International Journal of Hybrid Information Technology
Vol.9, No.2 (2016), pp. 299-310
http://dx.doi.org/10.14257/ijhit.2016.9.2.27

A Topic Community Detection Method for Information Network based on Improved Label Propagation

SHEN Gui-Lan[1,2] and YANG Xiao-Ping[1]

[1]*Information School, Renmin University, Beijng, China*
[2]*Business Collage of Beijing Union University, Beijing, China*
guilan.shen@buu.edu.cn; yang@ruc.edu.cn

Abstract

A large number of emerging information networks brings new challenges to the community detection. The meaningful community should be topic-oriented. However, the topology-based methods only reflect the strength of connection, and ignore the consistency of the topics; the content-based methods focus on the contents and completely ignore the links. This paper explores a topic oriented community detection method simLPA based on label propagation for information work. The method utilizes Latent Dirichlet Allocation topic model to represent the node content, and calculate the content similarity by the normalized Kullback–Leibler divergence. simLPA extended by LabelRank fuses the links and the contents naturally to detect the topic community. Extensive experiments on nine real-world datasets with varying sizes and characteristics validate the proposed method outperforms other baseline algorithms in quality. Additionally simLPA integrated into the content is equivalent to LabelRank in efficiency, which is easy to handle large-scale information networks.

Keywords*: information network, Latent Dirichlet Allocation topic model, topic community detection, label propagation, modularity, consistent topic*

1. Introduction

The evolution of Internet and information technology has produced a large number of complex networks with information content. According to the different types of information, we can divide these networks into Web network, email network, citation network, scientist cooperation network, social media network, etc., which can be named with information network. Here we formally define the information network is a kind of complex network of nodes with distinct information content. Considering the content information of nodes brings new challenges to the community detection of information network. The meaningful detected community of the information network should be topic-oriented, which has two characteristics: the nodes inside one community should have dense connections and they should have consistent or similar topic. The topic community detection is a good expansion for information network analysis, which helps for analyzing the topology, understanding the function, finding the hidden rule and predicting the behavior. Moreover, it has a profound effect on many practical applications, such as intelligent information retrieval, personalized service, social recommendation, *etc*.

At present, the topic community detection in information network has become a research trend. Two key issues need to be solved. One is how to represent the content of the nodes; two is how to effectively combine two heterogeneous information sources, including the links of the network and node content attributes without increase the complexity of computing.

In this paper, we present simLPA, a highly efficient label propagation algorithm combining network topology and content to detect the topic community in information

ISSN: 1738-9968 IJHIT

network. In order to reduce the dimension of the node's content attributes, we apply the LDA topic model to represent the node content. Our approach fuses the links and the contents naturally in the process of normalizing the label propagation probability defined by content similarity between node and its neighbors. We adopt the modularity and the purity to evaluate the quality of the topic community. Through extensive experiments on real-world datasets drawn from WebKB, Cora and Wikipedia, we demonstrate the effectiveness and efficiency of our method. We find that simLPA often detects topic communities of comparable or superior quality on most these datasets.

This paper is organized as follows: Section 2 discusses the related work; Section 3 presents the implementation details of simLPA; Section 4 introduces the datasets and reports quantitative experiments results and Section 5 concludes the findings and identifies the future research.

2. Related Works

Topic Community Detection using Topology: Topology-based methods are most popular for topic community detection in information network. All of these are based on the basic assumption that the nodes linked tightly tend to have the similar interest or topic, which are intended to detect the dense connections. According to the definition of community and its potential principle, these methods can be divided into the following categories: (1) those based on optimizing the community quality. The current widely used measure is the modularity function Q proposed by Newman [1], There are many emerging optimization strategies to maximize the modularity, such as greedy algorithm [2], iterative heuristic inspired optimization [3], general optimization [4], simulated annealing [5]. (2) Those relying on cluster the nodes. Considering community is the cluster with the high similarity nodes, those methods typically adapt the spectral graph theory [6] or structural similarity [7] to cluster the nodes in network. (3)Those based on graph partitioning. In order to detect community, different graph partition approaches apply different strategy to remove some special edges or nodes in network, which includes select those edges with the largest betweenness [8], or select those edges with the smallest clustering coefficient [9]. (4) Other methods. Some researchers adapt the dynamic methods to detect community, such as label propagation [10] and Markov random walk [11]. In addition, [12, 13] transformed the community detection into the problem of statistical inference.

Despite the use of different techniques, the above methods can always detect dense connections in network. However, they only focus on the topology information and ignore the content information contributing to improve the quality of the community.

Topic Community Detection using Content: Content-based methods can be regarded as the problem of text clustering. These methods are based on similarity matrix or distance matrix, which edges are not real connection rather than the similarity of two nodes. For example, Newman [14] apply single linkage hierarchical clustering into similarity matrix to detect communities. In addition, text clustering also can be resolved by the probability model, such as Latent Semantic Analysis (LSA) [15], the probabilistic LSA (pLSA) [16] and Latent Dirchlet Allocation (LDA) [17]

Although those methods take the node content into account, they completely discard the topology information, which lead to the detected results only related to the topic, but had nothing to the community structure.

Topic Community Detection using Topology and Content: Based on the assumption that the content information can improve the quality of community discovery, various approaches have been combined the links and contents for community detection. One of them is generative probabilistic model, which regards both contents and links as being dependent on one or more latent variables, and then estimates the conditional distributions to find community assignments. The representation in this category is Link-PLSA-LDA

[18]. The other is attribute graph clustering which consider the contents into the network structure clustering. The representation in this category is SA-Cluster [19]. Different from those methods using topology, these methods account for the content information of the nodes, so the division results for the network is more cohesive in the topics. However, considering the content of nodes, the complexity of the algorithm is greatly increased, because the contents are mostly based on text analysis, which will lead to some new challenges, such as how to deal with the high dimensional sparse for node attributes, how to deal with the possible emergence of the dimension disaster.

Community Detection using Label propagation: Label propagation algorithm (LPA)[10] is a fast community detection algorithm with near linear computational complexity, so it can handle large-scale network. However, the algorithm has the random factors, and the results are not stable. Subsequently, many researchers put forward a series of improved methods. Among them, LabelRank [20] proposed by Xie resolve the randomness issue in traditional LPA mainly owing to initialize the label distribution of each node defined by the probability of see each neighbor. Meanwhile, LabelRank introduces a set of operators to control and stabilize the propagation dynamics. The detailed operators include propagation, inflation, cutoff and conditional update, which extend Markov Cluster Algorithm [21]. LabelRank stores, propagates and ranks labels in each node. During each execution, LabelRank relies on four operators applied to the labels. Taking modularity as the objective function, the method loops each execution, until the labels are updated to stability. Finally, nodes with the same label can be divided into the same community.

3. Methodology

As mentioned in the previous section, there are two challenges to detect high quality topic communities at present. One is how to represent the node content attribute; the other is how to combine links and contents without increasing the complexity of the algorithm. simLPA proposed by us can be a good solution to these two challenges. In this section, we present our method simLPA to detect the topic community in information network. Firstly, we introduce how to represent the node content attributes and how to calculate the similarity for pairs of nodes; Secondly, we present algorithmic details of simLPA, a novel method like label propagation algorithm, extended from LabelRank.

3.1. The Content Representation

For most information networks, such as citation network, web network, the content information for the nodes is usually presented in the form of text. So, here we simplify, focus on how to deal with the text information. The classical text representation method is the text vector space model, because it is based on the bag-of-words easy to lead to high dimensional attributes, not suitable for large-scale information network. In order to solve it, we use LDA to model the node contents.

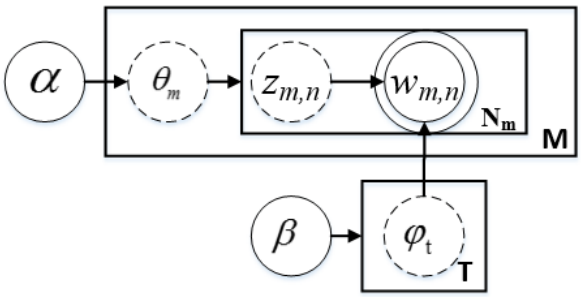

Figure 1. Latent Dirichlet Allocation Topic Model

LDA proposed by Blei, *et al.*, in 2003[17], is a Bayesian probabilistic model, including document-topic-word three layer structure, document-topic obey Dirichlet [22] distribution, topic-word obey the multinomial distribution. Owing to its clear hierarchical structure, LDA model parameter space is independent of the number of training text documents, which is suitable for processing large-scale text corpus in information network.

The basic idea of LDA is that each document in the text corpus represents a probability distribution of potential topics, and each topic is represented by a probability distribution of a lot of words. Figure 1 illustrates the working principle of LDA, where, M is the number of the document set, N_m is the number of words in the m document, α and β represent the prior knowledge for the document-topic probability distribution and topic-word probability distribution respectively,T is the number of the topics, $Z_{m,n}$ is specified as the topic of the N word in the M document. $W_{m,n}$ is the N word in the M document. In addition, in Figure 1, a single solid circle represents a fixed value specified by the user in advance; double circle represents the observed data; single dashed circle is hidden parameter. The hidden parameters of documents need to be solved by probability inference, and the Gibbs sampling method is usually used to solve the hidden parameters θ and ϕ.

As a topic model, LDA can greatly reduce the dimension of attributes by means of representing the content of the document as a topic probability distribution. Meanwhile, previous research has shown using the LDA topic model is better than the vector space model in the quality. We extract K topics for each node as the dimension of the content attribute. Therefore, the content in form is expressed as the probability distribution for topics. We use the Kullback–Leibler(KL) divergence, which is a measure of the difference between two probability distributions, to caculate the similarity for the pair of node contents. The smaller KL divergence means the greater similarity.Let P be the topic probability distribution for node u, and Q be for v. And then, $P(t_i)$ is the probability for the i topic for node u, and $Q(t_i)$ is the probability for the j topic for node v. We use formula 1 to calculate the normalized similarity of two nodes u and v.

$$sim(u,v) = \exp(-\frac{d_{PQ} - Min_d}{Max_d - Min_d}) \quad (1)$$

It is not symmetric in P and Q. We use formula 2 to calculate d_{PQ}.

$$d_{PQ} = \frac{1}{2}(KL(P,Q) + KL(Q,P)) = \frac{1}{2}(\sum_{t_i \in K} P(t_i) \lg \frac{P(t_i)}{Q(t_i)} + \sum_{t_i \in K} Q(t_i) \lg \frac{Q(t_i)}{P(t_i)}) \quad (2)$$

Where, Min_d is the minimum value of all pair nodes for the KL divergence, and Max_d is maximum.

3.2 simLPA Algorithm

In this Section, we present the details of simLPA algorithm for fusing the links and the contents in nature when node labels are propagated to detect the topic community. We mentioned above that LabelRank as a label propagation algorithm can produce stable communities' structure. It has two advantages to handle large-scale networks: one is the running time near to linear with the number of edges, the two is that the information needed to label propagation only rely on the local neighbors, therefore it is suitable for processing the network in parallel. However, the LabelRank algorithm ignores the fact that the content of nodes can promote the quality of the detected topic community and only focuses on the network topology. In which, the label distributions for nodes are determined by the probability of the observed neighbors. We argue whether a node receive the label from one of neighbors or not should not only depend on s the link

between the two nodes, but also should be determined by the content similarity of the two nodes. Therefore we design label propagation strategy of simLPA, which takes into account both the node's content attributes and the network topology.

In simLPA, we maintain these data structures, A is the $n \times n$ adjacency matrix defining the network structure, and P is the $n \times n$ label distribution matrix. In each node, an entire distribution of labels defined by vector $P_i$is maintained and spread to neighbors. Each element $P_i(c)$ holds the current estimation of probability of node i observing label $c \in C$. C is the a finite set and the each element is represented by the node id, that is $C = \{1, 2, \cdots, n\}$

In simLPA, to initialize P,each node is assigned the probability by the content similarity with each neighbor:

$$P_{i,j} = Sim(j,i) \Big/ \sum_{j \in Nb(i)} Sim(j,i) \; \forall j \; s.t. \; A_{ij} = 1 \quad (3)$$

Each node broadcasts the label distribution to its neighbors at each time step and calculates the new label distribution $P_i(c)$

$$P_i^{'}(c) = \sum_{j \in Nb(i)} Sim(j,i) \cdot P_j(c) \Big/ \sum_{j \in Nb(i)} Sim(j,i), \; \forall c \in C \quad (4)$$

In order to ensure the higher probability labels easy to spread to lower probability labels, P is applied to the Inflation operator Γ_{in} in the label distribution matrix:

$$\Gamma_{in} P_i(c) = P_i(c)^{in} \Big/ \sum_{j \in C} P_i(j)^{in} \quad (5)$$

The third step of our algorithm introduce the pruning operator Φ_r , applied to P, in order to remove those labels that probability threshold is lower than r value. It helps to effectively reduce the storage space of the labels, so as to reduce the space complexity.

Our method takes the modularity as the objective function. When the algorithm achieves convergence, it can only guarantee the performance of the detected community in dense connections; howeverit cannot guarantee the performance in content consistency. Hence, we propose the update operator Θ , which updates a node only when it is significantly different from its neighbors in terms of content.

At each iteration, the change is accepted only when meet the following equation:

$$\sum_{j \in Nb(i)} Sim(i,j) \le qk_i \cdot avgSim \quad (6)$$

Where $Sim(i,j)$ is the content similarity of the two node i and j, $Nb(i)$ is the neighbor set of node i, k_i is the degree of node i, $avgSim$ represents the average similarity of all nodes in the network. q is a parameter between [0,1].

If there are small difference in the label distribution between consecutive iterations, s steady state of a node can be defined. We determine whether the network reaches a stable state by judging the steady state of all nodes. When satisfy the stop criterion, the algorithm reach the best performance, and the detected communities are needed for us.

We firmly believe that the label propagation strategy based on content similarity of nodes can detect the more consistent topic community than simply based on the links of the nodes. Specifically including the following two reasons, first, the content similarity of nodes is used as the label propagation probability of nodes, so the content information of nodes is considered. Secondly, the topology of the network is also integrated in the process of normalizing the label propagation probability of neighbors.

3.3 Performance Metrics

Dense connections and consistent topic should be both considered into measure the quality of detected topic community. Hence, we use two metrics to evaluate the quality.

To evaluate the denseness of the structure, the classic modularity proposed by Newman [4] is adapted, which is defined as:

$$Q = \frac{1}{2m}\sum_{ij}[A_{ij} - \frac{k_i k_j}{2m}]\delta(C_i, C_j) \quad (7)$$

The value of the modularity Q is determined by the strength of division of a network into communities. Networks with high modularity have dense connections between the nodes within communities but sparse connections between nodes in different communities.

To evaluate the consistency in topics, the purity proposed by Strehl A etc. [23] is employed. Given the standard communities $G = \{G_1, G_2, \cdots, G_S\}$ andthe communities detected by the algorithms is represented by $C = \{C_1, C_2, \cdots, C_S\}$.The purity of C_i is defined as:

$$Purity(C_i) = \frac{1}{|C_i|}\max_j\{C_i \cap G_j\} \quad (8)$$

Usually, the detected community C_i includes nodes belong to other G_j in the ground-truth. For C_i we compute the intersection set with each standard community G_j ,then take the maximum as the result for it. So the purity of C is defined as:

$$Purity(C) = \frac{1}{K}\sum_{i=1}^{K} Purity(C_i) \quad (9)$$

The average purity of the detected communities is measured by the average purity of the each community. The higher purity means that results are closer to the ground-truths.

4. Experiments

In this Section, we use nine real datasets to carry out extensive experiments to verify the performance of the simLPA, including the impact of inflation factor on the community metrics, the impact of the label propagation strategy based on the content similarity, the running time of simLPA, and the comparative analysis of the simLPA and the classical algorithm in the community evaluation criteria.

4.1Datasets

We select 9 datasets from three real data sources, including Cora [24], WebKB [25], Wiki[25]. These datasets can represent citation network, web network and social network respectively. In order to simplify the operation, we handle all networks formed by these datasets as undirected network. The statistical information of specific datasets is shown in Table 1.

Table 1. The Statistical Information of Datasets

Dataset	*# Class*	*|V|*	*|Edges|*	*Average Degree*
Cornell	5	195	283	2.90
Washington	5	217	366	3.37
Wisconsin	5	262	459	3.50
Texas	5	185	280	3.03
Information Retrieval	4	343	821	4.80
Data Structure	9	951	1684	3.52
Database	7	802	2140	5.34
Programming	9	2107	5909	5.61
Wiki	19	2877	33300	23.15

4.2 Effect of Inflation Factor

To verify the value of inflation factor how to effect the modularity and the purity, We chooses information retrieval, data structure two datasets from Cora and Texas from WebKB to carry out the experiments. In simLPA, the inflation factor can accelerate the convergence of the algorithm, which is same as the LabelRank. The value of the inflation factor is larger, then the convergence of the algorithm is faster and the size of the detected communities is smaller. Taking into account the sensitivity between interval [1, 2], 12 values of the inflation factor are selected, including 1 to 10, 1.2 and 1.5 to validation. Experimental results are shown in Figure 2.

From the results, to the network with significant community structure, as shown in Figure 2 (a) and (c), the inflation factor in the smaller value, the modularity can achieve the optimal value. Moreover, with the increase of the inflation factor, the modularity decreases gradually. However, to the network with tar topology, as shown in figure (b), the modularity increases with the increase of the inflation factor at begin, when reaches to the peak, it decreases gradually. In general, the modularity will decrease with the increase of the inflation factor.

For another important metric, the purity is not dependent on the inflation factor whichever network structures. From the results as shown in figure (b)(d)(e), the value of inflation factor is small, the purity increases gradually, but when reaches to a specific value, the purity does not change. That is, our method has robustness in purity. In our experiments, the inflation factor is set to 4

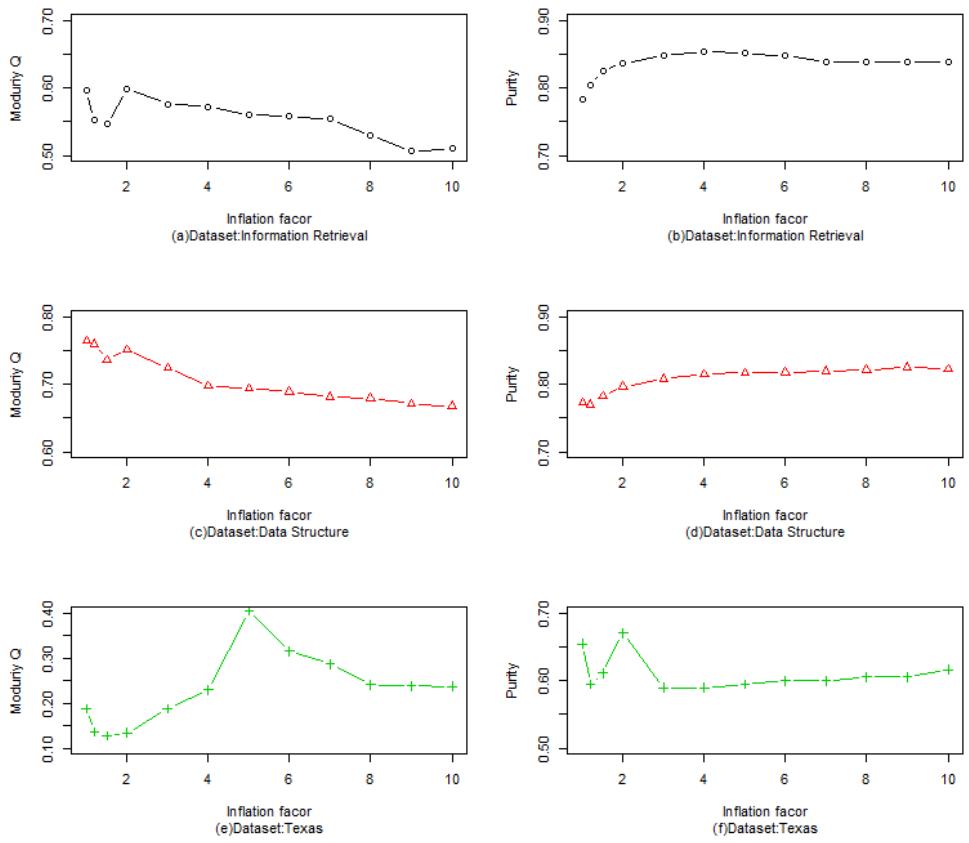

Figure 2. The Results of Three Datasets with Varying Inflation Factor

4.3 The Impact of the Different Label Propagation Strategy

LabelRank uses the label propagation strategy based on the probability of observing its neighbors, while, our approach simLPA uses the strategy based on the probability of content similarity between node and its neighbors. We select 4 datasets from Cora and Wiki dataset to illustrate the difference of the two label propagation strategy on the topic community metrics in Figure 3.It is evident from the results that since no content information is incorporated, LabelRank does not perform well in the purity on any datasets. However, simLPA performs better due to contents and links being combined together. Although, in some datasets, the modularity is dropped off a little, the denseness of the detected communities is still remained. This verifies that node content does help in modeling the topic community, and therefore the underlying the effectiveness of our approach simLPA.

4.4 Running Time

We select three datasets including Wisconsion, Data Structure and Programing to verify whether the incorporation of content information increase the computational complexity or not. We test the running times for each datasets in different choices of inflator factor. The comparisons between simLPA and LabelRank with inflator factor is illustrated in Figure 4. We can find the content information integrated into label propagation processing almost does not increase the computational complexity, and even in some cases, the convergence is accelerated.

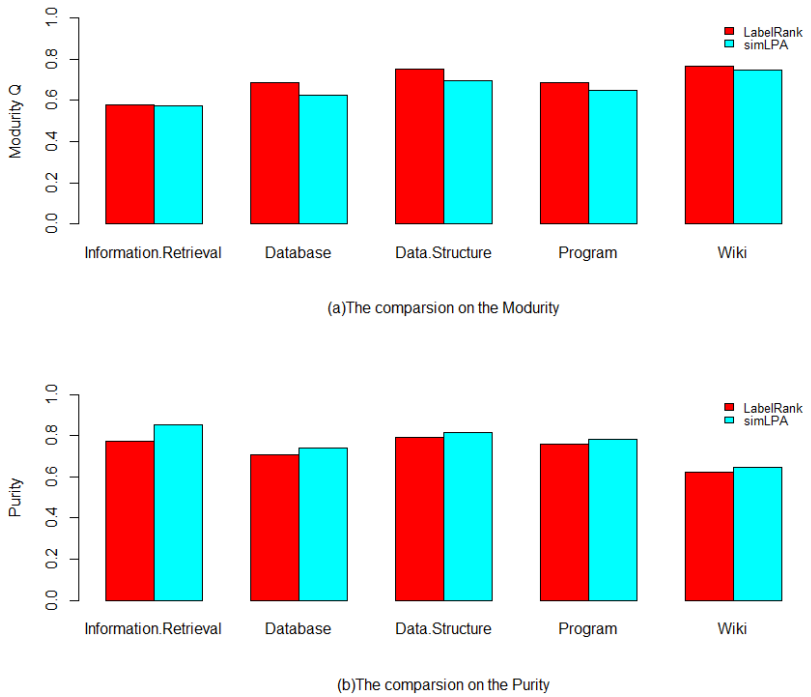

Figure 3. The Results of Different Label Propagation Strategy

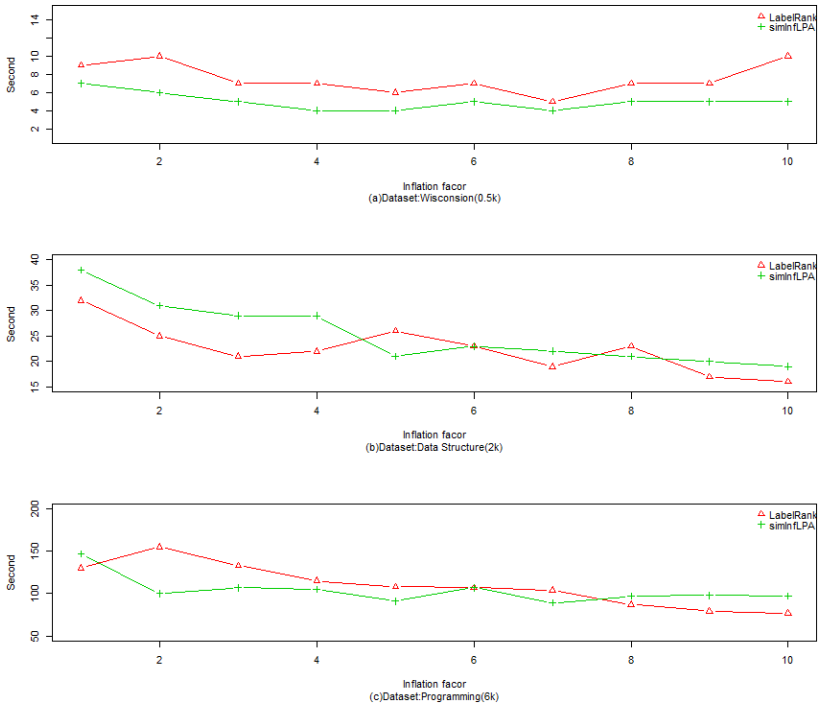

Figure 4. Running Time with Variation of Inflation Factor

4.5 Community Detection Results

To evaluate the effectiveness of simLPA, we compare our method with four baseline methods: LPA [10], Fastgreedy [2], LabelRank [20] and K-Means. We use the modularity and purity metrics to quantify the performance of each algorithm. LPA is run ten times and then get the average as the final result. The K value of K-Means algorithm is set to the number of ground-truth categories for each dataset; in addition, being a pure content-based approach, the cluster results are evaluated just only the purity. In LabelRank and simLPA, r is set to 1, inflation factor I is set to 4 and q is set to 0.6.The results for the different algorithms and data sets are illustrated in Table 2 and Table 3. These confirm when detecting topic community in information network, simLPA algorithm can achieve best performance than other baseline algorithms.

Table 2. Comparison of Different Community Detection Algorithm in the Modularity Metric

Dataset \ Algorithm	LPA		FastGreedy		LabelRank		simLPA
	Q	Ratio	Q	Ratio	Q	Ratio	Q
Cornell	0.562	-19.88%	0.644	-30.09%	0.528	-14.84%	0.450
Washington	0.261	-3.84%	0.554	-54.59%	0.314	-19.92%	0.251
Wisconsin	0.429	-0.27%	0.641	-33.19%	0.487	-12.07%	0.428
Texas	0.246	31.16%	0.552	-41.50%	0.343	-5.88%	0.323
Information Retrieval	0.613	-6.69%	0.635	-9.87%	0.577	-0.78%	0.572
Data Structure	0.807	-13.52%	0.852	-18.12%	0.754	-7.43%	0.698
Database	0.717	-13.01%	0.722	-13.65%	0.685	-8.93%	0.624
Program	0.708	-8.46%	0.714	-9.17%	0.685	-5.33%	0.648
Wiki	0.761	-2.40%	0.718	3.49%	0.764	-2.81%	0.7425

Table 3. Comparison of Different Community Detection Algorithm in the Purity Metric

Dataset	LPA		FastGreedy		WaikTrap		LabelRank		Kmeans		simLPA
	Purity	Ratio	Purity	Ratio	Purity	Ratio	Purity	Ratio	Purity	Ratio	Purity
Cornell	0.472	31.52%	0.477	30.11%	0.523	18.62%	0.503	23.46%	0.446	39.06%	**0.621**
washington	0.562	27.05%	0.535	33.61%	0.613	16.54%	0.657	8.79%	0.500	42.86%	**0.714**
wisconsin	0.500	25.20%	0.504	24.26%	0.542	15.50%	0.515	21.48%	0.536	16.83%	**0.626**
Texas	0.562	14.41%	0.562	14.41%	0.622	3.47%	0.632	1.71%	0.551	16.78%	**0.643**
Information Retrieval	0.795	7.36%	0.757	12.74%	0.772	10.61%	0.772	10.61%	0.564	51.30%	**0.854**
Data Structure	0.756	7.79%	0.701	16.18%	0.726	12.31%	0.790	3.13%	0.259	215.00%	**0.815**
Database	0.675	9.98%	0.549	35.24%	0.656	13.11%	0.710	4.57%	0.256	190.26%	**0.742**
Network	0.749	1.04%	0.715	5.78%	0.758	-0.15%	0.722	4.76%	0.456	65.91%	**0.756**
Program	0.741	5.57%	0.539	45.20%	0.661	18.32%	0.759	3.08%	0.252	210.40%	**0.782**
Wiki	0.590	8.19%	0.480	32.88%	0.598	6.74%	0.622	2.61%	0.516	23.60%	**0.638**

5. Conclusions

In this paper, we proposed a label propagation approach simLPA for topic community detection. Based on LabelRank algorithm, our approach exploits combine node contents and links together in nature during processing the label propagation. In order to reduce the dimension of the node content attributes, we applied the LDA topic model to represent the node contents and calculated the content similarity using KL divergence for node pairs. Then we replaced the link-based label propagation strategy by the content-based one, besides, we integrated the link information of the network in the process of the normalization for the content similarities between node and its neighbors. To evaluate the performance, we conducted experiments on nine real datasets. Compared with the link-based methods and the content-based method, our approach gained a better performance in topic community detection.It would be worth mentioning, taking content information into consideration does not increase the computational complexity, that is, simLPA can detect the topic community in nearly linear time. Furthermore, simLPA only depends on the local information of the node; therefore, it is easy to extend for parallel processing.

Our approach has many potential applications. It can be applied to many kinds of information networks, which nodes contain content. With the communities detected by our method, we are able to improve the efficiency of collaborative scientific research, discover experts for each topic, and analyze topic-oriented influence propagation.

Acknowledgements

This paper is supported by Natural Science Foundation of China (No.71572015), Scientific Research Project of Beijing Union University (No. Zk10201506), Scientific Research Project of Beijing Educational Committee (KM201511232016)

References

[1] M. E. J. Newman and M. Girvan, “Finding and evaluating community structure in networks”, Physical review E, vol. 69.2, **(2004)**.
[2] M. E. J. Newman, “Fast algorithm for detecting community structure in networks”, Physical review E, vol. 69.6, **(2004)**.
[3] V. D. Blondel, “Fast unfolding of communities in large networks”, Journal of statistical mechanics: theory and experiment, vol. 2008.10, **(2008)**.
[4] M. E. J. Newman, “Modularity and community structure in networks”, Proceedings of the national academy of sciences, vol. 103.23, **(2006)**.
[5] C. P. Massen and J. P. K. Doye, “Identifying communities within energy landscapes”, Physical Review E,

vol. 71.4, **(2005)**.
[6] L. Donetti, "Detecting network communities: a new systematic and efficient algorithm", Journal of Statistical Mechanics: Theory and Experiment, vol. 2004.10, **(2004)**.
[7] X. Xu, N. Yuruk, Z. Feng and T. A. Schweiger, "Scan: a structural clustering algorithm for networks", In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, **(2007)** August 12-15; San Jose, USA.
[8] M. E. J. Newman, "Modularity and community structure in networks", Proceedings of the national academy of sciences, vol. 103.23, **(2006)**.
[9] P. Zhang, J. Wang, X. Li, M. Li, Z. Di and Y. Fan, "Clustering coefficient and community structure of bipartite networks", Physica A: Statistical Mechanics and its Applications, vol. 387.27, **(2008)**.
[10] U. N. Raghavan, R. Albert and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks", Physical Review E, vol. 76.3, **(2007)**.
[11] P. Pascal and M. Latapy, "Computing communities in large networks using random walks", Computer and Information Sciences-ISCIS 2005, Springer Berlin Heidelberg, **(2005)**.
[12] M. B. Hastings, "Community detection as an inference problem", Physical Review E, vol. 74.3, **(2006)**.
[13] B. Karrer and M. E. J. Newman, "Stochastic block models and community structure in networks", Physical Review E, vol. 83.1, **(2011)**.
[14] M. E. J. Newman, "The structure and function of complex networks. SIAM review, vol. 45.2, **(2003)**.
[15] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by latent semantic analysis", Journal of the American society for information science, vol. 41.6, **(1990)**.
[16] T. Hofmann, "Probabilistic latent semantic indexing", Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, **(1999)** August 15 -19; Berkeley, CA, USA.
[17] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation", the Journal of machine Learning research, vol. 3, **(2003)**.
[18] R. Nallapati, A. Ahmed, E. Xing and W. Cohen, "Joint latent topic models for text and citations", In 14thSIGKDD, **(2008)** August 24-27, Las Vegas, USA.
[19] Y. Zhou, H. Cheng and J. X. Yu, "Graph clustering based on structural/attribute similarities", Proceedings of the VLDB Endowment, vol. 2.1, **(2009)**.
[20] J. Xie and B. K. Szymanski, "Labelrank: A stabilized label propagation algorithm for community detection in networks", IEEE 2nd International Workshop on Network Science, **(2013)** April 29- May 1, New York, USA.
[21] S. Van Dongen, "A cluster algorithm for graphs", Report-Information systems, vol. 10, **(2000)**.
[22] M. C. Galligan, R. Saldova, M. P. Campbell, P. M. Rudd and T. B. Murphy, "Greedy feature selection for glycan chromatography data with the generalized Dirichlet distribution", BMC bioinformatics, vol. 14.1, **(2013)**.
[23] A. Strehl, J. Ghosh and R. Mooney, "Impact of similarity measures on web-page clustering", Workshop on Artificial Intelligence for Web Search (AAAI), **(2000)**, Austin, Texas, USA.
[24] "Cora Research Paper Classification", http://www.datatang.com/data/28850, [EB/OL]. 2015-4-12.
[25] "LINQS", http://linqs.umiacs.umd.edu/projects/projects/lbc/index.html, [EB/OL]. 2015-10-12.

Authors

ShenGui-Lan, She is an associate professor in Business College of Beijing Union University.She got M. D. of Computer Science from Information and Computer College, Beijing University of Chemical Technology (BUCT) in 2006. she iscurrently aPh.D. candidate of Information College in People's University of China (RUC), She is interested in the following fields:Communities detection for information network; topical modeling for short-text.She has published articles in several professional journals and international conferences, participated in two textbook's writing, hosted and participated in several externally funded research projects.

Yang Xiao-Ping, He is a professor, doctoral supervisor of Information College in People's University of China (RUC), the major field is web data mining,information system engineering.

Prof. Yang is Vice President of Association of Fundamental Computing Education in Chinese Universities (AFCEC),Committee of China Computer Users Association and Vice President ofSystems Engineering Society of Beijing