Centrality Research on the Traditional Chinese Medicine Network*

Zhang Dezheng
School of Information Engineering University of Science and Technology Beijing Beijing, China 100083
zdzchina@126.com

Gao Lixin
School of Information Engineering University of Science and Technology Beijing Beijing, China 100083

Zhang Huansheng
Department of Computer Hebei Engineering And Technical College Cangzhou, Hebei 061001
why05118@sina.com

Liu Jianming
School of Information Engineering University of Science and Technology Beijing Beijing, China 100083

Abstract

Aiming at the complex data in the Traditional Chinese Medicine, a new way is proposed in this paper that data mining of complex relations to find out the potential information among different medicine objects. We turned the Traditional Chinese Medicine knowledge network into graph by using information from ontology, then adopted centrality algorithm to analyze and process this graph, and finally mined valuable medicine knowledge. As the result of the verification test, this algorithm shows very good practicability.

1. Introduction

The conventional method in data mining is based on simple structure, such as a simple relation data set. However, the relations among the Traditional Chinese Medicine objectives are so complex that the old technology is not suitable. As a result, it was very difficult to mine data in the field of the Traditional Chinese Medicine before our research. It is the key to sum up and pass on old famous the Traditional Chinese Medicine doctors' academic knowledge and their clinical experiences that analyzing the relationship among concepts of treatments in the traditional Chinese medical. These concepts have been associated with each other by the construction of ontology[7], making their relationship into a complex network. In this paper, a reasonable solution based on graph data mining is proposed, in which graphs are composed out of the Traditional Chinese Medicine networks by making the cases of distinct disease to match the ontology[8] in the network.

As a formal language, graph[5] is a kind of tool to describe network and its features, to formalize the network data, and to quantize characteristics of a real network. Here, graph is used to formally describe the traditional Chinese network, while spot is used to stand for medicine, symptom, evidence, and line is used to represent relationship among spots. Figs1 shows a first diagnosed symptom network. The centrality algorithm referred is based on graph algorithm[4] to analyze the Traditional Chinese Medicine network.

The method of describing the real network as graphs before data mining is embodied in many applications. For example, in sociology, scholars characterize social relationship by graph which has many vertexes and lines[1]. A vertex represents an actor that may be a person, an organization, or a group, etc. The line connecting two vertexes represents a social joint[1]. In the field of biology, the protein structure is described as graph in which a vertex is an atom and an edge is a valence. By mining data on the protein structure graph, we can find the internal relations of protein structure[3].

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if

* Supported by the National High-Tech Research and development Plan of China (863) under Grant (No.2007AA01Z170); the Beijing Natural Science Foundation of China under Grant (No.4062022).

there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

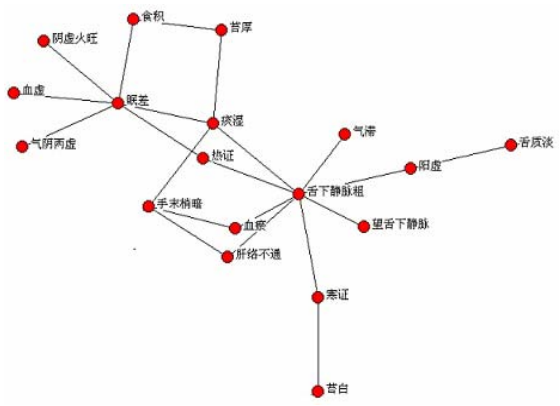

Figure 1. **A first diagnosed symptom network**

The centrality mark of the vertex plays an essential role in analysis on network structure which we can determine the key vertex of the whole network. In the graph of social network, centrality mark can define social rights according to relationship, which shows individuals or organizations are whether most important in social or not according to its position whether in a joining centre of not[3]. Similarly, in the Traditional Chinese Medicine, we can quantize the importance of a disease, a medicine or a symptom with measurement of its centrality. In a whole treatment, a higher centrality medicine plays more primary role, while a higher centrality symptom repeats more times in the diagnoses.

2. Basic thought and concept

2.1. Relevant concepts

The network can be conveniently described as a graph G = (V, E). In the Traditional Chinese Medicine networks, the vertex set V represents medicines, symptoms or evidences, while the edge set E represents relations between vertexes, medicine, and symptom.

As knowledge on graph is used in centrality study, the relevant concepts [5] are introduced first.

1) Path: a route in which each line or point doesn’t repeat. In network analysis, more attention is pay on paths than lines.

2) Length: the number of lines constructing the path.

3) Geodesic: the shortest path between two points. If there are some shortest paths between two points, they are all geodesic.

4) Distance: the length of geodesic of two points. So as to say, the distance between two points is the shortest length to connect them. d_G (s, t) is used to represent the distance between point s and t.

2.2. Betweenness centrality, centrality based on the shortest path

Classifying by different portray mark, centrality measure method is of different kinds as follows: degree centrality, closeness centrality and Betweenness Centrality. Degree Centrality uses the number of edges close to vertex; Closeness Centrality reflects the centrality degree of vertex in the network, because it defines the derivative of sum of the distance from this vertex to all the others; and Betweenness Centrality reflects the control of the vertex onto other ones. This paper analyzes the Traditional Chinese Medicine by studying of the Betweenness Centrality arithmetic[2].

Graph in this paper is undirected and no-weighting. Suppose ω is weighting function defined on the edge, then ω (e) = 1.

Suppose σ_{st} is the number of shortest path from vertex s∈V to vertex t∈V, and $\sigma_{st}(v)$ is the number of paths passing the vertex v in the shortest path. Here is a standard measurement formula of Betweenness Centrality.

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad \text{(Freeman, 1977)}$$

What gives above is the expression of Absolute Centrality Degree. In order to be easy to compare between centrality from different graphs, we give the mark of Relative Centrality Degree, which is the standard of absolute centrality degree mark. The method to calculate the Relative Centrality Degree of a point is to divide the Absolute Centrality Degree of the point by a sum result which is got by summing up each possible maximal value of all the other points in the same graph. The discussion below is on calculating of centrality.

Lemma 1: (Bellman-ford algorithm) Vertex v∈V lies in the shortest path from vertex s∈V to vertex t∈V, if and only if $d_G(s,t) = d_G(s,v) + d_G(v,t)$ [6].

Pair-dependence of vertexes s, t∈V of medium point v is $\delta_{st}(v) = \frac{\sigma_{st}(v)}{\sigma_{st}}$, that is the possibility of vertex v lie in the shortest path between s and t. In order to get the Betweenness Centrality mark of vertex v, it is need to calculate all the pair-dependence of the vertex $C_B(v) = \sum_{s \neq v \neq t \in V} \delta_{st}(v)$.

Therefore, the Betweenness Centrality should be calculated by two steps:

1) Calculating the length and number of the shortest path between each pair of vertexes.

2) Calculating each pair-independence.

3. Method to calculate Betweenness Centrality

3.1. Calculating the length and number of the shortest path between each pair of vertexes

The calculation begins from vertex s using Board First Search algorithm and Dijkstra's Single Source Shortest Path algorithm[4]. Dijkstra's algorithm is applied to get the shortest path from any vertex to any other one.

$$P_s(v) = \{u \in V : \{u,v\} \in E, d_G(s,v) = d_G(s,u) + \omega(u,v)$$

Lemma 2: (combinatorial calculate the shortest path) For s ≠ v∈ V, $\sigma_{sv} = \sum_{u \in P_s(v)} \sigma_{su}$.

3.2. Calculating each pair-independence

In order to elaborate calculation demands of all pair-independence, we first introduce the concept of independence of vertex s ∈ V on single vertex v ∈ V.

Theorem: The independence of s ∈ V on any vertex v ∈ V agrees with the relationship:

$$\delta_{s\bullet}(v) = \sum_{w:v \in P_s(w)} \frac{\sigma_{sv}}{\sigma_{sw}} \cdot (1 + \delta_{s\bullet}(w))$$

An edge {v, w} exists where vertex v lying on at least one shortest path from s to t, δ_{st} (v)>0 and v ∈ Ps(w). This is a little more complex, as Figs 2 shows.

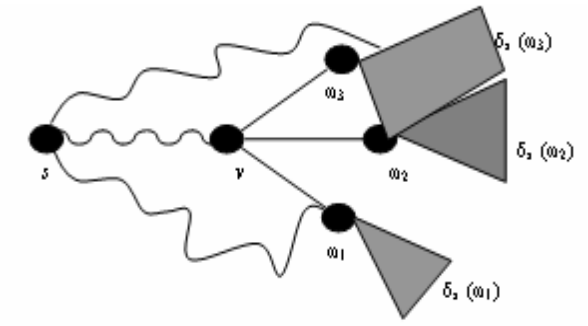

Figure 2. **Condition theorem**

4. Betweenness Centrality Algorithm

There are some variables and arrays at the beginning of the algorithm. Array CB stores centrality, σ[t] stores number of t on the shortest path. Chain array P[w] store sets of all vertexes w on the shortest path of vertex s.

(09-22) do BFS traverse, (13-16) judge whether w is visited for the first time, (17-20) judge whether v is one point on the shortest path to w, if so then add the path into array P[w]. And finally pop out the stack (24-29) and calculate the centrality.

```
01   CB[v] ← 0,v∈V;
02   for s∈V do
03      S ← empty stack;
04      P[w] ← empty list, w∈V;
05       σ [t] ← 0,t∈V; σ [s] ← 1;
06      d[t] ← -1, t∈V; d[s] ← 0;
07      Q ← empty queue;
08      enqueue s → Q;
09      while Q not empty do
10        dequeue v ← Q;
11        push v → S;
12        for each neighbor w of v do
13          if d[w] < 0 then
14            enqueue  w → Q;
15            d[w] ← d[v] + 1;
16          end
17          if d[w] = d[v] + 1 then
18              σ [w] ← σ [w] + σ [v];
19              append v → P[w];
20          end
21        end
22      end
23       δ [v] ← 0, v ∈ V;
24      while S not empty do
25        pop w ← S;
26        for v ∈ P[w] do
27         δ [v] ← δ [v] + σ [v].(1 + δ [w])/ σ [w];
28        if  w ≠ s then CB[w] ← CB[w] + δ [w];
29      end
30   end.
```

5. Example Analyses

There is the calculation result showing in table 1, which is made out by applying the above algorithm to analysis the Traditional Chinese Medicine network that has 32 vertexes. Betweenness Centrality is used to measure the ability of a symptom serving as the medium in first diagnoses. That is if there isn't the symptom, no relation exits between other symptoms. More times such roles a symptom plays, higher

Betweenness Centrality it is, and then more other symptoms are connected by it. In another word, this symptom is more important part during the treatment.

TABLE 1. **Result of analyzing centrality**

Vertexes	Betweenness	nBetweenness
1.Coarse sublingual vein	334.367	71.907
2.phlegm dampness	169.267	36.401
3.yang deficiency	168.000	36.129
4.whitish tongue	150.000	32.258
5.chill syndrome	130.000	27.957
6.White Tongue	114.000	24.516
7.The problems of Sleep Quality	105.700	22.731
8.thick tongue coating	91.967	19.778
9.heat syndrome	42.733	9.190
10.Gravy palm	38.967	8.380
11.obstruction of liver Collaterals	11.833	2.545
12.blood stasis	11.833	2.545
13.dyspepsia	9.333	2.007
14.qi-yin deficiency	0	0
15.qi stagnation	0	0
16.sublingual vein	0	0
17.blood deficiency	0	0
18.yin defic fire hyperact	0	0

The result of analyzing centrality algorithm is showed by data, as table 1. To make it more clear and understandable, a visual picture is given below, as Figs 3. Because the centrality of the symptom named coarse sublingual vein is the highest, we can conclude that it was the most possible medium of other symptoms. That is to say curing the coarse sublingual vein is of great importance during the treatment.

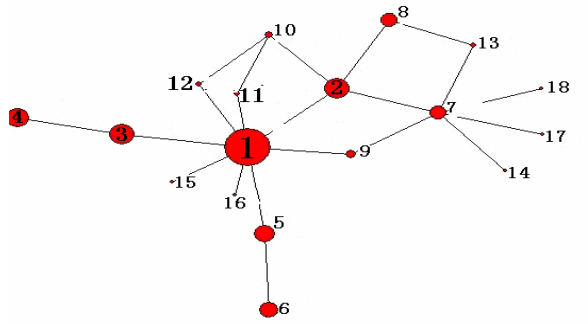

Figure 3. **The result of centrality algorithm**

6. Conclusions

In this paper, we discussed a new way of data mining in the area of the Traditional Chinese Medicine and solved the problem that complex relationship can not be searched by the data mining based on set. First make the concepts extracted from cases of diseases match pair with entries in ontology library and then compose a graph structure out of the Traditional Chinese Medicine network. Second, analyze the structure of the network on this graph by processing it with the centrality algorithm. Last, get the data mining result.

7. Reference

[1] Liu Jun. The Introduction of Social Network Analysis [M]. Beijing: Social Science document press,2004.

[2] Robert Sedgewick. Algorithm of C.[M]（Volume 2 Algorithm Graph）. Beijing : Posts & Telecom Press,2004

[3] S.Wasserman and K.Faust. Social Network Analysis: Methods and Applications[M]. Cambridge University Press, 1994.

[4] Yin Jianhong, Wu Kaiya. Graph Theory and Its Algorithm [M]. Hefei: University of Science and technology of China Press,2003.

[5] Xiao Weishu. Graph Theory and Its Application [M].Beijing: Aviation Industry Press, 2005.

[6] Dictionary of Algorithms and Data Structures, Paul E. Black, ed., U.S. National Institute of Standards and Technology. 4 March 2005.

[7] He Hai-yun, Yuan Chun-feng, Overview of Technology of Building Domain Knowledge Based on Ontology. Application Research of Computers.2005 (3):14-25

[8]NataLya F Noy and Deborah L McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford University, Stanford, CA, 94305