一种针对社会化导购的橙领推荐方法研究

谢晓芹 韩 帅 吕 斌 张志强 潘海为

(哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)

摘 要 社交网络和网络购物的发展普及导致了社会化导购的产生和发展,也催生了通过在社交网络中推荐产品从而获取利润的"橙领".通过对橙领相关技术的研究,能更透彻地了解基于社会网络的产品营销机制以及探索社会化导购的底层模式.目前国内外少有这方面研究.因此,文中针对橙领的自身定位问题和面向用户或商家的橙领推荐问题,提出一种针对社会化导购的橙领推荐方法,主要包括 3 个算法:橙领定位算法、面向用户的橙领推荐算法 OCRA4U(Orange Collar Recommending Algorithm for User)和面向商家的橙领推荐算法 OCRA4S(Orange Collar Recommending Algorithm for Shop).橙领定位算法依据橙领的推荐历史对橙领进行定位特征向量化描述,最终转化为一个聚类问题进行解决.OCRA4U 考虑了橙领在社交网络中的影响力和橙领与用户需求的匹配度,得到橙领推荐列表.OCRA4S 结合橙领在网络中的影响力以及橙领的历史推荐产品,推荐出最符合商家产品需求的橙领列表.基于新浪微博数据集和 DBLP 数据集,文中设计并实现了 3 个相关实验:橙领定位算法实验、橙领推荐实验以及社会化数据影响实验,实验结果验证了所提算法的准确性和可行性.

关键词 橙领推荐;橙领定位;社会化导购;网络影响力;社会媒体;社交网络;推荐系统

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2016.02114

Research on Social Sales Oriented Orange-Collar Recommendation Method

XIE Xiao-Qin HAN Shuai LV Bin ZHANG Zhi-Qiang PAN Hai-Wei

(*School of Computer Science and Technology, Harbin Engineering University, Harbin* 150001)

Abstract Social shopping guiders appear and develop with the development of the social network and online shopping. This also leads to the emerging of orange-collar who gets income by recommending products to other persons in SNS in China. Through studying orange-collar related technologies, we can not only learn the mechanism of online product marketing thoroughly, but also explore the deep level patterns of social sales. However, current researches seldom focus on this issue. Hence, to solve the problem of orientating the orange-collars and recommending them to users or shops, this paper proposes an orange-collar recommending method, which includes three algorithms that are orange-collar positioning algorithm, OCRA4U algorithm and OCRA4S algorithm. The orange-collar positioning algorithm describes orange-collar by orientation vector based on the varieties of the products in its recommending history and finally transforms to a clustering problem to solve. OCRA4U takes the orange-collar's network influence and the matching degree between the orange-collar and user's need into consideration and returns an orange-collar recommending list to user. OCRA4S combines the orange-collar's influence and its recommending history, and finally gains the most satisfied orange-collar recommendation for the product need. Based on Sina microblog dataset and DBLP dataset, this paper has designed and

收稿日期:2015-10-21;在线出版日期:2016-03-10.本课题得到国家自然科学基金(61202090,61370084,61272184)、教育部新世纪人才支持计划(NCET-11-0829)、哈尔滨市科技创新人才基金(2015RQQXJ067)、中央高校基本科研业务费专项资金(HEUCF100602)资助.谢晓芹,女,1973 年生,博士,副教授,中国计算机学会(CCF)高级会员,主要研究方向为社会网络分析与挖掘、智能信息处理和服务计算等.E-mail:xiexiaoqin@hrbeu.edu.cn.韩 帅,女,1991 年生,硕士研究生,主要研究方向为社会网络分析与挖掘.吕 斌,男,1989 年生,硕士研究生,主要研究方向为社会网络分析与挖掘.张志强,男,1973 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为信息检索、数据库和智能信息处理等.潘海为,男,1974 年生,博士,副教授,主要研究方向为数据库、医学图像挖掘.

implemented three experiments: the orange-collar positioning experiment, the orange-collar recommending experiment and the social data influence experiment, whose results have proved the correctness and effectiveness of the proposed algorithms.

Keywords orange-collar recommendation; orange-collar positioning; social sales; network influence; social media; social network; recommender systems

1 引　言

社交网络已经是 Web2.0 众多业务中使用用户数最多、对用户生活工作影响力最大且最有商业潜力的业务之一. 截至 2012 年 12 月为止，已经有 2.75 亿用户使用社交网络分享和获取信息. 以社交网络为代表的虚拟社交与现实生活中社交的交融日益加深.

与社交网络一样，网络购物也得到了飞速发展，且日益普及. 据 CNNIC 的《2013 年中国网购市场调查报告》显示[1]，2013 中国网络购物市场完成了 1.85 万亿元交易金额，同时在 2013 年底有 3.02 亿人在网络购物市场上有消费行为，网络购物的使用率高达 48.9%. 这些表明网络购物行为已经成为人们购物行为的重要组成部分. 同时，与网络购物息息相关的网购网站也得到飞速的发展，而流量是网购网站的生存基础，为了扩源、导流、增加流量，网购网站在社会网络中推荐自己的产品，这导致了社会化导购的产生和发展. 社会化导购是指社会个体在社会网络站点中推广产品的一种行为. 依据调查结果，用户接受的社会化导购 37.5%来源于微博，其次为蘑菇街、美丽说等[1].

社会化导购发展的同时诞生了橙领这个职业. 2011 年 8 月淘宝联盟“武林大会”首次提出“橙领”概念，用来指代“通过淘宝联盟赚取收入的人”，是一种电商相关从业者的新称谓. “橙领”的橙，取自淘宝主页的橙色系. 通常橙领指的是活跃在淘宝网和微博等 SNS 网络的一群人，他们有庞大的关注者或者粉丝群，通过在 SNS 发布关于购物消息的微博来牟利. 橙领每天的工作就是在获取商品信息之后，在诸如新浪微博等的社交网络上发布博文来推广该产品，而这些博文中通常包含商品的购物链接. 当用户点击商品购物链接购买商品时，橙领就会从商家那里获取一定报酬. 本文研究的就是橙领推荐的相关技术. 本文中，橙领是指在社交网络中作为社会化导购而存在的用户，包括基于淘宝联盟的常规橙领，还包括购物网站旗下的自媒体，以及第三方中立的团队. 本文探索了基于橙领的产品社会化营销模式，针对橙领的自身定位问题、面向终端用户的橙领推荐和面向商家的橙领推荐等问题展开了研究，提出了面向社会化导购的橙领推荐方法.

2 相关工作

2.1 社会网络的产品推荐

基于社会网络的产品推荐研究主要包含好友推荐、产品推荐以及热点社区发现等内容. 产品推荐常见的算法有购物网站使用的协同过滤算法、基于内容的推荐算法等，通过基于用户和产品基本信息、用户历史信息等数据向用户推荐产品. 好友推荐主要分析研究用户在社会网络中的关系，从而向用户推荐好友. Hannon 等人[2]在分析用户、用户好友、粉丝以及他们的 tweet 的基础上，研究实现了 Twittomender 系统，并使用该系统进行好友和 tweet 推荐. Chen 等人[3]认为基于用户在社交网络中产生的数据，能够向用户提供更好的好友推荐. Sakaguchi 等人[4]基于模糊概念集合在 Twitter 上向用户推荐与其有相似兴趣爱好的好友. Kim 等人[5]基于概率模型针对 Twitter 提出了一个推荐系统 TWITOBI，该系统分析了用户所发的 tweet 和用户好友关系链，向用户推荐最感兴趣的 K 个好友和 K 条 tweet. 陈克寒等人[6]提出了基于两个阶段聚类的推荐算法 GCCR，该算法解决了推荐算法的稀疏矩阵和冷启动问题，并通过实验验证了该算法能否向用户推荐感兴趣的主题. 王玓等人[7]使用社交圈划分社交网络中的用户，提出了基于社交圈的在线社交网络朋友推荐算法，该算法的主要思想是基于社交圈计算用户间的相似度，将与当前用户社交圈重合度高的用户推荐给当前用户. 高明等人[8]使用 LDA 主题模型来推断微博用户兴趣、社交网络中微博的主题分布，从而实现向用户实时推荐的目的.

2.2 社会网络影响力

在社会学中并没有对社交网络中的影响力做出明确的定义，但是大家普遍接受影响力是影响他人行为想法的一种能力，即在社会群体中某个体能够

通过语言或适当措施等行为改变他人想法或者对他人行为产生影响的一种能力.现实生活中,对影响力的研究分析可以通过问卷调查、跟踪实验人群等方式来获取相应数据.但是在社交网络中,如新浪微博中,若要研究某个个体的影响力,无论是采用问卷调查或者统计该个体的每位粉丝对他在社交网络中所发微博的认可度,进而来探究其对粉丝心理的影响,还是跟踪记录该个体粉丝看到该用户博文后的行为,都由于各种条件限制而难以实现.影响力的关键点是对他人想法或行为产生影响.用户在社交网络中浏览某条微博后若有以下行为:关注某人或者取消对某人关注、对微博发表自己的评论、转发该条微博、收藏该条微博、给发布该微博的个体发私信等中的一种或几种,则说明该用户受到了这条微博影响.收藏微博的行为相对于转发和评论的行为发生的概率低很多,所以本文不予考虑这种行为.发私信行为不但概率低而且该行为仅仅是针对于某条微博的原因的概率更低,因此本文也不予考虑.Cha 等人[9]从粉丝数、转发和提及(@某个体)3 个指标入手,对影响力进行评价,并且对三者做了对比研究.他们认为:个体被提及得越多,该个体微博被转发得也越多,但是该个体的粉丝数和其他两个指标之间关系不大.这表明用户影响力主要受用户平时对微博的管理的影响,而和粉丝数关系不大.评论行为是指微博用户对其看到的微博发表评论,与微博作者交流互动,这同时也表明其受到微博作者的影响.转发行为是指微博用户认可博文并将其转发给自己的朋友圈,其受到微博作者影响程度比评论行为发生时大,而且该行为更有意义,它能延长消息传播链并扩大消息传播范围.

2.3 用户画像技术

本文研究中进行的橙领的识别和定位,本质上也是一种用户画像问题.现有的用户画像技术可以分为三类,一类是通过分析用户的行为(如点击流和 post 行为)来对用户进行建模,一类是通过分析用户生成数据来对用户进行建模,还有一类是综合前面两种方法,如 Ikeda 等人[10]提出的基于文本挖掘和社区挖掘的画像技术.本研究中方法实际上是第三种用户画像技术,综合使用了用户生成的博客信息以及用户之间的社会关系,当然也可采用其它的用户画像技术.Pazzani 等人在文献[11]中提出了多种方法来实现用户画像,有决策树方法、最近邻方法、线性分类方法、朴素贝叶斯方法等.Wang 等人在文献[12]中提出了利用 LDA 方法获取用户主题模型进而最终得到用户画像的技术.

3 面向社会化导购的橙领推荐方法

3.1 问题分析

本研究中将参与社会化购物过程的用户分为 3 种类型:终端用户类 EUT(End User Type),商家类 SHOPT(Shop Type)和橙领类 OCT(Orange Collar Type).下面给出相关定义.

定义 1. 终端用户 EU(End User)是指只有购买行为的用户.所有终端用户可表示为集合 EU,$EU=\{eu|eu.type=\text{EUT}\}$.

定义 2. 商家用户 SU(SHOP User)是指提供产品给终端用户的人或群体.所有商家用户可表示为集合 $SHOP$,$SHOP=\{su|su.type=\text{SHOPT}\}$.

定义 3. 橙领 CU(Orange Collar User)是指既不购买商品,也不提供产品的个体,他们为终端用户推荐合适的商家,以帮助终端用户获得满意产品,并从商家获取利润.所有橙领用户可表示为集合 CL,$CL=\{cl|cl.type=\text{OCT}\}$.

定义 4. 社会化购物网络 $G=(V,E)$,其中 $V=EU\cup SHOP\cup CL\cup P$,$EU$ 表示终端用户集合,$SHOP$ 表示商家集合,CL 表示橙领集合,P 表示产品集合;E 是 V 中节点之间的关系集合,$E=BuyRe\cup SellRe\cup RecomRe$,其中 $BuyRe$、$SellRe$ 和 $RecomRe$ 中的 Re 是取单词 Relationship 的前两位,$BuyRe$ 表示终端用户对某产品的购买关系集合,表示为 $BuyRe=\{\langle eu,p\rangle|eu\in EU\cap p\in P\}$,$SellRe$ 表示商家对某产品的销售关系集合,表示为 $SellRe=\{\langle shop,p\rangle|shop\in SHOP\cap p\in P\}$,$RecomRe$ 表示橙领对某产品的推荐关系集合,表示为 $RecomRe=\{\langle cl,p\rangle|cl\in CL\cap p\in P\}$.

橙领是商家和终端用户(消费者)间的链接枢纽,产品信息通过商家流向橙领,再由橙领流向终端用户(消费者).商家通过橙领来推广自己的产品,终端用户(消费者)通过橙领来获取适合自己的产品信息.但是在橙领自身、橙领与终端用户关系链、橙领与商家关系链中都存在尚未解决的问题,具体如图 1 所示.下面具体分析图中的各个问题.

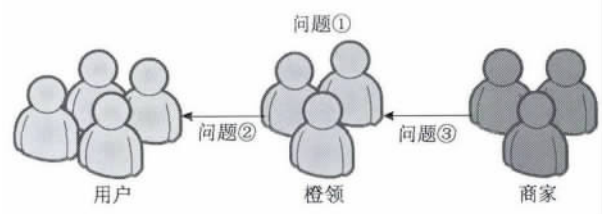


图 1 用户、橙领、商家关系

(1)图中问题①是指：橙领的自身定位问题. 橙领无法确切地知道自己最适合推荐什么产品，自身所在的社会网络中的用户对橙领推荐的哪些类产品比较认可. 针对该问题，本文的目标是：给定一个橙领，判定其是属于哪个类别的橙领.

(2)图中问题②是指：对于终端用户，如何选择最适合他们的橙领. 换言之就是：普通用户如何在形形色色的橙领中选择最适合自己的橙领，该橙领能给自己将要购买的哪类产品给出最中肯的建议. 针对该问题，目标是：给定一个用户，得到一个橙领的排序，并将排在前面的若干个橙领推荐给用户.

(3)图中问题③是指：对于商家，如何选择推广他们产品的合适橙领. 将最合适的橙领推荐给商家，商家可以基于这些橙领做更好更有效的产品推荐. 针对该问题，目标是：给定一个生产或销售某产品的商家，得到一个橙领的排序，并将排在前面的若干个橙领推荐给商家.

综上，面向社会化导购的橙领推荐问题，可以描述为以下两个子问题：假设给定所有用户所处的社会化购物网络 G 和所有用户所发的所有博文文档库 D(含导购和非导购两类博文)，对于终端用户 eu，如何向其推荐符合其需求的橙领或橙领列表；对于产品商家 su，如何向其推荐能使其销售额增长最快速的橙领或橙领列表.

3.2　算法基本思想

本文针对上述问题提出了一种面向社会化导购的橙领推荐方法(Social Sales Oriented Orange-Collar Recommending Method，SSOCRM). SSOCRM 算法首先要对原始文档进行预处理，去除停用词，抽取关键词. 然后调用 $CLClassify()$ 算法进行数据集的划分，实现橙领用户的定位. 最后，根据两种不同的目标用户：终端用户和商家分别调用不同的橙领推荐算法，获得排序在前面的一个橙领列表 $CLList$. 可以形式化描述为

算法 1.　SSOCRM 算法.

输入：社会化购物网络图 G，所有用户博文集合 D，用户 u，产品描述 p

输出：橙领列表

1. $D'\leftarrow$对 D 进行数据预处理；
2. $D''\leftarrow CLClassify(G,D')$；
   //调用算法对橙领进行定位，将橙领描述为定位特征向量集合
3. IF u 是终端用户 THEN
4. $CLList\leftarrow OCRA4U(D'',u)$；
   //调用面向用户的橙领推荐算法
5. ELSE
6. $CLList\leftarrow OCRA4S(D'',p)$；
   //调用面向商家的橙领推荐算法
7. END IF
8. RETURN $CLList$；

下面详细阐述各部分算法.

3.3　橙领定位算法

橙领定位是指依据用户在社交网络中发表的文章，判断其是否是橙领，并进一步判别其所属的类别，换言之，就是判断哪些橙领能推荐哪类产品. 只有明确了橙领的定位，才能基于橙领的定位进行橙领的推荐.

橙领定位算法包括数据集划分、橙领识别、橙领定位向量化三部分. 给定的输入包括两部分：(1)由橙领和非橙领用户的博文集组成的数据集 D；(2)某个用户 u 的所有博文. 输出为：若用户 u 被判定为橙领，则输出其定位向量，否则输出非橙领或半橙领标识. 橙领定位算法的形式化描述如下：

算法 2.　橙领定位算法 $CLClassify()$.

输入：(1) D={经过预处理后的所有用户的博文集}；
(2) B={待判别的某用户 u 的博文}

输出：若用户为橙领，则输出橙领标识及其定位向量；否则输出半橙领或非橙领标识

步骤：

1. $D'\leftarrow DataSplit(D)$；　//使用数据划分算法，将原始数据集 D 划分为橙领和非橙领数据集：U_{cl}={橙领用户博文集}，U_{ncl}={非橙领用户博文集}
2. $clflag\leftarrow CLIdentify(u)$；　//根据橙领识别算法，判断该用户类型，返回类型标记：橙领、半橙领、非橙领
3. IF($clflag$=1) THEN {　//用户为橙领
4. $IV\leftarrow CLFeatureLocate(u)$；　//利用橙领定位向量化算法，对用户进行定位
5. RETURN ($\langle IV,clflag\rangle$)；} //返回橙领标识和定位向量
6. ELSE RETURN $clflag$；//返回用户不是橙领的标识
7. END IF

其中主要包括以下 3 个步骤：

第 1 步. 基于原始数据集，使用橙领和非橙领数据集划分算法，将数据集划分为橙领数据集和非橙领数据集两部分，对应算法的第 1 步.

第 2 步. 利用橙领识别算法，判断该用户类型，类型为橙领、非橙领类，对应算法第 2 步.

第 3 步. 如果用户被判定为是橙领类，则利用橙领定位向量化算法，对用户定位，返回定位向量. 若用户不是橙领类，则返回其用户类型标识. 对应算法的第 3～7 步.

3.3.1 橙领和非橙领数据集划分算法

原始语料库 D 包含所有用户发布的所有博文文档，但并没有对博文数据进行橙领或非橙领的区分. 数据集划分算法的主要思想是先手工地选取一些橙领博文(即购物博文)与非橙领博文(非购物博文)，分别构成橙领数据集 U_{cl} 和非橙领数据集 U_{ncl}，然后进行迭代：对 D 中所有博文进行所属数据集的划分，就这样来迭代更新优化数据集. 这实际是一个数据预处理过程.

本文采用一个迭代更新过程对橙领和非橙领数据集进行优化. 优化的前提是能提取分别表示橙领数据集和非橙领数据集的关键字集. 每次迭代过程分为两部分：第一部分生成目前橙领数据集和非橙领数据集的加权关键字集合；第二部分基于加权关键字集合，计算 D 中每条博文分别归属于两个数据集的概率值，将博文归属到概率最大的类中，并从原类中删除，得到一个更新后的橙领数据集和非橙领数据集. 然后重复上述迭代过程. 下面详细介绍这两个部分.

(1) 加权关键字集生成. 本文使用 TFIDF 技术提取加权关键字集，关键字的权重为该字的 $tfidf$ 值. 我们将橙领或非橙领数据集中所有微博合并作为一个文档，因此该文档的内容为该集合所包含的所有博文，对文档进行预处理后能得到一个关键词的集合. 对于类 j(U_{cl} 或 U_{ncl})中的每个词 key_i，通过以下公式计算该词的 $tf_{key_{i,j}}$，$idf_{key_{i,j}}$，$tfidf_{key_{i,j}}$：

$$tf_{key_{i,j}}=\frac{n_{key_{i,j}}}{\sum_k n_{k,j}} \tag{1}$$

$$idf_{key_i}=\log\frac{|D|}{|\{key_i|key_i\in U_{cl}\bigcup key_i\in U_{ncl}\}|} \tag{2}$$

$$tfidf_{key_{i,j}}=tf_{key_{i,j}}\times idf_{key_{i,j}} \tag{3}$$

其中 $n_{key_{i,j}}$ 表示关键字 key_i 在类 j 中出现的次数，$\sum_k n_{k,j}$ 是类 j 中所有关键字出现次数之和. $|D|$ 表示橙领和非橙领文档集中所有词项的个数，$|\{key_i|key_i\in u_{cl_j}\bigcup key_i\in u_{ncl_j}\}|$ 表示橙领和非橙领集合中包含该关键字的个数.

从而可以得到类 U_{cl} 中所有关键字的 $tfidf$ 值，进而可将橙领数据表示为由关键字与其 $tfidf$ 值组成的二元组的集合，本文称之为加权关键字集：

$$V_{cl}=\{\langle key_1,tfidf_{key_{1,cl}}\rangle,\langle key_2,tfidf_{key_{2,cl}}\rangle,\cdots,\langle key_i,tfidf_{key_{i,cl}}\rangle\}.$$

同理可以把非橙领数据表示为非橙领加权关键字集合 V_{ncl}.

(2) 博文重新归类. 本文将数据集分为橙领数据集和非橙领数据集两大类. 一篇博文 t 归属于某数据集 j(U_{cl} 或 U_{ncl})的概率值可由式(4)获得

$$p_{t,j}=\sum_{key_i\in K_t\cap key_i\in K_j} tfidf_{key_{i,j}} \tag{4}$$

其中 K_t 表示博文 t 中的所有关键词，K_j 表示类 j 的博文文档中的所有关键词集合. 依据朴素贝叶斯分类原理，将博文归属于由式(4)获得结果最大的类. 迭代完成后，分别将橙领和非橙领数据集中的博文数据按照作者名称聚成一个个子类.

下面给出橙领和非橙领数据集划分算法. 输入是经过初步划分的原始数据集 U_{ncl} 和 U_{cl}.

算法 3. 橙领和非橙领数据集划分算法 $DataSplit()$.

输入：手动生成的非橙领数据集 U_{ncl}，橙领数据集 U_{cl}

输出：收敛后的非橙领数据集 U_{ncl}，橙领数据集 U_{cl}

```
1. DO
2.   依据式(3)计算类 U_ncl，U_cl 的关键字的 tfidf 值，
     将所有类表示为关键字和 tfidf 值二元组集合；
3.   FOR EACH 非橙领数据集 U_ncl 和橙领数据集 U_cl
     中的一条博文 t DO
4.     依据式(4)计算 p_{t,ncl}，p_{t,cl}；
5.     IF p_{t,cl}>p_{t,ncl} THEN
6.       t 归属于 U_cl 类，并且从原先类中删除；
7.     ELSE
8.       将博文 t 归属于 U_ncl 类，并且从原先类中删除；
9.     END IF
10.  END FOR
11. UNTIL 达到收敛；
```

在上述算法中，当非橙领数据集 U_{ncl} 和橙领数据集 U_{cl} 中的数据不再改变或不再重新分配时，则认为达到收敛状态.

3.3.2 橙领识别算法

判断发表微博的某个用户是橙领、非橙领、还是半橙领，一种有效的方法是依据在最近一段时间内该用户所发的购物博文的百分比，即推荐产品的微博占他所发微博总数的百分比来判断.

通过对上述划分后的数据集进行统计分析，可以得出橙领集合中橙领购物博文最低比例 β_1，非橙领购物博文最高比例 β_2. 因此若用户的购物博文占总博文百分比在区间 $[0,\beta_2)$ 上时，判断当前用户为非橙领，在区间 $[\beta_2,\beta_1)$ 判断当前用户为半橙领，区间 $[\beta_1,1]$ 上判断当前用户为橙领. 下面是橙领识别算法的形式化描述.

算法 4. 橙领识别算法 $CLIdentify()$.

输入：要判断的微博用户 u

输出：依据购物博文百分比所在区间输出橙领、非橙领或半橙领标签

```
1. 初始化微博总数(count_wb)和购物博文总数(count_
```

shopping)为 0；
2. FOR u 发的每一篇博文 t DO {
3. 依据式(4)计算 $p_{t,ncl}$，$p_{t,cl}$；
4. IF $p_{t,cl} > p_{t,ncl}$ THEN { //该博文是购物博文
5. *count_wb*++；
6. *count_shopping*++；}
7. ELSE *count_wb*++； //该博文是非购物博文
8. }
9. *ratio*=*count_shopping*/*count_wb*； //计算博文百分比值
10. IF *ratio* is between $[\beta_1,1]$ THEN RETURN *clFlag*；
11. ELSE IF *ration* is between $[\beta_2,\beta_1)$ THEN RETURN *half-clFlag*；
12. ELSE RETURN *nclFlag*；

3.3.3　橙领定位向量化算法

利用橙领识别算法可以识别出橙领、非橙领以及介于两者之间的半橙领，但这还远远不够. 尤其对于橙领，我们还需要进一步知道他到底是哪个方面的橙领. 尽管每个用户在最初选择加入橙领这个新兴职业时，对自己也做了一个定位，这个定位是对要推荐产品的大致定位，例如定位于服装类、小饰品类、高端大气、清新脱俗范等等. 但是，这些都不是量化的定位. 本研究提出一种橙领向量化定位算法，对橙领进行定量的定位. 同时，本文中的定位是通过分析橙领每一条购物类的博文来完成的. 我们用一个特征向量来描述一个用户的橙领定位. 该定位算法实际上是一个分类算法.

定义 5.　橙领定位类别. 本文根据产品类别将橙领定位分成 m 类，因此橙领数据集可进一步细分为这 m 类数据集. 接着与橙领数据集和非橙领数据集划分类似，将这 m 类数据集中每类数据集都可根据博文作者名称划分成一个个用户子类，然后使用式(1)～(3)计算生成这 m 类数据集中所有子类的加权关键字集，从而将每一子类描述为一个加权关键字的二元组集合.

定义 6.　单条微博的定位特征向量 ***LFV***(Locating Feature Vector). 假设给定一条博文 t，它归属于第 1 类数据集的概率值记为 i_1，归属于第 2 类数据集的概率值记为 i_2，归属于第 m 类数据集的概率值记为 i_m，由此我们得到博文 t 归属于 m 类数据集的特征向量 $\boldsymbol{I}=(i_1,i_2,\cdots,i_m)$，向量的每一维计算公式为

$$i_j=\max\{p_{i,j}\mid u_i\in U_j\} \tag{5}$$

其中 $1\leq j\leq m$，$p_{i,j}$ 表示博文 t 归属于某用户类 u_i 的概率，该概率的计算参考式(4).

对于一个橙领用户 u，我们也利用定位特征向量来描述其具体的类别. 橙领定位算法的核心思想就是对博文的向量集进行聚类，选取最大簇的中心向量作为橙领的定位向量. 簇越大包含的数据就越多，则该簇相对比其他簇更能反映输入数据的特征. 本文采用了 K-Means 聚类算法. 首先事先选取 k 个中心点，在博文的向量集上进行中心聚类(本文中 $k=8$)，将其他博文向量归类到这 k 个类中与其距离中最近的那个类，然后再依次计算这 k 个类中心点，进行更新，这样反复循环迭代直到这 k 个中心点趋于稳定. 算法的形式化描述如下：

算法5.　橙领定位向量化算法*CLFeatureLocate*().
输入：要进行定位的微博用户 u，橙领数据集细分的 m 类数据集
输出：u 的定位特征向量 $\boldsymbol{I}=\langle i_1,i_2,\cdots,i_m\rangle$
1. 输入给定聚类节点数 K；
2. 设 $\boldsymbol{U_I}$ 为 m 维向量的集合，初始化为空集；
3. FOR EACH u 所发微博 t Do
4. 由式(4)计算 $p_{t,ncl}$，$p_{t,cl}$；
5. IF $p_{t,ci} > p_{t,ncl}$ THEN
6. 该博文是购物博文，将博文 t 作为输入，由式(5)生成 $\boldsymbol{I}=\langle i_1,i_2,\cdots,i_m\rangle$，并且将该向量加入向量集合 $\boldsymbol{U_I}$；
7. END IF
8. END FOR
9. 在向量集合 $\boldsymbol{U_I}$ 按给定数值 K 进行 K-Means 中心聚类；
10. 获取最大簇的中心向量，进行归一化，并返回.

其中第 6 步是对该用户的每一条购物博文进行定位特征向量化，从而得到关于该用户微博的 m 维向量集合，第 9 步是在该集合上进行 K-Means 中心聚类，选取最大簇的中心点作为该用户的定位.

3.4　面向用户的橙领推荐算法

用户在购物时往往愿意选择有影响力的橙领进行产品的浏览. 如何查找最能满足用户需求的橙领，并将他推荐给用户，也是社会化导购的一个重要问题之一. 本文研究了面向用户的橙领推荐算法 OCRA4U(Orange Collar Recommending Algorithm for User)，该算法在综合考虑了橙领在网络中的影响力，以及橙领的自身定位与用户需求之间的匹配程度两方面因素. 所以本研究首先构造了橙领网络，以分析得到橙领的影响力程度.

3.4.1　橙领网络影响力

首先给出橙领网络、橙领间网络、橙领子网的定义，然后在该定义上，基于独立级联模型计算橙领在橙领网络中的影响力.

定义 7.　橙领网络. 是指由橙领、粉丝、橙领和

其粉丝间的关注和被关注关系，以及橙领间的好友关系构成的网络. 本文只考虑橙领间联系、橙领与粉丝间联系，而忽略与本文研究相关性不大的关联关系，比如粉丝间的联系等等. 橙领网络如图 2 所示，灰色结点表示橙领，白色结点表示其粉丝.

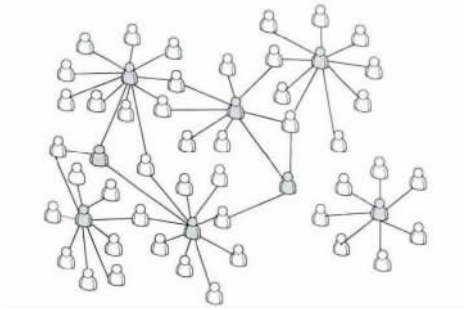


图 2 橙领网络

定义 8. 橙领间网络. 是指只由橙领以及橙领之间的关系组成的网络，关系可以是直接好友联系，也可以是有共同好友的间接联系. 橙领间网络可以由橙领网络演化而来，具体来说，橙领间网络是通过从橙领网络中提取橙领结点和这些结点之间的边(若存在)而得到的，若两个橙领之间不存在边，但是他们连接了共同的某个粉丝节点，则在这两个橙领间添加一条边. 这样就初步得到橙领间网络，网络中可能存在孤立的点，由于其代表的橙领影响不到其他橙领，且不受其他橙领影响，因此剔除该孤立点，获得最终橙领间网络. 图 3 为由橙领网络抽取出的橙领间网络.

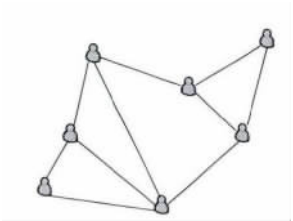


图 3 橙领间网络

定义 9. 橙领子网络. 该网是橙领网络的子网，是单个橙领与其粉丝组成的网络，如图 4 所示.

图 4 橙领子网络

本文分层次分析橙领在橙领网络中的影响力. 首先研究橙领在橙领子网中的影响力，接着研究橙领在橙领间网络中的影响力，最后给出橙领在橙领网络中的影响力.

(1) 橙领在橙领子网中的影响力

橙领子网是以橙领为中心的小团体，且以橙领为中心呈星型结构发散，是一个天然聚类结构，该结构中所有粉丝都能在同一时刻收到橙领所发的微博. 因为橙领子网络结构比较简单，影响力可以取值为橙领每条微博能够影响粉丝数的平均值，即平均每条微博被转发的次数.

(2) 橙领在橙领间网络中的影响力

本文使用独立级联模型计算橙领在橙领间网络中的影响力. 独立级联模型描述如下：社交网络中每个用户只有两种状态，激活状态和未激活状态，不存在中间态. 用户只有处于激活态时才能影响与他有联系的其他未激活的点，且如果被激活后将处于激活态不再被激活[13]. 每次激活过程作为一次影响力传播，且该激活过程作为独立事件. 一个用户历史激活另外一个用户的次数越多，则该用户激活另一用户的概率越大. 因此用户 a 激活另外一个用户 b 的先验概率可以由历史激活次数推断，计算公式为

$$P_{a,b}=1-(1-P_0)^n \tag{6}$$

其中，$P_{a,b}$ 表示用户 a 激活用户 b 的概率，n 表示 a 激活用户 b 的历史次数. P_0 是介于 0 到 1 的一个基础概率值，一般取 b 入度的倒数. 可以看到 n 越大，即历史激活次数越多，概率值越大，且始终小于 1.

用户的网络影响力具有传递的性质，用户 u 能以概率 $P_{u,v}$ 激活用户 v，则用户 u 以概率 $P_{u,v}$ 通过 v 扩大自己的影响力. 在网络中用户 u 首先处于激活状态，经过一定时间 t，网络中被激活的用户趋于定值，该值越大，用户 u 的影响力越大. 因此，本文中依据该模型计算用户的网络影响力公式为

$$Inf_u=Inf_0+\sum_{v\in N(u)}P_{u,v}\times Inf_v \tag{7}$$

其中 Inf_u 和 Inf_v 表示用户 u 和用户 v 的社会影响力，$N(u)$ 表示能被用户 u 影响的用户集合，$P_{u,v}$ 表示用户 u 激活用户 v 的概率，Inf_0 表示影响力的基值，考虑到社交网络中不可能存在没有影响力的个体，所以能保证影响力大于零.

(3) 橙领在橙领网络中的影响力

使用独立级联模型可以计算出橙领间网络中的橙领影响力，同时橙领在橙领子网中的影响力为橙领博文平均转发和评论百分比 Inf_p. 假设现在只考虑橙领用户 u 和橙领用户 v 组成的网络，他们之间有联系，橙领用户 v 能以概率 $P_{u,v}$ 激活 v，那么可以利用下列公式来计算 u 的影响力：

$$Inf_u=Inf_0+P_{u,v}\times Inf_v \tag{8}$$

现在考虑 v 的粉丝，假设用户 v 在他的子网中

的影响力为 Inf_{pv}，那么考虑用户 u 是否能够通过 v 影响 v 子网中的用户，答案是肯定的. 用户 u 能通过概率 $P_{u,v}$ 激活 v，那么用户就能以概率 $P_{u,v}\times Inf_{pv}$ 影响到 b 的子网. 因此，橙领 u 在整个网络中的影响力计算公式为

$$Inf_u = Inf_0 + Inf_{pu} + \sum_{v\in N(u)} P_{u,v}\times(Inf_v + Inf_{pv}) \quad (9)$$

其中，Inf_u 和 Inf_v 表示用户 u 和 v 的社会影响力，Inf_{pu} 和 Inf_{pv} 分别表示用户 u 和 v 在其子网中的影响力，$P_{u,v}$ 表示用户 u 激活 v 的概率.

下面给出橙领在橙领网络中影响力计算算法.

算法 6.　橙领在网络中的影响力计算算法 $CLInfluence()$.

输入：(1) 橙领间网络的图 G；

(2) 橙领所在橙领子网中他的博文的转发/评论(只有粉丝参与)的百分比集合 Inf_p. 其中 Inf_{pu} 表示橙领 u 在子网中的他的本文被转发和评论的百分比

输出：输出橙领 u 在网络中的影响力 Inf_u

1. 初始化橙领间激活次数集合 C 和 Inf_0；
2. u 标记为激活状态；
3. 设 Inf_u 为橙领 u 的影响力初始化为 Inf_0+Inf_{pu}；
4. 设 P 为节点 u 激活其他节点的概率集合，初始化为空集；
5. WHILE 图 G 中节存在节点 v 处于未激活状态且节点 u 和 v 之间存在最短路径 L DO
6. IF 路径 L 上只有 v 未激活 THEN
7. 路径 L 上离 v 最近的节点，记作 w，从集合 C 查询 w 激活 v 的次数 n，依据式(6)计算 w 激活 v 的概率 $P_{w,v}$；
8. 从集合 P 中查询节点 u 激活 w 的概率 $P_{u,w}$(若 u 和 w 是同一节点)则该值为 1；
9. 节点 u 激活 v 概率为 $P_{u,v}=P_{u,w}\times P_{w,v}$，并且将 $P_{u,v}$ 加入到集合 P 中；
10. $Inf_u = Inf_u + P_{u,v}(Inf_v + Inf_{pv})$；
11. 将节点标记 w 为已激活；
12. END IF
13. END WHILE

3.4.2　橙领的自身定位和用户需求的匹配程度

计算橙领的自身定位和用户需求的匹配程度分为两步. 第 1 步，用户需求向量化即将用户需求(关于用户需求产品的描述)按照与 m 类数据集的匹配值生成对应的向量 $\boldsymbol{I}_{\text{product}}=(i_1,i_2,\cdots,i_m)$. 第 2 步，计算橙领定位向量和用户需求向量的相似度，本文使用余弦相似度公式计算：

$$Sim(u,v)=\frac{\boldsymbol{I}_u\cdot\boldsymbol{I}_v}{|\boldsymbol{I}_u|\times|\boldsymbol{I}_v|} \quad (10)$$

其中，u 表示橙领，v 表示橙领，$\boldsymbol{I}_u$ 表示用户需求向量，$\boldsymbol{I}_v$ 表示橙领定位向量.

3.4.3　面向用户的橙领推荐算法

针对用户需求查找最能满足用户需求的橙领，必须要考虑橙领定位与用户需求的匹配和橙领在网络中的影响力. 前者保证橙领给出的推荐能够满足用户需求，后者保证橙领的可靠程度. 本文把橙领网络影响力值与橙领的自身地位和用户需求的匹配度值的乘积作为推荐置信值 T_u，如式(11)：

$$T_u = Inf_u\times\left(\frac{\boldsymbol{I}_u\cdot\boldsymbol{I}_{\text{product}}}{\|\boldsymbol{I}_u\|\times\|\boldsymbol{I}_{\text{product}}\|}\right) \quad (11)$$

其中 $\boldsymbol{I}_{\text{product}}$ 表示用户需求对应的定位特征向量. 下面给出橙领推荐算法的形式化描述：

算法 7.　面向用户的橙领推荐算法 OCRA4U.

输入：(1) 由橙领数据集细分的 m 类数据集；

(2) 用户 u 的需求描述 $Infomation$

输出：排序后的橙领列表 $clList$；

1. 设 T 为推荐橙领置信值的集合，初始化为空集；
2. 用户需求的描述 $Information$ 由式(5)计算生成向量 $\boldsymbol{I}_{\text{product}}$；
3. FOR EACH 橙领 u DO
4. 调用 $CLFeatureLocate()$ 算法得到橙领定位向量 $\boldsymbol{I}_u$；
5. 调用 $CLInfluence()$ 算法得到橙领网络影响力 Inf_u；
6. 依据式(11)计算推荐该橙领的推荐置信值 T_u，并将该值加入集合 T；
7. END FOR
8. 依据橙领在集合 T 中的值进行排序；
9. 将排在前面的 k 个橙领返回.

3.5　面向商家的橙领推荐算法

最终商家为了销售产品，也必须选择最符合其要求的橙领进行产品的推广，所以提供向商家推荐橙领的功能也很有必要. 本文研究了面向商家的橙领推荐算法 OCRA4S(Orange Collar Recommending Algorithm for Shop)，该算法综合考虑了橙领历史推荐产品与要推荐产品的匹配度和橙领在社会网络中的影响力两个因素.

计算橙领历史推荐产品与商家拟希望推荐的产品的匹配度借鉴了协同过滤算法和基于内容的推荐算法的思想. 基于产品的协同过滤考虑了用户的购买历史，计算产品的相似性是基于用户对产品的偏好. 而基于内容推荐中的产品相似性是基于产品自身属性来得到的. 我们了解到橙领所推荐的产品一般比较集中在某个范围内，例如做文具类的橙领推荐衣服概率不大，所以可以结合基于产品的协同过滤和基于内容的推荐，将要推荐产品自身属性与橙领已推荐过产品的相似度的期望作为向橙领推荐该产品的依据. 可以计算要推荐的产品和橙领已经推荐过的每个产品的匹配值：首先生成要推荐的产品

和橙领已经推荐过的产品在 m 类数据集上的对应向量，即由式(5)生成对应向量 $\boldsymbol{I}=(i_1,i_2,\cdots,i_m)$；然后通过向量间的余弦值来量化两个向量的相似度. 要推荐的产品与橙领推荐过的产品的相似度的期望(橙领历史推荐产品的概率都为 $1/N$，N 为历史推荐产品的数量)，作为该产品的推荐值. 计算公式如下：

$$Sim(u,pro)=\frac{\sum\limits_{item\_i\in L(u)}\frac{I_{pro}\cdot I_{item\_i}}{|I_{pro}|\times|I_{item\_i}|}}{N} \tag{12}$$

其中 pro 为要推荐的产品，u 为某个橙领，$L(u)$为橙领 u 已经推荐产品的集合，$item\_i$ 表示 u 已经推荐的第 i 个产品，N 为橙领 u 已经推荐产品的总数. 下面给出产品与橙领历史推荐相似度量化算法.

本算法考虑橙领历史推荐过的产品和橙领在社交网络中的影响力. 历史推荐产品与要推荐产品的相似度的期望越高，橙领接受该产品概率越大，同时橙领也越能把握该产品的推荐，因此橙领子网中橙领粉丝接受该产品的概率就越大. 橙领在社会中的影响力反映了橙领子网在某方面喜好的集中度，影响力越高，集中度越高，用户接受该产品可能性越大. 因此，向某个橙领及其子网推荐当前产品可行性值可以表示为当前产品与该橙领历史推荐相似度值和该橙领的网络影响力值的乘积，由以下公式计算：

$$T(u,pro)=Inf_u\times sim(u,pro) \tag{13}$$

依据该置信值对橙领进行排序，向排名在前的橙领或橙领的粉丝推荐产品，或者说是向该产品的商家推荐排名在前的橙领. 算法形式化描述如下：

算法 8. 面向商家的橙领推荐算法 OCRA4S.

输入：(1) 橙领集合；

(2) 要推荐的产品 pro

输出：橙领排序列表

1. 设集合 T 为橙领推荐产品 pro 的可行性值集合，初始化为空集；
2. FOR EACH 橙领 u DO
3. 通过算法 6 计算橙领在网络中影响力；
4. 依据式(12)计算产品 pro 与历史推荐产品的相似度；
5. 由式(13)计算让橙领 u 推荐该产品的可行性值，并将该值加入集合 T；
6. END FOR
7. 依据集合 T 对橙领进行排序；
8. 将排在前面的 k 个橙领返回.

依据该算法输出的橙领推荐列表，可以使商家有针对性地选择适合推广其产品的橙领，从而实现高效的产品推广效果.

4 实 验

本节对提出的 3 个算法，即橙领定位算法、面向终端用户的橙领推荐算法和面向商家的橙领推荐算法做实验进行验证，其中将面向终端用户的橙领推荐算法和面向商家的橙领推荐算法的实验统称为橙领推荐实验. 本文所有实验所采用的机器设备为 PC 机，内存 2.00GB，处理器为 Intel(R) Core(TM) DuoCpu T6400@2.00GHz，操作系统为 32 位 Windows 7 旗舰版. 算法代码全部由 Java 实现. 下面详细介绍具体的实验设计和结果分析.

4.1 实验数据

实验采用两个数据集，分别是实时的新浪微博数据集和仿真的 DBLP 数据集.

新浪微博数据集是通过新浪微博开发者平台提供的接口爬取得到，总共抓取 1000 位用户 2015 年 10 月 1 日至 2015 年 12 月 30 日之间所发的微博数据和相关属性数据. 经过对用户有关购物的微博的分析，我们把购物博文分为以下 6 类：衣裤类、裙子类、鞋子类、电子产品类、洗护类和其他类. 原始博文数据通过橙领定位算法可以被定位为上述 6 类的某个定位向量，然后通过橙领推荐算法针对用户、商家的需求，得到橙领的推荐排序，完成橙领的推荐.

另外，我们将 DBLP 数据集作为仿真数据集对提出的 SSOCRM 方法进行了实验验证. DBLP(DataBase systems and Logic Programming)[14] 数据集是计算机领域内以作者为核心的一个计算机类各个研究领域英文文献的集成数据库系统，数据集大小 1.17GB. 为了实验的需要，本研究只提取了 DBLP 数据集中有关知识发现、用户推荐、用户画像和链接预测这 4 个领域的文章来进行相关的实验. 对于每一位作者主要提取包含以下的信息：其发表文章的题目(title)、作者(authors)以及所属的领域(field)等. 在该数据集上使用 SSOCRM 方法，输入上述 4 个领域中某个领域的研究需求，得到学者的推荐排序，根据该排序，获得针对该研究需求的杰出学者，找到合适的合作学者. 学者的推荐和橙领的推荐方法本质上都是基于发布数据进行这些发布者的推荐和发现，所以用 DBLP 数据来验证 SSOCRM 的橙领推荐方法也具有借鉴意义和可行性.

4.2 评价指标

实验中采用查准率、MAP 和 $NDCG$ 指标评价推荐结果的质量. 由于在橙领推荐实验中，参与进行排序的都是橙领，所以计算的查全率始终为 1，所以没有采用查全率指标. 首先采用查准率来衡量实验结果的准确性. 但由于查准率无法衡量检索结果的优先顺序，所以还采用 MAP 和 $NDCG$ 指标来衡量实验排序结果与标准结果之间的匹配度，它们都是

有序检索结果的评价指标. *MAP*(Mean Average Precision)表示返回结果中在每篇相关文档位置上的正确率的平均值,针对前 k 个检索结果,且只针对二值情况,即相关为 1,不相关为 0. 假设 $k=6$,则标准排序结果的前 6 个认为是相关的,设为 1,其他为 0. *NDCG*(Normalized Discounted Cumulative Gain),针对非二值的情况,同样基于前 k 个检索结果进行计算,计算前需要给标准排序的前 k 个结果赋值.

4.3　实验设计

我们将本文提出的 SSOCRM 方法和朴素贝叶斯方法[12]以及 LDA 方法[13]分别在上述数据集上进行了实验比较. 对于橙领的识别定位以及推荐等,本质上也是一种用户画像问题. 朴素贝叶斯方法和 LDA 方法都曾被用于解决用户画像问题,以这两种方法作为对比方法,可以衡量本文提出的 SSOCRM 方法的可行性及有效性.

本研究分别针对橙领定位算法和橙领推荐设计了以下相关实验.

4.3.1　橙领定位算法实验

橙领定位算法由数据集划分算法、橙领识别算法、橙领向量化定位算法三部分组成. 数据集划分算法实质是数据的预处理过程,通过迭代方法划分为橙领数据集和非橙领数据集,数据集的划分会随着输入的迭代次数的增加而收敛. 本实验中,在新浪微博数据集上,橙领数据集与非橙领数据集在第 9 次迭代后收敛. 在 DBLP 数据集上,数据集在第 7 次迭代后收敛. 在此基础上进行橙领识别算法实验和橙领定位实验,以论证算法的准确性、可用性. 具体的实验设计如下:

实验 1. 橙领识别算法实验.

本实验目的是获取用于判定橙领、半橙领、非橙领的系数 β_1 和 β_2,并验证获取的系数是否能够正确地判断橙领. 实验数据集采用新浪微博数据集. 首先要使用训练集训练出判定系数 β_1 和 β_2. 训练集是由随机抽取的 100 位用户组成,人工阅读这 100 位用户的微博数据,按照常识判断他们是否为橙领. 然后从其余用户中抽取 100 位用户作为测试数据,每 10 位一组,运用系数 β_1 和 β_2 判定这些用户的类别. 再人工地判断这 100 位用户的类别作为标准答案,与实验结果对比,计算准确率. 橙领识别算法使用系数 β_1 和 β_2 对所有用户的类别进行了识别,为后续橙领定位以及推荐提供数据基础.

实验 2. 橙领定位实验.

本实验目的是验证橙领定位算法的可行性. 实验步骤分为以下 3 步:首先,根据橙领数据集划分算法的原理将实验 1 中生成的橙领数据集划分为 m 大类. 对于新浪微博数据集,将数据分为 $m=6$ 大类,分别是衣裤类、裙子类、鞋子类、电子产品类、洗护类以及其他类(包括饰品、包等);对于仿真的 DBLP 数据集,将数据分为 $m=4$ 大类,分别是:知识发现类、用户推荐类、用户画像类和链接预测类. 每大类中又含有若干用户小类. 然后,生成这 m 类数据集的关键字集及与关键字对应的 $tfidf$ 值集. 接着,对所有用户使用橙领定位向量化算法,中间得到每个用户每条购物博文(或发表论文)的定位向量,经 K-means 聚类,最终得到每位用户的定位向量. 最后,为了证明橙领定位的准确性,我们将抽取出 100 位橙领,平均分成 10 组,统计每一位橙领所有的购物微博中各类微博(或每一位作者所有论文中各类论文)的比例,构成比例向量,计算该向量与当前用户的定位向量的余弦相似度. 相似度越大表示定位越准确.

4.3.2　橙领推荐实验

该实验的目的将本文提出的面向终端用户的橙领推荐算法和面向商家的橙领推荐算法与基于 LDA 方法以及朴素贝叶斯方法的推荐算法进行对比,验证本研究提出的橙领推荐算法的可行性与有效性. 由于用户需求与商家需求都是对于商品的需求,所以实验中不区分这两种需求,统一两种橙领推荐算法的需求,称为商品需求. 调查周围人群发现,对于一位有购买需求的用户,返回给其的橙领无需太多,因为用户没有精力一一查看,并且为了减少计算量更有效地实现橙领的推荐,我们分别选取较合理的 k 值:$k=6$、$k=9$ 和 $k=12$.

依次在新浪微博数据集和仿真的 DBLP 数据集上进行上述 4 种橙领推荐方法的实验,计算在输出橙领个数 k 分别为 6、9 以及 12 时的各个评价指标值. 下面的 m 表示数据大类个数,在新浪微博数据集中 $m=6$,在 DBLP 数据中 $m=4$.

(1) 面向终端用户的橙领推荐算法 OCRA4U 的实验

首先分别输入不同类别的商品需求(对于微博数据,分别为衣裤类、裙子类、鞋子类、电子产品类、洗护类、其他类;对于 DBLP 数据,分别为知识发现类、用户推荐类、用户画像类和链接预测类),然后针对每类需求使用面向终端用户的橙领推荐算法得到该类上橙领排序的实验结果,将实验结果与标准答案对比,在只向用户推荐排名前 6、9、12 位橙领的条

件下，计算每类上橙领排序实验结果的查准率、MAP 值以及 $NDCG$ 值，最后进行结果的分析讨论. 对于 DBLP 数据集，关于影响力的计算方面，将橙领间的朋友关系定义为作者之间的合作关系，作者没有粉丝，所以影响力就是橙领间的影响力，即作者间影响力.

实验进行的具体步骤为：首先，获取标准答案. 调查了身边 20 位用户，对于每一类，依据输入的商品需求以及橙领的基本信息（朋友数、粉丝数等），记录他们对橙领的打分，橙领的分值越大表示用户认为该橙领越能给用户中肯的建议. 分值由 1 到 10，10 表示该橙领最能给用户中肯的建议，1 表示该橙领能给用户中肯的建议能力最低. 统计调查结果算出橙领的得分均值，并依据该得分从大到小对橙领进行排序，将排序得出的位次结果作为标准答案. 接着，计算橙领在网络中的影响力值，然后分别输入每类商品需求，依据 OCRA4U 算法给出橙领推荐的排序结果，即实验结果. 最后，对比实验输出的橙领排序和标准排序来评价算法的好坏. 针对不同的 k 值，分别计算前 k 个实验排序结果与前 k 个标准排序结果对比得到的查准率. 另外，还采用 MAP 和 $NDCG$ 指标来衡量橙领推荐实验排序结果与理想标准结果之间的匹配度. 对于两个排序都做这样转化：例如当 $k=6$ 时，将排序中的橙领依据排序号给以对应的分值（分值范围 1 到 6），排序第一的给以 6 分，排第二的给以 5 分，依次进行，比较标准排序结果（作为标准答案）与 OCRA4U 算法结果，计算每一类的 MAP 值与 $NDCG$ 值.

（2）面向商家的橙领推荐算法 OCRA4S 的实验

针对每一类的商品需求，使用面向商家的橙领推荐算法得到橙领排序的实验结果，分别将实验结果与标准答案对比，在只向用户推荐排名前 6、9 以及 12 位橙领的条件下，计算每类上橙领排序实验结果的查准率、MAP 值以及 $NDCG$ 值.

（3）基于 LDA 的推荐算法的实验

使用 LDA 方法计算每个橙领的主题向量（共 m 个主题（$m=6$ 或 4），与 m 大类数据一一对应），输入每类商品需求，得到商品需求与主题向量的余弦相似度，将相似度排序得到实验结果，与标准答案对比，分别选取 $k=6$、9 以及 12，计算实验结果的查准率、MAP 值与 $NDCG$ 值.

（4）基于朴素贝叶斯 NB 的推荐算法的实验

针对每个橙领，将其所有购物博文（对于 DBLP，是将每个作者发表的所有论文的题目）构成一个文档，使用朴素贝叶斯方法计算该文档分属于 m 大类数据的分类概率，将这 m 个概率构成分类概率向量并归一化. 同理对输入的每类需求求出其归一化的分类概率向量，然后计算橙领概率向量与需求概率向量的余弦相似度，将相似度排序得到实验结果，与标准答案对比，分别选取 $k=6$、9 和 12，计算实验结果的查准率、MAP 值与 $NDCG$ 值.

4.3.3 社会化数据影响实验

在本文所提的 SSOCRM 方法中，结合使用了橙领的微博内容数据及其社会关系两种社会化数据，最终实现橙领的推荐. 为了衡量这两种社会化数据的必要性，设计了社会化数据影响实验，目的在于验证社会化信息对于推荐结果的影响效率. 实验基本过程是：在两个实验数据集上，分别对只使用微博内容数据（或论文题目数据）、只使用社会关系、同时使用两种社会化数据 3 种情况来进行面向用户的橙领推荐算法和面向商家的橙领推荐算法的实验，并进行结果的对比分析. 分别选取 $k=6$、9 以及 12，计算各类实验结果的平均查准率、平均 MAP 值与平均 $NDCG$ 值.

4.4 实验结果与分析

下面分别给出各实验的结果并进行讨论分析.

4.4.1 橙领定位算法实验结果与分析

实验 1. 橙领识别算法实验.

在训练集中，非橙领中购物博文比例最高为 0.09. 在橙领中购物博文比例最低为 0.82. 因此，本实验中定义：如果用户所发博文中购物博文的比例值在区间 $[0, 0.09)$，该用户为非橙领；如果在区间 $[0.09, 0.82)$ 中，那么该用户为半橙领；如果在区间 $[0.82, 1]$ 中，那么该用户为橙领.

在测试集上验证了两系数的正确性，每组上的准确率如图 5 所示. 横轴是组别，纵轴是准确率，平均准确率为 0.84，这表明橙领识别算法可行性较高，具有一定实际应用价值. 实验结果也表明数据集划分算法可以较好地划分数据集，并进行归类，同时剔除无用数据. 目前实验结果误差主要来源以下三

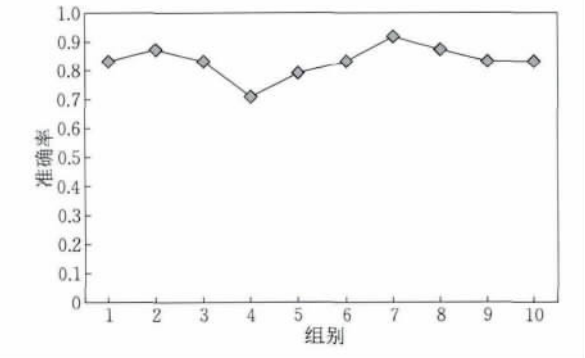


图 5 10 组用户的用户识别准确率

方面：首先，算法 3 中最初手动生成的非橙领数据集 U_{ncl} 和橙领数据集 U_{cl} 可能存在数据较片面、不全面问题，导致最终生成的橙领和非橙领数据集含有错误数据；其次，选取出的用于计算橙领和非橙领识别系数的橙领和非橙领代表性不是很强；另外，由于本实验是计算购物博文占总博文数的百分比，所以用户微博数目过少也会影响结果.

实验 2. 橙领定位实验.

在新浪数据集上的橙领定位算法对 10 组橙领的定位结果如图 6 所示. 在图中横轴为组别数，纵轴为橙领定位向量与购物博文类别比例向量相似度，10 组用户的相似度在 0.6 上下浮动，均值为 0.63. 在 DBLP 数据集上的对作者的定位实验结果如图 7 所示. 在图中横轴为组别数，纵轴为作者定位向量与论文类别比例向量相似度，10 组作者的相似度在 0.7 上下浮动，均值为 0.72. 结果证明橙领定位算法可以在一定程度上衡量橙领所发购物博文的类别比例，能够较真实地对橙领进行定位向量化.

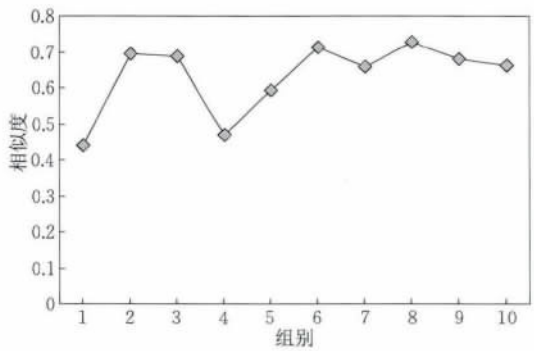


图 6　新浪数据集橙领定位向量与购物博文类比例向量相似度

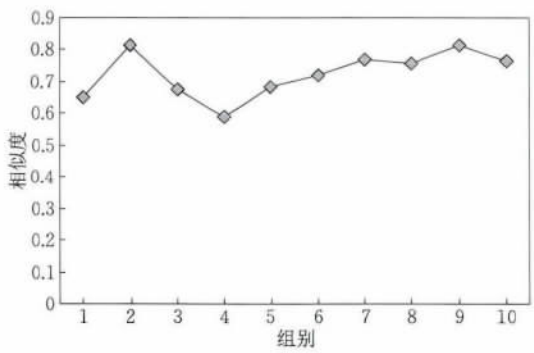


图 7　DBLP 数据集作者定位向量与论文类别比例向量相似度

4.4.2　橙领推荐方法的对比实验结果与分析

本实验将本文提出的 OCRA4U 和 OCRA4S 算法分别在两个数据集上的实验结果和基于 LDA 的推荐方法以及基于朴素贝叶斯的推荐方法的实验结果进行了比较，评价指标分别是查准率、*MAP* 值以及 *NDCG* 值.

(1) 新浪数据集

表 1、表 2 以及表 3 分别展现 4 种橙领推荐方法在不同 k 值情况下、针对 6 类不同商品需求得到的橙领推荐排序与标准答案对比的平均查准率、平均 *MAP* 值以及平均 *NDCG* 值，分别对应图 8 中的(a)、(b)以及(c)子图. 图 9 展现了 4 种橙领推荐方法在不同 k 值情况下、分别针对各类商品需求得到的橙领推荐排序的查准率、*MAP* 值和 *NDCG* 值.

表 1　新浪数据集上 4 种方法在不同 k 值下的平均查准率

查准率	LDA	NB	OCRA4S	OCRA4U
$k=6$	0.33	0.36	0.61	0.50
$k=9$	0.33	0.41	0.62	0.56
$k=12$	0.43	0.46	0.71	0.72

表 2　新浪数据集上 4 种方法在不同 k 值下的平均 *MAP* 值

MAP	LDA	NB	OCRA4S	OCRA4U
$k=6$	0.18	0.23	0.46	0.38
$k=9$	0.21	0.32	0.47	0.41
$k=12$	0.24	0.38	0.63	0.62

表 3　新浪数据集上 4 种方法在不同 k 值下的平均 *NDCG* 值

NDCG	LDA	NB	OCRA4S	OCRA4U
$k=6$	0.23	0.33	0.52	0.63
$k=9$	0.27	0.41	0.57	0.67
$k=12$	0.32	0.41	0.65	0.70

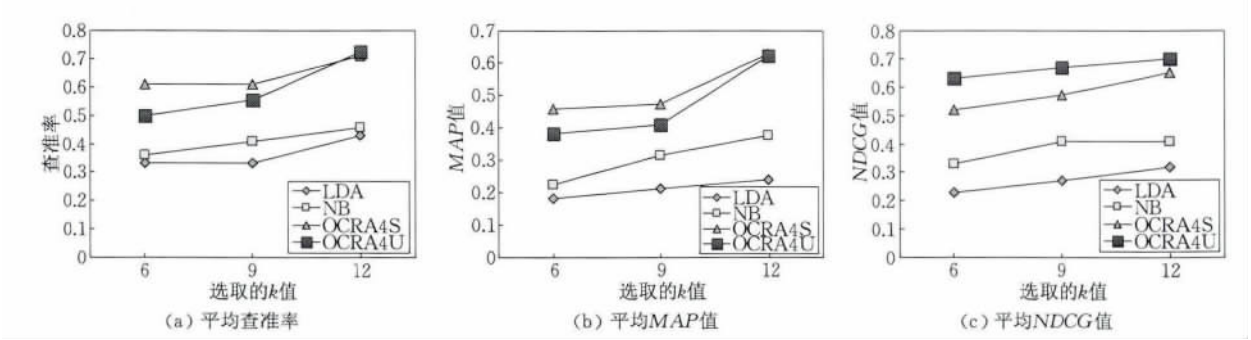


图 8　新浪微博数据集上数据集上 4 种方法在不同 k 值下的各评价指标平均值对比

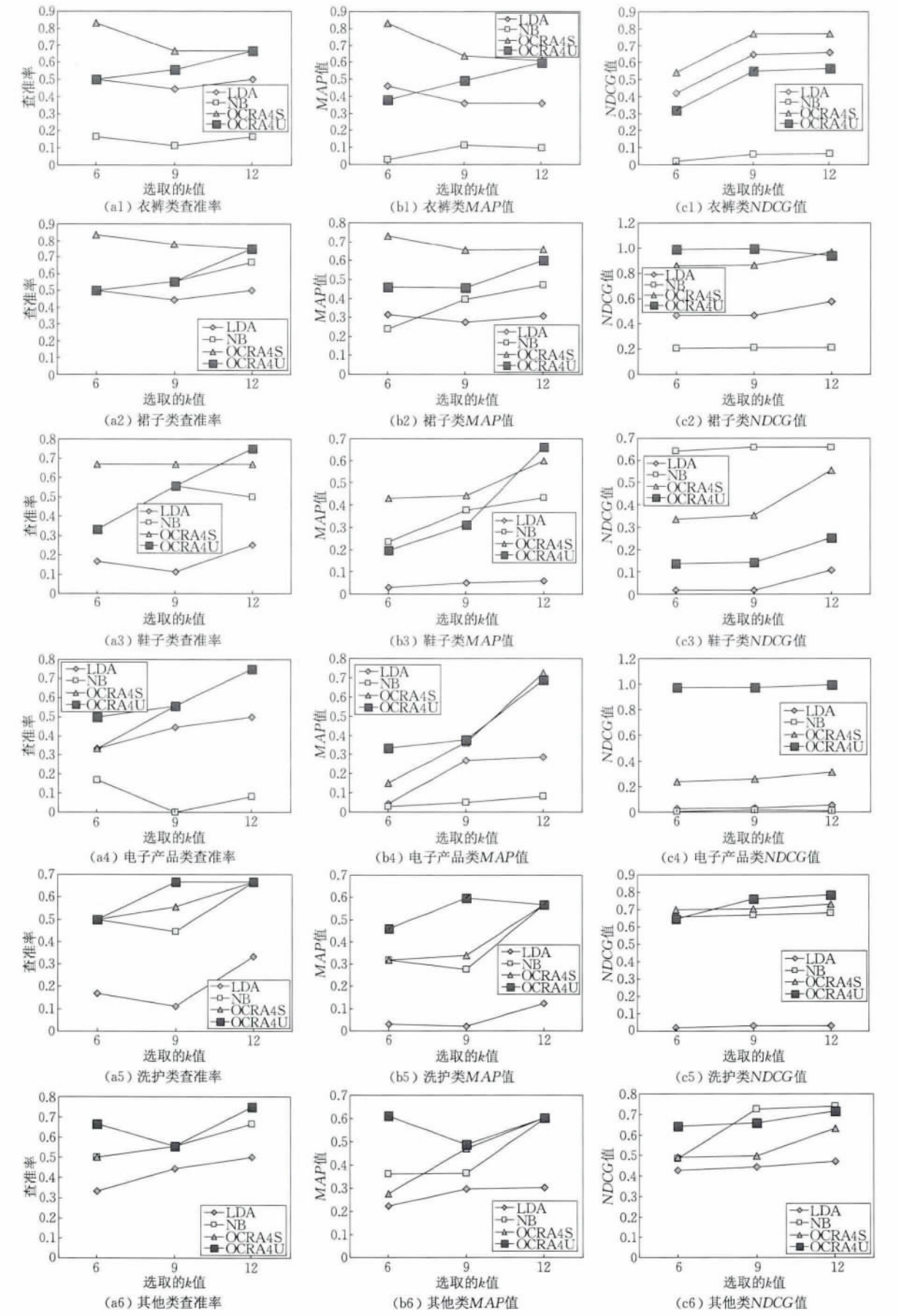


图 9 新浪微博数据集上 4 种方法在不同 k 值下的各类的不同评价指标对比

由图 8 可知，(1)针对同一种方法，随着 k 值的增大，平均查准率、平均 MAP 值、平均 $NDCG$ 值均大致呈递增趋势，这是由于因为 k 越大(即有效的橙领越多)，推荐难度就越小，推荐的命中概率就更大；(2)在相同 k 值下，平均查准率、平均 MAP 值、平均 $NDCG$ 大致有如下规律：面向用户的橙领推荐算法 OCRA4U 以及面向商家的橙领推荐算法 OCRA4S 效果优于朴素贝叶斯方法和 LDA 方法，例如，当 $k=12$ 时，在 $NDCG$ 值上，OCRA4U、OCRA4S、朴素贝叶斯以及 LDA 分别为 0.70、0.65、0.41 和 0.32，依次减小. 分析原因，朴素贝叶斯方法和 LDA 方法只考虑橙领发布的微博信息，而均没有考虑橙领的社会关系因素，没有衡量橙领在人群中的影响力程度，这会导致对橙领的认知较片面，进而推荐的橙领准确性不高；接着，LDA 方法的推荐水平最低是由于 LDA 方法可以将数据抽取为固定个数的主题，本数据集上选取主题个数为 6，由此得到的每一个主题大致与橙领数据集中的 6 大类一一对应，但是每一主题都会包含部分其他主题的内容，主题划分较不严格，由此得到的橙领主题向量不准确，最终导致推荐效果不理想.

由图 9 可知，在相同 k 值下，针对各个类别的商品需求得到的橙领排序的查准率、MAP 值、$NDCG$ 大致有如下规律：OCRA4U 以及 OCRA4S 效果优于朴素贝叶斯方法和 LDA 方法，例如，在裙子类上，当 $k=12$ 时，在 MAP 值上，OCRA4S、OCRA4U、朴素贝叶斯以及 LDA 分别为 0.66、0.60、0.47 和 0.31，依次递减. 在查准率和 $NDCG$ 值上也有同样规律，证明 OCRA4U 以及 OCRA4S 算法返回的橙领序列更匹配于标准答案，优于朴素贝叶斯方法和 LDA 方法. 但是存在部分例外的情况，在衣裤类上，当 $k=6$ 时，LDA 方法的 MAP 值为 0.46，大于 OCRA4U 的 0.38；并且当 $k=6$、9 以及 12 时，LDA 方法的 $NDCG$ 值分别为 0.42、0.65 以及 0.66，分别优于 OCRA4U 算法的 0.32、0.55 以及 0.57. 分析发现，LDA 方法针对衣裤类需求返回的橙领排序中正确橙领的数量较少，少于 OCRA4U 算法，但是这些橙领的排序序号均较靠前，导致顺序的匹配度较高，优于 OCRA4U 算法. 在鞋子类上，当 $k=6$ 以及 9 时，朴素贝叶斯方法的查准率与 OCRA4U 算法一致；并且 MAP 值分别为 0.23 和 0.38，分别高于 OCRA4U 算法的 0.19 和 0.31；当 $k=6$、9 以及 12 时，朴素贝叶斯方法的 $NDCG$ 值分别为 0.64、0.66 和 0.66，分别高于 OCRA4U 算法的 0.14、0.14 和 0.25，高于 OCRA4S 算法的 0.34、0.35 和 0.55. 分析原因可知，朴素贝叶斯方法针对鞋子类需求返回的橙领排序中正确橙领的排序序号均较靠前，导致橙领序列的匹配度较高，优于 OCRA4U 和 OCRA4S 算法. 在其他类上，当 $k=6$ 和 9 时，朴素贝叶斯方法的查准率与 OCRA4S 算法一致；在 $k=6$ 时朴素贝叶斯方法的 MAP 值为 0.36，大于 OCRA4S 算法的 0.28；并且在当 $k=9$ 和 12 时，朴素贝叶斯方法的 $NDCG$ 值分别为 0.73 和 0.74，分别高于 OCRA4S 算法的 0.50 和 0.63，高于 OCRA4U 算法的 0.66 和 0.72，分析原因可知，当 $k=9$ 和 12 时，朴素贝叶斯方法针对其他类需求返回的橙领排序中正确橙领的数量较少，少于 OCRA4U 算法和 OCRA4S 算法，但这些正确橙领的排序序号均较靠前，相关性较大，导致橙领序列的匹配度较高.

实验结果误差主要来源以下两个方面：首先，标准答案的获取有较大主观性，本文采取的是 20 个人评分的平均结果，调查人数较少，后续工作中可以增加调查人数，使获取的标准答案更具有普遍性. 另外，在橙领定位实验中存在的误差也会给橙领推荐实验中的结果带来不良影响.

在新浪微博数据集中，$k=6$、9 以及 12 时均证明，OCRA4U 以及 OCRA4S 算法效果良好，具有可行性及有效性，可以针对商品需求返回给用户或商家较满意的橙领排序. 而其他两种方法也具有可行性，但是整体上效果劣于 OCRA4U 算法以及 OCRA4S 算法.

(2) DBLP 数据集

表 4、表 5 以及表 6 分别展现 4 种方法在不同 k 值情况下、针对 4 类商品需求的橙领推荐的平均查准率、平均 MAP 值以及平均 $NDCG$ 值，分别对应图 10 中的(a)、(b)以及(c)子图. 图 11 是 DBLP 数据

表 4　DBLP 数据集上 4 种方法在不同 k 值下的平均查准率

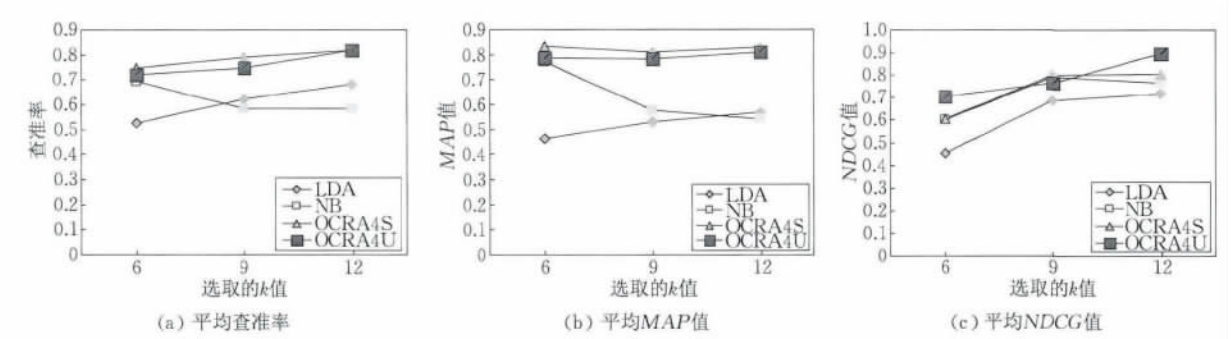
查准率	LDA	NB	OCRA4S	OCRA4U
$k=6$	0.53	0.69	0.75	0.72
$k=9$	0.63	0.58	0.79	0.75
$k=12$	0.68	0.58	0.82	0.82

表 5　DBLP 数据集上 4 种方法在不同 k 值下的平均 MAP 值

MAP	LDA	NB	OCRA4S	OCRA4U
$k=6$	0.46	0.77	0.84	0.79
$k=9$	0.53	0.58	0.81	0.78
$k=12$	0.57	0.54	0.83	0.81

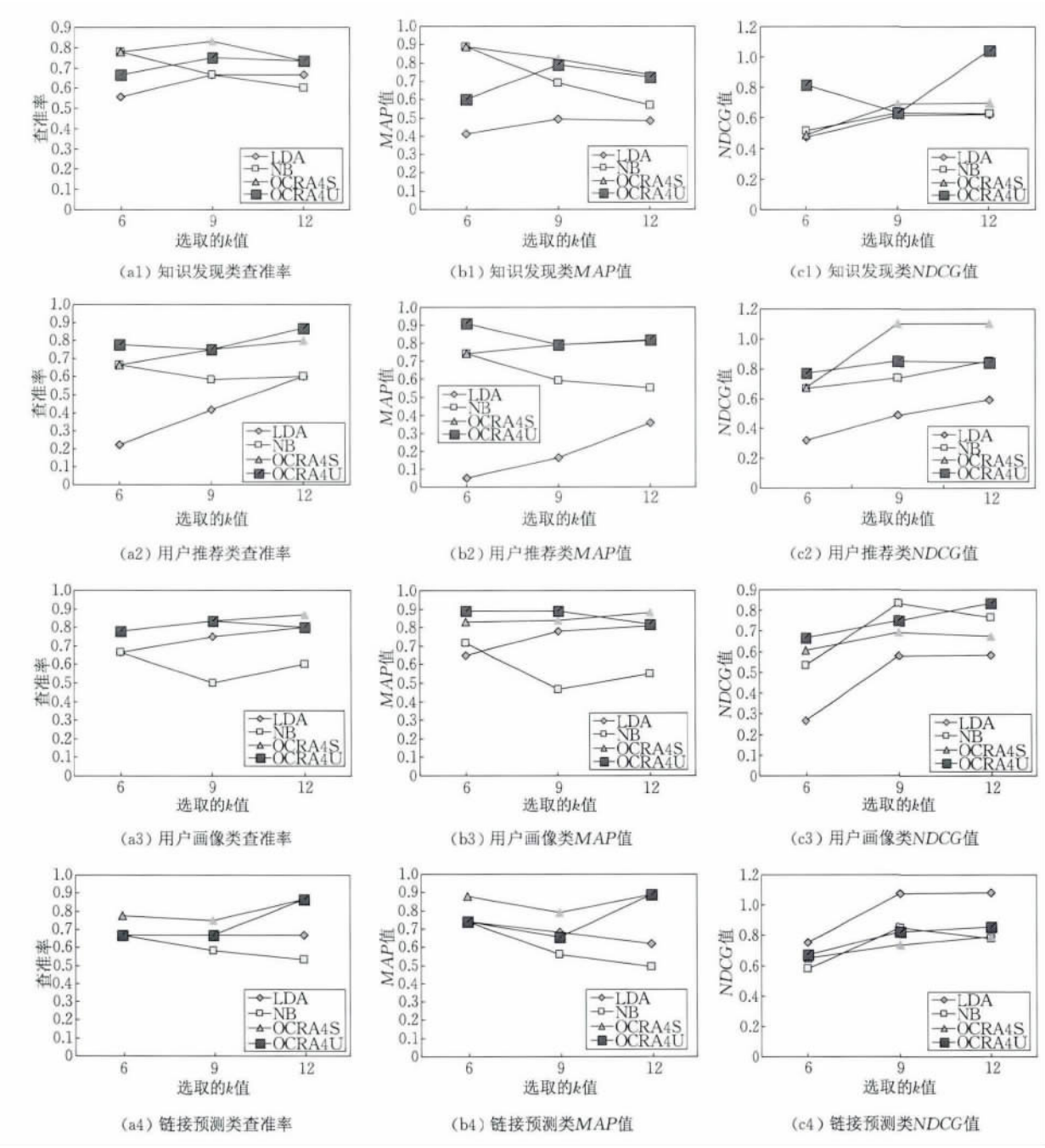
表 6　DBLP 数据集上 4 种方法在不同 k 值下的平均 $NDCG$ 值

$NDCG$	LDA	NB	OCRA4S	OCRA4U
$k=6$	0.45	0.60	0.61	0.71
$k=9$	0.69	0.79	0.80	0.77
$k=12$	0.72	0.77	0.81	0.89



(a) 平均查准率 (b) 平均*MAP*值 (c) 平均*NDCG*值

图 10 DBLP 数据集上 4 种方法在不同 k 值下的各评价指标平均值对比



(a1) 知识发现类查准率 (b1) 知识发现类*MAP*值 (c1) 知识发现类*NDCG*值

(a2) 用户推荐类查准率 (b2) 用户推荐类*MAP*值 (c2) 用户推荐类*NDCG*值

(a3) 用户画像类查准率 (b3) 用户画像类*MAP*值 (c3) 用户画像类*NDCG*值

(a4) 链接预测类查准率 (b4) 链接预测类*MAP*值 (c4) 链接预测类*NDCG*值

图 11 DBLP 数据集上 4 种方法在不同 k 值下的各类的不同评价指标对比

集上 4 种方法在不同 k 值情况下、分别针对各类商品需求的橙领推荐的查准率、*MAP* 值以及 *NDCG* 值.

由图 10 可知，(1) 针对同一种方法，随着 k 值的增大，平均查准率、平均 *MAP* 值、平均 *NDCG* 值均大致呈递增趋势，这是由于 k 越大(即有效的作者越多)，推荐难度就越小，推荐的命中概率就越大；(2) 在相同 k 值下，平均查准率、平均 *MAP* 值、平均 *NDCG* 大致有以下规律：OCRA4U 以及 OCRA4S 效果优于朴素贝叶斯方法和 LDA 方法. 分析原因，朴素贝叶斯方法和 LDA 方法只考虑学者的文章信息，而没有考虑学者的社会关系因素，没有衡量学者在众学者中的影响力程度，这会导致对学者的认知较片面，进而推荐的学者准确性不高；LDA 方法的推荐水平偏低的另外一个原因是，LDA 方法抽取出的每一个主题都会包含部分其他主题的内容，主题划分较不严格，由此得到的学者的主题向量不准确，最终导致推荐效果不理想. 但不排除个例，当 $k=9$ 时，朴素贝叶斯方法的平均 *NDCG* 值为 0.79，大于的 OCRA4U 的 0.77，说明此时朴素贝叶斯方法的橙领排序结果更加匹配标准答案.

由图 11 可知，在相同 k 值下，针对各个类别的商品需求得到的橙领排序的查准率、*MAP* 值、*NDCG* 大致有如下规律：OCRA4U 以及 OCRA4S 效果优于朴素贝叶斯方法和 LDA 方法. 但是存在部分例外的情况，在知识发现类，当 $k=6$ 时，朴素贝叶斯方法的查准率为 0.78，大于 OCRA4U 算法的 0.67，朴素贝叶斯方法的 *MAP* 值为 0.89，大于 OCRA4U 算法的 0.60，但是朴素贝叶斯方法的 *NDCG* 值为 0.51，小于 OCRA4U 算法的 0.82. 分析原因发现，朴素贝叶斯方法返回作者序列的正确作者的数量较多，但其中正确橙领的排序序号有些较靠后，导致顺序的匹配度较低，低于 OCRA4U 算法. 在用户画像类上，当 $k=9$ 时，朴素贝叶斯方法的 *NDCG* 值为 0.83，大于 OCRA4U 算法的 0.76 以及 OCRA4S 算法的 0.67，这是由于此时朴素贝叶斯方法返回的正确的作者普遍序号较靠前，虽然正确的作者数量不及 OCRA4U 算法和 OCRA4S 算法，但是与标准答案的匹配程度高于这两者.

实验结果误差主要来源以下两个方面：首先，标准答案的获取有较大主观性，本文采取的是 20 个人评分的平均结果，调查人数较少，后续工作中可以增加调查人数，使获取的标准答案更具有普遍性. 另外，在橙领定位中存在的误差也会给橙领推荐实验中的结果带来不良影响.

在仿真数据集 DBLP 数据集中，$k=6$、9 以及 12 时均证明，OCRA4U 算法以及 OCRA4S 算法用于推荐作者的仿真效果良好，具有可行性及有效性，可以针对研究需求返回给用户较满意的学者排序. 而其他两种方法也具有可行性，但是整体上效果劣于 OCRA4U 算法以及 OCRA4S 算法.

4.4.3　社会化数据影响实验结果与分析

社会化数据影响实验的结果分别如图 12～图 15 所示，它们分别是新浪微博数据集上 OCRA4S 算法以及 OCRA4U 算法的社会化数据影响实验对比结果，以及 DBLP 数据集上 OCRA4S 算法以及 OCRA4U 算法的社会化数据影响实验对比结果.

图 12 是新浪微博数据集上面向商家的橙领推荐算法的社会化数据影响实验对比结果，其中 influence 系列表示只考虑橙领社会关系(即影响力)、不考虑微博信息时的面向商家的橙领推荐算法，blog 系列表示只考虑橙领微博信息、不考虑其社会关系时的面向用户的橙领推荐算法. 由图 12 可知，无论 $k=6$、9 或 12，仅考虑微博信息或社会关系的 OCRA4S 算法各类的平均查准率、平均 *MAP* 值以及平均 *NDCG* 值均分别低于同时考虑两者的 OCRA4S 算法，例如，当 $k=6$ 时，OCRA4S 算法的平均查准率为 62%，比仅考虑微博信息的 OCRA4S 算法以及仅考虑社会关系的 OCRA4S 算法分别高

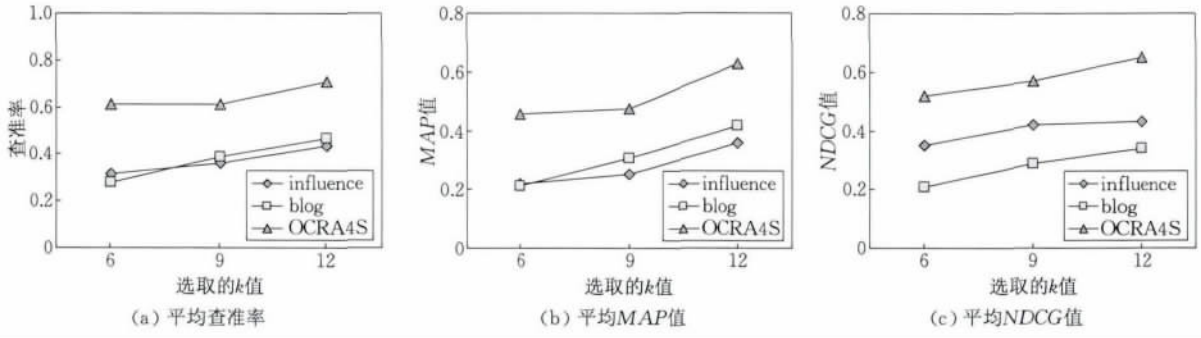


图 12　新浪微博数据集上面向商家的橙领推荐算法的社会化数据影响实验对比结果

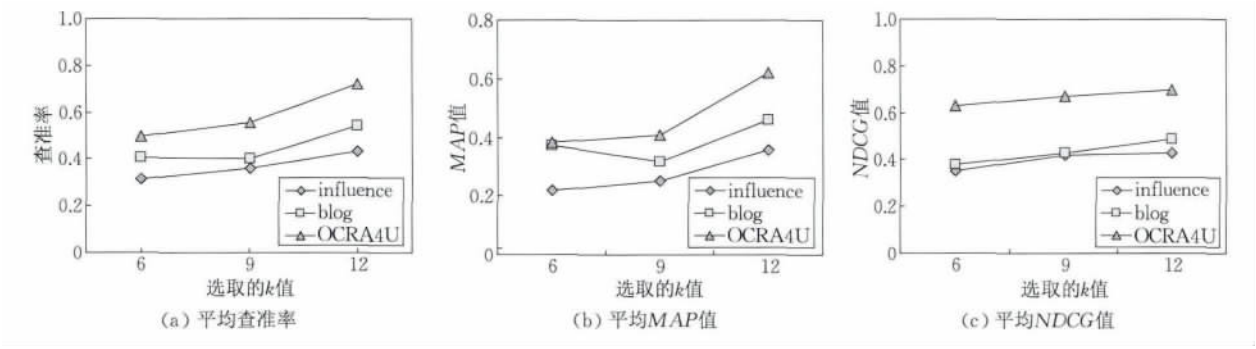


图 13 新浪微博数据集上面向终端用户的橙领推荐算法的社会化数据影响实验对比结果

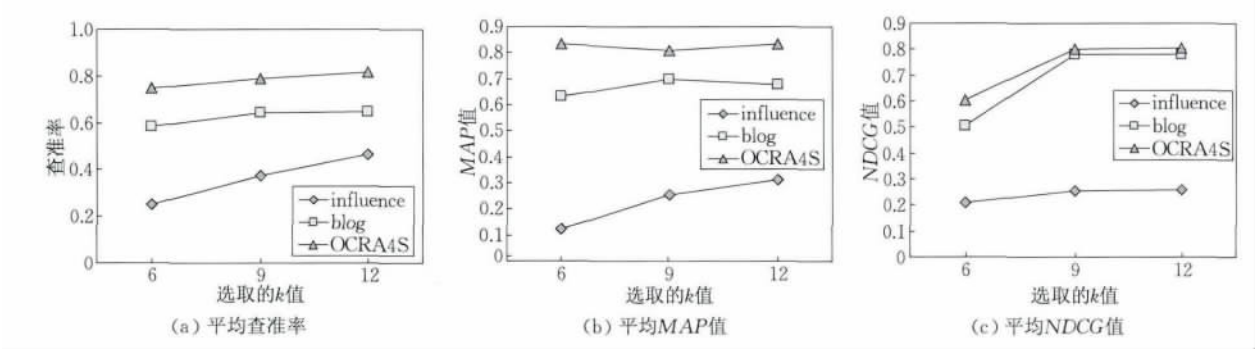


图 14 DBLP 数据集上面向商家的橙领推荐算法的社会化数据影响实验对比结果

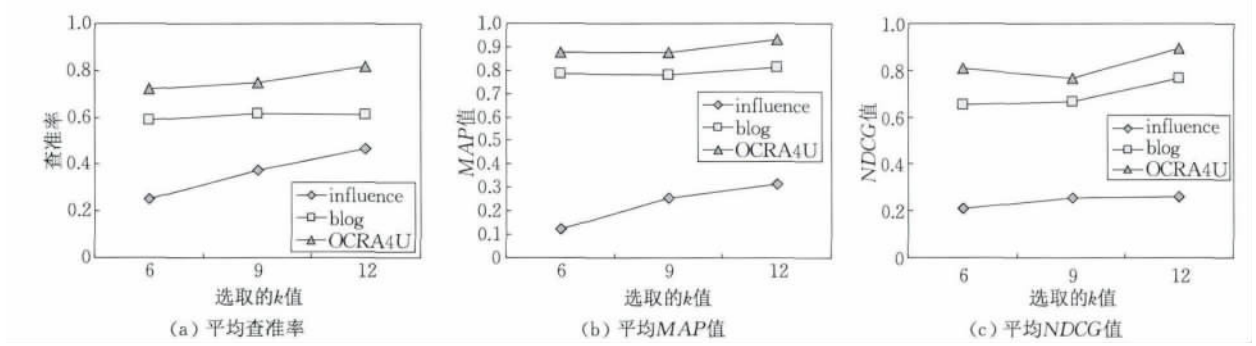


图 15 DBLP 数据集上面向终端用户的橙领推荐算法的社会化数据影响实验对比结果

出 34%和 30%. 由此证明,无论 $k=6$、9 或 12,仅考虑微博信息或社会关系的 OCRA4S 算法的推荐效果均劣于同时考虑两者的 OCRA4S 算法. 并且仅考虑社会关系的 OCRA4S 算法的效果略优于仅考虑微博信息的 OCRA4S 算法,当 $k=9$ 时,前者的平均 *NDCG* 值比后者高出 0.09. 图 13 是新浪微博数据集上面向终端用户的橙领推荐算法的社会化信息实验对比结果,同样可以看出,当 $k=6$、9 或 12,仅考虑微博信息或社会关系的 OCRA4U 算法的各类的平均查准率、平均 *MAP* 值以及平均 *NDCG* 值分别劣于同时考虑两者的 OCRA4U 算法,并且仅考虑微博信息的 OCRA4U 算法的效果优于仅考虑社会关系的 OCRA4U 算法,当 $k=6$ 时,前者的平均 *MAP* 值比后者高出 0.15. 图 14 是 DBLP 数据集上面向商家的橙领推荐算法的社会化数据影响实验对比结果,可以看出无论 $k=6$、9 或 12,仅考虑微博信息或社会关系的 OCRA4S 算法各类的平均查准率、平均 *MAP* 值以及平均 *NDCG* 值均分别低于同时考虑两者的 OCRA4S 算法,由此证明,无论 $k=6$、9 或 12,仅考虑微博信息或社会关系的 OCRA4S 算法的推荐效果均劣于同时考虑两者的 OCRA4S 算法. 并且仅考虑微博信息的 OCRA4S 算法的效果优于仅考虑社会关系的 OCRA4S 算法,当 $k=12$ 时,前者的平均 *NDCG* 值比后者高出 0.51. 图 15 是 DBLP 数据集上面向终端用户的橙领推荐算法的社会化信息实验对比结果,同样可以看出,当 $k=6$、9 或 12,仅考虑微博信息或社会关系的 OCRA4U 算法的各类的平均查准率、平均 *MAP* 值以及平均 *NDCG* 值分别劣于同时考虑两者的 OCRA4U 算法,并且仅考虑微博信息的 OCRA4U 算法的效果

优于仅考虑社会关系的 OCRA4U 算法，当 $k=6$ 时，前者的平均 MAP 值比后者高出 0.65.

由此，可以证明，使用社会化信息的橙领推荐效果优于不使用社会化信息的橙领推荐效果，本文提出的面向社会化导购的橙领推荐方法中的 OCRA4U 算法以及 OCRA4S 算法，均考虑了两种社会化信息，一个是橙领的微博信息，另一个是橙领的社会关系，利用这两种信息共同完成橙领的推荐，两者缺一不可，均对橙领的推荐起到重要作用.

5　应　用

本文提出的 SSOCRM 方法的应用性很强，不仅可以应用于面向社会化导购的橙领推荐，还可以应用于其他的应用场景. 使用 SSOCRM 方法解决相关问题的使用模式如图 16 所示. 图中，橙领可以看做富有经验的专家，而橙领的微博购物信息可以看做专家拥有的丰富经验，橙领的社会化关系可以看做专家的社会化关系，其中橙领之间的好友关系可以看做专家之间的合作关系，橙领与粉丝间的被关注关系，可以看做专家的经验被引用关系等. 于是，SSOCRM 方法适用于解决所有的向亟需专家经验的用户推荐可信专家问题. 用户提出用户需求，调用 SSOCRM 方法可以获取可信的专家序列.

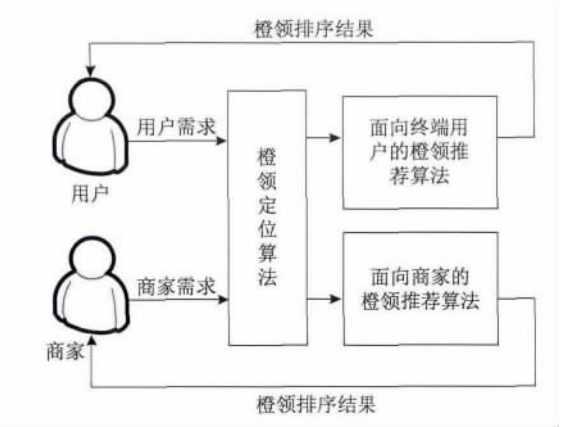


图 16　SSOCRM 方法的使用示意

SSOCRM 方法中关键的技术包括两大方面：专家的定位以及专家的推荐. 而专家的定位采用了较为标准的信息检索和打分加权的方法，其主要包含 3 步：数据集的划分、专家的分类以及最终的专家定位. 专家的推荐主要针对用户的需求，综合专家的定位情况以及专家在社会关系中影响力程度两者，进行专家的推荐，最终按照橙领可信度从大到小返回专家序列给用户.

6　结　论

社交网络飞速发展和网络购物日益普及导致了社会化导购的出现，也促使了橙领这一职业的出现. 但是由于橙领发展时间较短且每个橙领的成长模式不同，橙领自身以及普通用户与橙领之间都存在亟待解决的问题，如何利用橙领信息进行产品推荐也是社会化购物要解决的一个关键问题. 通过对橙领相关技术的研究，能使我们更透彻地了解基于社会网络的产品营销机制，也能有助于探索社会化导购的底层模式. 目前尚未见针对橙领的社会化导购相关研究. 因此，本文研究了一种面向社会化导购的橙领推荐方法 SSOCRM，其具体包括 3 个算法，分别用于解决橙领自身定位问题、面向用户的橙领推荐问题和面向商家的橙领推荐问题. 橙领定位算法依据橙领推荐过的历史产品类型对橙领进行定位特征向量化描述. 在此基础上分别提出的面向终端用户的橙领推荐算法 OCRA4U 和面向商家的橙领推荐算法 OCRA4S，解决了如何获取最能符合用户/商家需求的橙领并将其推荐给用户/商家的问题. 前者综合考虑了橙领的影响力和橙领与用户需求的匹配度，同时对橙领在社交网络中的影响力进行了量化. 后者参考了基于项目的协同过滤和基于内容的推荐算法的思想，结合橙领在网络中的影响力以及橙领历史推荐产品，面向商家推荐出最符合其产品需求的橙领. 另外，基于新浪微博数据集以及 DBLP 数据集，设计和实现了相关实验，实验结果验证了上述算法的可行性和正确性.

虽然实验结果证明了算法的可行性，但是算法仍然存在许多可以改进的地方. 例如，橙领网络影响力计算时使用的独立级联模型简化了影响力的计算；推荐橙领时，只考虑了橙领在社交网络中的影响力，以及用户需求与橙领定位的匹配，没有考虑用户、橙领自身的其他比较次要因素，例如教育背景、宗教信仰和地理环境等，这些都会影响算法在实际中的效果. 这些都是未来工作要解决的.

参　考　文　献

[1]　2013 年中国网购市场调查报告. China Internet Network

Information Center, Beijing, 2014, 04, 21

[2] Hannon J, Bennett M, Smyth B. Recommending Twitter users to follow using content and collaborative filtering approaches//Proceedings of the 4th ACM Conference on Recommender Systems. Barcelona, Spain, 2010: 199-206

[3] Chen J, Geyer W, Dugan C, et al. 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, USA, 2009: 201-210

[4] Sakaguchi T, Akaho Y, Takagi T, Shintani T. Recommendations in Twitter using conceptual fuzzy sets//Proceedings of the Conference of the North American Fuzzy Information Processing Society. Toronto, Canada, 2010: 1-6

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

[6] Chen Ke-Han, Han Pan-Pan, Wu Jian. User clustering based social network recommendation. Chinese Journal of Computers, 2013, 36(2): 349-359(in Chinese)
(陈克寒, 韩盼盼, 吴健. 基于用户聚类的异构社交网络推荐算法. 计算机学报, 2013, 36(2): 349-359)

[7] Wang Yu, Gao Lin. Social circle-based algorithm for friend recommendation in online social networks. Chinese Journal of Computers, 2014, 37(4): 801-808(in Chinese)
(王玙, 高琳. 基于社交圈的在线社交网络朋友推荐算法. 计算机学报, 2014, 37(4): 801-808)

[8] Gao Ming, Jin Che-Qing, Qian Wei-Ning, Wang Xiao-Ling, Zhou Ao-Ying. Real-time and personalized recommendation on microblogging systems. Chinese Journal of Computers, 2014, 37(4): 963-975(in Chinese)
(高明, 金澈清, 钱卫宁, 王晓玲, 周傲英. 面向微博系统的实时个性化推荐. 计算机学报, 2014, 37(4): 963-975)

[9] Cha M, Gummadi K P. Measuring user influence in Twitter: The million follower fallacy. Artificial Intelligence, 2010, 146(1): 10-17

[10] Ikeda K, Hattori G, Ono C, et al. Twitter user profiling based on text and community mining for market analysis. Knowledge-Based Systems, 2013, 51: 35-47

[11] Pazzani M J, Billsus D. Content-based recommendation systems//Brusilovsky P, Kobsa A, Nejdl W eds. The Adaptive Web: Methods and Strategies of Web Personalization. Berlin Heidelberg: Springer, 2007: 325-341

[12] Wang Chong, Blei D M. Collaborative topic modeling for recommending scientific articles//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, America, 2011: 448-456

[13] Wang Biao. Analysis of User Influence in Social Networks [Ph. D. dissertation]. Harbin Institute of Technology, Harbin, 2012(in Chinese)
(王彪. 社交网络中的用户影响力分析[博士学位论文]. 哈尔滨工业大学, 哈尔滨, 2012)

[14] Majumder A, Datta S, Naidu K V M. Capacitated team formation problem on social networks//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 1005-1013

XIE Xiao-Qin, born in 1973, Ph.D., associate professor. Her main research interests include social network analyzing and mining, intelligent information processing, information retrieval and service oriented computing.

HAN Shuai, born in 1991, M. S. Her research interest is social network analyzing and mining.

LV Bin, born in 1989, M. S. His research interest is social network analyzing and mining.

ZHANG Zhi-Qiang, born in 1973, Ph. D., professor. His main research interests include information retrieval, database, and intelligent information processing.

PAN Hai-Wei, born in 1974, Ph.D., associate professor. His main research interests include database, medical image mining.

Background

With the rapid development of social networks and E-commerce, the combination of the two has been more and more popular. Social shopping guiders appear and develop with the developing of the social network and online shopping. This also leads to the emerging of orange-collar who gets income by recommending products to other persons in SNS in China. Through studying orange-collar related technologies, we can not only learn the mechanism of online product marketing thoroughly, but also explore the deep level patterns of social sales. However, current researches seldom focus on this issue. As a result, given an orange-collar, users could not know what product he or she is skilled

at recommending. Faced with a shopping require, users don't know which orange-collar to choose to get some useful advice from.

Hence, to solve the problem of orientating the orange-collars and recommending them to users or shops, this paper proposes an orange-collar recommending method, which includes three algorithms that are orange-collar positioning algorithm, OCRA4U algorithm and OCRA4S algorithm. The orange-collar positioning algorithm describes orange-collar by orientation vector based on the varieties of the products in its recommending history and we transfer the positioning problem to a clustering problem. OCRA4U returns an orange-collar recommending list to user by taking into consideration the orange-collar's network influence and the matching degree between the orange-collar and user's need. OCRA4S combines orange-collar's influence and its recommending histories, and finally gains the most satisfied orange-collar recommendation for the product need of shops. The experimental results have proved the correctness and feasibilities of our proposed method.

This work is supported by the National Natural Science Foundation of China (Nos. 61202090, 61370084, 61272184), the Program for New Century Excellent Talents in University (NCET-11-0829), the Science and Technology Innovation Talents Special Fund of Harbin under Grant (No. 2015RQQXJ067), the Fundamental Research Funds for the Central Universities under Grant (No. HEUCF100602).