基于用户聚类的异构社交网络推荐算法

陈克寒 韩盼盼 吴 健

(浙江大学计算机学院 杭州 310027)

摘 要 相比传统的社交网络,基于弱关系的微博类社交网络具有显著的异构特征.根据特征可以将节点分为用户(消息订阅者)和主题(消息发布者)两类,面向用户推荐其感兴趣的主题成为了该类社交网络中推荐系统的主要目标之一,同时该类社交网络中普遍存在的数据稀疏性和冷启动现象成为了推荐系统面临的主要问题.文中提出一种基于两阶段聚类的推荐算法GCCR,将图摘要方法和基于内容相似度的算法结合,实现基于用户兴趣的主题推荐.与以往方法相比,该方法在稀疏数据和冷启动的情况下具有更好的推荐效果,此外,通过对数据集进行大量的离线处理,使得其较以往推荐方法具有更好的在线推荐效率.最后通过真实社交网络的数据对本方法进行了验证,同时分析了各参数对推荐效果的影响.

关键词 社交网络;推荐系统;聚类算法;图摘要;数据挖掘

中图法分类号 TP311 **DOI号** 10.3724/SP.J.1016.2013.00349

User Clustering Based Social Network Recommendation

CHEN Ke-Han HAN Pan-Pan WU Jian

(*College of Computer Science and Technology, Zhejiang University, Hangzhou* 310027)

Abstract Comparing to the ordinary social networks services (SNS), the twitter-like weak-relationship based social networks are observably heterogeneous. By classifying the nodes into users (subscriber) and subjects (publisher), the goal of recommendation systems over this kind of networks is basically recommending the subjects to the users for subscription. Moreover, the data sparseness and cold-start scene always exists in these microblog networks. In this paper, we propose GCCR, a hybrid method combining both graph-summarization and content-based algorithms by a two-phase user clustering approach, which can recommend subjects according to user interests. With respect to other methods, the GCCR algorithm could generate better recommendation result in sparse datasets and cold-start scenarios. In additional, by separating the task into offline and online parts, GCCR works more efficiently online by using the pre-processed offline results. We use real data set from existing social networks to evaluate GCCR along with base-line methods. Moreover, an analysis of the parameters is given for evaluating their impacts on recommendation results.

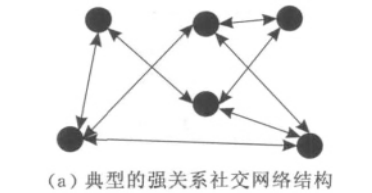
Keywords social network; recommendation system; clustering; graph summarization; data mining

收稿日期:2012-06-30;**最终修改稿收到日期**:2012-08-24.**本课题得到国家科技支撑计划项目基金**(2011BAH16B04)、**国家自然科学基金**(61173176)、**浙江省科技项目**(2008C03007)、**国家"八六三"高技术研究发展计划项目基金**(2011AA010501)**资助**.**陈克寒**,**男**,1987**年生**,**硕士研究生**,**主要研究方向为服务计算、数据挖掘**.E-mail:metalgear@zju.edu.cn.**韩盼盼**,**女**,1989**年生**,**硕士研究生**,**主要研究方向为服务计算、数据挖掘、社会计算**.**吴 健**,**男**,1975**年生**,**博士**,**副教授**,**主要研究方向为**Web**服务、语义**Web**、数据挖掘**.

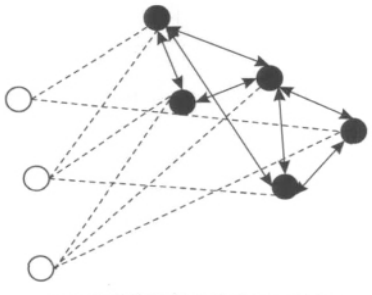
1 引 言

社交网络(Social Networks Services,SNS)随着 Internet 用户的普及呈现出飞速发展的趋势,不仅用户数量爆炸性地增长,其服务形态也在发生急剧的变化.近年来,大量新型的社交网络服务不断地涌现,其中以国外 Twitter 和国内新浪微博为代表的弱关系社交网络微博服务(Micro Blog)正成为一种主要的社交网络形态.

与传统的社交网络不同,由于弱关系的单向性,基于弱关系(即单向关注关系)的社交网络中的节点呈现出明显的异构性特征,包括大量以自然人为主体的用户节点(如"张三")和以媒体、机构以及各类消息源为主体的主题节点(如"北京天气"、"南方周末"、"热门视频"等).其中,用户节点,通常作为消息订阅者,单向关注大量主题节点,这些单向订阅关系,往往基于用户对于不同类型主题的兴趣倾向;同时用户节点常常与其它用户节点形成双向关注关系,这通常基于用户的真实社会关系.相反,主题节点,作为消息的发布者,被大量的用户节点订阅,而其主动关注和双向关注关系数量远远小于其被订阅的数量.图 1(a)展示了一个典型基于强关系的社交网络结构,网络中的节点呈现出同构性.图 1(b)为从新浪微博中提取出的一个典型异构弱关系社交网络(黑点为用户节点,白点为主题节点,虚线为单向订阅关系,实线为互关注关系).



(a) 典型的强关系社交网络结构



(b) 典型的弱关系社交网络结构

图 1 典型的强和弱关系社交网络结构

自然的,异构社交网络中的推荐系统所服务对象是用户节点,其推荐的内容主要分为两类:向用户推荐其它用户节点(如向"张三"推荐"李四")和推荐主题节点(如向"张三"推荐"北京天气"),即面向互关注关系的推荐和面向订阅关系的推荐.对于互关注关系和订阅关系的推荐需要基于不同因素:对于互关注关系的推荐,通过共同好友、联系人、通讯录等真实社交信息的方法通常就达到很好的效果[1];对于订阅关系,需要基于用户的兴趣进行推荐,这与推荐系统中常见的商品推荐、文档推荐等场景类似.关于这类推荐问题,有学者也进行了充分的研究[2-4],提出了协同过滤、基于内容等方法.

然而,社交网络上的推荐问题,特别是对订阅关系的推荐,不同于传统的推荐系统,其一大挑战在于它极端的数据稀疏性.Mislove 等人[5]指出,Internet 上的社交网络呈现出 Scale-free Network 的特点,极少量的用户拥有较多的关系连接,而大量的用户仅具有少量的关系连接.由于大量主题节点的存在,这一现象在弱关系社交网络中更加显著.如图 2 所示,根据对新浪微博中抽样的 500 个用户和 50 个主题的统计,仅有 20%的用户对 10%以上的主题有订阅关系,而关注主题数量低于总主题数量 5%的用户占了超过一半的比例.而对于如此稀疏的数据,协同过滤等单纯基于二元关系的方法不能达到理想的推荐效果.

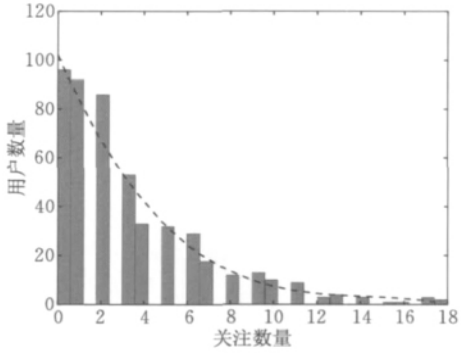


图 2 新浪微博用户关注度抽样统计

此外,社交网络随着新用户的不断加入,往往面临着冷启动(Cold Start)的问题.新加入的用户往往体现出很少的兴趣倾向,而基于内容的推荐方法往往不具有足够的多样性,使得推荐结果会很快地收敛于一个小范围的集合,从而丧失对更多用户感兴趣内容产生推荐的可能.

本文所解决的问题是在微博类的异构社交网络中对用户进行主题节点的推荐(即订阅推荐),并处理社交网络中普遍存在的数据稀疏性和冷启动场景.对此,本文提出了一种基于两阶段的用户聚类的

主题推荐的方法 GCCR(Graph-Content Clustering Recommendation). 首先，选取用户节点中关注数量较高的节点，从而抽取出稀疏数据中的一个密集子集，利用图摘要(Graph Summarization)的方法，对此密集子集形成关注兴趣相似的核心聚类. 然后，提取种子聚类的微博内容特征和整个数据集中其它用户的内容特征，基于内容相似度对整个用户群进行聚类，最后将聚类结果用于主题推荐.

通过对密集数据子集和全数据集的两阶段聚类过程，提高对极端稀疏数据集的聚类效果. 同时，由于图摘要聚类中的类模糊性，可以在对用户兴趣聚类的过程中保留一定的多样性，从而避免冷启动时收敛过快.

本文第 2 节介绍社交网络分析和推荐系统的相关工作；第 3 节阐述 GCCR 算法的总体架构；第 4 节介绍 GCCR 算法各环节的具体步骤；第 5 节介绍数据集的获取并且通过真实的数据集验证 GCCR 的推荐效果，通过多组对照实验，分析不同参数对推荐结果的影响；第 6 节是对本文工作的总结和展望.

2　相关工作

目前，对于推荐系统的研究很多，在推荐算法中，主要的研究方向包括协同过滤推荐、基于内容的推荐、聚类技术、Bayesian 网络技术、关联规则技术等.

协同过滤算法是目前最受欢迎的推荐技术，它利用用户爱好之间的相似性来进行推荐[3]，不依赖于物品的实际内容，而是需要用户对物品的偏好信息，通常以评价或者打分的形式[2]. 然而这种经典的协同过滤方法不能直接应用于社交网络的好友推荐，因为在社交网络中，没有物品和评分的概念. 此外，由于社交网络的数据稀疏性，协同过滤算法的推荐效果不好.

另一些研究利用物品的内容进行推荐，根据用户过去喜欢的物品，为用户推荐和他过去喜欢的相似的物品[4]. 基于内容相似性的方法可以很好地应用在社交网络的好友推荐中，文献[6]利用自然语言处理的技术对用户的 tweet 进行处理，提取出用户的兴趣点，从而推荐有相似兴趣的好友. Sakaguchi 等人[7]提出了一个基于概念模糊集(CFS)的系统，该系统识别 Twitter 用户的兴趣并推荐相关的好友，系统使用了基于模糊集的概念词典以及词向量来代表单个 Twitter 用户的兴趣，用向量余弦值衡量用户的相似度. 然而，基于内容相似度的推荐过于专一化，只能推荐出与用户兴趣相似的好友.

Facebook 上有一个功能是“你可能认识的人”，它是基于“Friend-of-friend”算法进行推荐的①. 该算法的思想是：如果 A 的很多好友是 B 的好友，那么 A 也可能会是 B 的好友. 这种算法只能帮用户寻找没有添加的强关系，经济社会学家马克·格兰诺维特提出：相对于强关系而言，弱关系有助于传递新信息[8]. 针对 Twitter 和新浪微博这种弱关系型社交网络，弱关系的推荐比强关系更有价值.

一些研究者把基于内容相似度的算法与社交网络的好友关系相结合. Hannon 等人[9]提出了 Twittomender 系统，根据用户发布的 tweet、好友、粉丝以及好友和粉丝的 tweet 对用户进行建模，利用 Lucene 的 TF-IDF 衡量关键词的权重. 文献[10]利用概率模型来进行协同过滤，可以为用户推荐最感兴趣的 K 个好友和 K 条 tweet. 概率模型综合考虑了 tweet 信息和用户之间的关系，Kim 等人[10]还提出了一个预测算法来推算概率模型的参数，并且使用 MapReduce 来处理大规模数据.

此外，在社交网络中，大型图数据集的研究和处理很重要. 图摘要技术可以用来发掘数据中隐藏的信息，现有的图摘要算法大多是基于统计学的，Tian 等人[11]提出了一种基于节点聚合的方法 k-SNAP 来进行图摘要计算，该算法可以自由地调整图摘要的聚合程度和迭代次数，得到对图不同粒度的分析结果. 文献[12]提出一种自动对数值属性值进行分类的算法，该算法通过发掘节点数据中隐藏的领域知识以及对图中边的结构的分析进行分类.

3　GCCR 总体框架

GCCR 算法旨在根据弱关系社交网络中用户对不同主题的兴趣程度，为用户推荐其可能喜欢的主题内容. 通过分析从用户-主题喜好矩阵和用户自身发表内容的中体现出的用户喜好信息，并将二者综合利用，提高在稀疏数据集上的推荐效果. 同时利用图摘要算法中的类模糊性，保证冷启动条件下推荐的多样性. GCCR 主要步骤包括预处理、核心聚类、全用户聚类、主题推荐阶段，主要流程如图 3 所示.

① Official Facebook Blog. http://blog.facebook.com/blog.php?post=15610312130.

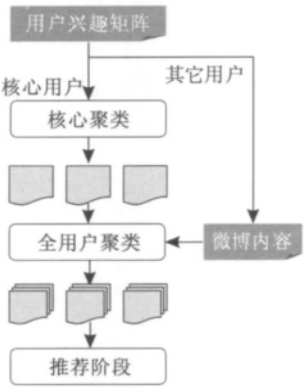


图 3 GCCR 主要流程

(1) 预处理阶段. 筛选出兴趣向量非零值比例大于密度阈值 λ 的核心用户集合. 根据核心用户对应的兴趣向量提取构造出原兴趣矩阵的密集子矩阵.

(2) 核心聚类阶段. 根据核心用户的兴趣矩阵构成的订阅关系图进行图摘要计算,利用摘要迭代的过程生成满足模糊度和独立性约束的核心聚类.

(3) 全用户聚类阶段. 利用上一步中生成的核心聚类,提取核心聚类内容特征向量,同时提取非核心用户所发表微博的内容特征向量,根据内容特征向量的相似度不断迭代,将非核心用户加入到已有的聚类集合中,直至完成对所有用户的聚类.

(4) 根据聚类结果以及每个用户在聚类内部的相似度和类兴趣特征,生成类成员内部的推荐向量,同时根据不同主题在不同聚类间的兴趣差异,形成跨类推荐向量,两者综合排序之后的结果作为最终推荐结果.

推荐算法执行过程中,对于训练数据集进行(1)~(3)离线计算生成聚类结果和类兴趣向量,对于任何新加入的用户,只需进行(3)中的聚类归属计算和(4)中的推荐过程. 通过将离线和在线处理运算尽可能的分离,可以使算法达到更高的在线推荐计算效率.

4 聚类推荐算法

4.1 问题建模

(1) 对于 N 个用户,M 个主题,可分别表示为用户集 $U=\{u_1,u_2,\cdots,u_N\}$ 与主题集 $S=\{s_1,s_2,\cdots,s_M\}$. 对于每个用户 u_i,有对应兴趣向量 $\boldsymbol{v}_i=(a_1,a_2,\cdots,a_M)$,所有用户的兴趣向量可构成 $N\times M$ 的兴趣矩阵 $\boldsymbol{m}$,对于存在订阅关系的用户 u_i 和主题 s_j,对应元素 $a_{ij}>0$,表示用户 u_i 对主题 s_j 兴趣度,如不存在订阅关系,则对应 $a_{ij}=0$.

(2) 基于兴趣矩阵 $\boldsymbol{m}$ 的兴趣图 G_m 可表示为有向图 $G(V,E)$,其中 V 为用户和主题节点构成的集合:

$$V=U\bigcup S,$$

E 为订阅关系构成的边集合:

$$E=\{e(u_i,s_j)\,|\,u_i\in U,\ s_j\in S,\ a_{ij}>0\}.$$

(3) 对于每个用户 u_i,定义其兴趣密度值 $des(u_i)$ 为兴趣向量 $\boldsymbol{v}_i$ 中非零元素所占的比例,那么对于 $des(u_i)$ 大于密度阈值 λ(通常取 10%)的用户 u_i 定义为核心用户. 那么核心用户集合可定义为

$$U'=\{u_i\,|\,u_i\in U,\ des(u_i)>\lambda\}.$$

由核心用户兴趣向量构成的兴趣矩阵为密集子矩阵 $\boldsymbol{m}'$,基于密集子矩阵可构造出核心兴趣图 $G_{m'}$.

4.2 核心聚类

图摘要算法通常用于从拥有大量节点的复杂图中提取隐含信息,发现主体结构和普遍规律. 不同于以往基于统计的图摘要方法,Tian 等人[11]提出了一种基于节点聚合的方法 k-SNAP 来进行图摘要计算. 该算法优势在于,图的摘要计算过程中不会丢失任何原始节点的信息,同时,可以自由地调整图摘要的聚合程度和迭代次数,得到对图不同粒度的分析结果. 我们注意到,k-SNAP 在图摘要迭代过程中同时完成了对节点的聚类,每个节点聚类是一系列与外部节点拥有相似连接度的节点的聚合. 因此,我们认为通过对用户兴趣图进行 k-SNAP 摘要,可以对用户节点实现兴趣聚类. 基于 k-SNAP,我们设计了 SNAP-Cluster 算法,使得(1) 聚合过程仅发生在用户节点之间;(2) 仅将 k-SNAP 算法应用于核心兴趣图,以保证能产生足够多有效信息的聚类;(3) 通过计算模糊度和差异性指数来表示聚类结果的特征,并以此来估计对最终推荐结果的影响.

4.2.1 聚类的度量

(1) 通过密集兴趣矩阵 $\boldsymbol{m}'$ 在核心用户集 U' 和主题集 S 上构造核心兴趣图 $G_{m'}(V,E)$,用户集 U' 上的一组聚类集 $Clus$ 可以表示为用户聚类 C_i 的集合,其中:

$$U'=\bigcup_{i=1}^{n}C_i,\ C_i\neq\varnothing,\text{对于 } i\neq j,C_i\cap C_j=\varnothing.$$

对于每一个主题 s_j,我们定义 C_i 的参与集:

$$P_{s_j}(C_i)=\{u\,|\,u\in C_i\text{ 且 }(u,s_j)\in E\},$$

那么参与度 p_{ij} 满足

$$p_{ij}=\frac{|P_{s_j}(C_i)|}{|C_i|}>\sigma(\sigma>0,\text{为强度阈值})$$

的 C_i 和 s_j 称为“聚类 C_i 强关注于主题 s_j”.

(2) 定义用户聚类 C_i 在主题 s_j 上的模糊度 Amb_{ij}：

$$Amb_{ij}=\begin{cases}|C_i-P_{s_j}(C_i)|, & p_{ij}\geq\sigma\\ |P_{s_j}(C_i)|, & p_{ij}<\sigma\end{cases},$$

由此可定义 C_i 对于主题集合 S 的模糊度：

$$Amb_i=\sum_{s_j\in S}Amb_{ij},$$

那么用户集 U' 上的 $Clus$ 对于主题集合 S 的全局模糊度为

$$Amb=\log\left(\frac{\sum_{C_i\in Clus}Amb_i}{|Clus|}\right),$$

此处取对数是为了保证全局模糊度随聚类增长呈线性变化的趋势.

(3) 定义用户聚类 C_i 在主题 s_j 上的兴趣度为

$$ca_{ij}=\begin{cases}\dfrac{\sum_{u_k\in C_i}a_{kj}}{|C_i|}, & p_{ij}\geq\sigma\\ 0, & p_{ij}<\sigma\end{cases},$$

则 C_i 在主题集 S 上的类兴趣向量为

$$\boldsymbol{cv}_i=(ca_{i1},ca_{i2},\cdots,ca_{iM}),$$

其中每个非零分量对应于一条强关注关系. 我们用不同用户聚类在主题集 S 上的兴趣向量的相互距离，来衡量聚类结果体现出的用户群体兴趣差异度. 两个聚类间的兴趣距离采用余弦距离：

$$diff(C_i,C_j)=\frac{\boldsymbol{cv}_i\cdot\boldsymbol{cv}_j}{|\boldsymbol{cv}_i|\cdot|\boldsymbol{cv}_j|},$$

那么用户集 U' 上的一组聚类 $Clus$ 对于主题集合 S 的差异性指数为

$$dvst=\frac{\sum_{C_i\in Clus,\ C_j\in Clus}diff(C_i,C_j)}{|Clus|}.$$

4.2.2 图摘要算法

基于之前的定义，我们给出图摘要聚类算法 SNAP-Cluster 的过程.

算法 1. SNAP-Cluster 图摘要聚类算法.

输入：核心用户集 U'，主题集 S，核心兴趣图 $G_{m'}$

输出：核心聚类 $Clus$

1. $Clus=U'$, $maxAmb=0$, $srcCi=$null, $arget=$null;
2. while $Amb=0$ 或达到目标迭代次数 k
3. for C_i in $Clus$ do
4. 计算对所有 s_j 的 Amb_{ij}，找到 $s_j=\text{argmax}(Amb_{ij})$;
5. 计算对主题集 S 的 Amb_i;
6. if $Amb_i>maxAmb$
7. $maxAmb=Amb_i$;
8. $target=s_j$;
9. $srcCi=C_i$;
10. end if
11. end for
12. 从 $Clus$ 中删除聚类 C_i;
13. $C_i',C_i''=split(C_i,target)$; //分裂聚类
14. $Clus$ 中加入 C_i';
15. 存储当前阶段的聚类结果 $Clus$ 和对应的 Amb，$dvst$ 值;
16. end while
17. return 最优的 $Clus$，使得 $Amb\cdot dvst$ 最大.

算法 2. Split 聚类分裂.

输入：用户聚类 C，主题节点 s_j，兴趣图 G

输出：聚类 C 对 s_j 的参与集 C_i' 和非参与集 C_i''

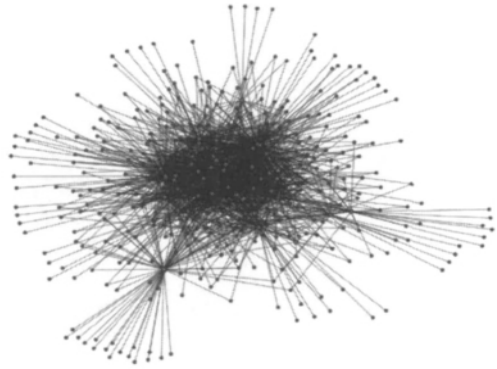
1. 初始化 C_i' 和 C_i'' 为空集;
2. for u_i in C
3. if $((u_i,s_j)\in E(G))$
4. 将 u_i 添加到 C_i' 中;
5. else
6. 将 u_i 添加到 C_i'' 中;
7. end if
8. end for
9. return C_i', C_i''.

对于最优聚类的选择，我们基于这样观察：当差异性指数越大时，表明类间的兴趣越不相同，这使得每一个聚类的兴趣特征越明显，增加了对兴趣预测的精确性. 相反，模糊值越大，则在一个聚类的内部保留了更大的差异性，因而增加了产生多样性推荐的可能. 因此我们考虑将二者综合考虑，在最后的实验中我们也将验证上述结论. 图 4(a) 展示了一个来自于新浪微博中 500 个用户和 50 个主题所构成的兴趣图（平均兴趣密度 7.2%），图 4(b) 是通过 SNAP-Cluster 计算之后的核心聚类图（平均密度 15%，小点为主题，大圆为用户聚类，中心数字为聚类大小）.

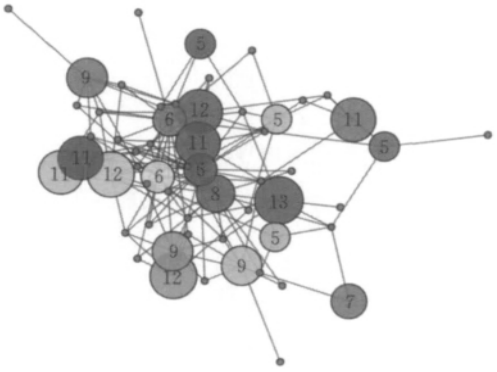
4.3 全用户聚类

得到用户核心聚类 Sim_{ik} 之后，我们需要提取核心聚类以及非核心用户的内容特征向量.

对于用户 u_i，其发表的微博为 $OriginTweets_i$，首先对原始微博数据进行预处理，比如去掉微博中的表情符号，去掉“@”某人的信息等等，得到用户发表的纯文本微博内容 $Tweets_i$.



(a) 500个用户和50个主题构成的兴趣图



(b) 由核心聚类算法生成的聚类关注图

图 4 聚类前后的兴趣图对比

定义用户 u_i 的特征向量为 $\boldsymbol{V}_{u_i}$，则 $\boldsymbol{V}_{u_i}=(Tweets_i)$. 核心聚类 $Clus_j$ 的特征向量为 $\boldsymbol{V}_{Clus_j}$，则有 $\boldsymbol{V}_{Clus_j}=(Tweets_m)$，$u_m\in Clus_j$.

我们采用改进的编辑距离算法来计算内容特征向量相似度[13]，编辑距离最初用于衡量字符串之间的相似程度，并以单个的字符作为基本的计算单位，为了使其更加适合具有语义的汉语句子相似度计算，算法采用对句子进行自动分词后的单个词作为基本的编辑单元. 此外算法还考虑编辑操作代价和句子长度对相似度的影响，提出了新的块交换操作，并根据词汇之间的语义相似度赋予不同的编辑操作不同的权重，在不用经过词义消歧和句法分析的前提下，兼顾了句子结构和词汇语义信息. 对于用户 u_i，我们用改进的编辑距离算法来计算他和所有核心聚类 $Clus_j$ 的相似度 Sim_{ij}，若最大值为 Sim_{ik}，则将用户 u_i 加入到聚类 $Clus_k$ 中. 将所有的非核心用户加入到相应的聚类之后就可以得到全用户聚类 $Gclus$.

4.4 推荐阶段

得到全用户聚类 $Gclus$，可计算出其中每个用户聚类 C_i 在主题集 S 上的类兴趣向量：

$$\boldsymbol{cv}_i=(ca_{i1},ca_{i2},\cdots,ca_{iM}),$$

所有聚类的类兴趣向量可构成类兴趣矩阵 $\boldsymbol{m}$，对于其中的零值，利用 Slope One 算法[14]进行预测. 定义主题 s_i 和 s_j 间的平均兴趣偏差值为

$$dev_{i,j}=\sum_{C_i\in Gclus}\frac{ca_{ki}-ca_{kj}}{|Gclus|}.$$

那么对于任意零值分量，均可以通过以下公式进行预测，其中 $\overline{ca}_i$ 为向量 $\boldsymbol{cv}_i$ 各分量的平均值，$M-1$ 是考虑 $i=j$ 时 $dev_{j,i}$ 值为零的情况：

$$ca_j=\overline{ca}_i+\frac{\sum_{i=1}^{M}dev_{j,i}}{M-1}.$$

将原向量 $\boldsymbol{cv}_i$ 中的零值用预测值填充，得到预测兴趣向量 $\boldsymbol{cv}'$，对每个分量表示的兴趣值进行排序，对每个用户，除开其已经关注的主题，对其余主题按照 Top-K 兴趣值进行推荐. 在实践中，我们通常取 K 值为用户已关注主题数或该数量的一半.

对于在线推荐的情况，首先对于需要进行推荐用户，可以提取其发布内容的特征向量，利用全用户聚类过程中的归类过程，将用户分配到合适的聚类之中，再利用该聚类的预测向量 $\boldsymbol{cv}'$，对其进行推荐. 可以看到，整个流程中，除用户归类的过程需要实时计算之外，用户聚类和兴趣值预测均可直接采用事先离线处理之后的结果. 在线推荐的计算复杂度，仅与用户聚类个数有关，而用户聚类个数在实际情况下是非常有限的，这也保证了本算法的在线推荐效率. 对于聚类和推荐的结果，需要在新用户增加到一定数量并对兴趣分布产生明显影响时进行调整.

5 实验分析

5.1 数据集

尽管本文的研究内容建立在已有的用户-主题兴趣的数值矩阵之上，然而我们在真实的数据集上却无法直接获得这一量化的兴趣指数. 因而在实验中我们需要建立用户兴趣指数的一个度量，尽管这与本文所描述的算法内容是无关的，但是为了表现实验的效果，这项工作是必要的.

我们用“用户期望转评率”来描述用户对主题的兴趣程度，它在微博系统中的意义可以理解为一个用户对某一主题所发表的内容进行评论或转发的潜在概率，这一指数与用户自身转评率规则化之后，可以用以下条件概率公式近似：

$$a=\frac{P(r|R)}{P(r)}=\frac{P(r)\cdot P(R|r)}{P(R)\cdot P(r)}=\frac{P(R|r)}{P(R)},$$

其中，$P(R)$为阅读到主题 R 的概率，$P(R|r)$为用户转发、评论的内容来自于主题 R 的概率，以上两个概率可以用对实验数据集的统计结果近似. 在接下来的讨论中，都以这一兴趣值度量为基础.

我们通过新浪微博的开放平台 API 抓取实验数据. 由于社交网络中存在海量的用户和信息，简单随机地抓取节点会造成实验数据过于稀疏，也不能体现弱关系社交网络中的结构特征. 因此，我们通过以种子用户为起始逐步在新浪微博的网络中生成兴趣图的方式来模拟基于弱关系的网络社区的形成过程，从而得到具有异构社交网络特征的局部样本. 主要过程为：① 以 5～10 个相邻或相近的用户节点作为种子节点. ② 对每一次迭代，采用深度优先的方式，抓取与当前用户相邻的用户节点；或者采用广度优先的方式，抓取当前用户所关注的主题节点. ③ 根据统计得到的用户关注对象中用户节点和主题节点的平均比值，来调整迭代中进行两种抓取的比例. ④ 根据抓取到的用户集合和主题集合，获取详细的关注、转发、评论的数据，根据前文的公式，计算“用户期望转评率”，从而得到最终的用户-主题兴趣矩阵.

在本实验中，我们抓取了多组各不相同的用户-兴趣关注矩阵，最终的实验结果为各种实验数据的平均值. 其中每一组均包含约 500 个用户、50 个主题和近 2 万条微博内容. 实验代码由 Python 和 Java 实现. 实验代码运行于 MacBook Pro MC990 上，Python 版本为 2.7，jdk 版本 1.6.

实验参照算法为：(1) 基于 Top-K 相似的协同过滤推荐算法(Collaborative-Filtering，CF)；(2) 基于主题内容相似度(Content-based)的 K 近邻推荐算法. 实验参照算法基于开源的机器学习库 Apache Mahout 而实现. 其中 Collaborative-Filtering 算法为基于用户的协同过滤，用户相似度的计算采用皮尔森相关系数，最终的推荐结果使用 Top-K 推荐. Content-based 算法中主题之间的相似度采用汉语句子相似度来计算[13]. 实验中，我们将用户的一半主题作为训练集，将另一半主题作为测试集进行实验.

5.2　推荐效果

我们在实验中观察到，几种算法在面对稀疏数据集的情况下，其产生推荐结果的能力是不同的，表 1 显示了不同数据稀疏度情况下，几种算法所能产生的最大推荐结果数量的对比(本实验中我们限制最大推荐数量不超过测试集的非零兴趣值数量).

表 1　产生推荐结果数量对比

稀疏度/%	需推荐数量	(实际推荐数量/需推荐数量)/%			
		CF，Top-4	CF，Top-10	Content-based	GCCR
5.50	725	7.1	39.8	16.0	99.3
6.75	639	8.1	41.9	18.1	99.7
8.65	520	8.4	43.8	20.7	100.0
11.75	169	8.8	45.6	18.9	100.0

显然，在数据极端稀疏的情况下，无论是 CF 还是基于内容的方法，都不能产生足够的推荐结果，而 GCCR 方法的推荐能力受数据稀疏性的影响非常小.

接着，我们对比几种算法在最优参数下的准确率和召回率. 我们总是取推荐个数等于训练集中的关注个数，而对于 CF 和 Content-based 算法，我们取 Top-k 个数为 10 个. 准确率可表示为推荐命中的数量与总推荐数量的比值，召回率可表示为推荐命中数量与测试集中总关注数量的比值. 图 5(a)显示了在不同数据稀疏程度下各算法的推荐准确率.

可以看到，在极端数据稀疏的情况下(密度低于 10%)，协同过滤算法具有最差的准确性，在密度超过 10%时逐渐提升. 而基于内容的推荐方法对于数据的稀疏性并不敏感，然而准确率维持在不高的水平. GCCR 算法在数据极端稀疏的情况下仍能保持较高的推荐准确率，并且随着数据密集度增大也能逐步提升，这与本算法在图摘要过程中实现了数据的密集化有关. 因而，在微博类异构社交网络的场景下，由于普遍存在的数据稀疏性(常常低于 10%)，GCCR 相比传统方法具有显著的提升.

图 5(b)显示了不同算法间召回率的对比.

可以看到，GCCR 算法在稀疏数据集中，召回率保持在稳定的较高的水平. 而 CF 和 Content-based 的方法，受数据稀疏性的影响较大，产生的推荐数量和质量都比较差，其中 Content-based 的方法由于产生推荐结果的数量较少而使得召回率非常低.

为了更好地比较算法的推荐质量，我们引入了 F_{measure} 和 MAP(Mean Average Precision)两项指数. 其中，F_{measure} 为准确率和召回率的调和平均值，该值越高则表明推荐算法的综合性能越好：

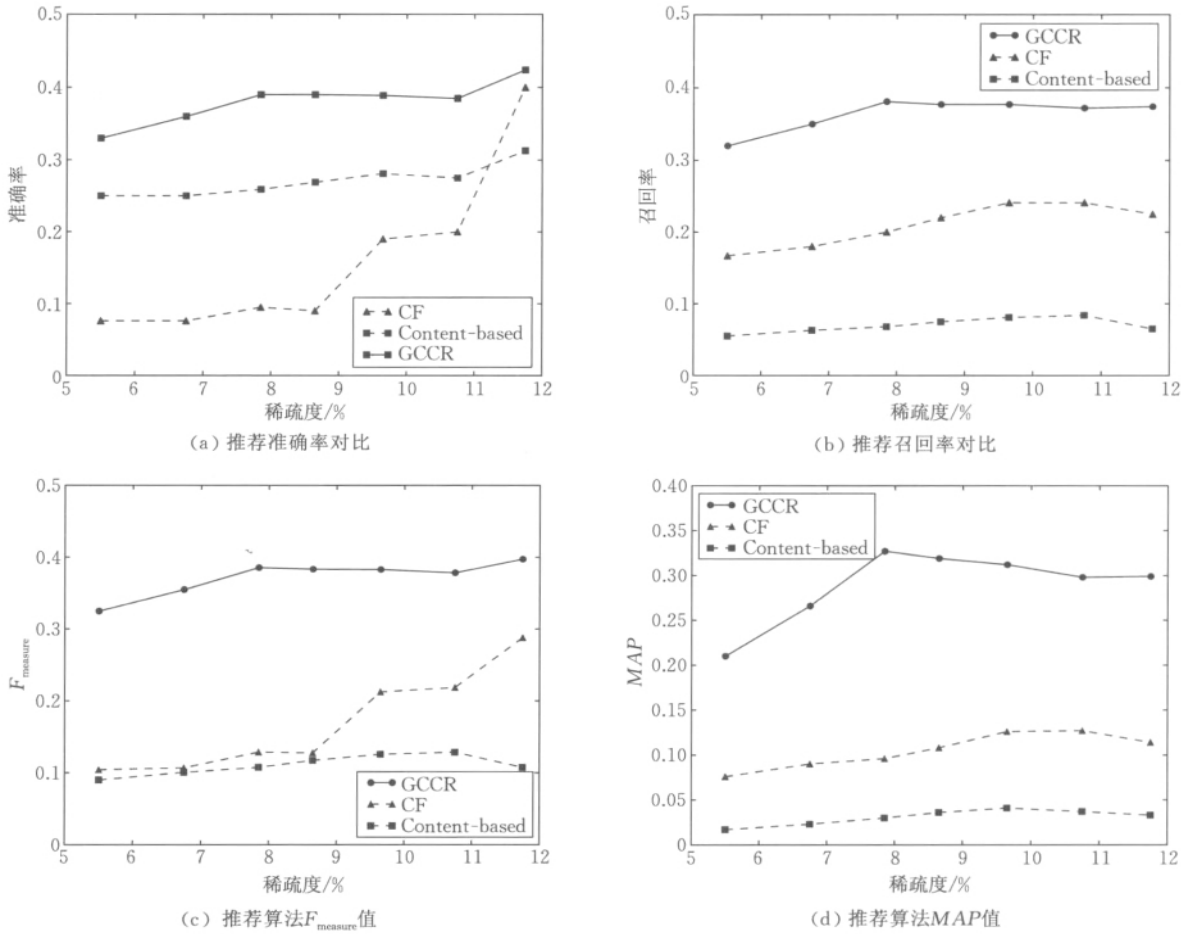
$$F_{\text{measure}}=\frac{2\cdot precision\cdot recall}{(precision+recall)},$$

同时，对一个用户群体产生的推荐结果的 MAP 可定义为对每个用户推荐结果的 AP(Average Precision)值的平均值，该值越高，则表明推荐算法的总体推荐

质量越好：

$$MAP=\frac{\sum_{k=1}^{U}AP(k)}{U}.$$

而 AP 值表示对某个用户推荐结果的平均准确率. 3 种算法的 F_{measure} 值和 MAP 值分别如图 5(c)和图 5(d)所示.



(a) 推荐准确率对比 (b) 推荐召回率对比

(c) 推荐算法F_{measure}值 (d) 推荐算法MAP值

图 5 3 种算法推荐效果的对比

可以看到，GCCR 算法无论在推荐综合性能和推荐质量上，都有较大的优势.

5.3 多样性

我们通过人工标注，将 50 个主题分为文字、时尚、笑话、资讯等 16 个类别. 由于在冷启动场景下，新加入的用户通常关注的主题有限，因而其兴趣范围可能仅体现在几个类别的主题之中，而推荐算法为了避免过快地收敛，我们希望能产生一定范围内更多样的推荐结果，从而提高跨类别推荐的可能性. 我们将训练集中不存在对某类别主题的关注而产生对该类别主题的推荐定义为“跨类别推荐”，同时对于此类推荐结果在测试集中的命中称为“跨类别命中”. 在实验中，我们通过针对性地删除在训练数据集中每个用户对于于某一类或几类主题的关注信息，来模拟冷启动场景下的数据特点，以此来测试算法对跨类推荐的能力. 表 2 显示了实验中的一组结果.

表 2 跨类推荐结果

算法	删除类别数	推荐数量	跨类别推荐数	命中数量	跨类命中数量
Content-based	1	271	71	59	21
	2	184	99	49	33
	3	133	88	39	27
GCCR	1	744	103	313	42
	2	980	197	327	79
	3	1150	214	268	83

可见 GCCR 算法在模拟冷启动的场景下，产生跨类推荐的数量和质量都明显较好. 在图 6 中，我们采用产生的推荐结果中跨类推荐的 F_{measure} 值和总体推荐结果的 F_{measure} 值的乘积来衡量推荐的多样性.

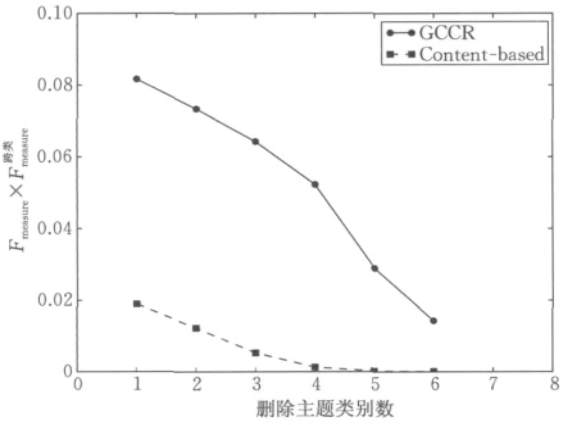


图 6　推荐多样性对比

可以看到，相比传统基于内容的推荐方法，GCCR 可以产生更高质量的跨类别推荐结果，这来自于 GCCR 中产生的聚类模糊性. 而基于内容的方法在主题类别缺失较多的情况下，具有极低的推荐多样性，导致推荐结果过快收敛.

5.4　各参数对推荐效果的影响

5.4.1　模糊度和差异性指数

模糊度是对一个聚类内部成员间，对于主题关注的差异程度的度量，用 Amb 表示. 在图 7 中我们可以看到，当前聚类结果的全局模糊度随着聚类个数的增加而减少，这是因为当聚类变小时，会更容易形成强关注关系. 同时，推荐算法的整体效果则随着模糊度的减少而提高，并且随着数据集密度的增加，这一差距显得更加明显. 然而当聚类数量过多时，推荐的准确率会出现降低的情况，这是由于过小的聚类使得兴趣矩阵变得稀疏.

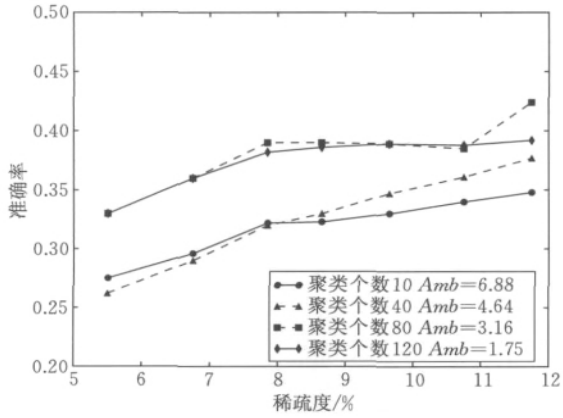


图 7　不同模糊度下的推荐准确率

差异性指数反应聚类之间的的兴趣平均差异程度，其随着聚类数量的增加而升高. 在图 8 中我们可以看到随着差异性指数的升高，用 $dvst$ 表示. 推荐效果的多样性逐渐降低. 聚类数为 10 时，$dvst$ = 0.524 取值最低，此时具有更强的推荐多样性，当聚类个数达到 80 时，$dvst$ 取值最小，此时较小的聚类使得此时推荐的多样性显著降低. 这可以理解为聚类间的兴趣差异增大，而聚类内部兴趣更加一致时，更难产生跨类别的推荐.

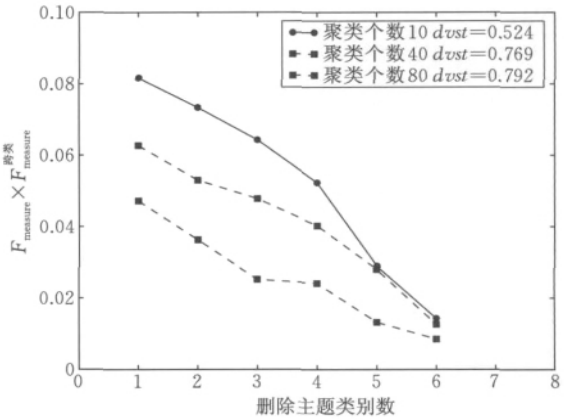


图 8　不同差异性指数下的推荐多样性

我们可以从上述实验结果中看到，选择最优的聚类数量时，需要同时考虑对推荐多样性和准确率的影响. 更多聚类个数使得每个聚类的模糊度降低，在提高了推荐的准确率的同时缩小了兴趣的预测范围，从而减少了产生跨类推荐的可能. 相反，较少聚类的个数，能够提供更广的推荐范围，从而提高在冷启动时的推荐效果. 因此，聚类数量的确定依赖于具体的推荐需求. 实践中，在没有明确倾向的情况下，我们选择使得差异性指数和模糊度乘积达到最大值时的聚类结果.

5.4.2　关系强度阈值 σ

σ 定义强关注关系在一个聚类中需要满足的最小覆盖度，σ 取值决定了在聚类过程中对聚类兴趣的置信程度. 当我们需要推定一个聚类对某主题是有兴趣时，若 σ 取值越大，则需要此类中更多的成员满足对该主题的关注关系. 而 σ 值越小时，对于聚类兴趣的判定条件则趋于宽松. 图 9 显示了 σ 取值对预测准确率的影响.

σ 定义强关注关系在一个聚类中需要满足的最小覆盖度，在 Tian 等人[11] 的论文中，将 σ 取值为 0.5，而 GCCR 的实现中，面对更加稀疏的数据集，相对宽松的强关系判断条件（σ=0.3 时达到最优），使得由图摘要形成的聚类具有更多的非零兴趣值，从而能达到更好的推荐效果. 而当强度阈值过低时推荐效果有所下降的原因是由对于类兴趣判断过于模糊所致.

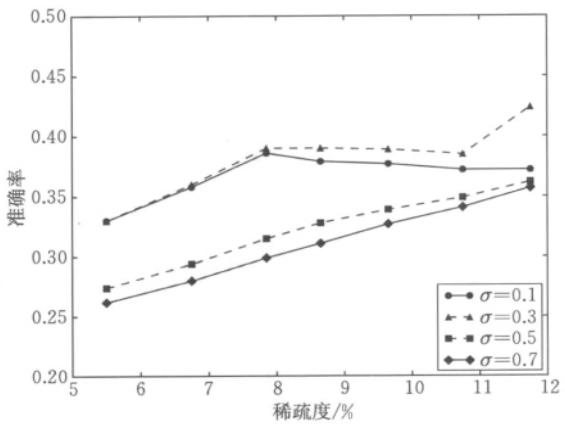


图 9 不同 σ 取值下的推荐准确率

6 总结与展望

为了解决微博类异构社交网络中存在的数据稀疏性和冷启动问题，本文提出了基于图摘要和内容相似混合聚类的推荐算法 GCCR. GCCR 在极端稀疏的数据集上具有较高的准确度，同时在冷启动的场景下能够提供多样性的推荐结果，从而避免推荐结果收敛过快的问题. 最后，我们通过真实的数据集验证了算法的效果，并且分析了各参数对推荐结果的影响.

在接下来的工作中，我们准备将系统实际部署实施，并且希望引入反馈机制，根据用户对推荐结果的实际兴趣反馈，实现推荐算法的动态优化. 同时希望能够将离线计算的部分并行化处理，以获得更高的算法执行效率.

参 考 文 献

[1] Chen J, Geyer W, Dugan C, Muller M, Guy I. Make new friends, but keep the old: Recommending people on social networking sites//Proceedings of the 27th International Conference on Human Factors in Computing Systems. New York, NY, USA, 2009: 201-210

[2] Sarwar B M, Karypis G, Konstan J A, Riedl John. Analysis of recommendation algorithms for e-commerce//Proceedings of the 2nd ACM Conference on Electronic Commerce (EC-00). Minneapolis, MN, USA, 2000: 158-167

[3] Linden Greg, Smith Brent, York Jeremy, Amazon. com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 2003, 7(1): 76-80

[4] Pazzani M J, Billsus D. Content-based recommendation systems//Brusilovsky P et al eds. The Adaptive Web. Springer Verlag, 2007: 325-341

[5] Mislove Alan, Marcon Massimiliano, Gummadi Krishna P, Druschel Peter, Bhattacharjee Bobby. Measurement and analysis of online social networks//Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. San Diego, CA, USA, 2007: 29-42

[6] Piao Scott, Whittle Jon. A feasibility study on extracting twitter users' interests using NLP tools for serendipitous connections//Proceedings of the 3rd IEEE International Conference on Social Computing (SocialCom-2011). Boston, MA, 2011: 910-915

[7] Sakaguchi T, Akaho Y, Takagi T, Shintani T. Recommendations in twitter using conceptual fuzzy sets//Proceedings of the 2010 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS). Toronto, Canada, 2010: 1-6

[8] Granovetter M. The strength of weak ties. American Journal of Sociology, 1973, 78(6): 1360-1380

[9] Hannon John, Bennett Mike, Smyth Barry. Recommending twitter users to follow using content and collaborative filtering approaches//Proceedings of the 4th ACM Conference on Recommender Systems (RecSys'10). Barcelona, Spain, 2010: 199-206

[10] Kim Younghoon, Shim Kyuseok. TWITOBI: A recommendation system for twitter using probabilistic modeling//Proceedings of the 2011 IEEE 11th International Conference on Data Mining (ICDM). Vancouver, Canada, 2011: 340-349

[11] Tian Yuanyuan, Hankins Richard A, Patel Jignesh M. Efficient aggregation for graph summarization//Proceedings of the SIGMOD Conference. Vancouver, Canada, 2008: 567-580

[12] Zhang Ning, Tian Yuanyuan, Patel Jignesh M. Discovery-driven graph summarization//Proceedings of the ICDE. Long Beach, California, USA, 2010: 880-891

[13] Xia Tian, Fan Xiao-Zhong, Luo Zheng-Hua, Liu Lin. Improved edit distance algorithm and Chinese sentence similarity computing//Proceedings of the 2th Excellent Doctoral Conference of China Science and Technology Association. Suzhou, China, 2004: 444-449(in Chinese)
(夏天，樊孝忠，骆正华，刘林. 改进编辑距离算法与汉语句子相似度计算//中国科协第 2 届优秀博士生学术年会. 苏州，中国，2004: 444-449)

[14] Lemire Daniel, Maclachlan Anna. Slope one predictors for online rating-based collaborative filtering//Proceedings of the SIAM Data Mining (SDM'05). Newport Beach, California, 2005: 471-480

CHEN Ke-Han, born in 1987, M. S. candidate. His main research interests include service computing, and data mining.

HAN Pan-Pan, born in 1989, M. S. candidate. Her main research interests include service computing, data mining and social computing.

WU Jian, born in 1975, Ph. D., associate professor. His main research interests include semantic Web, Web service, and data mining.

Background

Recent development of information technologies produces a lot of community services. SNS (Social Network Service) is one of the community services on the world wide webs. In the SNS, a user can register other users as friends and enjoy communication through a virtual message and a diary such as blog. Recommendation systems can help users find known, offline contacts and discover new friends on social networking sites. Facebook has recently launched a feature, called "People You May Know", which recommends people to connect with based on a "friend-of-a-friend" approach. However, data on the effectiveness of this approach is not available.

However, despite the difficulty, connecting with weak ties or unknown but similar people can be more valuable to users than merely re-finding existing strong ties. Comparing to the ordinary social networks services like Facebook, the twitter-like weak-relationship based networks is observably heterogeneous. By classifying the nodes into users (subscriber) and subjects (publisher), the goal of recommendation systems over this kind of networks is basically recommending the subjects to the users for subscription.

In this paper, we propose GCCR, a hybrid method combining both graph-summarization and content-based algorithms by a two-phase user clustering approach, which can recommend subjects according to user interests. With respect to other methods, the GCCR algorithm could generate better recommendation result in sparse datasets and cold-start scenarios. In additional, by separating the task into offline and online parts, GCCR works more efficiently online by using the pre-processed offline results.