Journal of Southeast University (English Edition) Vol. 24, No. 3, pp. 276 – 280 Sept. 2008 ISSN 1003—7985

Extracting and evaluating method of web dense cores

Yang Nan Gao Jie Xue Honghu Liu Xiude

(School of Information, Renmin University of China, Beijing 100872, China)

Abstract: This paper focuses on some key problems in web community discovery and link analysis. Based on the topic-oriented technology, the characteristics of a bipartite graph are studied. An x bipartite core set is introduced to more clearly define extracting ways. By scanning the topic subgraph to construct x bipartite graph and then prune the graph with i and j, an x bipartite core set, which is also the minimum element of a community, can be found. Finally, a hierarchical clustering algorithm is applied to many x bipartite core sets and the dendrogram of the community inner construction is obtained. The correctness of the constructing and pruning method is proved and the algorithm is designed. The typical datasets in the experiment are prepared according to the way in HITS (hyperlink-induced topic search). Ten topics and four search engines are chosen and the returned results are integrated. The modularity, which is a measure of the strength of the community structure in the social network, is used to validate the efficiency of the proposed method. The experimental results show that the proposed algorithm is effective and efficient.

Key words: dense cores; link analysis; hierarchical clustering; modularity measure

The rapid growth of Internet has made the web a very important resource which the human being relies on more and more. But the distribution in structure, the convenience of page publication and the huge scale have made the web difficult to control. Information of diverse format, style, source and geographical area are contained in the web. The web has become an ocean of information and the total pages have reached 2.9 billion[1]. How to find useful information from the ocean is a long and challenging job. Although the evolution of the web looks chaotic, it reflects features of self-organization[2] locally. Among them, the web community is a very important one and it reflects the social activities in the web[3]. In general, the web community shows a collection of pages with dense hyper-links, usually on several topics. Discovery of these communities will help us to understand the evolution of the web, guide the crawler, see inside the organization of the web and arrange portals efficiently.

There are many researches on community discovery. All these algorithms roughly fall into three categories: HITS-based methods[4-5], bipartite graph-based methods[3,6] and flow-based methods[7-8]. HITS and flow are topic-oriented methods and bipartite graph is the non-topic method. In this paper, we focus on the topic-oriented methods and analyze the trawling method, which is a bipartite graph method. To indicate some shortcomings derived from the complete bipartite graph defined in trawling, we propose a new signature of cores of communities, which is the x bipartite cores set. Then we also give a method to construct the cores sets by a scan topic subgraph and an (i, j) pruning algorithm. We prove that the forward and backward expansion method can include all x bipartite cores and we design an algorithm. The web subgraph, which is also called a topic subgraph, is collected by typical HITS way and we modify it by integrating results from several search engines. We define the x bipartite cores sets as the basic components of communities and based on these sets, apply the hierarchical clustering algorithm to obtain the dendrogram of communities. To evaluate the communities, we use modularity measurement. The experimental results show that the new method is effective and efficient.

1 Related Work

Along with the scale expansion and informative diversity, the web has become a huge resource of information. The problem is how to find useful information on the web effectively; this is also called resource discovery. There are many methods to deal with the problems. The typical ones fall into three categories:

HITS-based methods[4-5] use eigenvectors of an adjacency matrix of a web subgraph to find communities. The collection of subgraphs is based on the results from a search engine when one submits a term or keywords. From the search results, about the first ones($C = 200$) are gathered into a root set and then the root set expands to be a basic set by adding the pages with in-links (only first 50 in-links, $K = 50$) and out-links of pages in the root set. To calculate the main eigenvector and several non-main eigenvectors, the main eigenvector corresponds to the main community and non-main eigenvectors correspond to non-main communities respectively. This method can find many communities. But it has several problems. First, it does not guarantee all the possible communities[3]. Second, the meaning of positive and negative eigenvalues is not clear[9]. Third, it is difficult to make sure what the topics of the non-main communities are and how many pages should be chosen into communities[10].

The bipartite graph-based method[3,6] is also called the trawling method. According to this method, a community at least contains a core, which is also a complete bipartite graph. Therefore, the discovery of communities is done by extracting cores. The dataset collection is from a web crawler. By (i, j) enumeration, it scans a dataset to extract all the possible (i, j) cores, and expands these cores to obtain the communities.

The flow-based method[7-8] defines a community as a cluster and the number of intra-cluster links exceeds that of inter-cluster links. Before the dataset collection, a topic is predefined and a seed set of pages is needed that is genera-

Received 2008-04-15.

Biography: Yang Nan (1962—), male, associate professor, yangnan@ruc.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 60773216), the National High Technology Research and Development Program of China (863 Program) (No. 2006AA010109), the Natural Science Foundation of Renmin University of China (No. 06XNB052), Free Exploration Project (985 Project of Renmin University of China) (No. 21361231).

Citation: Yang Nan, Gao Jie, Xue Honghu, et al. Extracting and evaluating method of web dense cores[J]. Journal of Southeast University (English Edition), 2008, 24(3): 276 – 280.

ted from human effort or from a search engine. By calculating the edge flow from the pages in the community to a sink page, we can find the minimum cut and remove it to decide the borders of the communities. With this method, the sink page is determined manually and the discovery of hierarchical or multi-communities is impossible.

According to whether or not to predefine a topic before processing, all the methods can be divided into two types. One is the topic-oriented and the other is non-topic-oriented. HITS-based and flow-based methods belong to the topic-oriented type. This type of method must give a topic to a search engine and a dataset is collected from the search results. The trawling method is non-topic-oriented. It does not give a topic before processing and the dataset collection is from a web crawler.

This paper focuses on the topic-oriented method. Because HITS and flow cannot work well in a hierarchical organization of communities, we think the trawling method can. We analyze the features of trawling and the complete bipartite graphs are elements of a community. But the extracting method of trawling is not very efficient and it is also not complete. We propose an x bipartite cores set as an element of a community. And we prove that the x bipartite cores set can be complete and that it can cover all possible x complete bipartite graphs. The dataset collection is the same as HITS. From the returned results of a search engine, we can construct a topic subgraph. Based on this subgraph, we apply our new algorithm to it.

2 Extracting and Clustering Algorithm

This section introduces a new method, which is analogous to the bipartite core extracting method. We modify the definition of the signature of a community and it can cover more cores than that of trawling. We propose an x bipartite cores set as an essential element instead of a complete bipartite graph and employ forward and backward expansion criteria to construct subgraphs from page x. Then we apply (i, j) pruning algorithm to gain the x bipartite cores set. We also prove that the new method can cover all possible x cores. In the end, we use hierarchical clustering extracted cores sets to obtain a dendrogram.

2.1 Preliminaries and notations

The web can be abstracted to a directed graph. A directed graph is denoted as $G = (V, E)$, where V is a set of vertices and E is a set of edges. If there exists an edge(u, v) included in G, we have $u, v \in V$ and $(u, v) \in E$. If an edge $(u, v) \in E$, u is the predecessor of v and v is the successor of u. $N_G^-(u)$ and $N_G^+(u)$ denote the set of predecessors and successors of u in G. $|N_G^-(u)|$ and $|N_G^+(u)|$ are the in-degree and the out-degree of u in G, respectively.

Definition 1 Bipartite graph(BG)

BG $= (F, C, E)$ is a G, where V is divided into two disjoint sets F and C, and E is the set of edges. In the graph, vertex u and v of an edge (u, v) belong to both sets $(u \in F, v \in C)$.

Definition 2 Complete bipartite graph(CBG)

CBG $= (F, C)$ is a BG $= (F, C, E)$, where $E = \{(F \times C)\}$ and can be omitted. So for a CBG, we have $\forall u \mid u \in F$, $N_{\mathrm{CBG}}^+(u) = C$ and $\forall v \mid v \in C$, $N_{\mathrm{CBG}}^-(v) = F$. CBG is also called a bipartite core, denoted as $C(n, m)$, where $n = |F|$ and $m = |C|$.

2.2 Topic subgraph

A topic subgraph is a graph constructed from a topic. The first step is to get a root set from a search engine. We submit a topic to a search engine, for example, www. google. cn, then collect first K($K = 200$ in HITS) pages returned to construct a set S_{root}. The second step is expansion. We expand S_{root} to obtain a basic set S_{base}. The expansion criterion is that for every page x in a root set, we expand any pages pointed out by and pointing to the page x. Because the number of in-links is enormous, only the first C($C = 50$ in HITS) pages are considered. To get the number of in-degrees and out-degrees of a page easily, we use two datasets to present a subgraph, which are EOS (edge on source) and EOD (edge on destin). These datasets are a collection of edges, with the form $\langle$source, destin$\rangle$. The EOS is sorted by source and the EOD by destin.

2.3 The xCores-set

The essence of our algorithm is the set of x bipartite cores, which is denoted as xCores-set. The xCore is a bipartite core CBG $= (F, C)$ with $x \in F$. If there are several xCores, CBG_1, CBG_2, …, CBG_n, the union of them is an xCores-set.

Definition 3 xCore

If there are a page x and a CBG $= (F, C)$, and $x \in F$, the CBG is an xCore.

Definition 4 xCores-set

If there are a page x and $n(n \geqslant 1)$ cores, CBG_1, CBG_2, …, CBG_n, and $x \in F_1, x \in F_2, \ldots, x \in F_n$. The union of them is a BG(F, C, E), where $F = F_1 \cup F_2 \cup \ldots \cup F_n$, $C = C_1 \cup C_2 \cup \ldots \cup C_n$ and $E = \bigcup_{i=1}^{n} \{F_i \times C_i\}$.

2.4 Extracting xCores-set as web dense cores

The new algorithm is based on scanning the dataset EOS. Usually, we choose the first edge of EOS. From the source node x of the edge, we construct xBG $= (F, C, E)$. We use forward and backward expansion methods to construct xBG. Given a node x and a $G = (V, E)$, a BG $= (F, C, E')$ can be constructed from x. Forward expansion, let $C = N_G^+(x)$. Backward expansion, for each $u \in C$, $F = F \cup N_G^-(u)$. $E' = \{(u, v)\} \mid (u \in F, v \in C, (u, v) \in E)$. After xBG is constructed, we then prune it with i and j, which is also defined as (i, j) pruning. If the result is not empty, it must be an xCores-set. At each scan, whether or not xBG contains an xCores-set, we always delete some of the edges related with x. This operation includes EOS and EOD. The following definitions and theorems are given to prove that an xCores-set can be gained from xBG and (i, j) pruning.

Definition 5 xBG

Given a graph $G = (V, E)$ and a node $x \in V$. xBG $= (F, C, E')$ is derived from G and x by forward and backward expansion. The xBG has the following features: $x \in F$, $C = N_G^+(x)$, $F = \bigcup_{u \in C} N_G^-(u)$ and $E' = \{(u, v)\} \mid (u \in F, v \in C, (u, v) \in E)$.

Definition 6 (i, j) pruning

Given a BG $= (F, C, E)$ and i, j are two integers $(i, j \geqslant 2)$. The (i, j) pruning is to scan BG repeatedly. Through each scan, every node in EOS is pruned if its out-degree is

below i and every node in EOD is pruned if its in-degree is below j. This operation is iterated until no edges are pruned.

Theorem 1 xBG contain all possible xCores.

Proof Let $G = (V, E)$ be a directed graph and x is a node. $x\mathrm{BG} = \mathrm{BG}(F, C, E')$ is derived from G by forward and backward expansion. If there exists any $x\mathrm{Core} = (F', C')$, we have $\forall u \mid (u \in F', N_G^+(u) \supseteq N_{\mathrm{BG}}^+(u) \supseteq N_{\mathrm{CBG}}^+(u))$ and $\forall v \mid (v \in C', N_G^-(v) \supseteq N_{\mathrm{BG}}^-(v) \supseteq N_{\mathrm{CBG}}^-(v))$. So for any $x \in F'$, we have $N_{\mathrm{BG}}^+(x) \supseteq N_{\mathrm{CBG}}^+(x)$, it is also $C \supseteq C'$. Then $\bigcup_{u \in C} N_G^-(u) \supseteq \bigcup_{v \in C'} N_G^-(v)$ holds, it is also $F \supseteq F'$. Due to $E' = \{(u, v)\} \mid (u \in F, v \in C, (u, v) \in E)$ and $C \supseteq C'$ and $F \supseteq F'$, $E' \supseteq \{F' \times C'\}$. Therefore, xBG contains all possible xCores.

Lemma 1 If $x\mathrm{BG}^* = (F^*, C^*, E^*)$ is $(2, 2)$ pruned from $x\mathrm{BG} = (F, C, E)$, then $x \in F^*$ and $N_{x\mathrm{BG}^*}^+(x) = C^*$.

Proof Because of $N_{x\mathrm{BG}}^+(x) = C$, this means that each node in C has a link from x. So $x\mathrm{BG}^*$ is pruned always with $x \in F^*$ and $N_{x\mathrm{BG}^*}^+(x) = C^*$ unless $x\mathrm{BG}^*$ is empty.

Theorem 2 If $x\mathrm{BG}^* = (F^*, C^*, E^*)$ is $(2, 2)$ pruned from $x\mathrm{BG} = (F, C, E)$, then $x\mathrm{BG}^*$ is the x Cores-set.

Proof Assume that $x\mathrm{BG}^* = (F^*, C^*, E^*)$ is $(2, 2)$ pruned from $x\mathrm{BG} = (F, C, E)$. From lemma 1, $x \in F^*$ and $N_{x\mathrm{BG}^*}^+(x) = C^*$. Because any node $u \in F^* \mid u \neq x$ has $\mid N_{x\mathrm{BG}^*}^+(u) \mid \geqslant 2$, $\mid N_{x\mathrm{BG}^*}^+(u) \cap N_{x\mathrm{BG}^*}^+(x) \mid \geqslant 2$, that is to say u and x can form a $C(2, m)$, and $\mathrm{CBG} = (F', C')$, where $F' = \{x, u\}$ and $C' = N_{x\mathrm{BG}^*}^+(u) \cap N_{x\mathrm{BG}^*}^+(x)$, $m = \mid C' \mid$.

According to theorem 1 and theorem 2, we can obtain xCores-set from xBG.

2.5 Deleting strategy of subgraph

We can construct an $x\mathrm{BG} = (F, C, E)$ from any node x of G and then apply $(2, 2)$ pruning to it. If xBG is not empty after pruning, it must be an xCores-set. After every scan, whether or not xBG contains an xCores-set, we always delete some of the edges related with x. This operation includes EOS and EOD. But for the deletion of other nodes $u (u \in F, u \neq x)$, we should make a further decision. If $N_{\mathrm{BG}}^+(u) = N_G^+(u)$, this means that xBG has included all the possible cores which contain node u. Therefore, u and related edge can be deleted from G. Otherwise, u is not deleted. As a result, every scan of EOS will cause some edge deletion from G and the algorithm will converge to empty G.

2.6 Algorithm

From previous definitions, we can construct xBG from any node x of a topic subgraph and then apply $(2, 2)$ pruning to xBG. If xBG remains non empty, it is clear that an xCores-set is found and we output the xCorse-set. Whether xBG is empty or not, we always delete some edges from the topic graph. The complexity of the algorithm depends on n and m, where n is the number of source nodes in EOS and m is the number of source nodes in EOD. If xBG has sets F and C, $\mid F \mid = n$ and $\mid C \mid = m$, the worst time for $(2, 2)$ pruning is $O(n^2 + m^2)$. The pseudocode for an xCores-set extraction is shown as follows:

Algorithm 1 xCores-set extracting

```
Extract_xCoers-set(G){
  While G is not empty{
    node x = first item of EOS(G)
    Construct_xBG from x; //construct xBG
    Apply (2, 2) pruning to xBG; //Prune xBG with i =
2 and j = 2
    if xBG is not empty
    output xBG; //output xCores-set
     Update G according to xBG or x; //delete edges
from G
    }
  }
```

2.7 Hierarchical clustering of cores

After applying our algorithm to the topic subgraph, we obtain a set of xCores-sets. Due to overlap between xCores-sets, we need to combine the section of overlap again. According to hierarchical clustering, two objects will merge if the similarity is above a threshold. We employ down-top method to clustering xCores-sets into a dendrogram.

2.8 Evaluation

In order to evaluate the quality of communities, we use the measure of the quality of a network division defined by Newman and Girvan, which is called the modularity[11]. In many complex networks, some nodes may belong to more than one community. Each vertex can be assigned a membership in a community to solve the problem. In the overlapping community structure, define Q as that in Ref. [12].

If there are k communities in total, define a corresponding $n \times k'$ soft assignment matrix $\boldsymbol{U}_k = \{u_1, \ldots, u_k\}$ with $0 \leqslant u_{ic} \leqslant 1$ for each $c = 1, 2, \ldots, k$ and $\sum_{c=1}^{k} u_{ic} = 1$ for each $i = 1, 2, \ldots, n$. With this, define the membership of each community as $\bar{V}_c = \{i \mid u_{ic} > \lambda, i \in V\}$, where λ is a threshold that can convert a soft assignment into final clustering. A new modularity function $\tilde{Q}$ is defined as

$$\tilde{Q}(U_k) = \sum_{c=1}^{k} \left[\frac{A(\bar{V}_c, \bar{V}_c)}{A(V, V)} - \left(\frac{A(\bar{V}_c, V)}{A(V, V)} \right)^2 \right]$$

where U_k is a fuzzy partition of the vertices into k groups and

$$A(\bar{V}_c, \bar{V}_c) = \sum_{i \in \bar{V}_c, j \in \bar{V}_c} \frac{u_{ic} + u_{jc}}{2} w(i, j)$$

$$A(\bar{V}_c, \bar{V}) = A(\bar{V}_c, \bar{V}_c) + \sum_{i \in \bar{V}_c, j \in V \setminus \bar{V}_c} \frac{u_{ic} + (1 - u_{jc})}{2} w(i, j)$$

$$A(V, V) = \sum_{i \in V, j \in V} w(i, j)$$

This can be thought of as a generalization of Newman's Q function. Modularity is widely used in social networks; in a non-social network, for example, a web graph, is also useful[11].

3 Experiment and Result Analysis

3.1 Setup

The collection of datasets is the same as that of a typical procedure of HITS. We expand it to somewhat. We choose 10 topics for consideration. For each given topic, we submit it to four popular search engines: www. baidu. com; www. google. cn; www. yahoo. com; and www. altavista. com.

The union of four returned results from these engines is S_{root}. Then we expand S_{root} to topic subgraph by including the pages pointing to and pointed out by the nodes of S_{root}. Tab. 1 shows the data we prepared and the *x*Cores-set we found. Fig. 2 shows three phrases of extracting *x*Cores-sets on "Audi".

Tab. 1 The dataset and *x*Cores-set extracted

Topics	$\|S_{root}\|$	Nodes	Edges	*x*Cores-set
Linux	499	3 4970	46 218	982
Algorithm	495	6 204	6 332	97
Movie	577	25 933	28 117	163
Music	659	41 506	46 649	346
Football	425	13 302	15 464	232
Ajax	503	22 875	24 601	61
Audi	413	12 404	13 867	133
Software	540	40 924	44 135	196
Wallpaper	576	23 252	26 644	318
Stocks	577	27 066	34 172	791

3.2 Result analysis

In Fig. 3, we show the dendrogram of community "Audi". In the figure, we also show the modularity, which is aligned with the dendrogram so that one can directly read modularity values for different divisions of the network. The figure presents a single peak and the height of the peak is around 0.35, which indicates that the network is a typical community structure. In Fig. 3, we also show the communities where the modularity has a single peak. Because the leaves of the dendrogram are bipartite core rather than vertices, in the figure modularity is shown by a broken line after the leaves. We can see from the results of manual classification, Audi homepage, Audi sale, Audi forum and so on have good divisions in the communities where the modularity has a single peak.

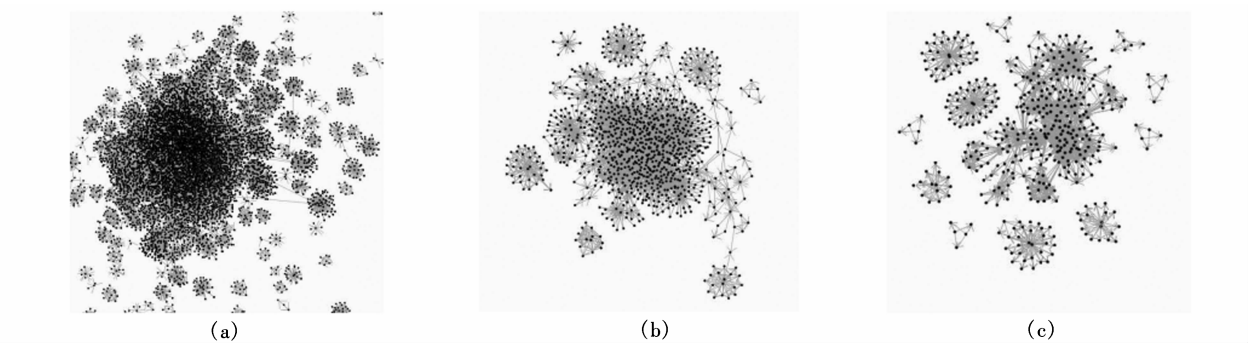

Fig. 2 Three phrases of extracting *x*Cores-set on "Audi". (a) Topic subgraph; (b) Pruned subgraph; (c) *x*Cores-set

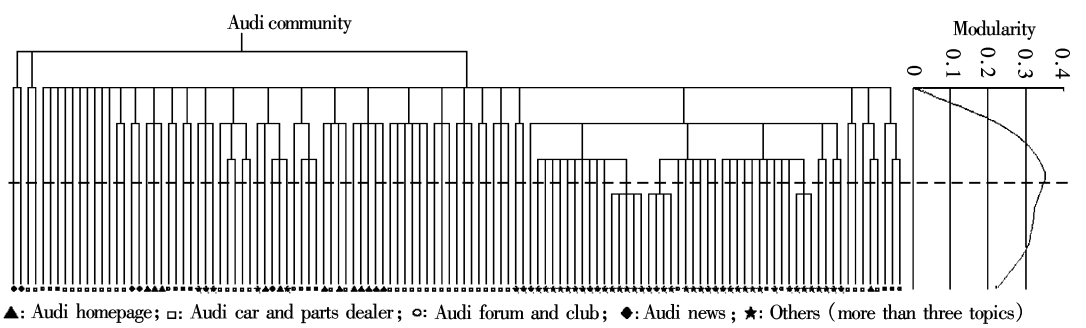

Fig. 3 The dendrogram of community "Audi"

4 Conclusion

The web, as an enormous information resource, is not only abundant in content, but also diverse in format. The web is changing the life-styles of people and it includes more and more knowledge for human beings. For this reason, searching on the web is an increasingly urgent job. The web community is a very important characteristic in the structure of the web and will help people in search of the information that interests them. This paper focuses on some key problems in the web community's discovery and link analysis. Based on the topic-oriented technology, the characteristics of the bipartite graph are intensely studied. An *x* bipartite core set is introduced to define extracting ways more clearly. By scanning a topic subgraph to construct an *x* bipartite graph and apply (2, 2) pruning to it, an *x* bipartite core set, which is also the minimum element of a community, can be found. Finally, a hierarchical clustering algorithm is applied to many *x* bipartite core sets and a dendrogram of community inner construction is obtained. We prove the correctness of the constructing and pruning algorithm and implement them in the end. The typical datasets in the experiment are prepared according to the way they are in HITS. Ten topics and four search engines are chosen and the returned results are integrated. And we use modularity, which is a measure of the strength of the community structure that is often used in social networks, to validate the efficiency of our method. The experimental results show that our algorithm is effective and efficient. Future work will engage in the contents of a page, on both html text and semantic meaning between hyperlinks.

References

[1] BOUTELL. COM. WWW FAQs: How many websites are there?[EB/OL]. (2007-02-15)[2008-03-15]. http://www.boutell.com/newfaq/misc/sizeofweb.html.

[2] Kleinberg J, Lawrence S. The structure of the web[J]. *Science,* 2001, **294**(30): 1849 - 1850.
[3] Kumar R, Raghavan R, Rajagopalan S, et al. Trawling the web for emerging cyber-communities[C]//*Proc of the* 8*th Intl WWW Conf*. Toronto, Canada, 1999: 403 - 415.
[4] Gibson D, Kleinberg J, Raghavan P. Inferring web communities from link topology[C]//*Proc of the* 9*th ACM Conf on Hypertext and Hypermedia*. Pittsburgh, PA, USA, 1998: 225 - 234.
[5] Kumar R, Raghavan P, Rajagopalan S, et al. Extracting large-scale knowledge bases from the web[C]//*Proc of the* 25*th VLDB Conference*. Edinburgh, Scotland, 1999: 639 - 650.
[6] Dourisboure Y, Geraci F, Pellegrini M. Extraction and classification of dense communities in the web[C]//*Proc of the* 16*th Intl WWW Conf*. Banff, Alberta, Canada, 2007: 461 - 470.
[7] Flake G W, Lawrence S, Giles C L. Efficient identification of Web communities[C]//*Proc of the* 6*th ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining*. Boston, MA, USA, 2000: 150 - 160.
[8] Flake G W, Lawrence S, Giles C L, et al. Self-organization and identification of web communities[J]. *IEEE Computer*, 2002, **35**(3): 66 - 71.
[9] Wang Xiaoyu, Lu Zhiguo, Zhou Aoying. Topic exploration and distillation for Web search by a similarity-based analysis [C]//*Proc of the* 3*rd Intl Conf on Web-Age Information Management*(*WAIM'*02). Beijing, China, 2002: 316 - 327.
[10] Montfort N. Discovering communities through information structure and dynamics: a review of recent research. MS-CIS-04-18 [R]. Philadelphia, USA: University of Pennsylvania, 2004.
[11] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. *Phys Rev E*, 2004, **69**(2): 026113.
[12] White S, Smyth P. A spectral clustering approach to finding communities in graphs[C]//*SIAM International Conference on Data Mining*. Philadelphia, USA, 2005: 274 - 285.

Web 紧密核的抽取和评价方法

杨 楠 高 洁 薛鸿鹄 刘秀德

(中国人民大学信息学院,北京 100872)

摘要:针对 web 社区的发现和链接分析技术的一些关键问题,基于面向主题的技术,重点研究了二分图的特征,引入了 x 二分核集来更为明确地定义抽取的方法. 通过扫描主题子图构造 x 二分图,对该子图的 (i,j) 裁剪后得到 x 二分核集,这也是社区的最小元素. 最后,对所抽取的所有 x 二分核集应用层次聚类的方法得到社区内部结构的树状图,证明了构造和裁剪方法的正确性并设计了算法. 实验采用 HITS(hyperlink-induced topic search)算法中的典型数据集获取方法,选择了 10 个主题和 4 个搜索引擎并综合返回的结果. 采用社会网中测量社区结构强度的模块化度量来验证所提方法的有效性,实验结果表明所提方法是有效并可行的.

关键词:紧密核;链接分析;层次聚类;模块化度量

中图分类号:TP311