Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 15 (2011) 5536 – 5540

Procedia Engineering

www.elsevier.com/locate/procedia

Advanced in Control Engineering and Information Science

A novel soft clustering algorithm

Ruixin Ma [a,*], Xiao Wang [a], Fancheng Meng a*

[a] *School of Software, Dalian University of Technology, Dalian 116621, China*

Abstract

Paper clustering problems in citation network is one of the hottest spots in data mining. However, traditional paper clustering algorithm stresses on the keywords analysis while ignores the "refer-to" relationship, which results in the problem of high time complexity and low accuracy. In this paper, we come up with a novel soft clustering algorithm in accordance with the complex priority and thegrouth theorem, and classify our algorithm into two steps: refer-to relationship analysis and keywords comparison. Experimental results show that our algorithm is able to greatly improve the search accuracy and efficiency.

Keywords: citation network; complex priority; soft clustering; thegrouth theorem.

1. Introduction

Paper clustering is one of the most important methods for scientific paper arrangement, management and organization, with the development of science and technology, it is becoming one of the most effective ways to mass storage, manage and inquire scientific papers. The difference between traditional hard clustering(HC) and soft clustering(SC) is that individuals in HC are supposed to belong to one cluster at one moment while SC allows individuals to belong to several ones[1]. Studies on science often involves multiple disciplines' intersection and penetration, however, HC limits the effective analysis for

* Ruixin Ma. Tel.: 15140601605; fax: 0411-87571567.
E-mail address: teacher_mrx@126.com.

1877-7058
doi:10.1016/j.proeng.2011.08.1027

interdisciplines, as a result, SC is gradually replacing HC and becoming the main stream in paper clustering algorithm.

Our algorithm is classified into two parts: basic structure mining and refined content analysis. The first procedure is inspired by the complex priority and thegrouth theorem[2], it is based on the paper's referred times to dynamically simulate the network's formation and evolution process, at the same time, carries out the data structure clustering; the purpose of the second procedure is to improve the cluster's accuracy, it in accordance with the results of the structure clustering to build a cluster theme for each cluster, and compares the similarity between the target paper and its cluster theme to refine the cluster results.

2. Soft Clustering Algorithm Based on Role Assorted Thoughts

2.1. Algorithm Thinking

The thoughts of role assorted come from advertising practitioner Fang Shouxing, he said that there is a "law of special"[3], that is to say, there are three kinds of people play very important roles during the procedure of website promotion: experts, liaisons and recommenders. This paper is based on these ideas to look for the clustering center, and optimize the search accuracy.

Although most networks are different from each other, they do have something in common: growth. They come from one vertex to thousands of vertices. With the growth of vertices, the scale of network becomes larger and larger, and eventually gets the current scale. During the procedure of growth, vertex constantly builds connections with other vertices, the complex priority says that on the process of link construction, if there are two vertices for C to choose, A has twice link numbers as much as B, the probability for C to choose A is twice as high as choosing B. This paper is inspired by the theory above, puts forward the idea that vertices with more referred times appears much earlier than vertices with few referred times.

Traditional paper clustering often ignores the citation network's characteristic of growth, and chooses fixed cycle to cluster and re-cluster while to some extent ignores the new comers, which results in the new comers can not appear at the right place as soon as possible. This algorithm constructs a time axis by counting and analyzing the referred times of each paper to simulate the formation of the citation network, at the same time to cluster papers. Experimental results show that our algorithm is helpful to improve the search accuracy.

2.2. Detailed algorithms

2.2.1 Structure analysis

Firstly, we construct a directed refer-to matrix in accordance with the reference relationships, $P_A \to P_B$ represents that paper P_A quote paper P_B .

The detailed structure dividing is described as below.

First of all, all papers are sorted in decreasing order of referred times which constitute a list L_{refer}, the clustering center set is initialized as empty; second, papers in L_{refer} are checked in turn from the beginning to the end of the list. Take out of the first member in L_{refer} and make it the first clustering center *C1*. Calculate the relatedness between *i* and *Ci* and if relatedness *(i, Ci)*$<\delta$, *i* become a new center and be added to *C*; if there are many existing clusterings, and only relatedness*(i, Ci)*$>\delta$, *i* becomes a member of *C*, else if both relatedness (*C1,i*) $>\delta$ and relatedness (C2,i) $>\delta$, we mark *i* as liaison and put *i* into the clustering with which it has closer relatedness. As the papers with high referred times are checked first,

every clustering center becomes the best one in their clustering. The clustering center leads the rest ones to locate at the best place with it[5].

We use the relatedness between paper and the existing clustering center to judge the belongings of the papers. *{C}* is the clustering center set, the relatedness between paper *i* and center *C* is calculated as formula (1) shows.

$$R_C^i = \frac{\sum_{j \in C} [E_{i,j} \times d_j]}{\sum_{k \in \{C\}} [E_{i,k} \times d_k]} \quad (1)$$

$E_{i,j}$ is a bipolar threshold[6], if there is a link from *i* to *j*, $E_{i,j} = 1$, else $E_{i,j} = 0$. The more of *i's* referred times, the better of the paper's position. Formula (1) tells us that if most of paper *i's* referred papers belong to clustering C, and then *i* also belongs to *C*. One paper can belong to several clusterings, for example, if both the relatedness between paper *i* and Clustering *A*, paper *i* and clustering *B*($R_A^i > \gamma \&\& R_B^i > \gamma$), we say that paper *i* is a liaison and belongs to both *A* and *B*.

2.2.2 Clustering refinement

The refer-to attribute is one of the papers' attributes, but not the only one, as we all know, citation network is always growing, so does the reference. Therefore, it is not suitable to use the refer-to relationship as the only standard for paper clustering. As a result, we put forward that on the basis of structure analysis to further study the paper's keywords similarity. We construct a clustering theme in accordance with the distribution of papers' keywords, and use space vector model[7] to help us with the clustering refinement.

The detailed refinement steps are as below.

Step one: Pick up the keywords in each paper and construct a paper vector;

Step two: Pick up D keywords with the highest appearance frequency in the clustering to represent the clustering's theme. The clustering theme is represented by vectors, for clustering C, $theme_C =< keyword\,1, keyword\,2, \ldots keywordD >$. $X_C = (X_{C,1}, X_{C,2}, \ldots X_{Ci,D})$ shows the appearance times of different keywords, $theme_C \cap theme_D = \varnothing$.

Step three: Calculate the similarity between each paper and their clustering theme, if the similarity between paper *i* and clustering *C* is 0, then take *i* out of *C* and look for the right clusterings for *i*.

Clustering theme is the reflection of the whole clustering; therefore, it is able to reflect the clustering's content and function. In the citation network, *D* represents the dimension of clustering theme, as for time *t*, the growth speed of clustering *C* can be represented by $V_C^t = (v_{C,1}, v_{C,2}, \ldots v_{C,D})$, it shows the situation of change for theme *C* during time *t*.

The similarity between paper i and clustering C is calculated as formula (2) shows.

$$Similarity(X_C^i) = \frac{\sum_{j=1}^{D} keyword_j \times X_{C,j} \times x_{i,j}}{\sqrt{\sum_{j=1}^{D} X_{C,j}{}^2}} \quad (2)$$

$keyword_j$ is a bipolar threshold, if both paper i and theme C have $keyword_j$, $keyword_j = 1$, else $keyword_j = 0$. According to formula (2) to judge the belongings of each paper, if similarity between paper i and theme C is 0, then look for the right clustering for i and re-cluster the clustering.

For new comers, we firstly compare the relatedness between it and the existing clusterings then calculate the similarities to find the best locations for it. At the same time, clustering theme is changing with the partition of new comers; formula (3) shows the change of theme *C*. Besides, formula (3) shows the whole clustering theme with uncertain length while the refined theme length is D.

$$X_C^{T+1} = X_C^T + R_i^{T+1} \tag{3}$$

3. Analysis and Comparison of Experimental Results

To test the efficiency of our soft clustering, we separately download 100 papers in computer, math, physics, biology and politics as the test data, to analyze the refer-to relationship and keywords of all those papers. This paper adopt the usual assessment method: recall rate and precision[9] to compare the performance of our algorithm and S2FCM[10], improved S2FCM[11]. Table 1 shows the call rate comparison and table 2 shows the precision comparison.

Compare table 1 and table 2, we find that our algorithm behaves much better than S2FCM and the improved S2FCM both in recall rate and precision. It is worth notice that computer science is some kind of cross-discipline; that is to say, papers about computer science are also connected to math, physics or some other frontiers. Therefore, the precision is to some extend limited. In contrast, papers in politics are easy to find, that's because politics has hardly any cross point with physics, math and biology. As results, the clustering about politics becomes a high cohesion and weak coupling community.

Table 1. Comparison of recall rate

	Math	physics	politics	biology	computer
S2FCM	79.4%	81.3%	83.5%	76.2%	72.5%
Improved S2FCM	84.6%	84.2%	88.7%	82.4%	79.6%
Our algorithm	91.2%	92.1%	95.3%	91.3%	87.3%

Table 2. Comparison of precision

	Math	physics	politics	biology	computer
S2FCM	82.5%	83.4%	88.6%	81.4%	76.8%
Improved S2FCM	86.4%	87.6%	90.7%	84.8%	81.8%
Our algorithm	93.7%	94.5%	96.8%	91.5%	86.6%

4. Conclusion

This paper comes up with a novel soft clustering algorithm, by analyzing the refer-to relationship to divide roles, and then effectively search the liaison-papers. Experimental results show that the structure analysis decreases the time complexity of clustering algorithm; the keywords analysis efficiently contains the topic drift phenomenon. Therefore, our algorithm is able to greatly improve the search accuracy and efficiency with high practical values.

References

[1] Meng Haitao, Chen Xiaorong. Fuzzy similarity based document clustering algorithm [J]. Journl of Guizhou University(Natural science). 2007, 24(2): 175-178.

[2] G.Bianconi, A.Barabsi. Competition and Multiscaling in Evolving Networks[J]. Europhysics Letters. 2001,5:436-442.

[3] Wang Xian, Xie Chi, Rong Xue, Fan Wen. “Research of the Present Situation of Operation and the Future Tendency of SNS Website.” http://media.people.com.cn/GB/22114/119489/140165/8454258.html（2009）

[4] Albert L.Balabasi. Linked[M]. Chang Sha: Hunan Science and Technology Press,2007.

[5] X. Li. “Adaptively Choosing Neighborhood Bests using Species in A Particle Swarm Optimizer for Multimodal Function Optimization,” Proceedings of Genetic and Evolutionary Computation Conference. 2004, 105-116.

[6] FAN Cong-xian XU Ting-rong “Research and Improved Algorithm of HITS Based on Web Structure Mining.” Computer Information. 2010, 26:160-162.

[7] DUAN Huai-chuan, HU Ping. Improved PageRank algorithm based on topic character and time factor [J]. Computer Engineering and Design. 2010,31(4):866-868.

[8] Hu Jian, Dong Yuehua, Yang Bingru. Community Structure Discovery Algorithm in Large and Complex Network [J]. Computer Engineering. 2008,34(19): 92-93.

[9] FAN Cong-xian, XU Ting-rong, FAN Qiang-xian. Research and Improvement Algorithm of HITS Based on Web Structure Mining [J]. Micro Computer Information, 2010, 26（1-3）:160:162.

[10] Pei Jihong, Fan Jiulun, Xie Weixin. A New Effective Soft Clustering Method: Sectional Set Fuzzy C-Means(S2FCM) Clustering [J]. ACTA ELECTRONICA SINICA. 1998, 26(2): 83-86.

[11] Bai Sixue，Lu Ping. A Feature Selection Method Based On Text Classify[J]. Journal of Nanchang University(Engineering & Technology). 2008, 30(1): 87-90.