http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (12): 2252-2259, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.2252.2259

Adaptive Social Network Construction using Gaussian Mixture Model

[1,2]Xin Guo, [1,2]Yang Xiang, [1,2]Qian Chen and [1,2]Wei Wei
[1]Department of Computer Science and Technology, School of Electronic and Information Engineering, Tongji University, Shanghai 201804, China
[2]The Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 200092, China

Abstract: A Social network graph shows social interactions and relationships between individuals in a specific social environment, which is very helpful for analyzing social relationships, activities, structures, etc. The author quantized the strengths of social objects' relationships in social environment using an improved vector space model. Gaussian mixture model was employed to set the threshold for identifying social relationships adaptively and divide social subgroups automatically. According to the threshold, social network graph would be constructed based on performance measures. It is concluded that hidden social relationships can be discovered effectively by using this approach which is very flexible and adaptive for dynamic information feedback mechanism.

Key words: Gaussian mixture model, social network, adaptability

INTRODUCTION

Social network analysis is the mapping and measuring of relationships and flows between people, groups, organizations, computers or other information processing entities. Social network was first promulgated by John Barnes (Freeman, 2006). Compared with other data analyzing and mining technologies, social network analysis focuses on analyzing various relations quantitatively. Social network graph, a graphic representation of social networks, can express relation information content, including relationship existence, relationship direction, relationship weight and etc. It can help us to understand the way social individuals interact with each other and discover social information and chances.

Nowadays, there are various types of documents including news stories, scientific literature, blogs, conversations, etc. One primary challenge is how to obtain social relationships from these documents. Lots of contributive work had emerged for this. Mika (2007) took advantage of ontology to construct social relations based on semantics. Diesner and Carley (2005) employed meta-matrix model to reveal the social structure in text. Magnini *et al.* (2002) recognized entities from the text based on WordNet. However, there are some problems to extract social relationships in texts using methods above. (1) It is a tedious and difficult work to build ontology libraries and dictionaries for special organization or social environment. (2) When the knowledge library is very huge, consulting dictionaries and ontology libraries are time-consuming and inefficient. (3) As social members, relationships and environment usually change over time, knowledge library has to change consequently. (4) The above methods can not solve problems such as ambiguous reference, subject ellipsis and context dependent. (5) Social relationships in text corpus have sociological, anthropological and psychological properties, no interactions with users always bring low rate of accuracy. (6) To obtain optimal results once and for all is not possible without considering any performance feedback.

SOCIAL RELATIONSHIPS QUANTIFICATION

Social relationships quantification is the preliminary work of social relationships extraction. In order to extract and quantify social network relations from corpus, we had built a social vector space based on vector space model which is one of the most widely used models in information retrieval (Salton *et al.*, 1975). The social vector space model is an improved vector space model as it is not about words and documents but about cases and affiliations (Guo *et al.*, 2011). A case c is an actor, including social individual and social group, organization or other collective social unit. An affiliation

Corresponding Author: Xin Guo, Department of Computer Science and Technology, School of Electronic and Information Engineering, Tongji University, Shanghai 201804, China

a is a collective social unit, events or activities which social individuals, groups, companies or other smaller collective social units are subordinate to. The T_k represents documents containing affiliation a_k. Case frequency is denoted by cf_{ik} which imply the number of times a case c_i occurs in T_k. The $avecf_{ik}$ refers to average case frequency:

$$avecf_{ik} = cf_{ik} / |T_k| \quad (1)$$

where, tf_{ik} represents text frequency, the number of documents containing both affiliation a_k and case c_i. The average text frequency $avetf_{ik}$ is calculated by:

$$avetf_{ik} = tf_{ik} / |T_k| \quad (2)$$

where, af_i is explained as the affiliation frequency containing cases c_i, where m, A[is the collection of affiliations:

$$af_i = \sum_{k=1}^{|A|} avetf_{ik} \quad (3)$$

The probability of selecting an affiliation containing case c_i from all the affiliations can be given by $af_i/|A|$. Thus, log $(|A|/af_i)$ is the inverse affiliation frequency which is denoted by IAF_i:

$$IAF_i = \log(|A|/af_i) \quad (4)$$

The weight or the strength of a social relation between case c_i and affiliation a_k is denoted as w_{ik} which is given by:

$$w_{ik} = avecf_{ik} \cdot IAF_i \quad (5)$$

Each dimension corresponds to a separate case. If an affiliation involves a certain case, its value in the vector is non-zero and equals to weight. Affiliation vector a_k can be represented by $ak = (w_{1k}, w_{2k}, ..., w_{|C|k})^T$, where, Cv is a collection of cases. A vector space of social relations Vv is given by $V = (a_1, a_2,..., a_{|A|})^T$.

The social vector space can be represented as an incidence matrix which presented the relationships between cases and affiliations, i.e., case-by-affiliation matrix. Then incidence matrix can be decomposed into case-by-case adjacency matrix B which reflects the relations between cases and cases. Suppose $|A| = m$, $|C| = n$, the matrix B can be given by:

$$B_{ij} = B_{ji} = (b_{ij}) = \begin{cases} 0, & \text{while } i = j; \\ \sum_{k=1}^{k=m} \min(w_{ik}, w_{jk}), & \text{while } i \neq j. \end{cases} \quad (6)$$

ADAPTIVE SOCIAL NETWORK CONSTRUCTION

Settings and background: In this study, the word adaptive means an algorithm or system can serve the user by learning requirements or understanding performance feedbacks, thus information delivered to the user can adapt to dynamic environment automatically. By constantly updating the feedback, compute the performance automatically till the results are satisfactory. It's important and useful in such situation when there is little training data in the initial stages of relation mining.

As far as we know, there is not much research on adaptive threshold setting for relation mining. Existed approaches are mainly focused on learning a user's profile on whether the user thinks a document should be retrieved or not while interacting with the user in the field of information filtering system, such as Rocchio, language models, Okapi and pseudo relevance feedback (Allan *et al*., 1998; Callan, 1996; Ma *et al*., 2002; Anghelescu *et al*., 2002; Collins-Thompson *et al*., 2002; Srikanth *et al*., 2002). They usually score retrieval performance and set threshold for filtering, then deliver the results to the user in order to obtain user's feedback. Authors researched adaptive threshold setting for novelty mining and they think novelty documents and non-novelty documents follow Gaussian distribution. They constructed the optimization criterion for searching the best threshold. Actually, how to choose the threshold for social relation extraction is a sociological issue. It involves social cognitive knowledge and relates to special individual's social consciousness and psychology. Thus, threshold setting is a more difficult task in social relation extraction than that in information retrieval. In reference, Guo *et al*. (2011), initial threshold is determined by expert in some special social field. According to the accuracy of social relation identification in training documents when the threshold changes around the initial value, select the optimal threshold. It is a weak approach obviously. For this reason, we designed an adaptive threshold setting approach by employing Gaussian mixture model.

Gaussian mixture model-based adaptive threshold setting

Gaussian mixture model: In our experimental study, we found that the strengths of social relations distribution from every cohesive subgroup can be approximated by Gaussian distributions. This is intuitive for social relationships are always concentrated in a cohesive subgroup, while most relations with extremely high

relation strengths could not mean the two objects are related but they are the same thing. There may be several subgroups involved in documents and every relation can be classified into some subgroup so we suppose the relation strength distribution in documents follow a mixture of K Gaussian distributions. Here, K refers to the number of subgroups in documents. The mixture distribution of social relation strength can be represented as a linear superposition of K Gaussian distributions in the following form:

$$p(x)=\sum_{k=1}^{K}\pi_k N(x|\mu_k,\sigma_k)=\sum_{k=1}^{K}\pi_k \frac{1}{\sqrt{2\pi\sigma_k}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} \quad (7)$$

Random variable x represents social relation strength, follows a Gaussian distribution with mean μ_k, variance Σ_k. The parameters $\{\pi_k\}$ satisfy $0\leq\pi_k\leq 1$ and $\Sigma_k\pi_k = 1$.

Expectation-maximization (EM) for Gaussian mixture model:

- Initialize $\{\mu_k\}$, $\{\sigma_k\}$ and $\{\pi_k\}$ and evaluate the initial value of the log likelihood

First, user need to set the threshold θ of social relation strength. The element value of matrix will be set to 0 while social relation strength b_{ij} below the value of θ, then matrix $B_{n\times n}$ is updated to D_θ. Temporarily, we set all the elements of matrix D_θ to 1 when the elements are not equal to 0, then matrix $D_{n\times n}$ is updated to $M_{n\times n}$. Based on the theory of graph connectivity, a case c_i can access any other case c_j though L cases when $m_{ij} = 1$, where, $M^{L+1} = (m_{ij})$. Thus, θ can be determined when:

$$\eta(\sum_{l=1}^{L}(M^L) < \eta(\sum_{l=1}^{L+1}M^{L+1})$$

and

$$\eta(\sum_{l=1}^{L+1}(M^{L+1}) = \eta(\sum_{l=1}^{L+2}M^{L+2})$$

where,

$$\eta(H)=\sum_{j=1}^{n}\sum_{i=1}^{n}\delta(h_{ij})$$

and

$$\delta(x)=\begin{cases}1, x\neq 0\\ 0, x=0\end{cases}$$

$H_{n\times n} = (h_{ij})$ is any matrix and x is any number.

Second, we adopt the idea of k-means algorithm for reference to allocate remaining cases into k clusters. Select k cases as initial cluster centers randomly and allocate case c_i to $cluster_k$ while the accessibility a_{ij} between the case c_i and the center c_i of $cluster_k$ is the highest. The accessibility a_{ij} comes from the element value of the following matrix M_a:

$$M_a=\sum_{L=1}^{L}\frac{1}{L}(M)^L=\begin{pmatrix}a_{11} & \cdots & a_{1n}\\ \vdots & \ddots & \vdots\\ a_{n1} & \cdots & a_{n\times n}\end{pmatrix} \quad (8)$$

After k clusters are formed, re-compute cluster centers by selecting k cases which have highest centrality in each cluster. The centrality cc_i of case c_i can be computed by:

$$cc_i=\sum_{c_j\in cluster_k} a_{ij} \quad (9)$$

Then, reallocate cases into k clusters. Repeat above process until cluster centers no longer change.

At last, we need to revert the element value of matrix $M_{n\times n}$ to matrix $D_{n\times n}$. The initial parameters including are set by:

$$\mu_k^{init}=\frac{1}{r_k}\sum_{\substack{c_i,c_j\in cluster_k\\ i<j}} d_{ij} \quad (10)$$

$$\sigma_k^{init}=\frac{1}{r_k-1}\sum_{\substack{c_i,c_j\in cluster_k\\ i<j}} (d_{ij}-\mu_k)^2 \quad (11)$$

$$\pi_k^{init}=\frac{r_k}{\sum r_k} \quad (12)$$

$$r_k=\sum_{c_i,c_j\in cluster_k, i<j}\delta(d_{ij}) \quad (13)$$

where, r_k is the number of social relations in cluster k.

Suppose, we have a data set of observations $X = \{x_1, x_2,..., x_N\}$, the log of the likelihood function $P(X|\pi, \mu, \sigma)$ is given by:

$$\ln p(X|\pi,\mu,\sigma)=\sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k N(x_n|\mu_k,\sigma_k)\right\} \quad (14)$$

The initial value of the log likelihood can be obtained by substituting the initial variables into the function. Observations $X = \{x_1, x_2, ..., x_N\}$ are from $\{d_{ij}\}$, where $i<j$ and $dij \neq 0$.

- E step and M step of EM algorithm

EM algorithm includes, estimate the expected values and re-estimate parameters.

E step: A K-dimensional binary random variable z is introduced. The value of z_k satisfy $z_k \in \{0, 1\}$ and $\Sigma_k z_k = 1$. The marginal distribution over z is specified as $p(zk = 1) = \pi_k$:

$$\gamma(z_k) \equiv p(z_k = 1 | x) = \frac{p(z_k = 1)p(x | z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x | z_j = 1)} = \frac{\pi_k N(x | \mu_k, \sigma_k)}{\sum_{j=1}^{K} \pi_j N(x | \mu_j, \sigma_j)} \quad (15)$$

where, $\gamma(z_k)$ is the responsibility that cluster k takes for explaining the observation x.

E step is to evaluate $\gamma(z_k)$ using the current $\{\mu_k\}$, $\{\sigma_k\}$ and $\{\pi_k\}$.

M step: Setting the derivatives of $\ln p(X | \pi, \mu, \sigma)$ with respect to $\{\mu_k\}$, $\{\sigma_k\}$ and $\{\pi_k\}$ to 0, respectively, we obtain:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n \quad (16)$$

$$\sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)^2 \quad (17)$$

$$\pi_k = \frac{N_k}{N} \quad (18)$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \quad (19)$$

M step is to re-estimate $\{\mu_k\}$, $\{\sigma_k\}$ and $\{\pi_k\}$ using the current $\gamma(z_{nk})$.

- Evaluate the likelihood and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied, then step returns to E step.

Threshold setting by evaluating performance measures: The performance in our task was measured by calculating the ratio of extracted relations by Gaussian mixture model to all the relations above the threshold when varying the threshold. The ratio is closer to 1, the performance is better:

$$F = \left| \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \delta(b_{ij}) \times p(x > \theta)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \delta(d_{ij})} - 1 \right| \quad (20)$$

$$p(x > \theta) = \int_{\theta}^{+\infty} p(x) dx = \int_{\theta}^{+\infty} \sum_{k=1}^{K} \pi_k N(x | \mu_k, \sigma_k) \, dx \quad (21)$$

The best threshold θ can be obtained by $\theta = \arg\max F(\theta)$. The new threshold can be substituted into EM for Gaussion Mixture Model, then a new round of adaptive threshold setting starts. Repeat this process till the threshold no longer change.

Social network construction: The element values of matrix $B_{n \times n}$ are set to 0 while social relation strength b_{ij} below θ, then we can draw a social network graph by draw a line between case c_i and c_j when the social relation strength b_{ij} is nonzero. In order to identify social subgroups, $\gamma(z_{nk})$ need to be re-computed. We can assume that $x_n \in cluster_k$ when $k = \arg\max(\gamma(z_{nk}))$. Now the social network is constructed and subgroups are divided clearly.

EXPERIMENT

Datasets: We built a set of case-level data about 12 popular industries. The business news provider of the document set was China Daily Website. There were a total of 537 effective .txt files used in our experiments. In order to obtain the text corpus, we employed web crawler tool Heritrix to combine the sentences into a txt. file for each news event. Then, we performed our experiments on the text corpus. A threshold too low can result in most relations between cases in the dataset are considered to be positive and subgroups are indistinguishable. A threshold too high may lead to some important relations lost and subgroups are distinguished excessively. By setting proper threshold using distribution of relations, we can obtain the higher performance of relation mining on datasets and subgroups are distinguished properly. In this experimental study, the focus was on relation mining rather than text categorization. Therefore, our experiments started with all given industries (cases).

EXPERIMENTS AND RESULTS

We obtained the initial threshold by satisfying formulas 10 and 11 when L = 2. Then, 3 cases were selected randomly as initial cluster centers and cases were allocated to clusters according to formula 12. The result is shown in Table 1.

After several iterations of re-selecting cluster centers and reallocating cases, the initial parameters were determined, as shown in Table 2.

After several iterations of EM for Gaussian mixture model, the new parameters were determined, as shown in Table 3.

We evaluated performance measures F and its value is 0.762296. Then we re-computed F with the best threshold whose value is 3.8 and obtained its value 0.158801. Now, a round of adaptive threshold setting was finished. Then, repeat the whole process above till the threshold no longer change. At this moment, the threshold is 26.3. The parameters with the threshold are shown in Table 4.

The responsibility $\gamma(z_{nk})$ was obtained by substituting the parameters into the formula 19. Then, relations were allocated to clusters according to the value of γ (z_{nk}), the social network was formed and subgroups are divided.

DISCUSSION

Now compare the social network graphs (1) when obtained the initial threshold, as shown in Fig. 1 and 2,

Table 1: Clusters when initial threshold is 30.8

Cluster	Centers	Members
1	7	7
2	1	1 2 5 6 8 12
3	10	3 4 9 10 11

Table 2: Initial parameters of clusters after several iterations of selecting cluster centers

Cluster	Centers	Members	Standard deviation	Average	Proportion
1	7	7	0.0	0.0	0.0
2	8	1 2 3 4 5 6 8 9 10 12	59.032430	67.790891	1.0
3	11	11	0.0	0.0	0.0

Table 3: New parameters after several iterations of EM for Gaussian mixture model

Cluster	Standard deviation	Average	Proportion
1	0.0	0.0	0.0
2	34.025684	64.358892	1.0
3	0.0	0.0	0.0

Table 4: The parameters with the threshold

Cluster	Standard deviation	Average	Proportion
1	41.586131	101.738563	0.244185
2	15.226218	46.772283	0.755815
3	0.0	0.0	0.0

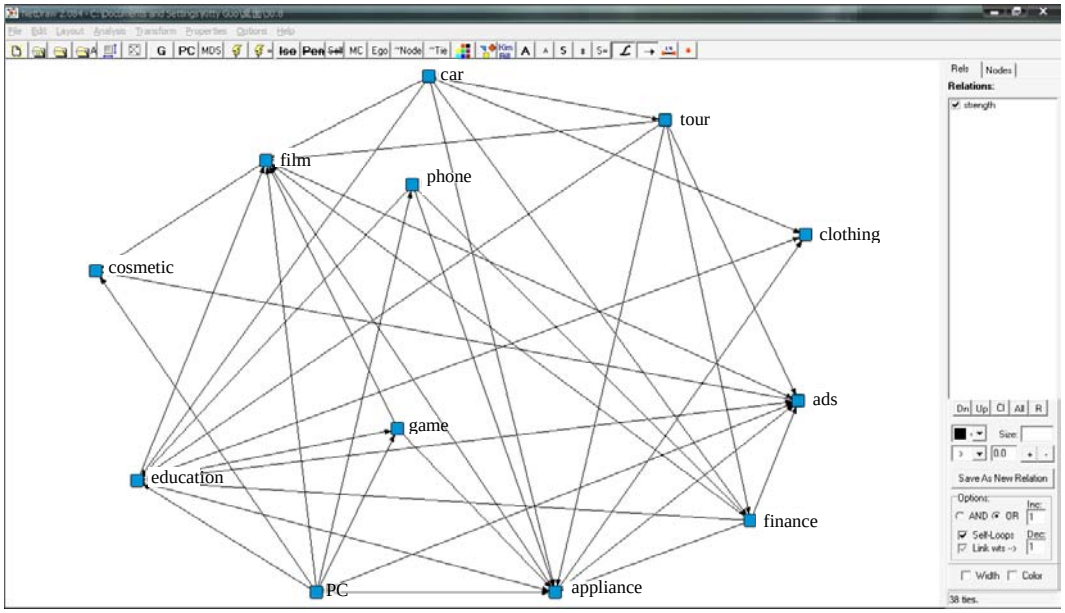

Fig. 1: Social network graph when deciding relations only using the initial threshold

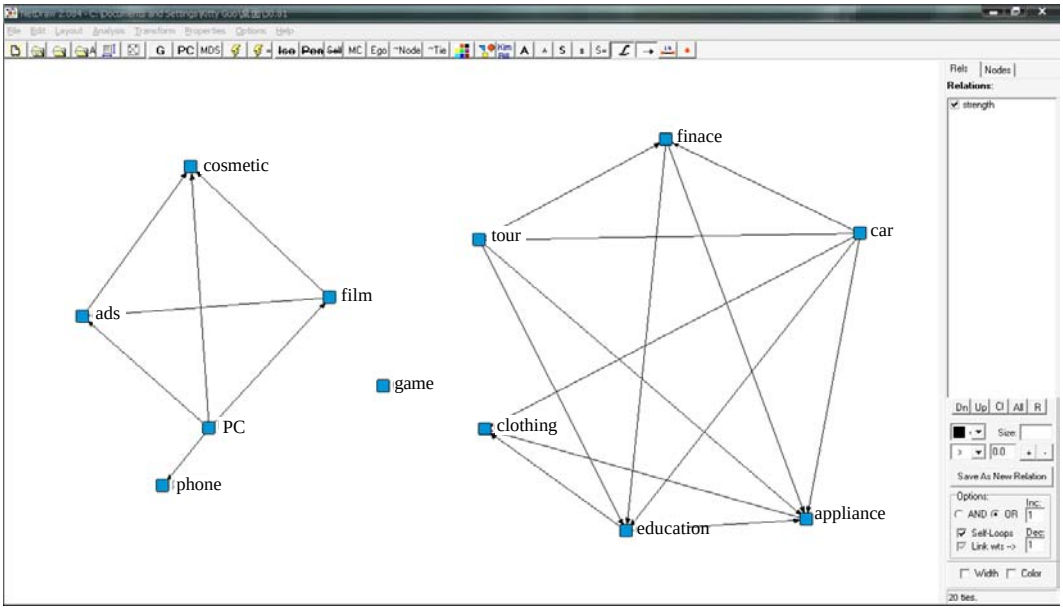

Fig. 2: Social network graph when 3 cases were selected randomly as initial cluster centers

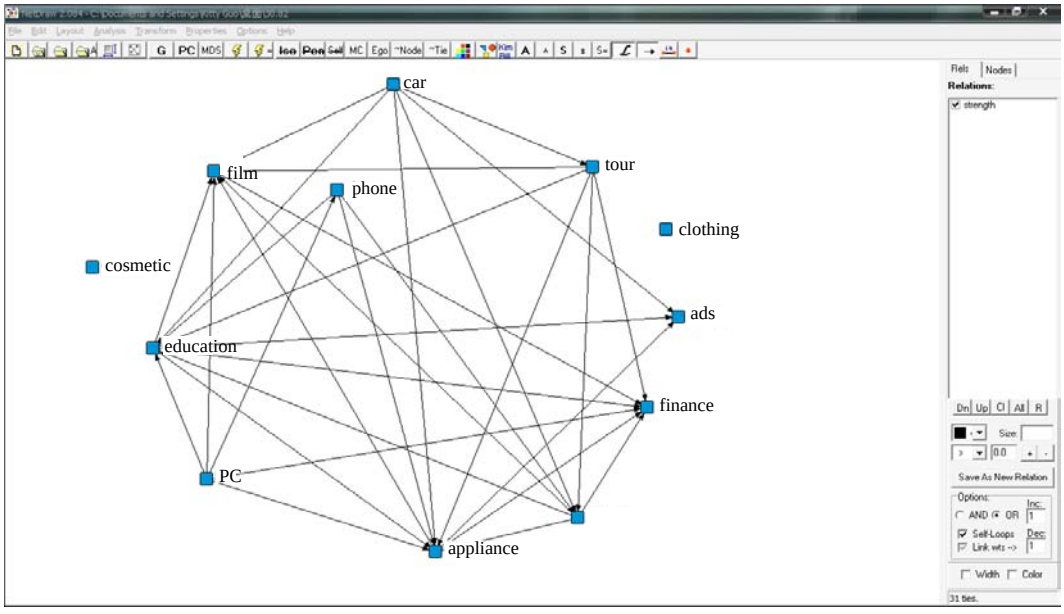

Fig. 3: Social network graph after several iterations of selecting cluster centers

when 3 cases were selected randomly as initial cluster centers, as shown in Fig. 2 and 3, after several iterations of selecting cluster centers, as shown in Fig. 3 and 4, when obtained the final threshold and relations were allocated to clusters according to the value of γ (z_{nk}), as shown in Fig. 4.

By experimental verification, we can find that (1) constructions of social networks are all based on social

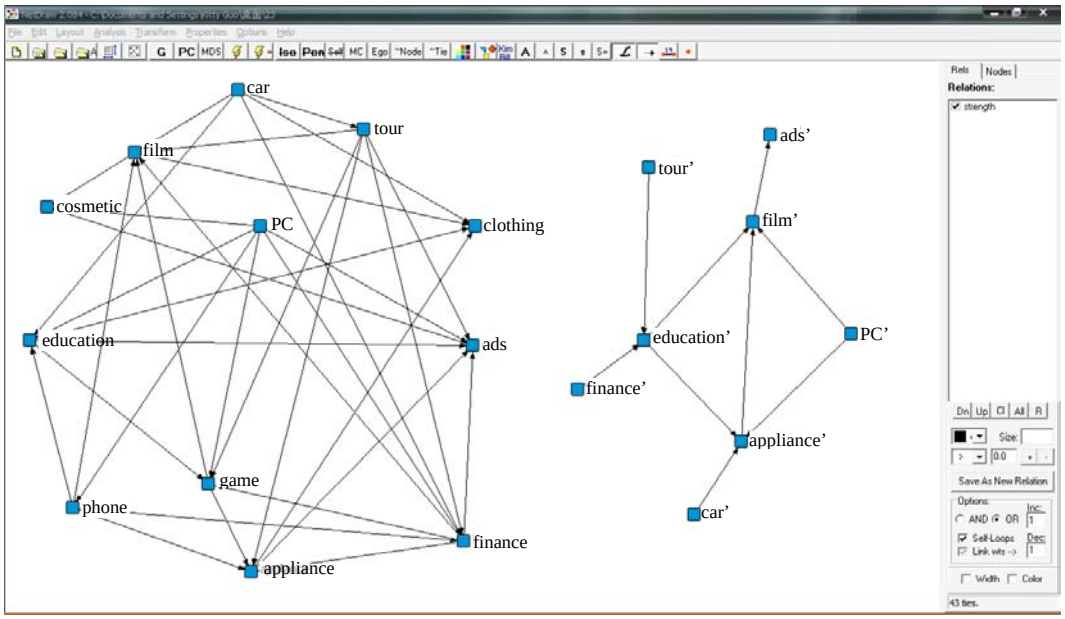

Fig. 4: Social network graph using our approach

individuals in Fig. 2 and 3, while construction of social network is based on social relations in Fig. 4. The later one is more in line with the needs of relation mining. (2) Outliers may occur after several iterations of selecting cluster centers. It can be seen sporadically in Fig. 3 through trials. (3) The number of clusters must be fixed in Fig. 2 and 3 while relations are divided into a reasonable amount of clusters in Fig. 4. A case can only be allocated into one cluster in Fig. 2 and 3, while a case can be allocated into several clusters in Fig. 4. As a social individual may be affiliated with one or more affiliations, Fig. 4 is more reasonable than Fig. 2 and 3. Thus, it is concluded that social network can be constructed effectively by using our approach and performance measure ensures that the final threshold is the best one in accord with Gauss distribution.

CONCLUSIONS

In this study, social relationships were retrieved and social network was constructed. In our experiment, the author observed how social network graph was drawn with our process model. Using the Gaussian mixture model to set threshold, the approach sifted out enough precision meaningful social relations. Besides, user was also capable of obtaining social structure in stead of reading all documents line by line. This approach is only applied to the text corpus which involves social relation and structure information and uses formal language, can be news stories, intelligence data, business information, and etc. Besides, how to determine k (the number of clusters) and L (access interval) will obviously require further investigation.

ACKNOWLEDGMENTS

This study is supported by the National High-Tech Research and Development Plan of China under Grant No.2008AA04Z106 and the NSFC under Grant No. 70771077. The work is also supported by the Project of special funds for the informatization development of Shanghai Municipality under Grant No. 200901015 and the Project of Science and Technology Commission of Shanghai Municipality under Grant No. 08DZ1122300.

REFERENCES

Allan, J., R. Papka and V. Lavrenko, 1998. On-line new event detection and tracking. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (RDIR'98), ACM., Inc., pp: 37-45.

Anghelescu, A., E. Boros, D. Lewis, V. Menkov, D. Neu and P. Kantor, 2002. Rutgers filtering work at trec 2002: Adaptive and batch. Proceedings of the Eleventh Text REtrieval Conference (TREC-11), 2002. http://comminfo.rutgers.edu/~cgal/CV%20PDFs/Trec02.pdf

Callan, J., 1996. Document filtering with inference networks. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (RDIR'96), ACM, New York, USA., pp: 262-269.

Collins-Thompson, K., P. Ogilvie, Y. Zhang and J. Callan, 2002. Information filtering, novelty detection and named-page finding. Proceedings of the 11th Text REtrieval Conference (TREC-11), 2002. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.7.9836&rep=rep1&type=pdf

Diesner, J. and K.M. Carley, 2005. Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a Novel Method for Network Text Analysis. In: Causal Mapping for Information System and Technology Research: Approaches, Advances, and Illustrations, Narayanan, V.K. and J. Deborah (Eds.). IDEA Group Publishing, Hershey, pp: 81-108.

Freeman, L., 2006. The Development of Social Network Analysis. Empirical Press, Vancouver.

Guo, X., Y. Xiang and Q. Chen, 2011. A vector space model approach to social relation extraction from text corpus. Proceedings of 8th International Conference on Fuzzy Systems and Knowledge Discovery, 2011.

Ma, L., Q. Chen, S. Ma, M. Zhang and L. Cai, 2002. Incremental learning for profile training in adaptive document filtering. Proceedings of the 11th Text REtrieval Conference (TREC-11). http://trec.nist.gov/pubs/trec11/papers/tsinghuau.filtering2.pdf

Magnini, B., M. Negri, R. Prevete and H. Tanev, 2002. A WordNet-based approach to named entities recognition. Proc. Build. Using Semantic Networks, 11: 38-44.

Mika, P., 2007. Ontologies are us: A unified model of social networks and semantics. J. Web Semantics, Elsevier, 5: 5-15.

Salton, G., A. Wong and C.S. Yang, 1975. A vector space model for automatic indexing. Commun. ACM., 18: 613-620.

Srikanth, M., X. Wu and R. Srihari, 2002. UB at TREC 11: Batch and adaptive filtering. Proceedings of the 11th Text REtrieval Conference(TREC-11), 2002. http://trec.nist.gov/pubs/trec11/papers/unybuffalo.pdf