A Context Analytical Method Basing on Text Structure

Yi Huang

Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia,
Beijing University of Posts and Telecommunications, Beijing
Email: huangyi@tseg.org

Jianbin Tan, Lei Zhang

Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia,
Beijing University of Posts and Telecommunications, Beijing
Email: tanjianbin}@tseg.org, zlei@bupt.edu.cn

Abstract—**In this paper the research techniques of complex network are introduced into the complement of missing data in text and a new method of text mining is put forward basing on the text structure of large-scale texts. First the GRE word net is constructed by using lots of relative articles specially for experiment, then the static characters of this network are analyzed and the context relationships of words are obtained in it according to the community discovery algorithm of complex network, next an complement algorithm is designed to judge whether it is the right complement words by following relationships among these words. In the experiment, we take the examination questions of GRE as test set and use this method to do the sentence completions in verbal sections, the result demonstrates the availability of this text analyzing method which focuses on topology information of network. It can not only apply to the imputation of missing data, but also the complement of full sentence after skeleton's forming in machine dialogs.**

Index Terms— **text mining, word net, community discovery, complement, missing data**

I. INSTRUCTION

During the preservation, filing and transportation of documents, various reasons would result in the loss of text. This instance, which brings lots of difficulties in using and analyzing data, is one of the most prominent causes for uncertainty of Information systems. In order to fully take advantage of the collected data, many scholars at home and abroad proposed their original opinions on the processing of missing data to save the incomplete data and guarantee the research work carrying on smoothly [1].

Aiming at general incomplete data set, researchers proposed many methods to solve the missing data [2,3]. These methods were applied to different situation basing on the different theory, specifically speaking such as directly discard the record including missing data[4], the most commonly used average value filling, the maximum

class filling and the complete example analysis, Bayes estimation[5], statistics method, like the combinatorial method, the condition combination method, the regression analysis, the HotDeck filling method, the Expectation maximization, the multiple imputation, the maximum likelihood, the C4.5 and so on, Quinlan[10] also suggested to forecast the value of missing data basing on other known attribute values and the category messages; Automatic associate neural network method and dynamic path generating method; Many scholars also studied knowledge discovery methods basing on the incomplete system in rough set theory[8]; Wang Qingyi[9] put forward a method of guessing missing data basing on the principal component analysis method.

Multiple imputation (MI) appears to be one of the most attractive methods for general purpose handling of missing data in multivariate analysis. The multiple imputation framework suggested by Rubin (1978, 1987a, 1996) is an attractive option if a data set is to be used by multiple researchers with differing levels of statistical expertise. This approach involves imputing several plausible sets of missing values in the incomplete data set resulting in several completed data sets. Each completed data set is analyzed separately, say by fitting a particular regression model. The resulting inferences – point estimates and the covariance matrices – are then combined using the formula given in Rubin (1987a, Chapter 3) and refinements thereof (Li, Raghunathan and Rubin 1991; Li, Meng, Raghunathan and Rubin 1991; Meng and Rubin 1992; and Barnard 1995).

A new method Unlike above is put forward in this article, which reviews context similarity in text mining basing on the context analysis suspicion of existing text structure. This method first analyzes the context of remain words in article, then choose and fill the missing data from the options, and thus arrive the final objects that using computer to accomplish the imputation of massive incomplete data. The first step of this method is

to study the large-scale articles (almost 500 English articles and the sum of word is close to 3,000,000) and seek for GRE words appeared in the articles. Second, take the core glossary by these words, and form a big graph using relationships between them. Besides, divide the large graph into several subgraphs, as called community, by using the Community Detection algorithm. For each word in the community, other words of the same community are united as its analytic context, while words in other communities take the convert distance between two communities as context to evaluate its importance on analysis result.

The rest of this paper is organized as follows: Section 2 presents the formation rules of word network used in the experiment. Section 3 particularly introduces community discovery method. Section 4 is the description of filling algorithm, which is designed by the theory of complex network. Section 5 is the experiment process of this new method on data imputation by taking GRE text questions as test sets and section 6 is the final conclusion of the paper.

II. RELATED WORK

Hidalgo C A(2007) analyzed the relationship between different kinds of product and the network is constructed by the similarity. It's concluded from his paper that adjusting the construction of the import and export merchandise according with the relationship between different kinds of merchandise might improve the situation of trading and economy of the country[2]. Li Munan(2007) analyzed the relationship between companies producing the similar kind of products[9]. He also did analysis of the basic topology characters of the products competition network.

The algorithm for community detection is involved in this paper in order to seeking for the essence structure of the word network. Girvan and Newman have introduced a betweenness based algorithm by iteratively cutting the edge with the biggest betweenness value to partition the whole network into small communities, by using a proposed Q modularity measure, it can generate the optimized division of the network. To improve the computation efficiency, Newman has also proposed a fast clustering algorithm very recently. Other proposed algorithms include METIS [14], spectral clustering [15], flow simulation, as well as co-clusterfing [5]. Nan Du (2007) proposed a clique-based algorithm which performs well in large graph [11].

The idea that vertexes in the same community should be tantamount approximately on the topology status of the network is introduced in the concept of community in this paper. Pagerank [13] is used here to evaluate the topology status of each vertex of the network. The edges between the vertexes which are far different on the ranking are weakened due to they are thought to be generated occasionally rather than necessarily. Contrastively, the edges within the communities are thought to be the necessary edges or the edges which do exist most possibly, and they are enhanced to reflect the essence of the structure of the real network.

III. GRE WORD-NET

Constructing network model is of a great help on analyzing characteristics of every entity and their relationships in practical analysis. Researchers had ever used the method of network construction and the minimum spanning tree algorithm to visualize scientific research characteristic and evolving process [7]. For non-social network, researchers of US once construct product network via entity similarity[11]. For text data, we construct and analyze GRE word net, which is constituted by words appear in GRE relative articles with highly frequency. Meanwhile, the research methods of complex network were introduced into the imputation of missing data to help the accomplishment of accurate imputation for incomplete data by analyzing the text structure and the relationships between words.

In traditional sense, Word-Net is an English dictionary basing on cognition linguistics, which is designed by the combination of psychologists, linguists and computer engineers in Princeton university. It not only arrange the words alphabetically, but also compose a network of words according to the word's meanings. In this Word net the relationships between words are considered as synonymy, antonym, and the frequency difference , but the word net mentioned in this paper is different with the that one in traditional sense.

When we encounter a missing data in the text, the most usual way is to guess or confer the word in missing location according to context. Two words can be considered relation tightly on the condition that they have high frequency of being appear at the same time, muti-words situation can also be generalized from this way. The word in the option, if has very close relations with words in context, will be chosen in priority as correct word. Therefore our task is transferred to finding word pairs which would coordinate appear with tightly association in massive articles.

As we know, the words appear in GRE test are obscure and most of them will not be used in spoken language. Materials including these words usually belong to fields like humanities, social science, natural science and the contents related to fields like history, ethic, religion, moral and etc. Therefore the data of word-net in our experiment will be distinguish with common word net which contains large quality of daily words , it comes from various papers of all kinds of field in Nature and Science magazine, CBN news, papers on biology, physics, history of Royal Society Publishing, Bible and <<THE SONNETS>> written by Shakespeare, the total amount of words is nearly 3,000,000. It is relatively large scale though not cover all over the words and satisfy experiment demand. We now need to define the rules of this GRE-Word Net here. The test in our experiment aims at GRE test questions, so we need to snatch GRE words as its core word set.

Take GRE word as node V of network, if these two words appear in a content with length L at the same time, this contemporary appearance relationship could be considered as an edge e

between them, and the weight of edge is the time of their appearance.

In this rule, length L could be the length of a nature sentence, and could also be the length of a nature paragraph or the total passage. The experiment on word sense disambiguation in reference [2] demonstrated that taking nature paragraph to carry on disambiguation could get the best effect, but in our experiment, the network is not a sparse graph and characteristic of the small world would also lost when taking nature paragraph as length of L, so L is defined as the length of nature sentence and the end symbol of sentence is used as standard to partition relation range. This conclusion was drawn from the following test experiment.

L was looked as a glide window and the value of it could be changed from 1 to the size of whole network. Here we took the representative value to calculate the topological information of Word Net when L is defined separately as 10, 50,100,200,and more than 200, since 10 represent the shortest length of a natural sentence, 50 represent the longest length of a natural sentence, 100 represent the average length of a natural paragraph and 200 represent the long paragraph words, more than 200 represent other situations. The result of test experiment is shown in table 1.

Table 1. Basic topological information compare of different L lenth

L length	10	50	100	200	>200
Nodes	53	738	2402	2667	2803
Average degree	3.53	5.2	8.31	10.65	20.78

The relationship graph of L , network nodes and the average degree could got from above experiment data, as shown in fig.1.
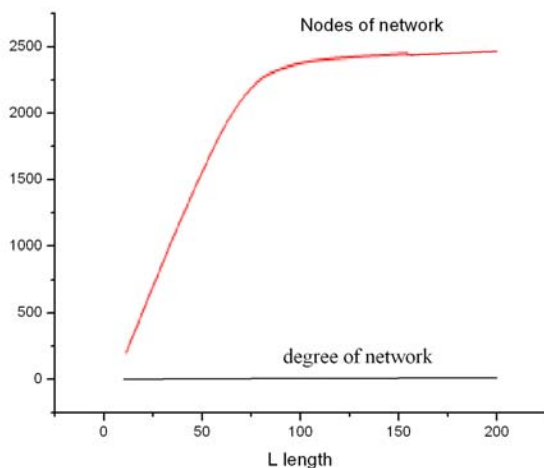

Figure 1. This picture shows contact relationships between nodes in the whole network and the length of L, between average degree of the whole network and the length of L.

From fig.1 it is easy to see the scale of network change drastically when L is increasing from 0 to 100, after 100, the curve change gently and mild. The average degree also turn to a better value according to our common experience when L is 100. Therefore, we define L is 100 to construct the whole network which could be saw in

fig.2 and the basic topological information of GRE-Word Net extracted by the above   rules is as follows:

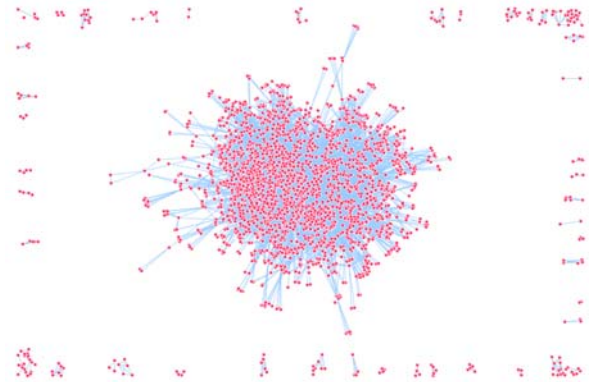|	
Nodes number of network	|V|=2,402
weight edge number	|E|=19,958
average weighted degree	wd=13.84
average degree	d=8.31
Density	Density=0.0069
mean clustering	Cluster=0.3454

Figure 2. This picture shows whole network with the length of L is 100..

Two questions exist on choosing and constructing the relationships:

1. In this experiment, we take the word that has different means as the same node in word-net without distinguishing and this would effect result to some extent, in other words, two relatively unattached communities would amalgamate together into a large one with big granularity because of multi-vocal words. In this way the accuracy of result would decline and cause the possibility that appropriate word is not be able to be chosen and fulfilled into the missing location. Our next work is to divide words of different means according to present research of Word Sense Disambiguation and thus get a word-sense meta network called Word-sense net, which takes word-sense pair as nodes. It is believed that this processed result would bear higher accuracy.

2.The preorder word and subsequence word of any one word were treated without difference, as to how to sound express the different relationship of preorder and subsequence word and make the word distinguish between time and spatial distribution, it is the key point of our next research.

IV. COMMUNITY DISCOVERY

Newman proposed the concept of community structure in complex network in 2002. Community is the set of the nodes in network, and the nodes in community connect closely to each other while not between different communities. This set is also called cluster, cohesion group. Although there is not an abroad acceptant, uniform and measurable definition of community, it is the real symbol of the arrangement and model structure in complex systems. The edge between two communities play an important role in communicating two different communities and the Edge Betweeness is a topological

number, weighing the quality of transferred information through it.

Edge betweeness : The time when this edge appears in the shortest path among all nodes of network.

From the definition, it is easy to see that edges located between communities bear higher betweeness. The community discovery of network bases on Newman theory[6], which keep reducing the edge with highest betweenness until the number of final gotten community is appropriate or the number of edges is less than a scheduled threshold value. However, because of the requirement that recalculating the shortest path and betweeness of each edge   after reducing edge each time, the operand of traditional GN algorithm is so large that it will take a quite long time to operate. Therefore, besides adopting GN algorithm as standard, a FAST algorithm ameliorated from GN also be used to fast calculate the article superposition result of each time. It is essential to use FAST adjusting the redundancy of network because the mining model is not incremental.

Since Weighted result of edges are not considered in FAST algorithm, the divided communities are slightly rough, but the operation time is rather short(11.4s) and 59 communities can be found in result, the average scale of each community is 40.7 nodes. While the communities got from GN algorithm, which nearly cost 18 hours to finish the operation, are fine-grained and the number of community containing single node is 1443, more than two nodes is 959. The reason why form so many single node communities is many GRE words appear only few time in few articles and merging them into feat community is difficult. A new method on dividing community is proposed by Nan Du[12] , its first step is to find the maximum community in network, then merge left nodes to the nearest divided maximum community and finally the communities are formed one by one. Taking the Q value as standard to evaluate effect of community partition, Nan Du algorithm is less than GN about 0.1 but more than FAST an order of magnitude.

The compare of those algorithm is illustrated in table.3. From the data of table , it is easy to draw the conclusion that the best choice is to choose FAST GN algorithm to discover communities.

Table 2. Basic performance compare of different algorithm

	Time	Communities	Average scale of each communities	Q
GN	18h	2402	1.2	0.29
FAST GN	11.4s	59	40.7	0.12
Nan Du	6h	3.43	8.31	0.19

The largest community got from Fast GN is show as follows, not all the point of this community but only a few ones with high appear frequency were drawn in this graph .
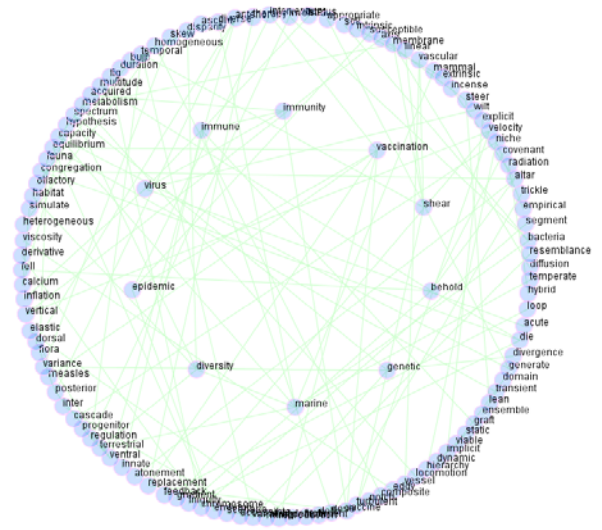

Figure 3. This picture shows contact relationships among nodes in the largest community, the nodes located in center are those whose degree ranks top 10. It is able to see that all words in this community are about medical biology field.

V. FILL ALGORITHM

Since most related key words were included in the GRE Network, when we have a passage with one vacancy, the first thing to get is the key words set in known texts and the words set in options. From the opinion of graph and complex network, all we need is to compare the tightness degree of words in options and known texts, then choose the proper word according different degrees. Basing on the result of community classification, we are convinced that words in the same community will contact more closely and vice versa. How to quantify the tightness degree of their relations? A topological distance is defined to evaluate it, which differed with semantic distance in general text mining.

The distance between nodes defined before is the length of the shortest path among nodes. The graph got after dividing community is called entropy chart, we need to define the distance among nodes in entropy chart over again.

1. **Distance among nodes in the same community is 0, D = 0；**
2. **For those nodes located in different communities, a ∈ C1 , b ∈ C2 , the distance between them is the distance between the communities they belong , D_{ab} = DC（C1，C2）.**

If a community in entropy chart is regarded as a node, the distance between communities could be defined as the same with the distance between nodes. The edges among the communities are edges that contact with the nodes inside community and those outside the community. In other words, for

$$a_1 \in C_1, a_2 \in C_2, a_3 \in C_2, e_1 = (a_1, a_2), e_2 = (a_1, a_3), e_1, e_2 \in E$$

Edge e1 and e2 between communities C1 and C2 are emerged into one edge e and the sum weight of e1 and e2 is taken as the weight of new edge e. Edge with high weight between two communities demonstrates the tight relations between them. In this experiment, they are

highly possible to have the relationships of former and latter sentence. On this basis the distance between communities is defined as follows:

Take the sum of weight values of all weighted edges between two neighbouring communities as the weight value W of that emerged single edge. Distance between communities is the reciprocal of weight value of all edges included in the shortest path. $DC_{inderect-connect} = \sum_{i=1}^{d} \frac{1}{W_i}$ **, d is the number of edges in the shortest path.**

Table 3. Filling algorithm

Fill Algorithm:
1. Extracting words Ve={v | v exists in GRE-WordNet AND v exists in given L words } from given sentence;
2. Reflecting each vertex in Ve to its Community Index, forming set of CIE;
3. Vf = {v | v is one of the possible words to be filled in given paragraph};
4. Reflecting each vertex in Vf to its Community Index, forming set of CIF;
5. Computing distance of each Community-index in CIF to all the Community-indexes in CIE

$$sumD_i = \sum_{j=1}^{n} D_{ij} \ ;$$

6. Choosing minimum value sumDx;
7. Fill the word, which reflects previously mentioned sumDx, into the vacancy of the paragraph;
8. OVER.

The time complexity of this algorithm is O(nm), n is |Ve|，m is |Vf|.

In order to implement this algorithm effectively, two key points need to be considered.

One problem s how to find the key words in question and given options efficiently. Many string search algorithms could be used in this step. Notable algorithms to perform the search have been devised by Knuth, Morris and Pratt and by Boyer and Moore. During a search, the pattern may be considered aligned with part of the text. Algorithms vary in the order in which character-character comparisons are made and in the way that they determine the distance the pattern can be shifted if the current alignment does not match.

In the straightforward method, corresponding pairs of characters are compared left-to-right. If a mismatch is found then the pattern moves one place down the text, i.e. so that P_1 is aligned with T_{k+1}. In the Boyer–Moore algorithm, character-pairs are compared right-to-left, that is the first comparison is between P_M and T_{k+M-1}. If a comparison fails, e.g. P_j with T_{k+j-1}, then both the index of the failure point (j) and the identity of the text character (T_{k+j-1}) are used to determine the distance that the pattern can be moved down the text for the next potential match. Two tables of distances are computed from P before searching starts: table delta1 is indexed by T_{k+j-1} and table delta2 is indexed by j.

Sunday's insight into the string matching problem was to observe that if the pattern does not match the currently aligned text fragment, the character T_{k+M} must be aligned with the pattern in its next position unless the pattern is shifted right past it. He uses this character to determine the shift amount, again using a table (DELTA1) recomputed from P. If T_{k+M} is in P then it is aligned with the rightmost instance. Using T_{k+M} to determine the shift distance has two consequences. First, if the character is not in P we can shift the pattern right past it. Thus, the greatest distance we can shift the pattern is one greater than the greatest shift possible with Boyer–Moore. Secondly, we are free to compare character pairs in any order. In particular, we can test first the pair that gives the best shift if there is a mismatch or test first the pair that seems most likely to be different. For example, we can compare characters in the pattern in ascending order of their a priori frequency in the language. Sunday termed this approach the optimal mismatch (OM) technique. His experiments showed that it results in fewer character comparisons than the Boyer–Moore algorithm (though the percentage improvement declines with increasing pattern length). There are two aspects of a fast string searching algorithm. The first is the quick detection of a mismatch between the pattern and the aligned text fragment; checking n character frequency order is effective here. The second is the largest possible shift of the pattern. From the discussion above, we use Sunday's algorithm to search string, the comparison is as follows:

Table 4. Pattern–text comparison

```
i ← 1
while i ≤ pattern length
    do if Pattern[Ordering[i]] = Text[offset+ Ordering[i]]
        then i ← i +1
        else { if i> 1 then Adjust (Ordering, i)
            i ← pattern length + 2
            }
```

Another problem is the calculation of distance between any two nodes in the network，that is, given a weighted graph and two vertices u and v, we want to find a path of minimum total weight between u and v. Classical algorithm solving this problem involve Dijkstra's Algorithm and Bellman-Ford Algorithm. Considering the network in our experiment is undirected, it is better to choose Dijkstra's Algorithm. The description is formally as follows:

Table 5. Dijkstra's Algorithm

```
Algorithm DijkstraDistances(G, s)
    Q ← new heap-based priority queue
    for all v ∈ G.vertices()
        if v = s
            setDistance(v, 0)
        else
            setDistance(v, ∞)
```

```
        l  ←  Q.insert(getDistance(v), v)
        setLocator(v,l)
   while ¬Q.isEmpty()
      u  ←  Q.removeMin()
      for all e  ∈  G.incidentEdges(u)
        { relax edge e }
        z  ←  G.opposite(u,e)
        r  ←  getDistance(u) + weight(e)
        if r < getDistance(z)
            setDistance(z,r)
            Q.replaceKey(getLocator(z),r)
```

In this algorithm, method incidentEdges is called once for each vertex, distance and locator labels of vertex are set z O(deg(z)) times, setting or getting a label takes O(1) time.Each vertex is inserted once into and removed once from the priority queue, where each insertion or removal takes O(log n) time. The key of a vertex in the priority queue is modified at most deg(w) times, where each key change takes O(log n) time. Dijkstra's algorithm runs in $O((n + m) \log n)$ time provided the graph is represented by the adjacency list structure, so the running time can also be expressed as $O(m \log n)$ since the graph is connected.

VI. EXPERIMENT RESULT AND ANLYSIS

The sentence completions in verbal sections of GRE test questions require the students to analyze, infer and guess the even definitely understand the meaning of uncompleted sentence according to its context and filling the missing data. The test set includes 16 sets of GRE test questions, about 244 questions in total. 19 questions of them meet following conditions: more than 2 options in the question can be looked up in GRE-Word Net and these options contain the correct answer. Using the communities resulted from FAST algorithm, the correct filling questions are 12 , with the right ratio 5.36% in whole 224 questions and 63.16% in 19 WordNet containing questions. It is believed that the correct filling questions and right ratio would be increased after aggrandizing the quantity of text set and the words checked.

Taking the second question in Section2 of GRE test on June 2006 as an example, the original question is as follows:

> In linking geographically disparate people, the Internet is arguably helping millions of spontaneous communities to bloom: communities defined by common interests rather than by the accident of------------------.

- (A)   compatibility
- (B)   affluence
- (C)   reciprocity
- (D)   contemporaneousness
- (E)   proximity

The key words of question checked from GRE-WordNet are:

disparate: Fundamentally different or distinct in quality or kind;

helping: the activity of contributing to the fulfillment of a need or furtherance of an effort or purpose;

spontaneous: happening or arising without apparent external cause;

The words of options checked from GRE-Word Net are:

compatibility ： capability of existing or performing in harmonious or congenial combination；

reciprocity：a relation of mutual dependence or action or influence；

According to Fill-Algorithm to calculate the distance between two words in option and the keywords in question:

compatibility 0.04785
reciprocity 0.2876

From the difference of the distance, it can be referred that option A has huge possibility to be the correct answer. The contact relations of the communities containing these 5 words can be described as figure 2 and 3.
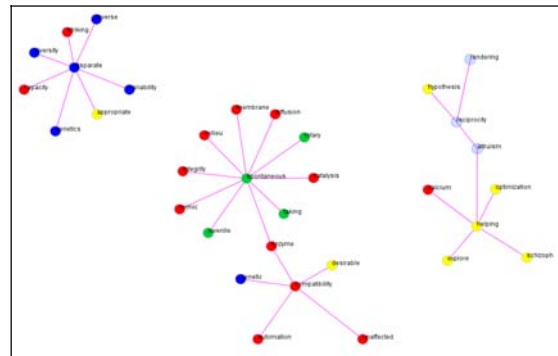

Figure 4. Different color stands for different community, it is observable that every keyword surrounds some red points which come from the community containing compatibility. The sky-blue points belong to community which reciprocity located. All the points drawn here may be the keywords of question we concerned, or perhaps the join points contacting these keywords with other communities. These join points belong to its corresponding communities directly or the communities of these keywords. This figure shows the points that can reach object community within 2 leaps.
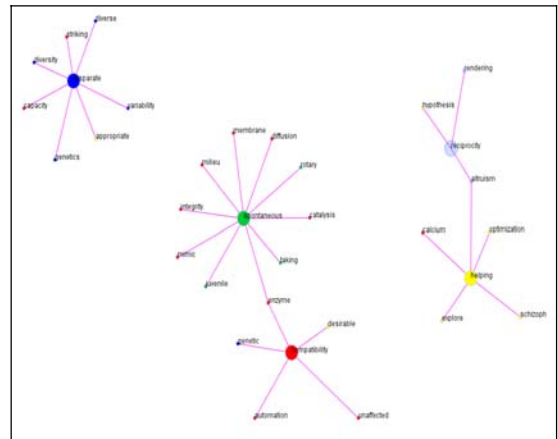
Figure 5. This figure shows the enlarged picture of 5 words we care. It is easy to see that reciprocity only contact with helping, the feeblish point in question.

Other questions in experiment can get similar effect. From the experiment results, it is considered that some missing data of the same structure can also be filled by studying the sentence structure of known article, as everybody knows that sense does a great help in choosing correct answers.

VII. DISCUSSION

This paper uses relative texts as training samples and constructs a GRE word-net for GRE test questions. The data set come from various papers of all kinds of field in Nature and Science magazine, CBN news, papers on biology, physics, history of Royal Society Publishing, Bible, etc and the total amount of words is nearly 3,000,000. We first construct the network according to following rule. Taking GRE word as node of network if these two words appear in a content with length L at the same time and the contemporary appearance relationship could be considered as an edge between them, and the weight of edge is the time of their appearance. The effect would be best when L is defined as the length of a sentence. Then the structure characteristic and community partition are also be analyzed by theory and method of complex network. We calculated the some topological information such as the nodes number of network , weight edge number, average weighted degree, average degree, Density and mean clustering.

After that, according to the community adscription and the contact degree of each word in word-net, a method of predicting the missing text data is put forward in this paper. It make use of the community detective methods proposed by Newman and other scholars to divide the whole network into several separated communities to stand for the words with closely relationships. This method takes the furthest advantage of known data and has prominent effect through the experiment verification. Massive text missing data can be repaired by operating this method in computer and the imputation of incomplete sentence, which was formed by skeleton words in machine dialog, can also be well done.

Our next work will be improved on three aspects:

1. Basing on the present research of word sense disambiguation, we will correspond every word in article with its mean and use these "word-sense pair" as nodes in network to distinguish different word sense and form a more clear word sense network.

2. It is relatively blur to process relations among the words just by drawing edges between words in the same sentence without difference as existing relations. For one word, there must be some words locating before or after it at most of time. All we need is to treat the preorder word and subsequence word differently in order to express the otherness of word-net in time and spatial distribution, this is our main job.

3. We will farther collect more text data, sort it, manage it with programming when its scale has reach enough quantity, and to form a corpus available in our research finally.

REFERENCES

[1] Rubin D. Inference and missing data [J]1 Biometrika , 1976 ,63 (3) : 581 - 592.

[2] Ghahramani Z,Jordan M I.Mixture models for learning from incomplete data [C]. Cambridge, MA: Computational Learning Theory and Natural Learning Systems,Volume IV: Making Learning Systems Practical,The MIT Press, 1997.67-85.

[3] Jiawei Han, Micheline Kamber，Data Mining: Concepts and Techniques, Morgan Kaufmann Publisher, 2000.

[4] Kantardzic M.Data Mining Concepts,Models,Methods,and Algorithms. Tsinghua University Press,2003

[5] Marco Ramoni. Learning Bayesian networks from incomplete databases [EB/OL].Technical report kmi-97-6,Knowledge Media Institute,The Open University, 1997.http://kmi.open.ac.uk/publications/index.cfm?trnumber=kmi-97-6.

[6] Newman M E J. The structure and function of complex networks [J]. SIAM Rev, 2003, l.45 (2): 167 - 256.

[7] Samoylenko I, Chao T C, Liu W C, et al. Visualizing the scientific world and its evolution [J]. Journal of the American Society for Information Science and Technology, 2006, 57 (11): 1461 - 1469.

[8] Chmielewski M R,Grzymala-Busse J W,Peterson N W,et al.The rule induction system LERS-a version for personal computers [J].Found Computer Decision Science, 1993,18(3/4):181-212.

[9] Wang Qingyi, Cai Zhi, Zou Xiang. An Approach for Knowledge Discovery under the Environment of Incomplete Data [J]. JOURNAL OF SOFTWARE, 2001,12(10):1516-1524.

[10] Quinlan J R.C4.5: Programs for machine learning [M].SanMate, CA: Morgan Kaufmann Publishers Inc, 1993.

[11] Hidalgo C A, Klinger B, Barabási A.-L, et al. The Product Space Conditions the Development of Nations [J]. Science, 2007, 317: 482 - 487.

[12] X. P. B. W. L. X. Nan Du, Bin Wu. Community detection in large-scale social networks. *submitted to SNAKDD2007.*

[13] Guo Chi ,Chen Jiajun, WangQixiang. An Approach Corpus-based Word Sense Disambiguation. Computer Engineering and Applications

[14] Sun Shuang, Zhang Yong .Clustering Method Based on Semantic Similarity. Journal of Nanjing University of Aeronautics & Astronautics

[15] Wang Lin, Dai Guan-zhong. Community Finding in Complex Networks—Theory and Applications . Science & Technology Review

[16] Richard V.Sole, Bernat Corominas Murtra, Sergi Valverde, and Luc Steels. Language Networks: their structure,function and evolution.

[17] William W.Cohen, Pradeep Ravikumar, Stephen

E.Fienberg. A Comparison of String Distance Metrics for Name-Matching Tasks. American Association for Artificial Intelligence.

[18] L.da F.Costa , F.A.Rodrigues, G.Travieso. Characterization of Complex Networks:A Survey of measurements.

[19] JIA Yan, WANG Yong-heng, YANG Shu-qiang. Survey of text mining based on ontology. Computer Applications Vol.26. No.9 Sept.2006

[20] WANG Jing-hua, LIU Jian-yi, WANG Cong.Word Sense Disambiguation with Semantic Graph Structure.Journal of Beijing University of Posts and Telecommunications Nov.2006 Vol.29 Sup

About the Author

Yi Huang is a graduate student major in Computer Science and Technology, with the study direction on next generation of Business and Operation   Support System and Enterprise Intelligence. At present, she works in related research issues on text mining. Currently, she is focusing on problems of model building and validation with survival data, including semantic studies; on parametric modeling of survival data; and on novel trial designs.

Jianbin Tan is a graduate student major in Computer Science and Technology, with the study direction on Data Mining and knowledge discovering. At present, he works in related research issues on complex network. Currently, he is focusing on problems of model building and validation with survival data, including semantic studies; on parametric modeling of survival data; and on novel trial designs.

Lei Zhang is a professor with 20 years of research experience, his research field involving the next generation of Business and Operation   Support System and Enterprise Intelligence. At present, he works in related research issues in text mining. Currently, he is focusing on problems of model building and validation with survival data, including semantic studies; on parametric modeling of survival data; and on novel trial designs.