2012 International Conference on Computing, Measurement, Control and Sensor Network

Paper Classification by Topic Grouping in Citation Networks

Yi-Jen Su
Dept. of Computer Science and Information
Engineering, Shu-Te University
Kaohsiung City, Taiwan
iansu@stu.edu.tw

Jian-Cheng Wun, Wei-Lin Hsu, Yue-Qun Chen
Dept. of Computer Science and Information
Engineering, Shu-Te University
Kaohsiung City, Taiwan
{s11639116, s97113258, s11639113}@stu.edu.tw

Abstract—**The enormous popularity of Web 2.0 social network services has led to much research on social network analysis (SNA). These studies focus on analyzing the complex interactive activities between users in the world of virtual networks. SNA has shown great potential in automatic document classification, especially in identifying citation networks of research papers and the references among them. This research adopts the Clique Percolation Method (CPM) to identify all overlapping subgroups in a citation network. In the grouping process, research papers with similar topics will be grouped into the same topic group. Two papers are regarded as having a relationship when the common citation rate between them is higher than the threshold. A modified TF-IDF calculates the weight of each keyword in the topic groups. The keyword-weight vector represents the main features of each group, while the category of a new-coming document is determined by a novel similarity function. All the papers under study are collected from the journal IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) published from 1979 to 2011.**

***Keywords-** Social Network Analysis; Citation Network; CPM; TF-IDF*

I. Introduction

In an age of information overload, effective information management has become a major task for users confronted with large amounts of data. Automatic document classification sorts data primarily by means of topic categories, content areas, or other features that users care about. An efficient document classification system can help locate and retrieve pertinent data from the repository and facilitate information reuse. In the journal or conference review process, for example, a document pre-classification process can be instrumental in matching submitted papers to suitable reviewers, and by doing so, promote objectivity and fairness in the review process. This research project is dedicated to developing a highly efficient and reliable automatic document classification method.

All citation network data analyzed in this research are derived from papers published from 1979 to 2011 in the journal IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). This study employs a web agent to extract information of the TPAMI papers that includes titles, keywords and reference lists. The citation network consists of papers (nodes) and undirected relationships (edges) with similar research focuses. The extent to which two papers have similar focuses depends on whether the common citation rate exceeds the threshold.

In view of the possibility that a research paper might involve several focuses, this study adopts an overlapping community identification algorithm, CPM[3], to identify all papers belonging to the same citation networks due to sharing similar topics. The TF-IDF (term frequency–inverse document frequency) [4] calculates keyword weight by the number of occurrences of each keyword in a topic community. The weight indicates a keyword's importance while the keyword-weight vector represents the specific features of each topic group.

A novel similarity function is devised to calculate the similarity between a paper not yet categorized and all the topic groups. The paper will be classified into the same topic group with the highest similarity. Both a partition grouping method and an overlapping grouping method will be used to discover topic groups from multi-scale citation networks. Finally, the grouping quality will be evaluated by the Q value and the classification rate of new coming papers.

II. Background

A. Social Network Analysis

The concept of social network analysis [1][2] was first proposed by Georg Simmel[5] in 1908. In 1934, Moreno created the idea of sociograms, in which *nodes* represent people whereas *edges* represent relationships among these people. The Sociometric technique maps social dynamics and indentifies structural properties of different social networks in order to assess social cohesion and group pressure. In recent years, social network analysis has been widely applied to different research domains, such as anthropology, psychology, and sociology. Due to big advances in computer technologies, SNA can be adopted to analyze complex networks. There are some famous SNA software systems such as UCINET[6] and Pajek[7].

B. Citation Analysis

Paper citation plays an important role in academic research. Eugene Garfield created the first citation index in 1953 that allowed scholars to quickly and accurately find references. In 1955, he proposed using cited references to

978-0-7695-4738-1/12 $26.00 © 2012 IEEE
DOI 10.1109/CMCSN.2012.53

track scientific developments [8]. This study aims to promote citation analysis by developing an automatic document classification process based on citation network features.

C. *Overlapping Communities*

One of the most well-known methods for discovering communities [9] in a complex network is the GN algorithm [10]. The algorithm iteratively removes the edge with the largest betweenness to partition the network into smaller independent connected graphs (subgroups). The grouping quality will be evaluated by the modularity Q value [11]: the higher the Q value, the better the overall grouping quality. Thus, after the edge removing process is completed, the grouping result with the highest Q value will be selected. The major drawback of the GN algorithm is that it is extremely time-consuming. Every time an edge is removed, the algorithm iteratively computes the betweenness values of all remaining edges.

Suppose there are two communities, c_i and c_j, in an undirected graph. The Q value can be derived by (1).

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \qquad (1)$$

When $A_{ij}=1$, there is a similar-topic relationship existing between paper i and paper j. But $A_{ij}=0$ indicates no relationship between the two papers. m is the total number of edges in the network. k_i and k_j represent the degrees of node i and node j, respectively. When node i and node j belong to the same community, $\delta(c_i,c_j)$ equals 1. If not, $\delta(c_i,c_j)$ is 0.

The Clique Percolation Method (CPM) [3] searches for densely connected node sets called cliques to identify overlapping communities in a complex network. The method incrementally merges all connected k-clique into Maximal Subgraphs. When two adjacent 3-cliques are merged, there will be 2 nodes belonging to two 3-cliques at the same time.

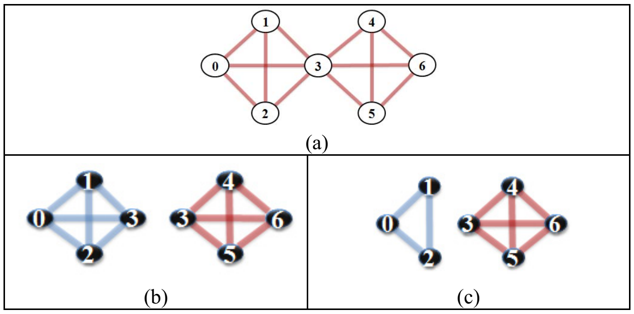

Figure 1. Grouping examples. (a) 7 nodes in the network (b) result of the CPM algorithm (C) result of the GN algorithm.

The CPM algorithm is an overlapping community [12] identification method to group paper nodes in a citation network. It is adopted because it satisfies the hypothesis that the content of each research paper will cover several topics. In contrast, a partition method like the GN algorithm, when adapted to do community finding, does not lend support to this multi-focus hypothesis. Moreover, because CPM does not need to compute the betweenness of each edge iteratively, it substantially speeds up the grouping process. In Figure 1(a), the result of the CPM algorithm is contrasted with that of the GN algorithm. In Figure 1(b), CPM identifies two overlapping groups {0,1,2,3} and {3,4,5,6} by a 3-clique with a Q value of about 0.5. The node {3} is present in two groups. In Figure 1(c), the group finding results are {0,1,2} and {3,4,5,6} with a Q value of about 0.219. It is clear that CPM results in better grouping quality than GN.

D. *TF-IDF*

TF-IDF [4] calculates the weight of a word or a term in a corpus, as shown in (2). The importance of each term to a document corresponds to the term frequency in the document, but is inversely proportional to the number of documents in which the term occurs. Generally, most search engines use this method to measure the similarity between documents and keyed-in-words when providing search services. This research modifies the TF-IDF approach for the citation network domain in (3),

$$tfidf_{ij} = tf_{ij} \times idf_i \qquad (2)$$

where tf_{ij} is the occurrences of term i in document j, and idf_i is the number of documents in the corpus divided by the number of documents having term i.

III. Research Menhod

The processes of automatic document classification are illustrated in Figure 2. First, a web spider is in charge of extracting paper-related information from the TPAMI journal website. The citation network is constructed by such information as paper titles, authors, and citation lists. When the common citation rate of two papers exceeds the default value of threshold, the two papers are regarded as having a citation relationship. Then the CPM algorithm is applied to discover overlapping topic communities in the network. The TF-IDF method computes the weight of the keywords in each topic group discovered in the previous step. Finally, a similarity function measures the similarity between a new document and the topic communities. The document will belong to the topic community with the closest similarity value.

In (3) OLp_ip_j represents the number of overlapping citations between two papers P_i and P_j. $|R_{P_i}|$ is the number of all citations in P_i. To enhance the information quality of the citation network, the Common Citation Rate (CCR) between two papers needs to be higher than the the threshold TH_{CCR}.

$$CCR = \frac{\left(OLp_ip_j\right)^2}{\left|R_{P_i}\right| \times \left|R_{P_j}\right|} \qquad (3)$$

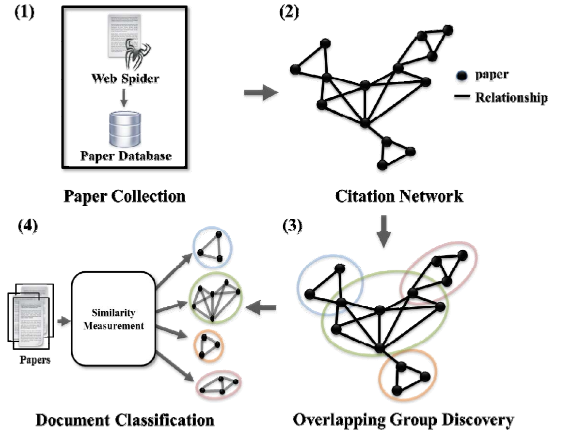

Figure 2. The operation flow of automatic document classification

This research modifies the original TF-IDF equation for the research domain of citation networks, as shown in (4).

$$tfidf_{ij} = tf_{ij} \times idf_i = \frac{n_{ij}}{\sum_{q=0}^{k-1} n_{qj}} \times \log \frac{|C|}{|\{j : kw_i \in C_j\}|} \quad (4)$$

n_{ij} is the occurrences of keyword kw_i in the community C_j. Σn_{qj} represents the total occurrences of K keywords in community C_j. |C| is the number of discovered communities. $|\{j{:}kw_i{\in}C_j\}|$ is the total number of keywords present in the communities.

Figure 3 shows an example of the TF-IDF calculation. There are three topic groups C_1, C_2, and C_3, with keywords and their occurrences indicated in Figure 3(a). The keyword-weight vectors of the three groups are $C_1(kw_A,kw_B,kw_C)=\{0.2041,0.05,0\}$,$C_2(kw_B,kw_C)=\{0.132,0\}$, $C_3(kw_C,kw_D)=\{0,0.1908\}$, respectively, as shown in Figure 3(b). Suppose there is a new paper P_1 with keywords $\{kw_A,kw_B,kw_D\}$, and its similarity with the three topic groups are $Sim(C_1,P_1)$ =0.2541, $Sim(C_2,P_1)$ =0.132, and $Sim(C_3,P_1$ =0.1908. Because $Sim(C_1,P_1)$ is the highest similarity value of the three, $P_1$is classified into group C_1.

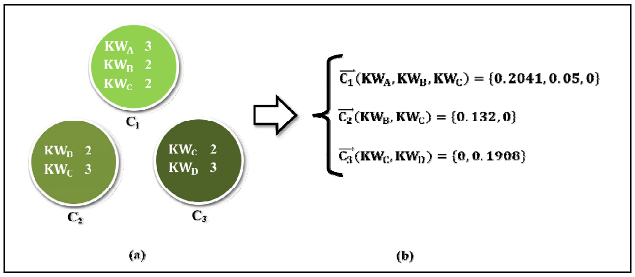
KW_A 3
KW_B 2
KW_C 2
C_1
KW_D 2
KW_C 3
C_2
KW_C 2
KW_D 3
C_3
(a)
$\vec{C_1}(KW_A, KW_B, KW_C) = \{0.2041, 0.05, 0\}$
$\vec{C_2}(KW_B, KW_C) = \{0.132, 0\}$
$\vec{C_3}(KW_C, KW_D) = \{0, 0.1908\}$
(b)

Figure 3. Keyword weight vectors generated by TF-IDF computation

IV. Experiment Results

The dataset in this research contains 1,153 papers published from 1979 to 2011 in the TPAMI and made available on the journal website. In the citation network, the nodes represent the papers while edges stand for citation relationships, i.e., indicating two papers sharing similar research topics. A citation relationship exists when the common citation rate is higher than the default threshold. ◦

A. An Experiment with a Single Journal

In order to compare the grouping results of the GN algorithm and the CPM algorithm, this research uses different numbers of papers to construct the citation network. There are ten test datasets of various sizes:100, 200,…,1000. In this experiment two types of algorithms are compared: GN is a standard partition-based grouping method whereas CPM is a representive algorithm for overlapping groups. The smallest modularity in CPM is a 3-clique; that is, the smallest group needs to have at least 3 nodes.

The grouping results of GN and CPM are compared in Figure 4. CMP discovers more groups than GN when the number of papers exceeds 600. An interesting result is that the topic groups discovered by CMP stays at 35 when the number of papers exceeds 700. With only 100 papers, there are not enough papers to form groups based on similar topics. This is true for both algorithms.

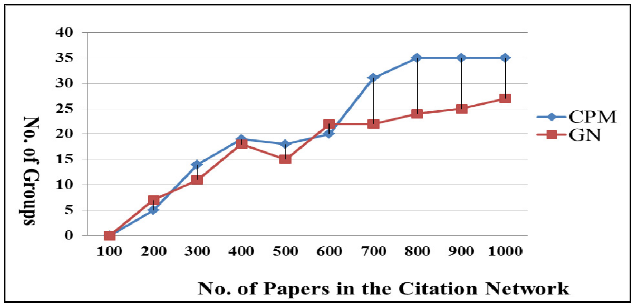

Figure 4. Grouping results of GN and CPM with a single journal

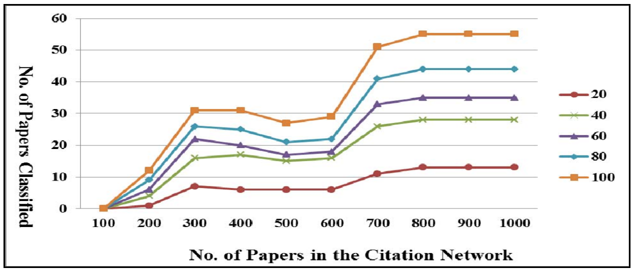

Figure 5. Document classification results of CPM with a single journal

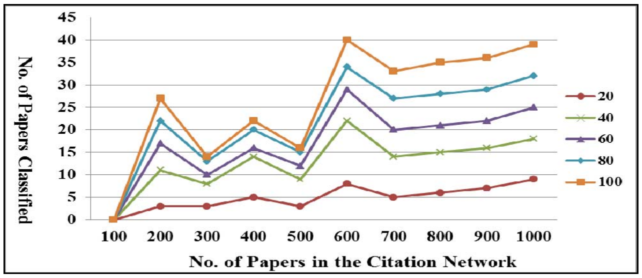

Figure 6. Document classification results of GN with a single journal

B. *An Experiment with Multi-journals*

In the experiment with multi-journals, the papers selected to compose the citation network are collected from three journals. These three journals are IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Computers, and IEEE Transactions on Parallel and Distributed Systems. To avoid the situation of not having enough inter-connected papers to form topic groups, the number of papers collected from three journals starts from 600 to 2,100. These papers are evenly distributed in the three journals.

The number of topic groups identified by GN in the three journals is consistently larger than that of CPM. This result is totally different from that of a single journal in Figure 4. However, when the modularity Q value is used to evaluate grouping quality, the CPM result with datasets from the three journals has better quality than that of GN. This experiment result is almost the same as that of a single journal.

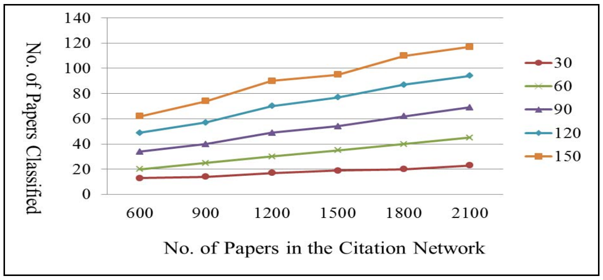

Figure 7. Document classification results of CPM with multi-journals

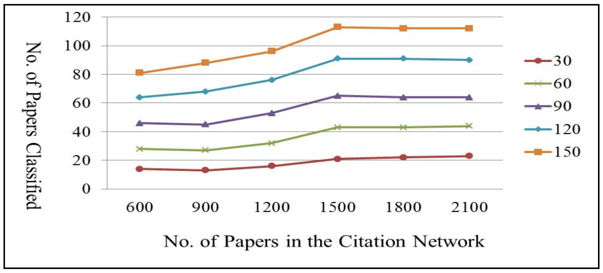

Figure 8. Document classification result of GN with multi-journals

Figures 7 and Figure 8 show the automatic document classification results of, respectively, CPM and GN with three journals. The papers in the test sets are randomly and evenly selected from the three journals. The number of papers in the test sets ranges from 30,60…to 150. In the two-stage experiment with multi-journals, the CPM classification rate is lower than that of GN when the citation network consists of 1,500 papers. CPM however performs better than GN when there are 1,800 and 2,000 papers. Significantly, while the GN classification rate remains relatively stable after the number of papers in the citation network reaches 1,500, the CPM classification rate rises continually. Based on the experiment result, the predominance of CPM algorithm will be more striking as the size of the citation network increases.

Conclusion

This research highlights the importance of automatic document classification in information management. When this technology becomes more mature, it can help to substantially save time and energy handling massive amounts of information. Especially in the academia, an effective automatic document classification system not only can assist scholars in speeding up the search for the most representative and influential works in a special research domain, but can also help journal publishers and conference organizers to enhance the efficiency of the review process.

This research includes two sets of experiments, one on a single-journal citation network while the other on a multi-journal citation network. The CPM algorithm and the GN algorithm are adopted to determine whether the overlapping grouping method or the partition-based grouping method is more effective. Both in the single-journal or multi-journal experiments, CPM has better performance in terms of grouping quality and the classification rates. Especially when the number of papers in the test sets grows larger, the difference in the experiment results of the two grouping methods becomes more obvious. To develop more effective grouping algorithms, future research needs to apply the overlapping grouping method to other domains, and explore alternative grouping methods.

References

[1] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann Publishers, Massachusetts, 2011.

[2] J. Scott, Social Network Analysis: A Handbook, Sage Publications, London, 2000.

[3] G, Palla, I. Derenyi, and T. Vicsek, “The Critical Point of k-Clique Percolation in the Erdos–Renyi Graph,” J. Stat. Phys., vol. 128, July 2007, pp. 219–227.

[4] H.C. Wu, R.W.P. Luk, K.F. Wong, and K.L. Kwok, "Interpreting tf–idf term weights as making relevance decisions," ACM Trans. Inf. Syst., vol. 26, no.3, 2008, pp. 1–37.

[5] J. Scott, “Social network analysis: developments, advances, and prospects,” Soc. Netw. Anal. Min., vol. 1, no. 1, 2011, pp. 21–26.

[6] UCINET, https://sites.google.com/site/ucinetsoftware/home

[7] Pajek, http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[8] E. Garfield, "Citation Indexes in Sociological and Historical Research," Am. Docum., vol. 14, 1963, pp. 289–291.

[9] N. Gulbahce and S. Lehmann, “The art of community detection,” BioEssays, vol. 30, 2008, pp. 934–938.

[10] M. Girvan and M.E.J. Newman, “Community structure in social and biological networks,” Proc. Natl. Acad. Sci. USA 99, vol. 99, 2002, pp. 7821–7826.

[11] M.E.J. Newman, “Modularity and community structure in networks,” Proc. Natl. Acad. Sci. USA 103, vol. 109, no. 12, 2006, pp. 8577–8582.

[12] G. Palla,“Uncovering the overlapping community structure of complex networks in nature and society,“ Nature, vol. 435, 2005, pp. 814–818.