基于标签系统中聚类分析的个性化推荐算法①

杨 墨[1,2], 李 炜[1,2], 王 晶[1,2]

[1](北京邮电大学 网络与交换技术国家重点实验室, 北京 100876)
[2](东信北邮信息技术有限公司, 北京 100191)

摘 要: 随着 YouTube、Flickr 和 Last.fm 等社会化网络的兴起, 标签系统在日常生活中扮演着越来越重要的作用. 为了给用户提供更优质的推荐, 分析用户为不同资源打标签的行为就显得尤为重要. 本文将主要的社区发现算法应用到标签系统中的聚类分析中, 并比较它们在不同数据集上的表现, 设计出针对标签系统的个性化推荐算法. 实验结果表明, 本文提出的算法能很好的发现不同用户的兴趣, 提高推荐系统的质量.

关键词: 标签系统; 聚类分析; 个性化推荐; 推荐系统; 图算法

Personalized Recommendation Using Clustering Analysis in Tagging System

YANG Mo[1,2], LI Wei[1,2], WANG Jing[1,2]

[1](State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)
[2](EB Information Technology Co. Ltd, Beijing 100191, China)

Abstract: With the rise of YouTube, Flickr, Last.fm and other social networks, tagging systems play an increasingly important role in our everyday life. Analyzing user's tagging behavior of different resources is very important in providing high quality services. In this paper, major community structure detection algorithms are implemented in clustering analysis in tagging system. By comparing their performances on different datasets, a personalized recommendation algorithm for tagging system is designed. Experimental results indicate that the proposed algorithm performs well in detecting different user interests and thus enhances the quality of the recommendation system..

Key words: tagging system; clustering analysis; personalized recommendation; recommendation system; graph algorithm

标签系统(Tagging System)是一种允许用户自由对自己或其他用户上传的资源打标签的系统. 近年来, Youtube、Flickr 和 Last.fm 等一大批基于标签系统的网站取得了巨大的商业成功和社会影响力. 基于标签系统的个性化推荐, 作为标签系统的重要功能, 也逐渐成为研究热点. 基于用户的推荐和基于内容的推荐是个性化推荐系统中最重要的两种技术手段, 本文采用基于用户的推荐思想, 即在用户群中找到指定用户的相似(兴趣)用户, 综合这些相似用户对某一信息的评价, 形成系统对该指定用户对此信息的喜好程度预测.

本文将不同的社区发现算法应用到标签系统中, 并比较它们在不同数据集上的表现, 以设计出高效的标签聚类生成算法. 将社会网络方面的研究引入标签系统的个性化推荐中, 有助于理解和解释用户行为方式, 为用户提供更加优质的推荐服务. 从 2003 年世界上第一个基于标签系统的网站 del.icio.us 出现到现在短短的不到十年时间里, 自由标记这一概念得到广泛普及并显示出旺盛的发展势头. 个性化推荐是标签系统的重要功能, 对这一功能的改进无疑可以产生出巨大的经济和社会效益.

1 标签系统中的聚类生成

1.1 标签系统介绍

标签系统是一个由用户、标签和资源组成的三元系

① 基金项目:国家自然科学基金(61072057);长江学者和创新团队发展计划(IRT1049);国家科技重大专项(2011ZX03002-001-01).
收稿时间:2013-04-07;收到修改稿时间:2013-05-20

统, 其中标签是连接用户和资源的纽带, 对整个标签系统进行聚合分析可以使语义相近的标签形成标签聚类(tag cluster)从而改善个性化推荐过程. 如图 1[1]所示, 首先从标签系统中提取出的用户兴趣信息(User Profile)和标签聚类, 然后将两者对比, 可以产生出个性化推荐. 在图 1 中, R 表示资源(resource), T 表示标签(tag), Rn、Tn 成对出现, 表示用户对某一资源的一次标注.

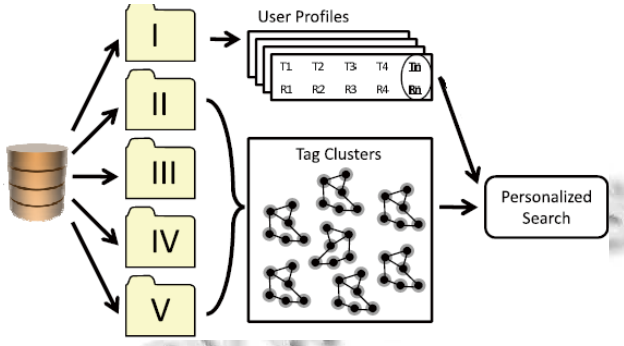

图 1　标签系统

1.2 标签系统与社区结构

标签系统中的标签聚类与社会网络(Social Network)中的社区结构(Community Structure)极为相似, 图 2[2]所示.

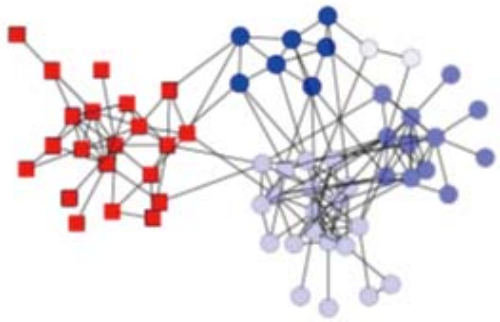

图 2　典型的社会网络

图 2 中的节点既可以看成是典型的社会网络中的成员, 也可以看成是标签系统中的标签; 图 2 中的边即可以表示社会网络的两个成员间有紧密联系, 也可以表示标签系统两个标签被用来描述同一个资源. 标签聚类是语义间联系紧密的标签形成的聚类, 它和社会网络中的社区(community)的概念十分相似. 而社会网络中的社区发现算法已经很成熟, 并且成功的应用在了包括流行病学[3]、新陈代谢系统[4]和生态系统[5]的研究中.

1.3 常用的社区发现算法

在 1.2 中已经论述了标签聚类的作用以及社区发现和标签聚合的相似性, 因而寻找标签聚类等价于社区发现. 在本文中主要比较以下四种社区发现算法在寻找标签聚类方面的表现.

(1) 最小切割法(Minimum-cut method)[2]: 最小切割法是将一个网络分割成多个部分的传统算法(其变形包括比例切割和标准切割). 在最小切割法中, 网络通常被切分为数量预先设置好的大小基本相同的部分, 使得连接不同部分的边的数量最小.

(2) Girvan-Newman[6]算法: 该算法的核心思想是移除连接不同社区的边, 剩余的部分即为社区. 被移除的边是根据中介性(betweenness)这一度量指标来确定的. Girvan-Newman 算法所获得的结果有较高的品质, 但是运行的复杂度较高.

(3) Louvain[7]方法: 该方法是一种贪心优化算法, 该算法包含两个阶段, 一是寻找最小的社区, 二是将第一步中每个社区的节点连接到一起形成一个新的节点. 通过反复执行上述两步, 形成相对较大的社区结构.

2 推荐算法

标签系统系统可以用一个三元组表示, D=<U, T, R>. 其中 D 是协同标记系统, U 是用户的集合 $U=\{u_i|i=1,2,\cdots,I\}$, T 是标签的集合 $T=\{t_j|j=1,2,\cdots,J\}$, R 是资源的集合 $R=\{r_k|k=1,2,\cdots,K\}$.

2.1 非个性化的推荐算法

在本文中用 U 表示用户兴趣信息, 用户在每一个维度上的兴趣为 ti 表示在该兴趣维度上的相关度权值用 $w(t_i)$表示.

$$U=\left\langle w_u(t_1), w_u(t_2), \cdots\cdots, w_u(t)_{|T|}\right\rangle \quad (1)$$

$w(t_i)$也可以理解成用户使用此标签进行标的频率.

本文中将用户已有的标签作为查询 q, 其中 q 可能包含多个关键词. 对于每一个 t_i, 当它为用户查询的一个关键词时 $w_q(t_i)$为 1, 反之则为 0. 这时计算资源 r 和查询 q 的相似性就转换为计算两个向量的相似性, 本文中采用广泛应用的余弦相似系数 cos(q,r)来表示这种相似性.

$$\cos(q,r)=\frac{\sum_{ti\in T} w_q(t_i)*w_r(t_i)}{\sqrt{\sum_{ti\in T} w_q(t_i)^2}*\sqrt{\sum_{ti\in T} w_r(t_i)^2}} \quad (2)$$

如果用户只进行过一次标注, 那么查询向量中只有对应位置的 t_i=1, 其他分量都为 0, 此时 cos(q,r)表示为

$$\cos(q,r)=\frac{w_r(t_i)}{\sqrt{\sum_{ti\in T} w_q(t_i)^2}} \quad (3)$$

2.2 产生个性化推荐

与非个性化推荐不同, 个性化推荐将首先离线产生聚类信息, 然后根据用户兴趣信息和聚类结果重新对按照普通非个性化推荐方法产生的结果排序. 具体步骤如下:

步骤 1: 应用公式(3)计算标签 q 和每个资源的余弦相似性.

步骤2: 在本步中, 标签聚类作为用户和资源的纽带把用户和资源联系起来, 用两个集合相似性的 Jaccard 系数来衡量用户标签集合、标签聚类以及资源的标签集合这三类集合之间的相似性, 其中前两者之间的相似性为用户对聚类的兴趣度, 后两者之间的相似性为资源和聚类的相关度.

步骤 2.1: 计算用户与不同聚类的相关程度, 记为 $J(T_u,C_i)$. 即:

$$J(Tu,Ci)=\frac{|Tu\cap Ci|}{|Tu\cup Ci|}$$

其中 T_u 为用户 u 的标签集合, C_i 为第 i 个标签聚类.

步骤 2.2 计算资源和聚类的相关度, 记为 $J(T_r,C_i)$. 即:

$$J(Tr,Ci)=\frac{|Tr\cap Ci|}{|Tr\cup Ci|}$$

步骤 3: 计算用户对每个资源的兴趣度, 记为 I(u,r). 该值为对于所有聚类计算用户对该聚类兴趣度与聚类对资源的相关程度的乘积, 公式为:

$$I(u,r)=\sum_{i=1}^{k} J(T_u,C_i)*J(T_r,C_i)$$

获得每个资源的 I(u,r)后就可对所有资源排序, 最后把前 n 项资源推荐给用户. 由于聚类和资源之间的相关度跟用户无关, 所以可以离线计算; 与非个性化的推荐相比, 个性化推荐算法将用户对某一类型的资源的喜爱程度加入到对资源的推荐评分标准中, 因而可以想见所得到的推荐结果更加有针对性. 然而这两种方法都只考虑用户对资源的喜好程度, 如果对资源内容进行更加细致的分析, 可以产生出更好的推荐结果.

3 实验结果

本文采用 BibSonomy.org 和 del.icio.us 这两数据集验证算法的有效性. 这两个数据集都是英文资源, 但是本文的算法没有进行语义分析, 而是间接根据标签的共同标记关系推断其相关关系, 因此所提算法同样适用于中文资源.

本文采用四重交叉验证法来测试算法的有效性, 即把每个数据集中的用户等分为四个小数据集, 其中三个用来执行聚类算法, 另外一个作为测试用例. 将用户已有的标签作为查询, 获得与此标签相关度最高的资源, 如果该用户已经标注过的资源出现在推荐结果中且排名靠前, 则说明算法是有效的.

采用两种方法进行对比实验, 第一种是 2.1 中的非个性化推荐方法, 第二种是采用 k 均值聚类并通过协同过滤算法[8]产生推荐. 本文采用前 n 项返回结果(Top-N)的 Recall 值(召回率)作为衡量指标.

实验结果如图 3 和图 4 所示. 图中 Top-n-Recall 值是 5 次测试结果的平均值. 可以看出, 对于这两个数据集, 采用社区生成算法的个性化推荐算法优于普通非个性化查询算法和基于 k 均值聚类的系统过滤算法.

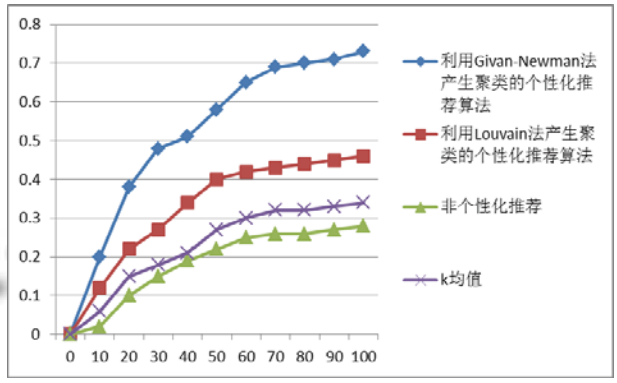

图 3 del.icio.us 数据集上四种算法的比较

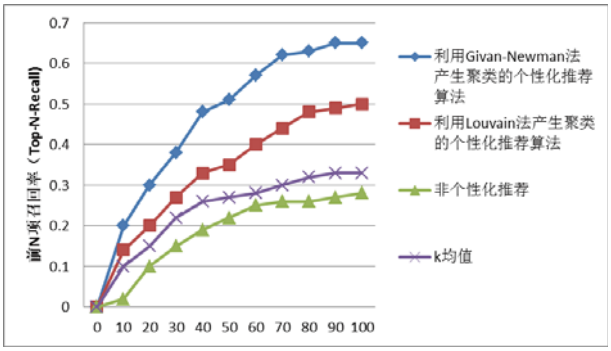

图 4 BibSonomy 数据集上五种算法的比较

4 结语

本文通过在 BibSonomy 和 del.icio.us 两个数据集上进行实验, 证明了社区发现算法可以提高推荐结果的准确性, 从而证实社区发现算法可以在一定程度上将语义相近的标签聚合到一起. 社区生成算法有广泛应用, 下一阶段的研究会将本文中没有涉及到的社区发现算法应用到标签聚类的生成中, 并和之前的结果比较, 以进一步提高个性化推荐的效果.

参考文献

1 Andriy S, Jonathan G, Bamshad M, Burke R. Personalized recommendation in social tagging systems using hierarchical clustering. Proc. of the 2008 ACM Conference on Recommender Systems. Lausanne, Switzerland, ACM, 2008: 37–48.

2 Newman MJ, Girvan M. Finding and evaluating community structure in networks. Physical Review, E 74, 036104, 2006.

3 Kleczkowsi A, Grenfell BT. Mean-field-type equations for spread of epidemics: The‘small world’model. Physica A 274, 1999: 355–360.

4 Wagner A, Fell D. The small world inside large metabolic networks. Proc. R. Soc. London B 268, 2001: 1803–1810.

5 Camacho J, Guimera R, Amaral LAN. Robust patterns in food web structure, Phys.Rev. Lett. 88, 228102, 2002.

6 Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA, 2002, 99 (12): 7821–7826.

7 Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of community hierarchies in large networks. J. Stat. Mech. 2008, 10: 10008.

8 万里,廖建新,王纯.基于社会网络信息流模型的协同过滤算法.吉林大学学报(工学版),2011,41(1):275–28.

(上接第 131 页)

5 结束语

本文将完全存储模型和逆增量模型相结合, 提出基于双向链表的产品协同设计版本存储模型, 通过设置临界值和版本相对基板本的长度来进行比较, 确定完整存储还是逆增量存储, 节约了存储空间, 提高了存取效率和安全性.

参考文献

1 于海斌,朱云龙.协同制造.北京:清华大学出版社, 2004:142.

2 Dix A, Rodden T, Sommerville I. Modeling versions in collaborative work. Software Engineering IEEE Proceedings, 1997, 144(4): 194–206.

3 曹祥,郑国勤,胡毓宁.协同设计环境下的版本管理模型.计算机工程与应用,2001,37(15):61–63.

4 李祥,周雄辉,阮雪榆.注塑模协同设计过程中的版本管理研究.模具技术,2000,(5):18.

5 徐保民,徐爱琴,李峰.协同编辑器中版本管理的设计与实现.计算机工程与应用,2002,38(5):134–136.

6 王红丽,孙长嵩,李钟隽.一个混合式版本管理存储模型.2006 年北京地区高校研究生学术交流会——通信与信息技术会议论文集.北京.北京邮电大学出版社.2006,1644–1647.

7 付喜梅.基于 STEP 的协同设计版本存储控制策略.计算机工程,2008,34(24):61–66.

8 付喜梅,陈家新.协同设计中版本存储控制策略的研究.微计算机信息(管控一体化),2006,22(4-3):105–108.

9 Westfechtel B, Munch BP, Conradi R. A layered architecture for version management. IEEE Trans. on Software Engineering, 2001, 27(12).