2009 IEEE International Conference on Data Mining Workshops

Hybrid Clustering by Integrating Text and Citation based Graphs in Journal Database Analysis

Xinhai Liu
K.U. Leuven, Dept. of Electrical Engineering, ESAT-SCD
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
WUST, CISE& ERCMAMT, 430081, Wuhan, China
xinhai.liu@esat.kuleuven.be

Shi Yu Yves Moreau Frizo Janssens Bart De Moor
K.U. Leuven, Dept. of Electrical Engineering ESAT-SCD
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{shi.yu, yves.moreau,frizo.janssens,bart.demoor}@esat.kuleuven.be

Wolfgang Glänzel
K.U. Leuven, Center for R&D Monitoring, Dept.MSI
Dekenstraat 2, B-3000 Leuven, Belgium
Wolfgang.Glanzel@econ.kuleuven.ac.be

Abstract

We propose a hybrid clustering strategy by integrating heterogeneous information sources as graphs. The hybrid clustering method is extended on the basis of modularity based Louvain method. We introduce two different approaches, graph coupling and graph fusion. The weights of these combined graphs are optimized with the criterion of maximizing the Average Normalized Mutual Information(ANMI). The methods are applied to obtain structural mapping of large scale Web of Science (WoS) journal database by integrating attribute based textual information and relation based citation information. From the experimental, the proposed graph combination scheme is compared with individual graph clustering, spectral clustering and Vector Space Model(VSM) based clustering methods.

1. Introduction

Grouping journals by clustering is a fundamental task for journal database analysis. There are two kinds of heterogeneous information sources in a journal database: textual content and citation link. However, both information sources are closely correlated and supplement each other. Clustering solely based on either textual or citation information might have some known shortcomings: on one hand, textual similarities are often affected by the ambiguities of vocabularies; on the other hand, citation based links might be biased by the personal relations in scientific research. So hybrid clustering by integrating textual and citation information is a promising approach. Hybrid clustering has been implemented in Vector Space Model(VSM) and gained good result [9]. But due to the heavy computation and memory requirement of VSM based methods, it is hard to extend to large scale journal database and immediate clustering task.

On the hand, many graph partition algorithms have appeared, such as spectral clustering and modularity optimization based algorithms. They are simple to implement and often outperform the traditional clustering such as KMeans algorithm. Louvain method, the fast approximation algorithm of modularity optimization, has some properties: efficient to implement and scalable to the huge database[2]. Since there is huge citation link data in a journal database, graph partition method which fits spare link features is a convenient way. Meanwhile in above graph partition methods, they often consider only the link structure and ignore attribute similarities[6]. However, if each journal is taken as a vertex and the textual similarity between two journal as their edge strength, graph can also be modeled from textual attribute.

So based on Louvain method, we propose a new hybrid clustering strategy to integrate citation relation and textual attribute from a graph view. We introduce a computation framework of combining multiple graphs by two basic approaches: graph coupling and graph fusion. Since various graphs have different relative importance, we present a novel weighting scheme based on information theory named Adaptive Mutual Information Weighting(AMIW).

2. Related work

This work is first related to a family of work on graph partition based on modularity optimization. Newman[11] proposes an effective graph partition method by optimizing modularity. Some fast approximate modularity optimization methods have appeared[2][3]. But those modularity based graph partition methods usually focus on link structure and ignore attribute similarities[6].

This work also relates to the category of work that multiple graphs are combined for partition [4] [13]. Whereas these clusterings are not based on modularity method and the heavy computation on optimization makes them not fit a large scale database.

978-0-7695-3902-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ICDMW.2009.65

IEEE computer society

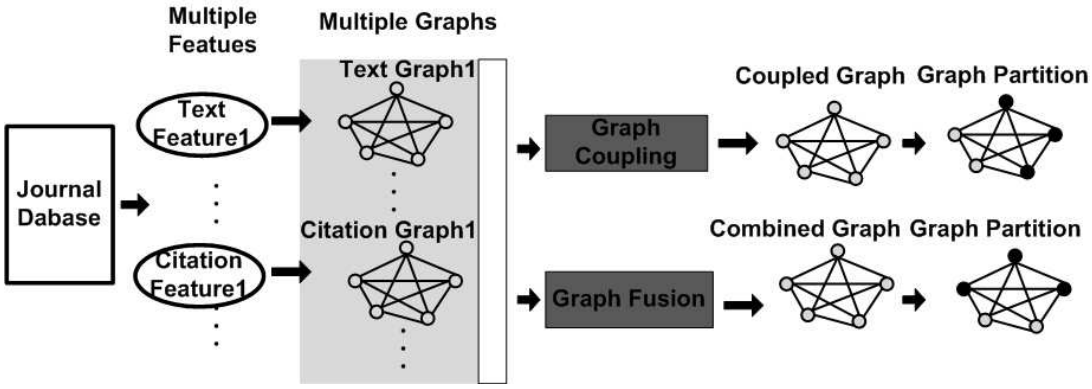

Figure 1. Hybrid clustering from a graph view

Finally, this work shares the idea of related work on combining text mining data with bibliometric data to map the structure of the scientific publication database[8][9]. But these clustering work was implemented in Vector Space Model(VSM).

Several graphs are extracted from heterogeneous information sources and multiple graphs are combined by graph coupling and graph fusion as shown in Figure 1. The difference between graph coupling and graph fusion is that during graph coupling, only the link relation information of citation is utilized while during graph fusion, both link relation and link information of citation are utilized together.

3. Modularity based Louvain method

Modularity is a benefit function used in the analysis of networks or graphs [11]:

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

where A_{ij} represents the weight of the edge between vertex i and vertex j; $k_i = \sum_j A_{ij}$ is the sum of the weights of the edges attached to vertex i; c_i is the community to which vertex i belongs; the δ function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise and $m = \frac{1}{2}\sum_{ij} A_{ij}$. The fast approximation algorithm on large graphs was proposed by Clauset *et al.*[3] by recurrently merging communities that optimize the production of modularity:

$$\triangle Q = \begin{cases} \frac{A_{ij}}{2m} - \frac{k_i k_j}{(2m)^2} & \text{if } i, j \text{ are connected} \\ 0 & \text{otherwise} \end{cases}. \qquad (2)$$

But this greedy algorithm tends to produce super-communities that contain a large fraction of the vertices. To improve the Clauset's algorithm, Blondel *et al.* [2] propose Louvain method by balancing optimization of modularity with running time and sensitivity to local structure.

The efficiency of Louvain method is obtained by extending Equation (2) to Equation (3) as follow so that the gain in modularity $\triangle Q$ obtained by moving an isolated node i into a community C can be easily computed.

$$\triangle Q = [\frac{k_{i,in}}{2m} - \frac{\sum_{tot} k_i}{(2m)^2}] = \frac{1}{2}[\frac{\sum_{in}+2k_{i,in}}{2m} - (\frac{\sum_{tot}+k_i}{2m})^2] - [\frac{\sum_{in}}{2m} - (\frac{\sum_{tot}}{2m})^2 - (\frac{k_i}{2m})^2] \qquad (3)$$

where $\sum_{in}$ is the sum of the weights of the links inside C; $\sum_{tot}$ is the sum of the weights of the links incident to nodes in C; k_i is the sum of the weights of the links incident to node i; $k_{i,in}$ is the sum of the weights of the links from i to nodes in C and m is the sum of the weights of all the links in the network.

4. Graph coupling by integrating attribute and relation

The citation relation data could be naturally described with a graph model. At the same time, if each journal is taken as a vertex and the textual similarity between two different journals as their edge strength, a graph could be modeled from textual attributes of journals as well.

We investigate the citations among the selected publications in three aspects: cross-citation(CRC), co-citation(COC), bibliographic coupling(BGC)[8].

We only consider citations between papers from 2002 till 2006 and aggregate all paper-level citations into journal-by-journal citations. Based on these relational data, three weighted and undirected graphs are formed: CRC, COC and BGC journal graph. Given a link graph $G = (V, E)$,

we define the adjacency matrix A of that graph to be

$$A_{i,j} = \begin{cases} W_{ij} & \text{if } e_{ij} \in E \text{ or } e_{ji} \in E \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where W_{ij} denotes the edge (link) strength between vertex i and vertex j, and it refers to the citation frequency between two journals.

From the journal database, we also generate 3 text features: TF, IDF and TF-IDF. Then we can model 3 text based graphs by their textual similarity, which can be described as projecting a text world into a graph world.

When modeling a graph from text attribute, a problem arises: text similarities(edge strength) among different journals are very dense due to the high dimension of text feature (669,860). As drawn by Figure 2, every vertex pair has nonzero edge in the original text graph which will bring heavy computation load. So if we want the partition algorithms work normally on these text based graphs, we need make their link degree sparse enough in advance. According to [10], there are two ways to make dense pairwise similarities sparse: (1) ε-distance neighborhood method; (2) k-nearest neighbor method. But how to determine the parameters ε or k is an open issue. Usually it relies on the empirical knowledge of different applications.

However, due to the inner tie between citation information and textual information, some research has applied text based similarity to modulate the edge strength of citation graph [4]. So here, we can also use the citation relation constraint to make text based graphs sparse. For instance: if there is a citation link between two journals, we take the textual similarity as the edge strength, otherwise, we set the edge strength to zero. We call this kind of combination as graph coupling. Based on above adjacency matrix, we define relationship matrix R to describe the citation relationship among journals.

$$R_{i,j} = \begin{cases} 1 & \text{if } A_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Here we also define R^{crc}, R^{coc} and R^{bgc} to represent the concerned citation relations.

Given S, the textual similarity matrix before coupling, A, the adjacency matrix of text based graph after coupling, R, the relationship matrix from above citation graph, then we use the logical operation of "AND"($\wedge$) to couple the text based graph with citation relation.

$$A = R \wedge S \quad (6)$$

$$(R \wedge S)_{ij} = \begin{cases} S_{ij} & \text{if } R_{ij} = 1 \\ 0 & \text{if } R_{ij} = 0 \end{cases}. \quad (7)$$

Since citation link among various journals is very sparse, solely relying on individual relation will cause the overly link-centered problem[13], that is, many other useful similarities between pairwise vertices would be ignored. The vertex link degree of a text based graph coupled by one relation is shown in Figure 2. As it is shown, compared with that of original text graph, the link degree of coupled graph is rather low. So to overcome this problem, we employ multiple citation relations for graph coupling. Since

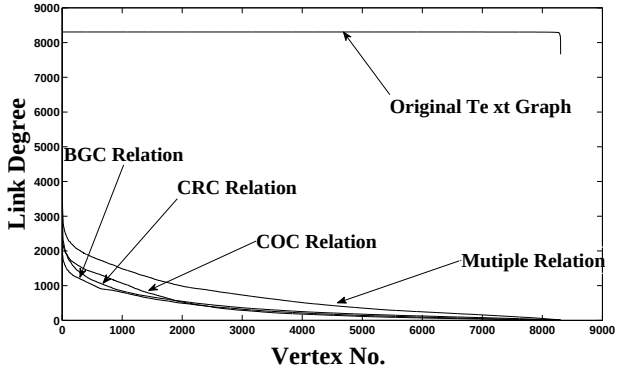

Figure 2. Link degree of vertices in text graph with various graph coupling schemes.

these citation relations are all sparse, we use "OR"($\vee$) operation among different relationship matrices, which means if any citation relationship exits between two vertices, there will be a relationship:

$$R^C = R^{crc} \vee R^{coc} \vee R^{cgc} \quad (8)$$

$$R^C_{ij} = \begin{cases} 1 & \text{if } R^{crc}_{ij} = 1 \text{ or } R^{coc}_{ij} = 1 \text{ or } R^{bgc}_{ij} = 1 \\ 0 & \text{otherwise} \end{cases}. \quad (9)$$

As plotted in Figure 2, the link degree of text based graph would be more dense than that with only one relation but still more sparse than that of original text graph.

5. Multiple graphs fusion for partition

Since there are two types of graphs existing: text based graph and citation based graph, we can take local modularity from both graphs into account to determine whether the two vertices should be merged:

$$\triangle Q = f(\triangle Q^{(T)}, \triangle Q^{(C)}) \quad (10)$$

where $\triangle Q^{(T)}$ and $\triangle Q^{(C)}$ represent the gain of modularity of vertex i in text and citation based graph respectively, $f(\cdot)$

is the graph fusion function. Here we only focus on the linear combination case. So the combined modularity gain could be reformulated as:

$$\triangle Q = w \triangle Q^{(T)} + (1-w) \triangle Q^{(C)} \tag{11}$$

where w represents the weight with a value between 0 and 1. We can also extend this graph fusion to several graphs combination, given N graphs

$$\triangle Q = w^{(1)} \triangle Q^{(1)} + ... + w^{(i)} \triangle Q^{(i)} + ... + w^{(N)} \triangle Q^{(N)} \tag{12}$$

$$\sum_i w^{(i)} = 1 \tag{13}$$

Putting Equation (12) in Equation (2), it could be formulated directly as the linear combination of adjacency matrices of various graphs. Due to the different measurements of these adjacency matrices, the normalized combination is defined as:

$$A = w^{(1)} \frac{A^{(1)}}{\| A^{(1)} \|_2} + ... + w^{(i)} \frac{A^{(i)}}{\| A^{(i)} \|_2} + ... + w^{(N)} \frac{A^{(N)}}{\| A^{(N)} \|_2} \tag{14}$$

In order to get the weights in Equation(14), based on information theory, we propose a weighting scheme named Adaptive Mutual Information Weighting(AMIW). Given N graphs, with the ith graph partition $P^{(i)}$, let the set of graph partitions $\{P^{(i)} \mid i \in \{1, \ldots, N\}\}$ be denoted by P and the optimal partition on the combined graph denoted $P^{(opt)}$, the Average Normalized Mutual Information(ANMI) between the optimal partition $P^{(opt)}$ and the set P is defined as:

$$J(P^{(opt)}, P) = \frac{1}{N} \sum_{i=1}^{N} M(P^{(opt)}, P^{(i)}) \tag{15}$$

It measures the common information shared between the optimal partition and the set of partitions. Where $M(\cdot)$ is Normalized Mutual Information(NMI)[12], which is used to measure the common information shared by two partitions. Our objective function can be formulated as finding an optimal partition with the maximum ANMI by optimizing the weight of each graph for combination as follow:

$$\arg\max_{w, P^{(opt)}} J(P^{(opt)}, P) \tag{16}$$

where the optimization parameter $w = \{w^{(1)}, ..., w^{(i)}, ..., w^{(N)}\}$ is the weight set of individual graphs. The pseudo code of our AMIW scheme is presented as follow.

Algorithm 5.1: AMIW($A^{(1)}, A^{(2)}, ..., A^{(N)}, C$)

All partitions are based on Louvain method
$P = \{P^{(1)}, ..., P^{(i)}, ..., P^{(N)}\} \leftarrow$ PARTITION($A^{(i)}, C$), $i \in 1, ..., N$
$A_0 = \sum_{i=1}^{N} \frac{1}{N} A^{(i)}$,
$\hat{P}_0 \leftarrow$ PARTITION(A_0, C),
comment: obtain an initial partition

while $<!convergence>$
do $\begin{cases} step1 : w_j^{(i)} \leftarrow \text{WEIGHING}(P^{(i)}, \hat{P}_j),\ i \in 1, ..., N \\ step2 : A_j = \sum_{i=1}^{N} w_j^{(i)} A^{(i)}, \\ step3 : \hat{P}_j \leftarrow \text{PARTITION}(A_j, C), \\ step4 : J(P, \hat{P}_j) \end{cases}$
comment: j is the counter of iteration

return $(\hat{P}_{j+1})$

In the later experimental Section, several empirical tests on the journal database have verified the converge of our AMIW weighting scheme as illustrated in Figure 3.

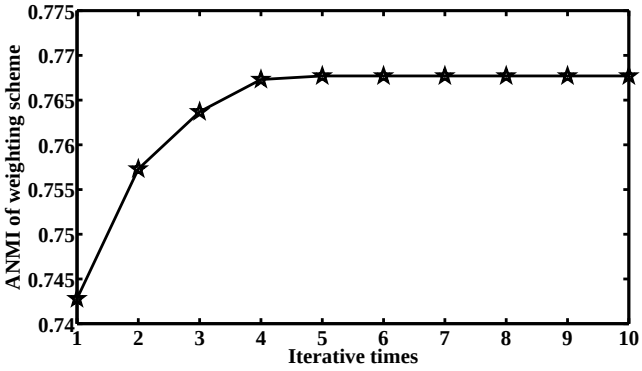

Figure 3. Iteration of the weighting scheme of AMIW for graph fusion. It is carried out by combining all graphs from our journal database with adaptive mutual information weighting(AMIW). After 5 times iteration, the ANMI value began to converge.

6. Experiment

6.1. Journal database

The journal database is from Web of Science(WoS) and after some text mining and citation analysis processing [9], we acquired 8,305 types of journals, which are assigned to 22-field categories according to Essential Science Indicators(ESI) [1] classification by Thomson Scientific[1]. We summarize our abbreviations in Table 1.

Abbreviation name	Full name
CRC	CRoss-Citation
COC	CO-Citation
BGC	BiblioGraphic-Coupling
ARI	Adjusted Rand Index
NMI	Normalized Mutual Information
KM	Kernel KMeans clustering
WL	hierarchical clustering with Ward Linkage
NJW	spectral clustering based on NJW
LM	Louvain Method
VSM	Vector Space Model
AMIW	Adaptive Mutual Information Weighting

Table 1. Description of abbreviations

We refer ESI subject classification to obtain the ground-truth labels of journal assignment. We fix the cluster number as equal to the number of standard category of ESI(22) to get a more precise partition evaluation.

6.2. Experiment analysis

In this part, we examine the performance of our graph coupling as well as graph fusion based partition schemes and compare them with other partition methods. We also investigate the validity of our AMIW scheme.

The quality of clustering result is assessed by three validation indices: NMI [12], ARI [5] and Modularity[11]. We compare our graph partition methods based on Louvain method(LM) with other three partition algorithms as follow: KM[9], WL [7] and NJW[10].

Firstly, our AMIW scheme obtains the best partition performance as shown in Table 3. The three validation measures all verify the effectiveness of our weighting scheme.

Secondly, graph coupling also accomplishes a nice performance as shown in Table 2. Although it is still not compatible to our weighted graph fusion scheme, it is easier to implement without iterative optimization.

Thirdly, as shown in Table 3, in graph fusion, we also compare other weighting schemes. Our AMIW scheme is still superior to modularity weighting scheme. That is because modularity solely measures the partition quality only based on independent graph itself while mutual information measure evaluates partition quality according to the informational relationship among graphs.

7. Conclusion

We introduce a computation framework of multiple graphs partition. We also present a weighting schemes named AMIW to combine multiple graphs for partition. Then we apply our methods to analyze the large scale journal database. The excellent partition performance of our method has been demonstrated by various validation measures. Moreover, this efficient scheme is a very general framework that enables a wide application with heterogeneous data sources, for instance, bio-information analysis and web searching.

8 Acknowledgement

Research supported by China Scholarship Council(CSC) and by grants and projects for the Research Council K.U.Leuven (GOA-Mefisto 666, GOA-Ambiorics, several PhD / Postdocs & fellow grants).

References

[1] Essential science indicators. Technical report, http://www.esi-topics.com/fields/index.html.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech.*, 2008.

[3] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70, 2004.

[4] X. He, H. Zha, C. Ding, and H. Simon. Web document clustering using hyperlink structures. *Computational Statistics and Data Analysis.*, 41(1):19–45, 2002.

[5] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[6] A. K. Jain. Data clustering : 50 years beyond k-means. Under review by the pattern recognition letters., Technical Report TR-CSE-09-11, 2009.

[7] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[8] F. Janssens, W. Glänzel, and B. De Moor. Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing and Management*, 2009,in press.

[9] X. Liu, S. Yu, Y. Moreau, B. De Moor, W. Glänzel, and F. Janssens. Hybrid clustering of text mining and bibliometrics applied to journal sets. In *Proc. of the SIAM Data Mining Conference 09 (SIAM DM 09)*. SIAM, 2009.

[10] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[11] M. Newman. Detecting community structure in networks. *Eur. Phys. J. B*, 38:321330, 2004.

[12] A. Strehl and J. Ghosh. Cluster ensembles-a knowledge reuse framework for combining multiple partitions. *JMLR*, 3:583–617, 2002.

[13] D. Zhou and C. J. C. Burges. Spectral clustering and transductive learning with multiple views. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 1159–1166, New York, NY, USA, 2007.

	Partition methods	TFIDF	IDF	TF	CRC	COC	BGC
NMI	WL	0.4977	0.5402	0.4953	0.4302	0.4879	0.3975
	KM	0.5554	0.5520	0.5234	0.4567	0.4334	0.3828
	NJW	0.5269	0.5516	0.5295	0.5407	0.4862	0.5341
	LM	–	–	–	**0.5544**	**0.5404**	**0.5504**
	LM(graph coupling)	**0.5577**	**0.5522**	**0.5469**	–	–	–
ARI	WL	0.2727	0.3384	0.2546	0.1182	0.2532	0.0912
	KM	0.3135	0.3446	0.2705	0.1875	0.1215	0.0938
	NJW	0.2794	0.3165	0.2915	**0.3240**	0.2923	**0.3215**
	LM	–	–	–	0.3187	**0.2933**	0.3112
	LM(graph coupling)	**0.3215**	**0.3543**	**0.2953**	–	–	–
Modularity	WL	0.3564	0.3116	0.3337	0.3626	0.3695	0.4177
	KM	0.4062	0.368	0.3507	0.4029	0.349	0.3808
	NJW	0.3848	0.3222	0.3578	0.3939	0.3475	0.3775
	LM	–	–	–	**0.4474**	**0.4216**	**0.4784**
	LM(graph coupling)	**0.4396**	**0.3815**	**0.4242**	–	–	–

Table 2. Performance of graph coupling. Validation measures: NMI, ARI and Modularity(MOD); The partitions of KM and WL clustering are repeated 20 times and the average results are reported. For LM based partition, the text based graphs(TFIDF, IDF and TF) have been coupled with multiple relations from citation based graphs. KM also performs well on partition of text based models since VSM based partition methods match the property of text feature. Concerning partition on citation models(COC,CRC,BGC), graph based partition methods(LM and NJW) outperform VSM based methods(WL and KM), which is consistent with the hypothesis that graph model more fits the sparse citation relation model than VSM.

Validations	Best individual graph	AF1	AF2	MWF1	MWF2	AMIWF1	AMIWF2
NMI	0.5554	0.5674	0.5591	0.5706	0.5574	0.5716	**0.5749**
ARI	0.3240	0.3726	0.3789	0.3803	0.3634	0.3839	**0.3850**
MOD	0.4784	0.5065	0.4951	0.5061	0.4885	**0.5076**	0.4923

Table 3. Performance of graph fusion. Due to iteratively running of our AMIW and the fast partition of LM, we only implement the AMIW by LM. Bold face in the table also denotes the highest performance in each row. Validation measures: NMI, ARI and Modularity; Graph fusion style: the combination of TFIDF graph and CRC graph(Fusion1), the combination of all text and citation graphs(Fusion2); Weighting methods: averagely weighting on two graphs fusion(AF1) and all graph fusion(AF2); Modularity wighting on two graphs fusion(MWF1) and all graphs fusion(MWF2); Adaptive Mutual Information Weighting on two graphs fusion(AMIWF1) and all graphs fusion(AMIWF2). The validation values of the best individual graph are extracted from the related highest value of individual graphs in Table 2.