基于信息熵的社区发现算法研究

王 刚[1] 钟国祥[2]
(安康学院电子与信息工程系 安康 725000)[1] (重庆教育学院科技处 重庆 400067)[2]

摘 要 针对现有社区发现依靠出度、入度、介数来进行社会划分的一些不足,研究了依靠信息熵来对社区进行度量,提出了基于信息熵的社区发现算法 CDBE(Community Detection Based on Entropy)。如果社区内部信息量大,熵就大。不确定事件发生的概率就大。社区具有凝聚力,信息的熵相对稳定,不会出现熵剧烈增加或减少的情况,根据节点集合熵的变化是否剧烈,可以判断节点是否是社区的成员,从而实现社区的发现。实验表明,CDBE 能够发现有价值的社区。

关键词 社区发现,信息熵,推荐系统,数据挖掘

Study on Algorithm of Community Detection Based on Information Entropy

WANG Gang[1] ZHONG Guo-xiang[2]
(Dept. of Electronic& Information Engineering, Ankang University, Ankang 725000, China)[1]
(Educational College of Chongqing, Chongqing 400067, China)[2]

Abstract There are some faults of present community detection algorithm, which is based on the in degree, out degree and betweenness of nodes, we presented a algorithm based on Entropy to detect community structure. A community includes many information and it's Entropy. Members of a community have some common gains or interests, we think that if a member want to join a community, it can't make the entropy of the community exceed a threshold, otherwise it can't be the member of a exist community. Our experiments show the processing and the efficiency of our algorithms.

Keywords Community detection, Information entropy, Recommendation system, Data mining

1 引言

基于复杂网络的社区发现成为当前研究的热点,现实世界中的很多系统都可以采用网络的形式来加以描述,复杂网络是复杂系统的抽象,网络中的节点是复杂系统中的个体,节点之间的边则是系统中个体之间按照某种规则而自然形成或人为构造的一种关系[1]。现实世界中包含着各种类型的复杂网络,如社会网络、万维网、生物网络。复杂网络的研究表明,复杂网络具有若干统计特征,包括如小世界性质、无尺度性质、泊松分布、幂律分布以及聚集性、传递性。复杂网络的另一个重要特征就是网络中所呈现出的社区结构。大量实证研究表明,许多网络是异构的,即复杂网络不是大批性质相同节点的随机连接,而是许多类型的节点的组合,其中相同类型的节点存在较多的连接,而不同类型节点的连接则相对较少。我们把同一类型节点以及这些节点之间的边所构成的子图称为网络中的社区。实际网络的社区代表着特定对象的集合,如,社会网络中的社区代表根据兴趣或背景而形成的真实的社会团体;引文网络中的社区代表针对同一主题的相关论文;万维网中的社区就是讨论相关主题的若干网站;而生物化学网络或者电子电路网络中的社区可以是某一类功能单元。发现这些网络中的社区有助于我们更加有效地理解和开发这些网络。

目前社区发现的算法主要有谱平分法[2]、KL 算法、层次聚类法、GN 算法以及相应算法的改进,这些算法有些利用了图的结构特征,如 GN 算法,利用介数来进行图像的分割,通过删除通信集中的边来得到两个社区;有些利用了相似度计算,如层次聚类法,它认为相邻节点应该等价或相似度很高,例如,如果两个人拥有完全相同的朋友,则两个人等价,可以通过邻接矩阵行向量均值和方差来计算其相关性;有些算法利用了节点矩阵的一些特点,如谱平分法,它认为矩阵特征值的正负对应两个不同社区;KL 算法利用了增益函数来对网络进行划分,通过使增益函数最大来确定社区。这些算法在取得成功的同时,通常也由于计算代价高、需要事先明确存在社区的数量以及需要明确网络的结构等,在实际应用中有一些局限[3]。由于这些算法注重于网络的结构特点,如节点的出度、入度、介数,而文献[4,5]很少注重于网络本身传递出的信息,如等价关系、从属关系、因果关系等、关联关系等,使得发现的社区结构特征很明显,它们通常认为,聚集结构上临近的节点应该属于一个社区,而实际应用中,考虑到一些具体的关系,上述结论不一定正确,如应用 GN 算法,可能会把边 ab 作为两个社区的分界线,如果 a 和 b 本身属于事件非常重要的成员,把它们分割到两个社区就不合适。信息熵理论认为,随着网络节点的扩充,信息量增加,网络蕴含不确定信息的概率增加。因此,一个社区内部,由于成员的增加,出现不确定

到稿日期:2010-03-21 返修日期:2010-06-19 本文受陕西省教育厅项目(09JK317),基于本体的服务研究(AYQDZR200916),智能信息处理技术关键问题及应用研究(2008akxy005)资助。

王 刚(1972—),男,副教授,主要研究方向为人工智能,E-mail:aktcdawang@163.com;钟国祥(1970—),男,博士,教授,主要研究方向为人工智能。

性信息的概率应该增加，熵就增加，反之，则减少。一个社区内部，由于信息交流频繁，以及有共同的利益和目标，表现出强的凝聚性，不确定信息出现的概率不会剧烈增加或减少，使得根据节点集合熵的变化来确定不同的社区成为可能。判断一个节点是否属于一个社区，可以通过判断节点加入社区后，社区熵的变化来确定。据此，本文提出了基于熵的社区发现算法，它不但考虑网络的结构，也考虑节点之间的关系，本文考虑的是关联关系，以推荐系统为例，研究了社区发现及应用。

2 信息熵及网络图的建立

通常认为，信息是认识主体所感受的或所表达的事物运动的状态和运动状态变化的方式。信息是人们在适应外部世界和控制外部世界的过程中，同外部世界进行交换的内容。1948年，香农第一次将熵这一概念引入到信息论中，从此，熵这一概念被广泛用于信息的度量，在自然科学和社会科学众多领域中得到广泛应用，并成为一些新学科的理论基础。信息的特征为：(1)接收者在收到信息之前，对它的内容是不知道的，所以信息是新知识、新内容；(2)信息是能使认识某一事物的未知性或不确定性减少的有用知识；(3)信息可以产生，也可以消失，同时信息可以被携带、贮存及处理；(4)信息是可以量度的，信息量有多少的差别。把熵从物理熵理论引伸到信息学、社会学等领域，称之为信息熵[6]。信息熵是信息论中用于度量信息量的一个概念。一个系统越是有序，信息熵就越低；反之，一个系统越是混乱，信息熵就越高。变量的不确定性越大，熵也就越大，清楚彼此之间关系所需要的信息量也就越大。

信息用熵来度量，对于 n 个事件构成的概率系统，每一事件 i 产生的信息量为 $p_i\log_2 p_i$，n 个事件或信息构成系统的熵定义为 H，$H=-\sum p_i\log_2 p_i\ (i=1,2,\cdots,n)$。信息熵的大小用于表示概率系统的不确定程度，信息熵越大，表示信息混乱的可能性就越大，信息的冗余度大，反之，表示信息混乱的可能性就越小，信息的冗余度小。

电子商务蓬勃发展，为顾客提供人性化服务成为迫切需要，基于社区的服务推荐系统提供了一条有效的途径。当前很多基于数据库的数据挖掘系统用于挖掘关联规则[7]，这些系统通过寻找频繁项目集来寻找符合满意度与支持度的序列模式，从而进行推荐，这些系统面临的一个问题是，确定支持度与满意度比较困难，可能产生大量的序列模式，对序列模式的解释有时也显得比较困难[8]。本文从社区发现的角度，通过发现处于一个社区的商品，从而进行推荐，一定程度上可以避免上述问题。要进行社区的划分，需对销售数据库进行处理，用算法1构建商品销售记录网络图，它以商品为节点，节点之间的边表示顾客同时购买该两种商品的概率，利用销售数据库，通过节点的不断合并，得到商品的销售记录网络如图1所示。

算法1 网络图的形成

Begin：

(1)一条记录包含多个商品，每个商品为一个节点，它们彼此互联，得到图 $g1$；

(2)对于下一条记录，每个商品为一个节点，它们彼此互联，得到图 $g2$；

(3)合并 $g1$，$g2$ 中相同的节点，原 $g1$，$g2$ 中节点间的关系得到继承，得到合并后的图 $g1$，重复(2)，直到所有的记录添加完毕。

End

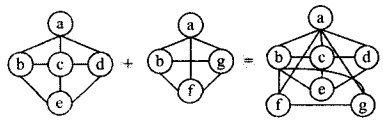

图1 节点合并示意图

由于信息熵需要事件发生的概率，显然各节点发生的概率不一样，通过分析销售数据库，本文把某事件发生的次数与总事件数的比作为该事件发生的概率。假设图中边 e 出现了 x_e 次，而数据库中两商品对应的关联关系有 y 个，则边 e 出现的概率定义为 $p_e=\frac{x_e}{y}$，例如上图中总的边数为16，每条边对应一个事件，如同时购买的商品为 (a,b,c,d,e)，其对应的边为 $(ab,ac,ad,ae,bc,bd,be,cd,ce,de)$，其对应的概率为 $(\frac{2}{16},\frac{1}{16},\frac{1}{16},\frac{1}{16},\frac{1}{16},\frac{1}{16},\frac{1}{16},\frac{1}{16},\frac{1}{16},\frac{1}{16})$；销售记录为 (a,b,f,g)，其对应的边为 $(ab,ac,ad,ae,bc,bd,be,cd,ce,de)$，对应的概率为 $(\frac{2}{16},\frac{1}{16},\frac{1}{16},\frac{1}{16},\frac{1}{16},\frac{1}{16})$。

3 社区结构的发现

一个社区内部，由于信息交流频繁，以及有共同利益和目标，表现出强的凝聚性，不会出现不确定信息剧烈增加或减少这样的情况，使得根据节点集合的熵的变化来确定不同的社区成为可能。判断一个节点是否属于一个社区，可以通过判断节点加入社区后，社区熵的变化来确定。算法先找出最大的一条边，得到集合，在增加另外一条边的时候，集合的熵必然增加，给定一个增加的阈值，如果新边的熵增加的量不符合阈值要求，该边就不应该加入社区，否则就加入社区。下面进行几个定义。

定义1 熵的增量 ∇H 定义为：$\nabla H=|H_1-H_2|$，H_1，H_2 为熵。∇H 为熵差的绝对值。

定义2 平均熵 $\overline{H}$ 定义为：$\overline{H}=\frac{\sum_{i=1}^{i=n}H_i}{n}$，$n$ 为节点个数。

定义3 利益函数定义为：$F=\nabla H-\overline{H}$。

定义4 如果某节点加入社区后系统熵的增量 ∇H 大于社区的平均熵 $\overline{H}$，则该边入社区，否则不加入社区。

算法2 基于熵的社区结构发现

Begin

(1)$i=0$

(2)建立一个堆栈 s，确定一个初始节点 a_i，a_i 为网络中概率最大边对应的节点

(3)计算 $\overline{H}$

(4)把 a_i 压入堆栈

(5)判断栈顶的邻节点 b_j，如果栈顶有邻节点，且满足定义4，则边 ab_j 标识为 true，b_j 压入堆栈，$j=j+1$

(6)如果栈顶没有邻节点，栈顶出栈，重复(5)，直到栈为空，输出 f_i 对应的社区

(7)$i=i+1$，对于没有标识的边，重复(2)，直到节点数不符合社区要求

End

该算法对所有的边进行了标识，可能存在很多孤立的点，它们不属于某个社区，这与现实生活中的情况也是相吻合的，

如某个孤立的事件、某个突然出现的情况等，实际应用时，可通过判断社区的大小来确定。若某剩余节点数不符合社区要求，就不认为它们构成一个社区，只有规模达到一定程度，才认为其构成社区。

4 示例

如图3所示，销售网络的最大熵为0.529，网络的总熵为7.48，最小熵为0.113，熵变化的平均值为0.356，与最大熵的差为0.463，图中边(ab,bf,fa,fg,gb,ga,bc,cj,jb,ce,eb,eh,hl,lk,ki,ih,im,cd,db,da,ca,fk)及其对应的熵分别为(0.36,0.113,0.152,0.113,0.152,0.332,0.442,0.529,0.006,0.152,0.292,0.113,0.152,0.113,0.113,0.152,0.006,0.387,0.113,0.113,0.006,0.006)，其值的变化及趋势如图2所示。GN算法发现的社区包含节点$abcdefg$,$hijklm$。而CDBE算法发现的社区为$abcjdg$,$efhlkmi$，如图3、图4所示，可见两种算法发现的社区有差异，CDBE着重于节点关联关系，而GE算法着重于节点连接的介数，在实际的运行过程中，把节点熵与最大节点熵的差作为熵的变化，如果该熵变化小于所有节点熵的平均值，则认为该节点属于社区，否则不属于该社区，本示例只运行算法一次，得到两个社区，如果把没有标识的节点继续细分，可以得到多个社区，当然应该考虑节点数是否符合社区要求，如果节点数很少，不认为构成一个社区。而CDBE算法结合商品销售的实际情况，每个社区内成员关联性很强，可以通过熵来度量，传统KDD算法一次运行其支持度、满意度是固定的。另外，假设节点中节点数为n，K-N算法的运行时间复杂度为$O(n^3)$，K-L算法的运行时间复杂度为$O(n^2)$，层次聚类法的时间复杂度为$O(n^2\mathrm{Log}^n)$，GN算法的时间复杂度为$O(m^2n)$。本文算法的时间复杂度最好的情况为$O(n^2)$。

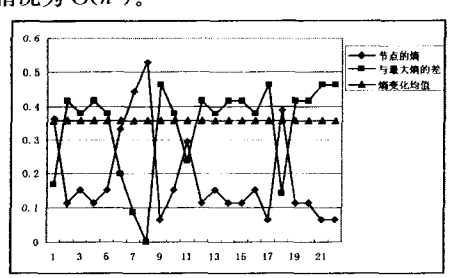

图2 熵，平均熵及熵的变化

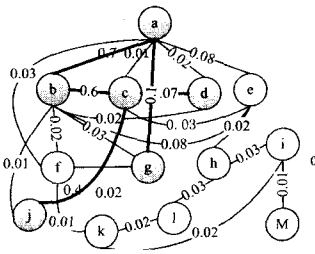

图3 本文发现的社区

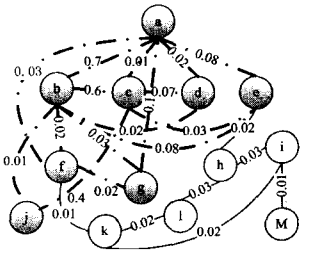

图4 GN算法发现的社区

结束语 根据商品销售记录，构建商品销售记录网络，提出了基于熵的社区划分方法，该方法既注重社区网络的结构，也注重社区网络连接的含义，即社区内熵的变化不会剧烈增加或减少。以前其它算法没有关注到这一点，与基于数据库的关联规则挖掘相比，该方法避免了指定支持度和满意度的不足，通过发现销售商品社区，从而进行推荐。将来的研究将进一步研究信息熵的理论，研究熵信息变化的规律并应用，以求发现更有意义的社区。

参考文献

[1] 丁元竹. 社区研究的理论与方法[M]. 北京：北京大学出版社，1995

[2] 丁连红，时鹏. 网络社区发现[M]. 北京：化学工业出版社，2008

[3] Girval M, Newman M. Community Structure in Social and Biological Network[C]// Proc. natl. acad. Sci. USA, 2002: 8271-8276

[4] Freeman L. A Set of Measure of Centrality Based Upon Betweens[J]. Sociometry, 1997(40): 35-41

[5] Albert R, Barabasi A L. Statistical Mechanisms of Complex Networks[J]. Reviews of Modern Physics, 2002(74)

[6] 唐鹏，张自力. 基于信息熵的多 Agent DDoS 攻击检测[J]. 计算机科学，2008(3)

[7] 唐敏. 关联规则挖掘算法在超市销售分析中的应用[J]. 计算机科学，2006(2)

[8] 范明，孟小峰. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2002：223-243

[9] 高博，周脩，崔英志. Web2.0 网站的特点与社区化模式[J]. 重庆工学院学报：自然科学版，2009，23(6)：102-106

(上接第190页)

[5] Fuchs N E, Kaljurand K, Kuhn T. Attempto Controlled English for Knowledge Representation[J]. Reasoning Web, Fourth International Summer School. Springer, 2008: 104-124

[6] Cregan A, Schwitter R, Meyer T. Sydney OWL Syntax-towards a Controlled Natural Language Syntax for OWL 1.1[C]// 3rd OWL Experiences and Directions Workshop (OWLED 2007). CEUR Proceedings. volume 258, 2007

[7] Hart G, Johnson M, Dolbear C. Rabbit: Developing a Controlled Natural Language for Authoring Ontologies[C]// ESWC 2008. 2008: 348-360

[8] Bernardi R, Calvanese D, Thorne C. Lite Natural Language[C]// IWCS-7. 2007

[9] Miller G A. WordNet: A Lexical Databas[J]. Communication of the ACM, 1995, 38(11): 39-41

[10] Fellbaum C. WordNet: An Electronic Lexical Database [M]. Cambridge, MA: MIT Press, 1998: 5-23

[11] van E J, Kamp H. Representing Discourse in Context[D]. Handbook of Logic and Language, Elsevier, Amsterdam, 1997: 179-237

[12] Pullum G K, Gerald G. Natural languages and context-free languages [J]. Linguistics and Philosophy, 1982, 4(4): 471-504

[13] BalaSundaraRaman L, Ishwar S, Ravindranath S K. Context Free Grammar for Natural Language Constructs-An implementation for Venpa Class of Tamil Poetry[C]// International Forum for Information Technology in Tamil. 2003: 128-136

[14] Klein D, Manning C D. Fast Exact Inference with a Factored Model for Natural Language Parsing[C]// Advances in Neural Information Processing Systems 15. Cambridge, MA: MIT Press, 2003: 3-10