on medical information.
- Relation between diseases and symptoms appear in the co-occurrence graph.
- Get a possible disease of the symptoms by calculating centroid.

To prove this according to the above objective, the experiment will be described below.

B. Experimental setup

The experiment has been performed on a personal computer (PC) with an Intel Core i3 CPU 2.50 GHz and 8 GB of RAM, to show that the goals of the algorithms have been achieve. To build the co-occurrence graph of medical recommendation system, the 221 documents which an article of diseases from Wikipedia (https://en.wikipedia.org) were set and add words (nouns) into co-occurrence graph and create relation (edges) between words in the same sentence, the co-occurrence graph base on 221 documents consist of 9,109 nodes (221 diseases) and 609,929 edges. The system flow processes are show in Figure 2.

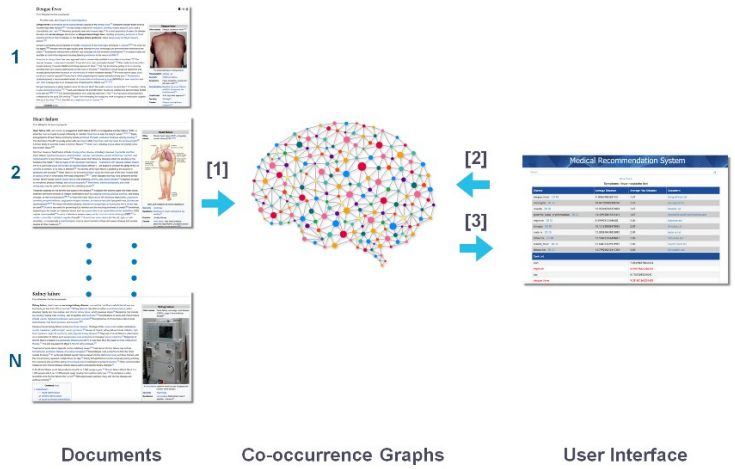

Figure 2 The system flow processes

In Figure 2, the medical recommendation system processes as follows:

1. The information from the document which a 221 article of diseases, is used to create the co-occurrence graph and represent to the brain or database of the system. The nodes in the co-occurrence graph must be classify to 3 cluster, which is has disease (“DS”), symptoms (“ST”) and normal word (“NN”), normal word is not either a disease or a symptom.
2. After the system has a co-occurrence graph, then the system able to receive data from user which is a set of symptoms that user want to find the disease of it, for instance, user may input three symptoms such as “cough”, “fever” and “itch”, after user input set of symptoms on the interface, then the system will send that data to next process of the system.
3. After the system received a set of symptoms from user, then the system will calculate for centroid to find the node that had the most relation to the symptoms, which can be known by average distance from disease node to symptoms nodes in the co-occurrence graph. The short distance indicates to high relation between connected nodes, and to calculate a centroid faster,

the system using spreading activation (hop) approach, which is a fast centroid calculation approach, then the result of centroid calculation will send back to user interface to show the diseases node that closest to all symptoms in the co-occurrence graph.

C. Results and discussion

Results of the experiment discussed herein presented that the medical recommendation system which used the co-occurrence graph base on medical information as a brain or database of the system. And the system using the co-occurrence graph to calculate the centroid, which node that closest to all symptoms in the co-occurrence graph. The output of medical recommendation system with centroid calculation of symptoms are presented below.

This is 5 sample of the system output to show the centroid and average distance of input symptoms.

1. Input symptoms = “fever”, “headache”, “itch”

Disease	Average Distance
dengue_fever	19.05297146367996
influenza	19.824042474832787
meningitis	22.38538264553797
chickenpox	24.931967825299456
lyme_disease	27.2727848953128

Figure 3 Output 1

2. Input symptoms = “cough”, “fever”, “headache”

Disease	Average Distance
influenza	8.62299101080542
common_cold	13.954501999359783
pneumonia	14.685066488789545
sinusitis	17.54155863654573
meningitis	19.332316268152333

Figure 4 Output 2

3. Input symptoms = “stomach”, “pain”, “heartburn”

Disease	Average Distance
peptic_ulcer_disease	16.723036123091607
stomach_cancer	24.641993704632302
myocardial_infarction	28.762273445680922
gastroesophageal_reflux_disease	31.597928968662703
cancer	34.93285565863841

Figure 5 Output 3

4. Input symptoms = “memory”, “loss”, “headache”

Disease	Average Distance
transient_ischemic_attack	14.88846816779665
traumatic_brain_injury	15.42730129773017
stroke	23.688536193744095
encephalitis	31.102807784462353
meningitis	31.349413044479647

Figure 6 Output 4

5. Input symptoms = “pain”, “breathlessness”, “cough”

Disease	Average Distance
pneumonia	24.808055971005192
heart_failure	30.989959976216927
sinusitis	39.38385439284792
asthma	41.71643822494338
meningitis	50.16120716743625

Figure 7 Output 5

All five output that show above present various input set of symptoms and show the result of centroid calculation, in each output from output 1 to output 5, it shows the top five diseases that closest to all input symptoms from user, in ascending order of average distance. In Figure 4 or output 2 presented the result of centroid calculation of the symptoms (“cough”, “fever”, “headache”), and the disease that closest to all symptoms is “influenza” with average distance is 8.6229910 which is very good value indicate that disease are very close to all symptoms or other meaning that “influenza” is the most possible to be a disease of these symptoms in the co-occurrence graph.

The set of symptoms that seem correlation to each other will return the good result of centroid calculation because if the symptoms have a high relation, these symptoms will also close to each other in the co-occurrence graph, then the centroid will also not far away from these symptoms, on the other hand, if set of symptoms that seem not correlation to each other, then these symptoms in the co-occurrence graph may be scattered and the result of centroid calculation may not relevant to the symptoms or bad average distance (too high).

CONCLUSIONS

The algorithm and implementation details of medical recommendation system using co-occurrence graphs have been studied. The co-occurrence graph base on medical information will create the graph that show the relations between diseases, symptoms and other relevant information that extracted from documents. Then calculating a centroid in the co-occurrence graph will return the disease that related to the set of symptoms, which from the experiment results, the disease that get from centroid calculation had a good result, the diseases was actual quite related to the symptoms, therefor the medical recommendation using co-occurrence graph will contribute the diagnosis more comfortable, moreover the co-occurrence graph is similarly to human brain, but the difference between learning of human brain and the co-occurrence graph is learning duration, human take time to learn but the co-occurrence graph can be automated.

REFERENCES

Kubek, Mario & Böhme, Thomas & Unger, Herwig. (2017). Empiric Experiments with Text Representing Centroids. Lecture Notes on Information Theory. 5. 23-28. 10.18178/lnit.5.1.23-28.

A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clutering: a review" ACM Computing Survey, pp. 264-323, 1999.

Kubek, Mario & Böhme, Thomas & Unger, Herwig. (2017). Spreading activation: a fast calculation method for text centroids. 24-27.

Kubek, M. & Unger, H., (2012). Detecting source topics by analysing directed co-occurrence graphs. In: Eichler, G., Wienhofen, L. W., Kofod-Petersen, A. & Unger, H. (Hrsg.), 12th International Conference on Innovative Internet Community Services (I2CS 2012). Bonn: Gesellschaft für Informatik e.V.. (S. 202-211).