面向 CRIC 的 Web 社区发现方法研究

李 翠
LI Cui

西安财经学院 信息学院,西安 710100
School of Information,Xi'an University of Finance and Economics,Xi'an 710100,China
E-mail:bxdlicui@163.com

LI Cui.Web community discovery research for cluster ranking of integrated cohesion.Computer Engineering and Applications,2009,45(25):129-131.

Abstract: Aimming at the deficiency of traditional Web community discovery algorithm and the problem of cluster strength measure,the object of Web community nodes and edges and structure is given.A new cluster strength measure method is researched,in order to settle the problems of community optimal partitioning and subject optimization.Object function of maximal community is presented,community discovery algorithm based on cluster ranking of integrated cohesion is described,and application system is built based on existing information searching kit.The result of experiment shows that the algorithm can fast, effectively search global optimum partition of network structure.This algorithm is highly effective and valuable in practice and academic study.

Key words: Web community;cluster ranking;integrated cohesion;partitioning

摘 要:针对现有 Web 社区发现方法存在的不足及其聚合程度的测量问题,以社区节点、边、结构为对象,研究 Web 社区聚合强度的测量方法,分析社区最大化目标函数,以解决社区最优划分及主题优化问题,并提出 CRIC 社区发现算法。在现有信息搜索软件工具包的基础上构建其应用系统,实验结果验证该算法的有效性及适用性,能快速、高效地完成对网络社区的划分,具有一定的理论及应用价值。

关键词:Web 社区;聚类等级;集成聚合度;划分
DOI:10.3778/j.issn.1002-8331.2009.25.039 **文章编号**:1002-8331(2009)25-0129-03 **文献标识码**:A **中图分类号**:TP391

1 引言

随着 Web 信息资源的急剧膨胀,如何从中抽取潜在的、有价值的信息,进而充分有效地利用 Web 信息资源,是当今信息领域及知识工程领域最具挑战性的研究内容之一[1]。Web 社区发现技术是提高网络搜索引擎检索质量的重要途径之一,发现 Web 社区对信息检索有很多好处, 不同的 Web 社区展示不同的内容,将搜索结果以社区的方式呈现并自动进行聚合,有助于用户迅速定位到感兴趣的内容, 还可辅助制作网页目录等。可见,对 Web 社区发现的研究具有重大的实际应用价值,其研究的关键问题在于如何利用较少的先验信息实现对网络的高效划分,进而提高信息搜索结果的质量。

目前对社区进行研究分析的方法大致分为两种:(1)网络结构化分析;(2)基于个体属性的代数分析。纵观社区分析算法,可分为四个方向:(1)迭代二分法;(2)层次聚类法;(3)G-N(Girvan-Newman)方法;(4)图结构分析法。文献[2-3]给出了具体的实现算法,但很多算法仍存在不同程度的缺陷。

主要利用一种集成聚类等级(Cluster Ranking of Integrated Cohesion,CRIC)算法进行 Web 社区划分,通过分析不同的聚合强度的度量及最大化目标函数解决社区划分优劣及主题优化问题,并基于原始数据获取工具及现有信息搜索软件工具包的基础上,构建了 Web 社区发现应用系统对算法进行验证,最后通过准确率及召回率对获取结果进行评价。

2 Web 社区聚合强度的度量

Web 社区的聚团程度是网络的复杂特性之一,选择合适的聚合强度测量方法至关重要,至今仍没有形成体系,定义方式也多样[4-5]。Web 社区的聚合程度可表示为映射函数 $\mu(c):c\rightarrow R^+, C\subseteq G, C$ 为社区,G 为网络。节点聚合程度只取决于 C 内节点连接度而与 C 和 G/C 的连接方式无关, 可见节点聚合程度本身不足以描述该社区即为所需, 进一步考虑边的聚合度、信息传递聚合度及集成聚合度十分必要。

2.1 Web 社区的节点聚合度

给定 Web 图 $G=(V_G,E_G)$,$n=|V_G|$,$u,v\in V_G$,p_{uv} 为节点 u 到 v 所有简单路径集合,$P=\cup_{u\neq v}p_{uv}$, 则从 u 到 v 的流量 f 为从 p 到

基金项目:陕西省自然科学基金(the Natural Science Foundation of Shaanxi Province of China under Grant No.2007F52);陕西省科技厅资助项目(No.2007F25)。

作者简介:李翠(1979-),女,讲师,研究方向:数据挖掘,知识发现。

收稿日期:2008-10-22 **修回日期**:2008-12-25

R^+的函数：$\sum_{p\in p_{uv}} f(p)=1$，$\forall_{u,v}\in V_G, u\neq v$，关于流量 f 所经节点 $v\in V_G$ 的聚合度记为 $\beta c_f(v)$，取决于通过节点 v 及以 v 为终点的流量的总和，即 $\beta c_f(v)=\sum_{p \text{ passes through } v} f(p)+\sum_{p \text{ starts or ends at } v} f(p)/2$，如果 $n\geqslant 2$，则对 $v\in V_G$，$\beta c_f(v)\geqslant 1$。

2.2 Web 社区边的聚合度

对于边 $e=(u,v)$，$u,v\in V_G$ 的聚合度可描述为：$\varepsilon cf(e)=\sum_{p \text{ passes through } e} f(p)/\varpi(e)$，$\varpi(e)$为边 e 的权值。令(U,U^c)为 G 中划分的非空、不相交、互补的两个集合，$<U,U^c>$是从 U 到 U^c 的有向边的集合，边的稀疏度为 $\rho(U,U^c)=\varpi(U,U^c)/(|U||U^c|)$，其中，$\varpi(U,U^c)=\sum e\in<U,U^c>,\varpi(e)$为从 U 到 U^c 的有向边的权值之和。

2.3 Web 社区信息传递的聚合度

为衡量社区内信息传递的聚合程度(Transitive Cohesion)，可用 $TC_i=\sum_{j=1}^{n}\sum_{k=1}^{n_i} p_{jk}/(n_i(n_i-1))$，$(j\neq k)$进行度量，其中 p_{jk} 为节点 k 与 j 之间的传递概率，n_i 为社区 C 内节点数，若将两节点之间的传递概率看作时节点之间的属性相关度，则 TC_i 主要用来衡量经过随机传递后，社区内节点发出的信息传播留在社区内的概率，即反映了社区内信息传递的聚合程度。

2.4 Web 社区结构聚合度与集成聚合度

令 S 为 G 中分离的节点集，由 S 可通过映射函数 $\chi(S,A,B)=<U,U^*>$，将 $V_G\backslash S$ 划分为 A 和 B，$S=\{s_1,\cdots,s_l\}(l\geqslant 1)$，$A=\{a_1,\cdots,a_q\}(q\geqslant 0)$，$B=\{b_1,\cdots,b_r\}(r\geqslant 0)$，$l+q+r=n$，$S$ 的稀疏性记为 $\eta_G=(S,A,B)=|S|/((2|A|+|S|)(2|B|+|S|))$，$U=\{A^{in}\cup A^{out}\cup S^{in}\}$，$U^*=\{S^{out}\cup B^{in}\cup B^{out}\}$，也即 $A=\{v:\{v^{in},v^{out}\}\in U\}$，$B=\{v:\{v^{in},v^{out}\}\in U^*\}$，$S=\{v:\{v^{in},v^{out}\}\in<U,U^*>\}$，$S\cap A=\phi$，$S\cap B=\phi$，$A\cap B=\phi$，$S\cup A\cup B=V_G$，即对于节点 $u\in A$ 和 $v\in B$ 之间无链接。

Web 社区结构聚合度可描述为：$cohesion(c)=\min_{(S,A,B)}|S|/(\min\{|A|,|B|\}+|S|)$，$cohesion(c)\in(0,1)$，其中$|S|/(\min\{|A|,|B|\}+|S|)$作为社区结构划分的稀疏度，当 S 很小且 A 和 B 均较大时它是最小的，也即当且仅当社区 C 不能通过消除部分节点而被进一步划分时它是聚合的。

进一步可得 Web 社区集成聚合度（integrated cohesion）：int $cohesion(c)=\int_0^{\infty} cohesion(c^\tau)\mathrm{d}\tau$，实际上 $cohesion(c)$是有限积分，当边界值 τ 大于最大边权值时，c^τ 为空图，$cohesion(c^\tau)=0$，即可通过累加最多不超过$|E_G|$个 $cohesion(c)$值而获得 int $cohesion(c)$。

通过以上度量及判定可知，在传统的 Web 社区分析中往往会忽略了其他特征，降低了社区发现的准确性，利用集成聚合度进行聚类是提高社区发现的准确性的有效方法，可有效地反映 Web 社区聚团程度即 Web 社区的概念及社区之间的复杂关系。

3 Web 社区的最优划分

3.1 Web 社区最大化目标函数

令 $G=(V_G,E_G)$，若所得 Web 社区 $C\subseteq V_G$ 不被其他社区包含，则为最大化。最大化标准为布尔函数，将划分的社区映射到 $\{0,1\}$，被映射为 1 的所有社区满足最大化，被映射为 0 的社区非最大化。可由边界值 ε 确定社区是否最大化，如果存在区间$[T_1,T_2]$，其中 $T_2\geqslant(1+\varepsilon)T_1$，则对于 $T\in[T_1,T_2]$，G^T 中的社区最大化。

Web 社区最大化目标函数 $Q=\sum_c[L_c/L-(D_c/2L)^2]$，其中 c 为社区，L 为链接总数，L_c 为社区 c 内节点间的链接数，D_c 为社区 c 的度，其值为社区 c 中节点度的总和，即 $D_c=\sum_n d_n Y_{nc}$，$\forall_c, d_n$ 为节点 n 的度。

3.2 Web 社区主题优化

对非空社区 c，有 $\begin{cases}\sum_l X_{lc}\geqslant\alpha E_c,\ \forall_c\\ \sum_l X_{lc}\geqslant\beta E_c,\ \forall_c\end{cases}$，$E_c=\begin{cases}1 & c\text{ 非空}\\ 0 & \text{否则}\end{cases}$。每个节点应准确分配到一个社区中，即 $\sum_n Y_{nc}=1$，$\forall_n$，$Y_{nc}=\begin{cases}1 & \text{if } n\in c\\ 0 & \text{otherwise}\end{cases}$，如果两节点 n,e 都在社区 c 中，则其 n,e 间的链接 l 也属于 c，即 $2X_{lc}\leqslant Y_{nc}+Y_{ec}$，$\forall_c, l=\{n,e\}$。$X_{lc}=\begin{cases}1 & \text{if}_{l\in c}\\ 0 & \text{otherwise}\end{cases}$，也可分解为 $X_{lc}\leqslant Y_{nc}$，$\forall_c, X_{lc}\leqslant Y_{ec}$，$\forall_c, l=\{n,e\}$，因此，$L_c=\sum_l X_{lc}$，$\forall_c$。

完成社区抽取后，需决定每个社区内的主题优化，可转化为社区同解问题。假设将所有节点划分到 M 个社区中，记为 AM_n 集，为了避免社区间同解，将每个节点允许分配到其中一个社区中，所有节点将基于连通性进行排序，假设 n_1 时度最大的节点，n_2 其次，则 AM_n 集构造过程为：n_1 划分到社区 1 中，n_2 可划分到社区 1 或 2 中，其他节点可划分到任意社区中，即 $\sum_{c\in AM_n} Y_{nc}=1$，$\forall_n$，节点 $n\in AM_n, e\in AM_e$，则 n 与 e 间的链接 l 有：$l\in AM_n$。令集合 $M_L=AM_n\cap AM_e$，其中 $l=\{n,e\}$，则 $X_{lc}\leqslant Y_{nc}$，$\forall_L=\{n,e\}, m\in ML_l, X_{lc}\in Y_{ec}$，$\forall_{l=\{n,e\}}, X_{lc}=0$，$\forall_{l,m\notin ML_l}$。令 AV_m 为社区 m 的节点集，B_n 为链接数大于 n 的节点集，假设 $m\in AM_n$，$e\in B_n\cap AV_{m-1}$，$\sum_{e\in(B\cap AV_{m-1})} Y_{e,m-1}=0$，则 $Y_{nc}=0$。

3.3 Web 社区内孤立点的柔性修复策略

在 Web 社区划分及优化过程中，存在孤立点的现象，所谓孤立点就是 Web 社区中其邻接的图节点全部或绝大多数与其相异社区的节点。孤立点的存在会影响 Web 社区发现算法搜索的效率，制定不同强度的修复策略可提高算法的收敛策略，同时能避免算法陷入局部最优。

处理孤立点的传统方法采用绝对修复策略，如在算法搜索过程中发现孤立点，将其强制转化为与其邻接点相同的社区。对孤立点更好的办法是采用柔性策略，计算与孤立点邻接且同社区的节点数目 α，与其邻接的异社区节点数目 β，令 $\eta=\beta/\alpha$，若 η 大于给定阈值 ε，则改变社区分组，否则保持社区分组不变。

4 CRIC 的 Web 社区发现算法

4.1 CRIC 算法描述

集成聚类等级的 Web 社区发现算法通过 3 步完成：(1)鉴别候选社区；(2)对候选社区进行等级评价；(3)消除非最大化社区。具体算法描述如下：

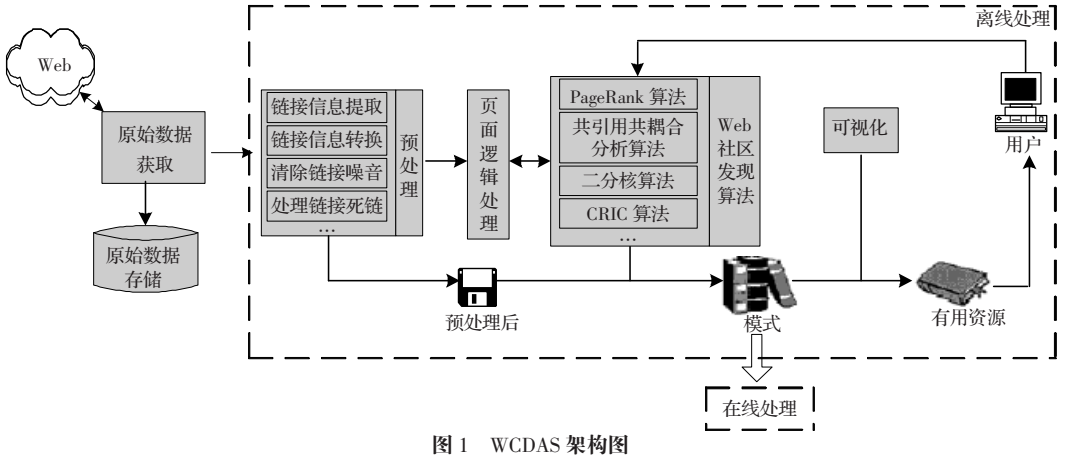

图 1　WCDAS 架构图

Input:G.　　// Web 图

Output:The communities in G.　　//划分的 Web 社区

Step1:identification of candidate communities L

（1）add G to L;

（2）if(G is a clique or a singleton) return;

（3）S=sparsest vertex set of G;

（4）$A_1,\cdots,A_k$=connected components of $G\backslash S$;

（5）for i=1 to k do

G_i=sub-network of G induced on $S\cup A_i$;

If Gi not already in L then,for Gi return（1）;

Step2:candidate ranking

（1）for each candidate community Gi do

int $cohesion(c)=\int_0^{\infty} cohesion(c^{\tau})\mathrm{d}\tau$;

（2）ranking the maximal communities in G by their strength;

step3:candidate elimination

for each community C_i do

comparing against supersets that belong to the list of candidates;

Eliminating non-maximal communities.

4.2　算法评价

输入 Web 图 G，经 CRIC 社区发现算法 A 返回的所有社区记为 $A(G)$,令 $I(G)$为 A 的理想输出,即 G 中用户所需的社区,算法质量可通过 A 的召回率(recall)及准确率(precision)进行评价。

$$recall(A,G)=\frac{|A(G)\cap I(G)|}{|I(G)|},precision(A,G)=\frac{|A(G)\cap I(G)|}{|A(G)|}$$

5　实验结果与分析

5.1　Web 社区发现应用系统架构

为测试 CRIC 算法能否发现或抽取所需的社区结构,并和著名的 PageRank 算法比较，设计了 Web 社区发现应用系统(Web Community Discovery Application System,WCDAS),主要包括原始数据获取、离线处理和在线处理等部分,具体架构如图 1 所示。其中:

(1)原始数据获取,是 WCDAS 进行算法分析的基础,即利用计算机生成有已知社区结构的网络。

(2)离线处理,对原始数据进行预处理,以便作为 Web 社区发现算法的输入,通过相关参数的调整进行社区划分,最后以易于理解的可视化方式呈现给用户。

(3)在线处理,目的是进一步应用与扩展离线部分中的 Web 社区发现算法,更好地与 Web 信息搜索技术相结合,以提高信息搜索精度,其具体架构如图 2 所示。

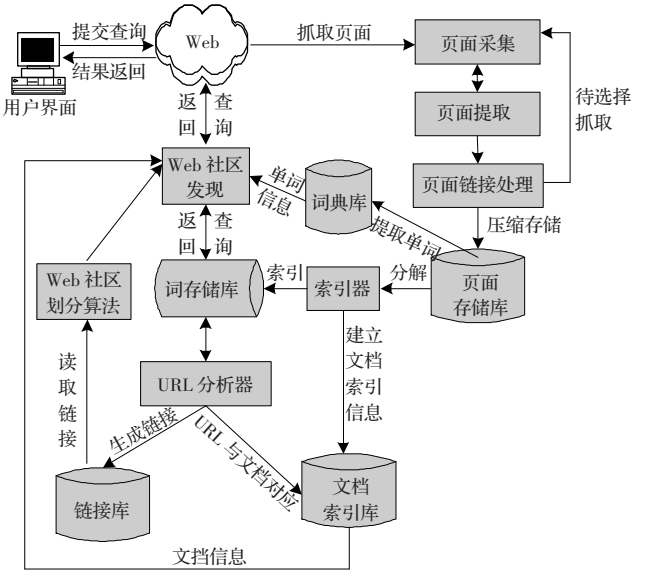

图 2　在线处理

5.2　Web 社区发现应用系统的实现

WCDAS 是基于原始数据获取工具及现有信息搜索软件工具包的基础上实现的，开发语言选择 C#，开发环境 IDE 选用 Visual Studio 2005,原始数据获取工具选用了 Offline Explorer 4.5.0.2532,检索软件工具包选用 DotLucene1.9。

5.3　Web 社区发现算法结果分析

通过实验希望解决 3 个问题:(1)采用 CRIC 算法是否能够提高 Web 社区发现的准确性;(2)社区主题是否能够得到优化;(3)社区内孤立点是否能得到修复。

将收集到的 Web 用户可能感兴趣的查询请求，提交给 WCDAS 系统,分别选择 PageRank 算法[6]和 CRIC 算法进行社区划分及主题优化,对返回的结果进行分析,由于很难估计文档库中与查询相关的社区文档数目，所以只计算针对前 8 个请求的相应准确率 $precision(A,G)$,并对其进行对比分析,如图 3 所示。

(下转 162 页)