基于在线百科全书的群体兴趣及其关联性挖掘

张海粟[1),2)] 陈桂生[3)] 马于涛[3),4)] 刘玉超[3)]

[1)](中国人民解放军理工大学指挥自动化学院 南京 210007)

[2)](中国人民解放军国防信息学院 武汉 430010)

[3)](中国电子系统工程研究所 北京 100141)

[4)](武汉大学软件工程国家重点实验室 武汉 430072)

摘 要 针对协同过滤、基于内容过滤等个性化推荐方法所存在的用户隐私数据收集、冷启动等问题,提出一种群体兴趣及其关联性的挖掘方法,并应用于推荐领域.以维基百科作为数据源,获取用户社团及其编辑的词条,设计了以词条及其所属类别为基础的泛树结构生长策略,使用泛树结构表征用户社团所对应的兴趣点.结合用户社团的结构特征和兴趣点的语义特征给出了用户社团对兴趣点的关注度及兴趣点间关联性的定义,用此群体兴趣取代个性化推荐方法中的个体兴趣,进行了人工直观评价、测试集对比以及视频点播中的新闻推荐等三种实验.结果表明,测试集上群体兴趣关联性的准确度达到了 50%,高于基准协同推荐方法的准确度;新闻推荐实验中,本方法比按热度推荐方法获得了高出近一倍的点击率,验证了群体兴趣及其关联性的合理性.

关键词 群体兴趣;兴趣点泛树结构;协同推荐;维基百科;社会网挖掘

中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2011.02234

Group Interests and Their Correlations Mining Based on Wikipedia

ZHANG Hai-Su[1),2)] CHEN Gui-Sheng[3)] MA Yu-Tao[3),4)] LIU Yu-Chao[3)]

[1)](*Institute of Automatic Commanding, PLA University of Science and Technology, Nanjing* 210007)

[2)](*Institute of National Defense Information, Wuhan* 430010)

[3)](*Institute of Electronic System Equipment Engineering, Beijing* 100141)

[4)](*State Key Laboratory of Software Engineering, Wuhan University, Wuhan* 430072)

Abstract Personalized recommendation technologies, such as collaborative filtering and content based filtering, face some problems. The obvious ones are the privacy history data collection and cold start. In this paper, we suggest a group interests mining method from Wikipedia. We also apply the group interests into the recommendation system, which avoid the cold start, and don't need any privacy data. Here, the group interest replaces the personalized interest in the traditional personalized recommendation technologies. In detail, we first suggest a general tree structure and a growing strategy to denote the interest of a users group, which includes the semantic relationship of each interest. Then we define the group interest based on the structure of users groups. At last, we measure the correlations of interests according to the general tree structure of interests. We further design three types of experiment to evaluate the reasonability of group interests, which is manual evaluation, test set evaluation and a news recommendation experiment in video service. The results show that, the accuracy of correlation between group interests can be more than 50%, and the news hits rate on the recommendation from group interests is 2 times larger than that on the recommendation from news popularity.

Keywords group interest; general tree of interests; collaborative recommendation; Wikipedia; social network mining

收稿日期:2011-08-29;最终修改稿收到日期:2011-09-15.本课题得到国家自然科学基金(69120912,61035004)、国家"九七三"重点基础研究发展规划项目基金(2007CB310804)、中国博士后科学基金(20090460107,201003794)资助.**张海粟**,男,1982 年生,博士研究生,中国计算机学会(CCF)学生会员,主要研究方向为社会网络挖掘.E-mail: zhanghaisu@139.com.**陈桂生**,男,1966 年生,博士,高工,主要研究方向为数据挖掘.**马于涛**,男,1980 年生,博士后,副教授,主要研究方向为复杂网络.**刘玉超**,男,1981 年生,博士研究生,主要研究方向为数据挖掘.

1 引　言

个性化信息服务技术，如个性化推荐等，通过对用户的互联网行为进行分析以发现其兴趣倾向，广泛应用于精准营销、社会关系挖掘和舆情分析等领域. 目前实现个性化信息服务的常见方法有协同过滤、基于内容过滤等，侧重于获取、表征和挖掘针对特定个体的特征信息，据此预测或引导个人行为. 与现有个性化推荐技术有所不同，本文提出一种新的从公开的用户合作行为数据中挖掘群体兴趣的方法，通过群体兴趣及其关联性给出推荐. 该方法不仅针对个体特征，而且更注重社团的共性特征. 如，群体兴趣挖掘可以回答此类问题：对于“泰坦尼克号”的观众来说，在观影前是播放香水广告还是汽车广告，会引起他们更大的兴趣？若能给出“泰坦尼克号”观众群的群体兴趣点表征及关联性，就可有针对性地推送更有效果的广告或相关信息，提高营销水平.

欲挖掘群体兴趣及其关联性，首先要有能够准确反映出群体兴趣的基础数据源. 近年来，包括博客、微博、社交网站和维基等形态在内的 Web 2.0 应用迅速普及，为进行此类挖掘提供了数据来源. 维基是一种群体合作模式，在创作需包罗万象的百科全书上取得了成功应用，如维基百科. 相对于其它几种 Web 2.0 应用形态，维基百科的数据全部公开，所包含的词条都是具有明确定义的概念，容易转化为兴趣点，而且其噪声数据相对更小，处理起来更为方便. 因此，本文选择维基百科作为挖掘群体兴趣的基础数据源. 基于维基百科进行群体兴趣挖掘的基本前提假设为：用户对词条做出的编辑表示其对该词条所涉及领域抱有兴趣，而大量用户的合作编辑过程则体现了用户群体的兴趣特征，且往往会呈现出用户社团结构.

群体兴趣挖掘首先对用户进行聚类以识别出社团，然后再通过社团的编辑行为特征来获取群体兴趣. 本文主要围绕群体兴趣挖掘中的两个核心问题展开研究. 其一，群体兴趣的表征. 由于用户社团的兴趣点不会局限在一个领域，往往较为广泛，且兴趣点之间也存在上下位、同反义等语义关联，因此不能简单地使用词条来表征兴趣点. 本文根据维基百科中词条的“类别”属性，通过基于类别树的生长策略构建泛树结构来表征兴趣点. 泛树结构能更准确地体现出兴趣点间的语义关系，得到更易于理解的挖掘结果. 其二，用户社团对兴趣点的关注度及兴趣点之间相关性的定义. 群体兴趣是社团所表现出的固有的、稳定的偏好特性，为了精确、有效地定义兴趣点，需要考虑兴趣点泛树结构特征. 本文给出一种融合语义与结构特征的兴趣点关注度及其关联性度量方法. 在测试集上进行的对比实验表明，群体兴趣获得了比基于物品的协同推荐方法[1]更高的精度.

群体兴趣可根据用户社团的兴趣点给出推荐结果，这一点和个性化推荐方法[1-2]（包括基于内容过滤的推荐、协同推荐和基于点击流的预测等）所达成的效果看似相同，但在实现思路上有很大不同. 基于内容过滤的推荐技术利用资源和用户兴趣的相似性来过滤信息，协同推荐利用用户动作历史之间的相似性，基于点击流的预测也是利用用户历史行为进行建模. 基于在线百科全书的群体兴趣挖掘则使用群体兴趣点来替代推荐对象的个体兴趣的表征，避免了内容过滤与点击流分析技术中新资源发现能力较弱、协同推荐技术中冷启动和稀疏数据等问题. 更突出的是，在推荐中，通过可公开获取的维基百科中的群体兴趣取代个性化推荐方法中用户的购买、观影记录等个人隐私数据，避免了大量收集用户数据时所遇到的隐私保护等社会问题.

本文第 2 节介绍维基百科数据源与合作编辑网的用户聚类方法；第 3 节针对兴趣点表征提出兴趣点泛树结构的构造方法；第 4 节给出了兴趣关注度和关联性的定义；第 5 节使用 3 种方法对实验结果进行了详细分析；第 6 节给出相关工作；最后总结全文.

2 数据源及用户聚类

2.1 维基百科数据源

截至 2011 年 9 月 15 日，维基百科已有用 280 余种语言编写的 1900 余万个词条（英文版的词条数量最多，已超过 373 万），提供包括词条历史版本在内的几乎所有数据（http://dumps.wikimedia.org）. 统计表明[3-4]，维基百科能在大部分领域保持较高的准确度和覆盖度. 文献[3]通过与美国国会图书馆的 3000 篇随机文章的对比发现，除了在法律和医学领域稍有逊色之外，其它方面维基百科几乎都有很好的覆盖；文献[4]统计了不同领域词条数目的分布情况，与传统知识库相比，维基百科在专有名词、新词、俚语、技术术语和新近事件等方面具有很大的覆盖优势.

本文以维基百科的中文版本作为数据源，收集数据的具体方法是：以“电影”类别下 390 个词条的 1165 名用户作为起始输入，再以这些用户所编辑的

所有其它词条向外扩充用户，然后通过新扩充的用户来扩充词条. 为获得一定规模的数据量，迭代了5次，共得到84 098个词条、179 023个用户，其中包括属于“电影”类别的词条1991个.

2.2 用户聚类

两个用户间如果有共同编辑的词条，则在两者间连边，即将用户-词条二部图转换为用户合作编辑网，据此合作编辑网进行社团划分，得到用户聚类. 对2.1节的数据集进行处理，最终得到的用户合作编辑网包括176 175个节点(删除了2848个孤立节点)、520 856条边.

此处合作编辑网络的节点规模达到了10^6数量级，为计算效率考虑，采用Mahout数据挖掘工具进行合作编辑网聚类. 通过部署在10台PC上的Hadoop框架，向Mahout中的k均值算法输入参数$k=1000$. 在划分为235个社区的时候，Mahout给出了衡量聚类效果的Silhouette指数[5]的最优值：0.3898. 因此取用户社团个数为235，其中最小社团包含的节点数为482，最大社团包含的节点数为2879.

在进一步进行兴趣点表征工作之前，在这里首先检查群体兴趣的区分度. 区分度是指用户社团所编辑词条的覆盖程度的差异性，通过区分度可以粗略地衡量不同用户社团之间的群体兴趣是否确有倾向性. 区分度Dis定义为

$$Dis=1-\frac{I}{K},\quad I=\sum_{\forall i,j\le k}|A_i\cap A_j|,$$

其中，K为词条总数，A_i表示用户社团i所对应的词条集合；k为用户社团的总数. 在$K=176\,175$，$k=235$，Silhouette指数为0.3898的情况下，Dis值达到了97.7%，这说明在最优社团划分的情况下，其词条的重合程度较低. 图1显示了区分度随Silhouette指数的变化趋势，可以看出，随着Silhouette指数增加，即用户社团划分趋于合理的时候，其所对应词条的区分度也呈现出增加趋势. 由于不同社团所编辑的词条具有良好的区分度，这就说明其兴趣点具有倾向性. 注意到这里的区分度直接以词条作为单位来计算，而词条往往包含大量非常细节的概念，因此在表征兴趣点时不能直接使用词条，否则将会由于高区分度而忽略掉关联性，此关联性往往是词条在更高层次的概念上所形成的. 下一节将引入词条所属类别来表征兴趣点以解决此问题.

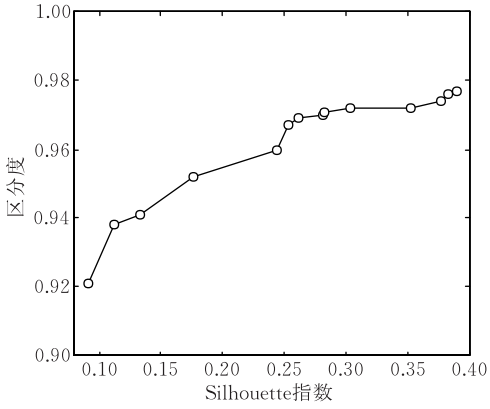
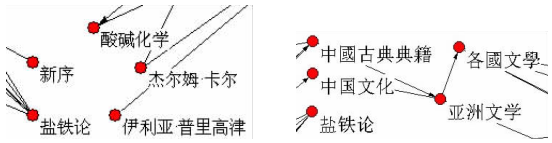


图1 区分度随社团划分结果的变化

3 兴趣点泛树结构

过于细节的词条除带来了前述关联性的问题外，还可能降低兴趣点表征与挖掘结果的可理解性与准确性. 如图2(a)所示，部分细节词条(如杰尔姆·卡尔、盐铁论等人名和专著等)，其具体含义不为大众所熟悉，作为挖掘结果返回的话将难以理解. 图2(b)示意了处于不同概念层次的词条，此时兴趣点之间存在包含关系，若挖掘结果仍将其看成同一层次，将会降低关联性的准确度. 如，中国古典典籍、中国文化和盐铁论之间具有包含关系，因此，若认为中国古典典籍和中国文化之间具有很强的兴趣关联性，则会干扰其与亚洲文学等更合理的关联关系.



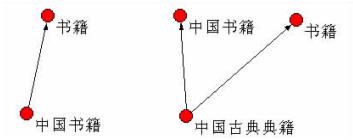
(a) 专有名词 (b) 包含关系

图2 词条不能直接作为兴趣点的例子

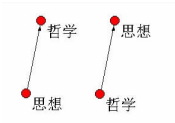
维基百科中每一词条都有很多所属类别的标注信息，对应于词条所属的上层概念. 据此，利用类别标注构造兴趣点泛树结构，通过兴趣点泛树将各个不同层次的概念联系起来，通过层次跃升给出更易理解的兴趣点表征、通过消歧义等给出更为准确的关联性.

下面说明兴趣点泛树的生长构造策略. 抽取词条所处的分类结构得到一层类别树，生长构造策略以一层类别树形成的森林作为输入，将森林合并成一个大的泛树结构. 以词条作为起始节点，其所属的类别标注也看为节点，两者间连边. 此过程中不断地合并具有同样名称的节点，从而将森林中原本多棵不连通的树合并为一棵整的泛树. 节点合并中的问题有：(a)不一致的标注层次关系(图3(a))，通过合并不一致的节点、删除捷径来调整. (b)矛盾的层次关系(图3(b))，通过合并节点、将任意一条边删除来调整. (c)标注的歧义(图3(c))，判断出不一致的

节点并合并.此外还要处理噪声数据,包括设立停用词、删除孤立节点等.



(a) 不一致的层次关系



(b) 矛盾的层次关系

(c) 歧义

图 3　类别树合并

针对图 3 所示的问题,类似于文献[6]的方法,设定节点之间的相似度,以节点间的相似度达到一定阈值作为节点合并的条件.节点 A 和 B 间相似度 $Sim(A,B)$的定义由语义相似度和结构相似度两部分组成:

$$Sim(A,B)=(1-\alpha)\times semanticSim(A,B)+\alpha\times structuralSim(A,B).$$

语义相似度 $semanticSim$ 主要由两部分组成,一是名字相似性,二是邻居相似性,计算方法为

$$semanticSim(A,B)=\beta\times nameSim(A,B)+(1-\beta)\times neighborSim(A,B)),$$

其中,$nameSim(A,B)$是 A 和 B 的名字相似度(可根据简单的字符串相似程度计算,也可以引入语义词典计算,本文的实验采用字符串相似程度),$neighborSim(A,B)$度量邻居相似度,定义为 A 和 B 所拥有的共同邻居的名字相似度的均值.

结构相似度 $structuralSim$ 根据一层类别树上词条 A 和 B 所处位置的相似性计算:

$$structuralSim(A,B)=k+(1-k)\ DiffSim(A,B),$$

其中,$k=neighbor(A,B)/\min(neighbor(A),neighbor(B))$度量的是结构特征上两个节点共同邻居的数目比例,其中 $neighbor(A,B)$为 A、B 的共同邻居数,$neighbor(A)$和 $neighbor(B)$分 B 表示 A 和 B 的邻居数;$DiffSim(A,B)=1-neighborSim(A,B)$,因此$(1-k)\times DiffSim(A,B)$衡量的是共同邻居节点之外的其它邻居节点的影响.$\alpha$ 和 β 作为权重调整因子,本文采用使得 $Sim(A,B)$结果序列的熵最大化的方法来估计其取值.

一个兴趣点泛树结构示例如附录所示(包含了图 2 中的词条).以类别为基础,可对兴趣点分类进行提升,使其更容易理解,如盐铁论被归类为文集、汉朝典籍、经济史;杰尔姆·卡尔被归类为物理化学家、诺贝尔化学奖得主.还可给出层次性的兴趣点表征,通过不同层次上兴趣点跃升与下降给出合理的关联性.例如,汉朝典籍上层的中国古典典籍和汉朝文化是具有相关性的,而汉朝典籍和中国古典典籍之间的相关性就可以弱化;中国古典典籍上层的中国文学和各国文学之间的相关性则可以相应地强化.

4　群体兴趣关注度及关联性分析

在兴趣点泛树结构基础上,本节结合用户社团的结构特征给出群体兴趣关注度的定义,然后再根据用户社团内部与社团之间两个影响因素来计算兴趣点关联性.

4.1　群体兴趣关注度

兴趣点关注度记为 $F(I,G)$,下面统一用 I 表示泛树结构的节点,它可能是根节点(原始词条),也可能是中间节点或叶节点(更高层次的类别标注),泛树结构上不再区分词条和类别.G 是某一个用户社团,G 对应的词条集合记为 I_G.本文主要从编辑行为和兴趣点在泛树结构中的位置这两点来定义 $F(I,G)$.

对于编辑行为而言,兴趣点 I 被 G 中用户编辑的次数 $NumInG(I,G)$与所有社团的作者全体集合对此兴趣点的编辑次数需要联立考虑.类似于评估单词重要性的 $tf\text{-}idf$ 方法,这里提出基于编辑特征的 $ef\text{-}isf$(edit frequency-inverse set frequency)方法:

$$ef\text{-}isf_{I,G}=ef_{I,G}\times isf_I,$$

其中,$ef_{I,G}=NumInG(I,G)/Num(I;I\in I_G)$,分母 $Num(I;I\in I_G)$表示在 I_G 中兴趣点出现的次数之和;$isf_I=\log(|G|/(1+|G;I\in I_G|))$,表示在所有的用户社团中,具有兴趣点 I 的社团所占比例的大小.$ef_{I,G}\times isf_I$表示的含义为:兴趣点 I 在社团 G 中得到编辑的概率乘上 I 在所有社团中得到编辑的比例.

若兴趣点处在泛树结构的一个回路中,由于可以认为回路中的兴趣点间具有强烈相关性,因此其对关注度的正面影响为

$$Z=|circle(I):circle(I)\in I_G|/|circle(I)|,$$

其中,$circle(I)$为 I 所处于的最短回路,$|circle(I)|$为其长度,$|circle(I):circle(I)\in I_G|$的含义为既属于回路 $circle(I)$又属于 I_G 的兴趣点数目,$|circle(I)|$为兴趣点 I 所处最短回路的长度.

I 处于泛树结构的层次特征主要使用其所覆盖的根节点数目、与根节点的距离来刻画.设 C_I 为兴趣点 I 所覆盖的根节点(词条)数目,C 为所有根节点(词条)的数目;H_I为距离所覆盖的根节点的平均

值，H 为根节点到叶子节点的距离之和(泛树的高度).关注度与兴趣点所覆盖根节点的数目、与根节点的距离成衰减关系，取高斯衰减函数刻画此关系：$\exp[-(C/C_I+H/H_I)]$.

综合考虑编辑行为和位置信息两个因素，用户社团 G 对兴趣点 I 的关注度为

$$F(I,G)=(1-\alpha)\times ef\text{-}isf_{I,G}+\alpha\times Z\times\exp[-(C/C_I+H/H_I)],$$

其中 α 为影响因子，用于调节编辑行为与位置信息两个因素的权重，本文按照使得 $F(I,G)$ 序列的熵最小化来确定 α 的取值，Z 为归一化因子.

4.2 兴趣点关联性

兴趣倾向是用户社团中稳定的偏好特性，而兴趣倾向之间的关联性通过用户社团联系起来.兴趣点间的关联性需考虑两个因素：(1)用户社团内部结构以及兴趣点关注度等所产生的关联；(2)用户社团之间所体现出的兴趣点关联.对于这两个因素处理的基本原则有：社团内部兴趣点的关联性要比社团之间的更强；社团内部具有更强关注度的兴趣点之间的关联性更强；处于类似位置的兴趣点之间的关联性强.

对于用户社团 G 内部的兴趣点 I 和 J 的关联性，转化为对点加权的网络频繁子结构挖掘问题.其中，点权重即为 G 对于相应兴趣点的关注度 $F(I,G)$，频繁子图为同时包含 I 和 J 的最小子图.首先暂不考虑节点权重，采用经典的 AprioriGraph 算法[7]来挖掘频繁子结构.在计算出频繁子结构 S_{ij} 之后，再使用简单的值和约束法则 $Sum_v\geq low$，其中 Sum_v 为子结构 S_{ij} 中所有节点的权重之和，low 为指定的至少应该满足的阈值.所挖掘出的满足指定支持度的频繁子图 S_{ij} 的数目记为 m，频繁子图 S_{ij} 所体现的关联性即可定义为 S_{ij} 在整个泛树结构 S 中所占的面积比(其中 $|S_{ij}|$ 和 $|S|$ 分别表示节点数目)：

$$C'_{IJ}=m\cdot|S_{ij}|/(|S|-m\cdot|S_{ij}|).$$

对于分别位于两个社团上的兴趣点 I 和 J 的关联性，关键是计算出其在各自所属的泛树结构上的共同祖先节点 $P_I=P_J$，然后根据 I 和 J 在两个泛树上分别与此公共祖先节点的距离来确定两者的相关性.具体方法为，假设存在此公共祖先节点 $P_I=P_J$，I 和 J 到 P_I 和 P_J 的最短距离分别为 $d(I-P_I)$，$d(J-P_J)$，则 I 和 J 的社团之间关联性为

$$C''_{IJ}=\exp-[1/d(I-P_I)+1/d(J-P_J)].$$

其物理含义是，相关性与公共祖先节点的距离成反比.若公共祖先节点不存在，则定义 $d(I-P_I)$ 和 $d(I-P_J)$ 为距离泛树结构叶子节点的最短距离.

兴趣点 I 和 J 的关联性 $C(I,J)$ 同时考虑社团内部和社团之间两个要素，对前述计算出的社团内部与社团之间的两个关联性结果联立考虑：

$$C(I,J)=C'_{IJ}\times C''_{IJ}.$$

4.3 群体兴趣挖掘算法步骤与复杂度分析

总结前述群体兴趣挖掘算法的主要步骤，包括用户聚类、兴趣点泛树构建、计算关注度与相关性等四步，如图 4 所示.其中，矩形框为算法步骤，椭圆形为数据，实箭头标示了数据的输入输出关系，空箭头标示了数据间的关系.

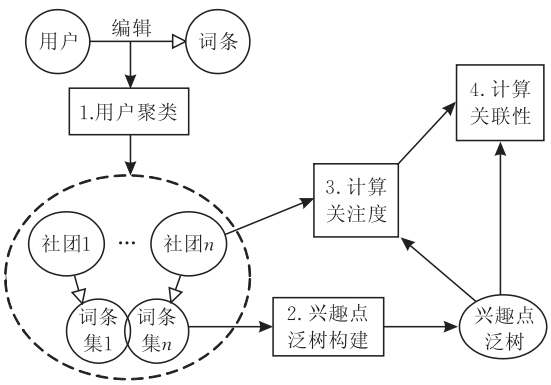


图 4 群体兴趣挖掘的主要算法步骤

其中，原本复杂度较高的大规模复杂网络聚类通过基于 MapReduce 的分布式计算平台，复杂度被降至 $O(N\cdot\log(N))$，其中 N 为网络的节点数目.兴趣点表征的复杂度主要在泛树结构的构建步骤，泛树结构的生长策略按照逐步处理一层类别树的方法，且要计算每一个类别的相似度，故复杂度为 $O(N'\cdot(1+\mathrm{sum}(O(l)))$，其中 N' 为所有类别树的总数，l 为一层类别树的平均节点数，记 l_i 为某类别树的节点数，$\mathrm{sum}(O(l))=(l_1l_2+(l_1+l_2)l_3+\cdots+(l_1+l_2+\cdots+l_{N'-1})l_{N'})$，因此 $O(N'\cdot(1+\mathrm{sum}(O(l))))=O(N'+N'^2/2)=O(N'^2/2)$.在兴趣关注度与相关性步骤，因为均要计算出整个矩阵的元素，故计算复杂度为 $O(|\boldsymbol{G}|^2)$，其中 $|\boldsymbol{G}|$ 为用户社团和兴趣点所构成的矩阵 $\boldsymbol{G}$ 的元素个数.泛树结构、兴趣关注度和相关性计算均没有很强耦合的步骤，因此可置于 MapReduce 分布式平台上，进而复杂度可以进一步降低.

5 实验结果分析

5.1 实验过程与评价方法

在前期抽取的 176 175 个用户节点、520 856 条边、235 个用户社团的基础上，按照图 4 所示的算法步骤，建立起群体兴趣及其关联的基础数据库，其中，针对每一个社团 G 和兴趣点 I 记录了关注度

$F(G,I)$，针对所有的兴趣点二元组$\langle I,J\rangle$记录了关联性程度$C(I,J)$. 此基础数据可以定期根据维基百科数据源的变动情况重新计算，达到实时更新的效果. 对群体兴趣的合理性和准确度采用 3 种方法进行评价：

(1) 人工直观观察结果.

实验以维基百科中“电影”类别下的词条为初始内容开始爬取. 因此，可以直接人工地输入一部电影作为兴趣点，再借助于常识，定性地评估其返回的关联兴趣点的合理性.

(2) 借助于测试数据集.

引言中已提及，根据用户点击流进行个性化行为预测会遇到新资源发现能力缺乏的问题. 但是，已记录下的点击流数据却是验证群体兴趣点关联性较为理想的测试数据集. 通过对比群体兴趣的关联性与真实点击流数据所体现出关联性之间的符合程度来验证挖掘结果的准确性. 本文的点击流数据源由两部分组成，一是抽取了 Amazon.cn 中电影书籍和 CD 类商品的点击浏览数据，二是校园网门户上新闻页的关键词. 这两部分的数据必须结合起来，以达到领域较宽(由新闻页关键词保证)和内容较准确(购物网站物品浏览)的目的，从而便于测试群体兴趣的结果.

通过点击流衡量群体兴趣的具体方法是，给定任意两个兴趣点 I、J 及其通过群体兴趣挖掘所得到的关联性$C(I,J)$，计算在点击流中兴趣点 I 和 J 之间的距离$k(I,J)$(若 I 或 J 在点击流中没有出现则定义为点击流序列的最长距离). 对于点击流中的相关性，采用简单的距离反比，即 $1/k(I,J)$越小，则 I 和 J 之间的关联性越大. 因此，如果 $C(I,J)$和 $1/k(I,J)$所构成的矩阵 $\boldsymbol{C}$ 和 $\boldsymbol{K}$ 呈现出很强的正相关，就可以说明群体兴趣关联性挖掘的有效性. 矩阵 $\boldsymbol{C}$ 和 $\boldsymbol{K}$ 的相关性通过每一行数据的皮尔逊相关系数相加得到.

作为基准算法进行类比，使用基于物品(兴趣点)的协同推荐方法进行了计算. 在这里，协同推荐采用文献[8]中的算法实现：兴趣点拥有的共同用户越多，推荐的关联性越强.

(3) 网站推荐的实际应用.

以在校园网的视频点播服务中添加新闻链接的方式进行测试，通过新闻链接被点击的次数作为对比. 实验过程为：在为期一个月的时间内，按照 Web 新闻页面点击的热度确定概率的大小、随机地分配新闻链接到各视频源上；然后在接下来的一个月时间内，以同样的视频源作为输入，但是将当前的新闻链接按照兴趣点相近的原则重新分配.

5.2　结果分析

(1) 直观结果.

输入测试用例：电影“建国大业”，“泰坦尼克号”，将返回的关联性最高的 9 个兴趣点及关联性程度绘制成雷达图，如图 5 所示. 这里返回的兴趣点分布于泛树结构各层次，主要依据是关联性程度的排序. 表 1 列举了另外 4 部知名电影作为兴趣点输入的测试结果.

表 1　兴趣点相关性

输入兴趣点	相关性(相关性按由前到后降序排列)
阿甘正传	第一滴血；野战排；奠边府战役；梅兰芳(电影)；巴顿将军；乱世佳人(电影)；国际象棋；大学；汤姆·汉克斯；约翰·肯尼迪
阿凡达	詹姆斯·卡梅隆；终结者；真实的谎言；京杭大运河；世界三大天然良港；戴维斯-蒙森空军基地；哈利·波特；比尔与美琳达·盖茨基金会；泰坦尼克号；泛美航空 214 号班机空难
唐山大地震	非诚勿扰(电影)；手机；秦皇岛；天津；iPhone；编辑部的故事；徐帆；国家地震局；我的机器人女友；泰坦尼克号
岁月神偷	永利街；新宝戏院；美国电影学院；上环；拔萃男书院；市区重建局；吴君如；资助学校；香港杰出学生

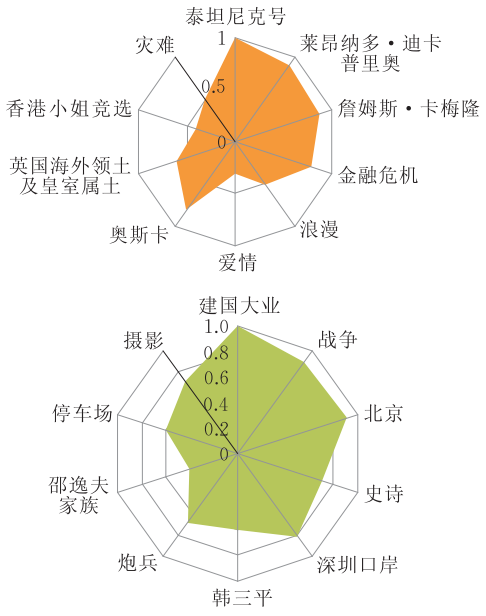


图 5　群体兴趣关联性示例(与电影“泰坦尼克号”和“建国大业”强关联的兴趣点)

从定性的角度分析，可以看出，对于图 5 中这两部电影来说，其导演和所反映的背景事件都是用户社团具有强烈兴趣的(如奥斯卡、战争等). 而其它的兴趣点中，“泰坦尼克号”社团更加关注一些时尚话题，如香港小姐竞选等；“建国大业”社团则对于摄影等有兴趣，而“停车场”的意外出现则可能与影院或

者生活服务相关.

进一步分析表 1 中所列举的案例,可以看出,每一部电影相关的背景、演员和导演等所涉及的信息往往都是关联性较高的兴趣点所在,同时一些“意料之外”的结果,如“阿甘正传”中的“国际象棋”、“阿凡达”中的“京杭大运河”等,往往更可能提示一些有趣的、隐藏的推荐信息.

总之,从直观上看,每一部电影的用户社团都带有较为强烈的群体兴趣的倾向性,这有助于挖掘互联网上大量用户的背景和关注点等有用信息.

(2)测试数据集的覆盖度.

在具体实验中,本文抽取的测试集数据规模为:Amazon.cn 影视分类下的 3980 部电影和 15 820 个相关的点击数据;校园网新闻(包括时政新闻和校园新闻)的 5692 个关键词和相关点击排名前 5 的 28 460 个关键词.最终所得测试集中一共包括 53 952 个关键词(其中包括重合的 9802 个关键词).

分 5 次随机抽取 5 个用户社团对应的兴趣点二元组$\langle I,J\rangle$集合,所得到的兴趣点在测试集中出现的比例分别为 10.71%,12.43%,6.17%,11.52% 和 9.47%.在此基础上,分别计算群体兴趣关联性程度、协同推荐的结果与点击流之间的相关系数,其结果对比如图 6 所示.需要说明的是,这里将 Amazon.cn 的数据看作点击流的真实数据而不是协同推荐的结果,因此协同推荐只计算了校园网中用户点击新闻网页的关键词这一部分.

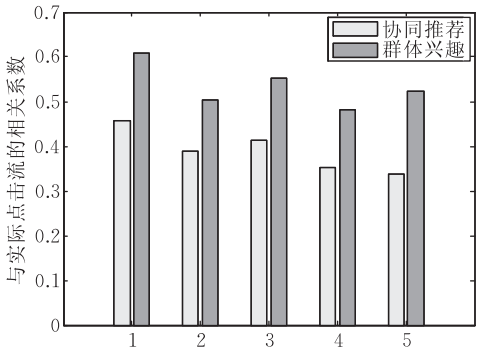


图 6 与实际点击流的相关系数

总体上看,两者的相关系数均维持在 30% 以上,说明都体现出了较为明显的推荐效果.但相对于协同推荐,群体兴趣的相关性还要高出 15% 左右.事实上,校园网用户点击行为所涉及的领域很宽,同时其数据较为稀疏(每一位用户收集到的点击关键词平均为 8.6 个).因此,在此种情况下,本案例中的群体兴趣推荐方法取得了更好的推荐准确率.

(3)视频点播中的新闻推荐实验

校园网的视频点播服务提供电影、新闻和教学视频的在线播放功能.我们在播放器窗口边设置了类似于广告的新闻链接来测试推荐效果.在为期两个月的测试周期中,共得到了 129 086 次播放和 9087 次有效点击(停留时间在 30 s 以上).其中,在按照新闻热度放置的一个月中,播放次数和有效点击次数分别为 67 340 和 5387 次;在使用群体兴趣推荐的一个月中,播放次数和有效点击次数分别为 61 746 次和 10 349 次,点击数量提高了近一倍.图 7(a)给出了进行实验的 2 个月内点击新闻链接的概率曲线.可以看出,使用了群体兴趣关联性推荐的效果,比起按照新闻热度排序放置的效果高出 10%~20%,相对于本来就约等于 10% 的点击率水平,高出接近一倍,而且基于群体兴趣推荐所得到的点击概率一直维持在相对较高的水平.

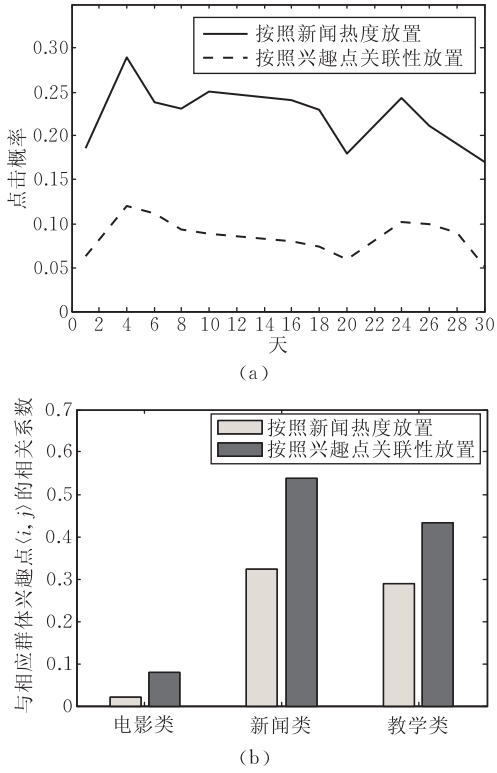


图 7 新闻推荐实验中的用户点击概率

另一方面,我们将视频播放中用户点击的新闻链接记录下来,对比其与群体兴趣关联性的相关系数(采用类似于图 6 的皮尔逊相关系数),结果如图 7(b)所示.这里将视频分成电影、新闻和教学三类.其中,播放电影时的用户点击新闻的概率相对较低,播放新闻和教学类视频时用户点击新闻的概率基本相当.与按照新闻热度放置的效果相比,在点击概率较低的电影类视频中,兴趣推荐能够很大地提高用户点击所推荐信息的概率.

6　相关工作

目前，用户兴趣挖掘的重点集中在动作行为日志分析[9-11]. 一种常用的方法是直接抽取用户访问文档集合中的关键词作为用户兴趣的表征项，借助聚类和学习算法，挖掘出用户的兴趣[12-13]. 通常，此方法同样面临着隐私数据的保护问题. 同时，由于所抽取的特征词在语义上的多义性[14]，单纯基于日志行为分析所发现的用户兴趣精度难以进一步提高，而且一般没有可用的本体或语义关系库用来实现兴趣在概念层次上的跃升与下降等.

还有很多其它技术可直接应用到群体兴趣挖掘中. 如，分类树和本体构建方面的研究方法，包括从文本中学习概念层次的单词聚类技术[15]，从标注和分众分类中获取层次结构的基于图的聚类技术[6,16]等，均可用于提高兴趣点泛树结构的精度. 针对维基百科数据源进行概念提取的工作，目前学术界也有涉及. 如 HeiNER[17] 和 YAGO[18] 等，将词条名字作为概念的候选，采取了一些语言学方法进行筛选，这些思路可用于进一步完善兴趣点的表征.

个性化信息服务技术的发展也在不断地深入. 文献[1]针对目前推荐算法仍然存在的特征提取困难、冷启动、数据稀疏等问题，总结了引入上下文环境来扩展“物品-用户”二维关系的多维度推荐、利用相关反馈提高精度、隐私保护以及推荐中的社会学因素等多个改进方向. 正如本文强调的，基于完全公平的在线百科全书的群体兴趣在数据收集、反馈、推荐新资源等方面采用了完全不同的思路，因此两者可以相互结合，在获取用户隐私数据难易程度不同的场景下发挥各自的作用.

7　总　结

本文以在线百科全书的数据源作为基础，通过提取合作编辑过程中的用户社团，给出了社团的群体兴趣及关联性，可针对非特定个体、但是具有共同属性的群体进行有效推荐. 其核心问题为群体兴趣的表征和用户社团对兴趣点的关注度及兴趣点之间相关性的定义. 本文提出了一种兴趣点泛树结构的生长策略，开发了结合结构特征和语义特征的群体兴趣关注度计算方法. 在人工评价、点击流真实数据测试集和校园网视频点播中的新闻推荐实验等三方面，都获得了较为明显的效果. 如，和点击流数据的相关性比传统的协同推荐算法要高出 15% 左右，新闻推荐上更比依热度放置链接的效果高出 1 倍. 这些实验证明了基于群体兴趣进行推荐的有效性. 随着互联网尤其是 Web2.0 的快速发展，我们能够更加方便地获取公开数据，但同时隐私数据也将得到更为严格的保护. 因此，推荐算法如何从收集私人历史行为数据转向利用庞大的公开数据，包括本文提及的维基百科合作编辑以及博客、微博等，应该值得进一步深入关注.

致　谢　感谢邓智龙对于实验数据处理的帮助！

参　考　文　献

[1] Xu Hai-Ling, Wu Xiao, Li Xiao-Dong, Yan Bao-Ping. Comparison study of internet recommendation system. Journal of Software, 2009, 20(2): 350-362(in Chinese)
(许海玲，吴潇，李晓东，阎保平. 互联网推荐系统比较研究. 软件学报，2009，20(2)：350-362)

[2] Stadnyk I, Kass R. Modeling users' interests in information filters. Communications of the ACM, 1992, 35(12): 49-50

[3] Halavais A, Lackaff D. An analysis of topical coverage of Wikipedia. Journal of Computer-Mediated Communication, 2008, 13(2): 429-440

[4] Kittur A, Chi E H, Suh B. What's in wikipedia: Mapping topics and conflict using socially annotated category structure//Proceedings of the ACM Conference on Human Factors in Computing Systems. Boston, USA, 2009: 1509-1512

[5] Lin Tsun-Chen, Liu Ru-Sheng, Chen Shu-Yuan, Liu Chen-Chung, Chen Chieh-Yu. Genetic algorithms and silhouette measures applied to microarray data classification//Proceedings of the 3rd Asia-Pacific Bioinformatics Conference. Singapore, 2005: 229-238

[6] Plangprasopchok A, Lerman K, Getoor L. Growing a tree in the forest: Constructing folksonomies by integrating structured metadata//Proceedings of the Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 949-958

[7] Han Jia-Wei, Kamber M, Pei Jian. Data Mining: Concepts and Technologies. 3rd Edition. Massachusetts, USA: Morgan Kaufmann Publishers, 2011

[8] Segaran T. Programming Collective Intelligence: Building Smart Web 2.0 Applications. USA: O'Reilly Media, 2007

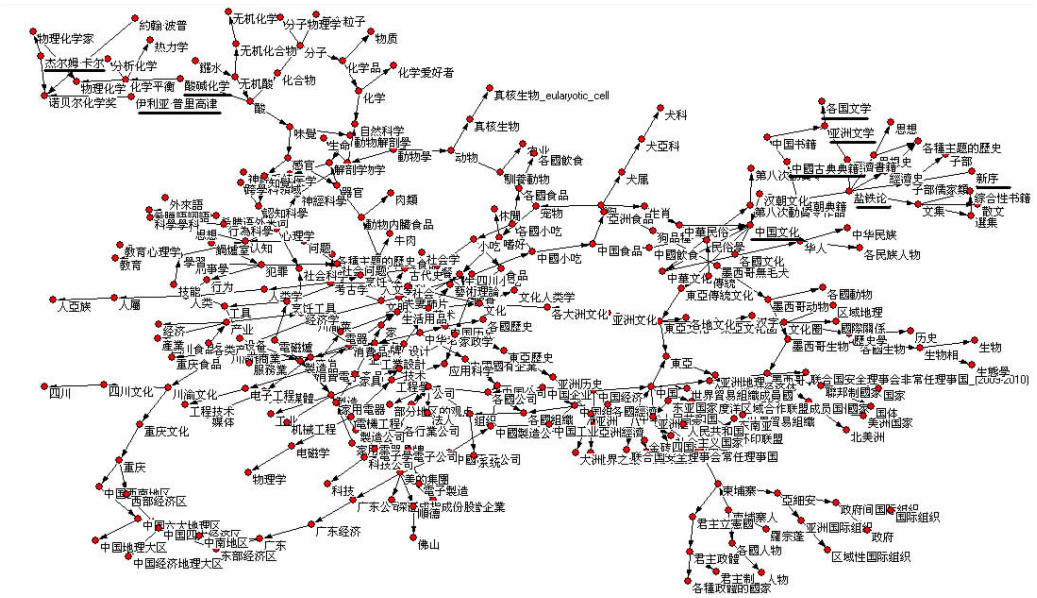
[9] Guo Yan, Bai Shuo, Yang Zhi-Feng, Zhang Kai. Analyzing scale of web logs and mining users' interests. Chinese Journal of Computers, 2005, 28(9): 1483-1496(in Chinese)
(郭岩，白硕，杨志峰，张凯. 网络日志规模分析和用户兴趣挖掘. 计算机学报，2005，28(9)：1483-1496)

[10] Brzozowski M J, Romero D M. Who should I follow? Recommending people in directed social networks//Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Barcelona, Spain, 2011: 458-461

[11] Liu Jia-Hui, Dolan P, Rønby E. Personalized news recommendation based on click behavior//Proceedings of the 15th International Conference on Intelligent User Interfaces. New York, USA, 2010: 31-40

[12] Mao Q-J, Feng B-Q, Li Y, Pan S-L. A novel users' interests prediction approach based on concept lattice. Journal of Shandong University(Engineering Science), 2010, 40(5): 159-163
[13] Zeng Qing-Wei, Jiang Jie. Research and implementation on personalized search engine. Key Engineering Materials, 2011, 467(2): 129-133
[14] Kim Hak-Lae, Breslin J G, Decker S, Kim Hong-Gee. Mining and representing user interests: The case of tagging practices. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 2011, 41(4): 683-692
[15] Snow R, Jurafsky D, Ng A Y. Semantic taxonomy induction from heterogeneous evidence//Proceedings of the 21st International Conference on Computational Linguistics. Stroudsburg, USA, 2006: 801-808
[16] Brooks C H, Montanez N. Improved annotation of the blogosphere via auto-tagging and hierarchical clustering//Proceedings of the 15th International Conference on World Wide Web. Edinburgh, UK, 2006: 625-632
[17] Wentland W, Knopp J, Silberer C, Hartung M. Building a multilingual lexical resource for named entity disambiguation, translation and transliteration//Proceedings of the 6th International Language Resources and Evaluation. Marrakech, Morocco, 2008: 3230-3237
[18] Suchanek F M, Kasneci G, Weikum G. YAGO: A large ontology from wikipedia and WordNet. Elsevier Journal of Web Semantics, 2008, 6(3): 203-217

附录. 兴趣点泛树结构示例.



ZHANG Hai-Su, born in 1982, Ph. D. candidate. His main research interest is social network mining.

CHEN Gui-Sheng, born in 1966, Ph. D., senior engineer. His main research interest is data mining.

MA Yu-Tao, born in 1980, Ph. D., associate professor. His main research interest is complex network.

LIU Yu-Chao, born in 1981, Ph. D. candidate. His main research interest is network mining.

Background

Mining users' interests for the recommendation systems is an important application in the e-commerce, such as precise marketing and public opinion analysis. Now the main popular methods for interest mining can be summarized as collaborative filtering and contentment-based filtering, which can be seen as a personalized tendency mining technology. Hence, there are some drawbacks of these methods, such as they must collect lots of users history data, sometimes it is difficult and illegal for the privacy protection policy. Other drawbacks, such as cold start and sparse data, more or less are related with the data collection problems. Thus, more and more researchers tend to find the methods without the privacy data. In this paper, we suggested the group interests mined from the public Wikipedia cooperation data, which can be used to replace the privacy data collection. We reduce the interests of a person into a similar and public group (in this paper, this group is the co-editors in Wikipedia).

This research is supported by the National Natural Science Foundation of China under Grant Nos. 69120912, 61035004, the National Basic Research Program (973 Program) of China under Grant No. 2007CB310804, and the China Postdoctoral Science Foundation under Grant Nos. 20090460107, 2010003794. All of these projects are around the problem of services on-demand, and how to characterize or model users' demands for the web services in Internet.

In these projects, a core problem is how to mining the huge data from Wikipedia, and how to apply the results into the social network and services on demand. It is obvious that the interests of users can improve the quality of services on demand.