情 报 学 报 ISSN 1000 - 0135
第 33 卷 第 5 期 520 - 529, 2014 年 5 月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 33 No. 5, 520 - 529 May 2014

doi:10.3772/j.issn.1000-0135.2014.05.007

基于社会网络关系的微博个性化推荐模型[1)]

蔡淑琴 袁 乾 周 鹏

(华中科技大学管理学院 武汉 430074)

摘要 信息过载是影响微博等社会化媒体平台消费者持续使用行为的重要原因。协同过滤推荐能有效解决信息过载问题,但既有研究未能在推荐系统中整合用户创造内容和社会网络关系,社会网络关系体现出了消费者的偏好。针对微博的用户创造内容和社会网络两要素,本文从关键词层次入手,引入向量空间模型描述用户对关键词偏好,设计社会网络修订系数修订用户相似矩阵,实现社会网络关系驱动的协同过滤推荐模型。实验结果表明,较于基准协同过滤推荐方法,本文所提出的基于社会网络修订的协同过滤推荐能更准确并有效地实现个性化推荐。

关键词 微博 推荐 协同过滤 社会网络

Personalized Microblog Recommendation Model Based on Social Network Relations

Cai Shuqin, Yuan Qian and Zhou Peng

(*School of Management*, *Huazhong University of Science and Technology*, *Wuhan* 430074)

Abstract Information overloading is one of important factor which impact users' continual usage behavior in microblog and other social media websites. Collaborative filtering recommendation is an effective approach to solve the problem. However, prior research did not consider the influence of both UGC(user generated content) and social networks at the same time. Social networks represent users' preference just as UGC. According to the feature of both UGC and social networks, the paper brings the vector space model into the user's rating model from the perspective of keywords. Method of calculating similar matrix between users is designed by combining the social networks structure characteristics. The collaborative filtering method driven by social networks relations has been realized. The experimental results have shown that the approach of collaborative filtering driven by social networks structural proposed in this paper outperforms standard ways.

Keywords microblog, recommendation, collaborative filtering, social networks

收稿日期:

作者简介:**蔡淑琴**,女,1955 年出生,华中科技大学管理学院教授,博士生导师,企业商务智能工程研究所所长,主要研究方向:信息管理、决策支持理论与方法;**袁乾**,男,1989 年生,华中科技大学管理学院管理工程与科学博士研究生,主要研究方向:信息管理、数据挖掘;**周鹏**,男,1981 年生,华中科技大学管理学院管理工程与科学博士研究生,主要研究方向:信息管理、数据挖掘。通讯作者:袁乾 电话:18702770857;邮箱:hustyuanqian@gmail.com

1) 国家自然科学基金项目——微内容生产加工模式及其支持平台的研究(71071066);移动社会化媒体中基于价值共创的在线负面口碑处理资源管理的方法及系统研究(71371081);教育部人文社科基金项目——基于互联网信息的企业危机事件识别研究(11YJA630098);高等学校博士学科点专项科研基金——基于价值共创的在线负面口碑处理知识推荐的研究(20130142110044)。

1 引 言

近年来,微博作为典型的社会化媒体应用得到高速发展,构建了以用户创造内容(User Generated Content, UGC)和社会网络关系为要素[1]的自组织的信息生态圈[2]。用户在拥有海量信息的同时也迷失于信息海洋,面临严峻的信息过载问题,碎片化的 UGC 以及复杂的社会网络关系使得用户无法及时有效获取感兴趣的内容。协同过滤推荐(Collaborative Filtering, CF)作为常用的推荐方法,能有效解决信息过载问题[3~5],大量应用于电影推荐[3]、音乐推荐[6]等多个领域,为企业带来巨大的经济价值[7],故而 Netflix 悬赏 100 万美金鼓励推荐研究①。并且,国家 863 计划将融合社会网络关系拓扑结构特征的推荐方法列为重点发展方向②。因此,面向微博的推荐系统迫切需要将社会网络结构特征整合到协同过滤推荐框架,其研究具有重大理论及实践意义。

协同过滤推荐假定相似用户对项目(Item)的评分相近,通过集成目标用户的近邻集(即最相似的若干个用户)对特定项目评分以预测用户的评分[8,9]。现有协同过滤推荐研究多依赖于用户-项目评分矩阵,通过结合人工智能、认知理论等多领域中的方法,实现准确预测目标用户对推荐项目的评分,如遗传算法[5],物质扩散[3,4]、随机游走[10]、遗忘函数[11]等。但是,整合社会网络关系的协同过滤推荐的研究尚处于起步阶段,Shang、Zhang 等利用用户和电影的标签标注关系,构建用户-电影-标签三部图的社会网络,采用物质扩散模型的方法有效实现协同过滤推荐[3,4];Ting 等利用对比实验论证了社会网络关系对于推荐的重要性,但未设计整合社会网络关系的方法[12];在国内,王丽莎等构建基于标签和项目的双游走推荐模型,提高推荐效率[10]。既有研究虽为网络拓扑结构下的推荐研究提供了一定思路和方法,但未将用户间的社会网络关系整合进来。社会网络关系蕴含了用户的相似性[13],能度量用户偏好。用户倾向于链接其偏好的用户,获得其关注的内容,相关实证研究证明了推荐模型中整合社会网络关系的重要性和可行性[14]。如何构建整合社会网络关系复杂化以及 UGC 碎片化和非结构化特征的推荐模型,研究网络拓扑结构对推荐的促进作用依旧是面向微博平台的推荐研究的难点。

面向微博的协同过滤推荐模型研究是管理科学、信息科学、计算机科学、社会网络科学等多学科交叉领域研究。鉴于研究的不足及其必要性,本文针对微博中 UGC 以及社会网络关系两大要素,基于基准协同过滤推荐模型,结合向量空间模型、社会网络分析等理论与方法,提出一个基于社会网络修正的协同过滤推荐模型(Social Networks Collaborating Filtering Recommendation Model, SNCF-RM),采用向量空间模型描述用户偏好关键词集,提出 3 种用户相似性的社会网络关系修订系数,实现对用户相似度的修订,从关键词层次实现用户信息推荐。从实验结果看,推荐效率因考虑社会网络结构特征而有所提高。

2 微博平台的推荐框架

2.1 基准协同过滤推荐模型

基准协同过滤模型基于用户历史行为信息(购买行为、评分行为)计算用户相似度,选择用户近邻集,通过对近邻集的评分进行加权以预测用户对项目评分,其输入是用户—项目评分矩阵 $R:U \times I$,其中 U 是用户集,I 是项目集,输出是目标用户 u_i 对项目 k 的预测评分 $\hat{R}(u_i,k)$,如公式(1)所示,并主要分为三个核心过程。

$$R:U \times I \rightarrow \hat{R}(u_i,k) \tag{1}$$

1)获取用户—项目评分矩阵

用户—项目评分矩阵 R 表示用户对项目的偏好程度(即评分)。现有研究多使用既有数据集(MovieLens 等)进行研究,如从时序视角讨论用户偏好的漂移[15],将用户对项目偏好转化为对属性偏好[16],多忽略用户评分获得及表达过程,尤其是用户偏好非结构化表达的情况。

2)计算用户相似度矩阵

用户相似度矩阵是基准协同过滤推荐的核心。既有研究提出不同方法计算用户相似度,如在相似度计算中考虑用户购买情境的相似度,提高推荐效率[5];或构建用户、项目及标签间社会网络关系,用社会网络相似度作为用户相似度[3];或构建基于学术概念网络层和研究者网络层的立体网络结构,利

① http://www.netflixprize.com/

② http://www.most.gov.cn/tztg/201304/t20130416_100843.htm

用价值扩散计算用户相似度[17]。其中皮尔森相关系数是最常用的用户相似度计算方法,如下所示。

$$S_{ij} = \frac{\sum_{k \in K'} (r_{ik} - \overline{r_i})(r_{jk} - \overline{r_j})}{\sqrt{\sum_{K'} (r_{ik} - \overline{r_i})^2} \sqrt{\sum_{K'} (r_{jk} - \overline{r_j})^2}} \quad (2)$$

其中,r_{ik}是 R 的一个元素,代表用户 i 对项目 k 的偏好程度,S_{ij}是用户 u_i 和 u_j 的相似度。基于用户相似度矩阵,常用两种方法选择用户近邻集:一是选取和用户相似度大于给定阈值的近邻;二是选择和用户最相似的给定数目 L 个近邻。本研究采用方法二。

3)预测用户对项目的潜在偏好

预测用户对被推荐项目的潜在偏好是基于协同的方法,整合近邻集对被推荐项目的评分,以相似度为权重进行加权求和,得到对潜在偏好的预测值。在微博下,用户对关键词或使用,或不使用,不存在类似电子商务环境中的使用倾向,且用户对关键词的使用频率范围波动较大,方差远超过电子商务环境下用户对产品的评分的方差,所以平均值调节的预测评分往往并不准确。因此,本文借鉴 kNN 常用预测方法预测用户偏好,如公式(3)所示。

$$\hat{r}_{ik} = \frac{\sum_{j \in \mathbf{N}_i} r_{jk} * S_{ij}}{\sum_{j \in \mathbf{N}_i} S_{ij}} \quad (3)$$

其中,$\hat{r}_{ik}$ 是预测的用户 i 对项目 k 的偏好水平,N_i 是用户 i 的近邻集。

2.2 面向微博的协同过滤推荐框架

在基准协同过滤推荐模型下,面向微博的推荐模型面临着三大困境:

首先,冷启动困境。现有面向 UGC 的研究大都将 UGC 视为最小信息单元,但高速新增的 UGC 带来了严重的冷启动问题。新增的 UGC 因缺乏和既有 UGC 及用户的关系,即无法对新进入的内容进行有效推荐,同时也是信息过载的重要原因之一。虽然已有协同过滤推荐方法尝试解决冷启动问题,但是无法应对高速新增的内容,需要研究针对高速新增的 UGC 特征的推荐方法;

然后,缺乏评分信息困境。以文本主的非结构化 UGC 中不包含用户对内容的评分数据,而既有推荐模型都是以评分矩阵作为输入。

最后,整合社会网络关系困境。社会网络关系是微博平台的重要信息。推荐系统得到的近邻集以及社会网络分析得到的朋友集相比较,目标用户更信赖朋友推荐的内容[18]。“同质”等现象使得社会网络结构与用户相似性强相关[14],强关系用户推荐的信息更容易被接受和采纳[19]。社会网络结构为推荐系统提供新型的数据,能提高推荐效率。

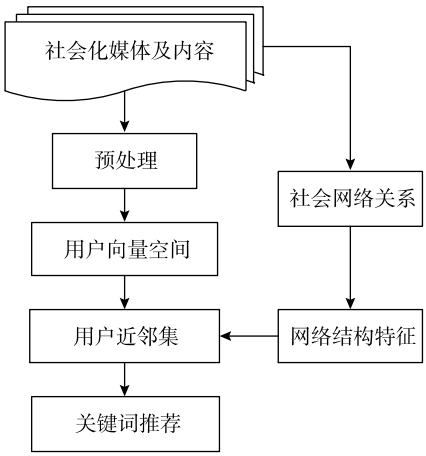


图1 SNCF-RM 框架

针对上述三个困境,本文提出如图1所示的基于社会网络结构、以关键词为推荐对象的 SNCF-RM 框架。在预处理的基础上,获取用户感兴趣关键词集,生成描述用户偏好模型的用户空间向量,通过社会网络结构特征,修订皮尔森相关系数,获取近邻集,最后预测目标用户对关键词的偏好程度。SNCF-RM 主要解决以下两个问题:①针对 UGC 的高速新增和非结构化特征,如何估计用户对关键词的评分,即偏好程度?②针对微博的社会网络结构数据,如何将之整合到协同过滤的推荐中,以提高推荐效率?

3 基于关键词的用户向量空间建模

为解决规避海量、不断新增的 UGC 给推荐带来的冷启动困境,本文借鉴文献[20]的思路,SNCF-RM 选择关键词作为推荐对象(即项目),将用户和 UGC 的交互关系转化为到对关键词的偏好,得到以关键词为中心的用户 - 关键词评分矩阵,预测目标用户对关键词的偏好,并基于获得高偏好关键词检索用户潜在偏好的相关内容,代替直接向用户推荐内容。为聚焦研究,本文关注于面向社会网络的关键词的推荐方法。

相比于高速新增的 UGC,关键词粒度更细,更新较少,类似于产品的属性信息,并且具有较好的统

计特征,能度量用户对关键词的偏好,解决了缺乏评分信息的困境。一般而言,用户使用关键词越频繁,用户的偏好越高,并且该关键词在用户群体中使用越不平均,用户偏好越高。本文利用自然语言处理工具,抽取关键词,度量其权重作为用户偏好评分,使用向量空间模型[21]构建用户偏好评分的向量模型,并计算用户间的相似关系。

向量空间模型基于代数模型刻画文本文档的内容,是信息过滤、检索以及关联规则等研究的关键。然而,向量空间模型的常用权重计算方法(如TF-IDF方法、信息熵法)无法直接应用于本研究。TF-IDF 方法常用于文档检索领域,计算关键词对文档的代表能力,强调关键词对文档的区分能力,但无法表示用户偏好,如关键词 k 只被用户 A 使用过一次,但其他用户没有使用过,虽然具有高 TF-IDF 值,但和本文假定相违背;信息熵虽度量关键词的使用不均匀程度,但无法体现用户个性化需求及特征。

词频(单个用户使用单个关键词的次数)是常用的用户偏好度量方法。假设用户常使用其偏好的产品或关键词,满足本文对用户偏好的假定,故词频用作偏好度量指标。然而,频繁出现的常用词并没有为推荐提供有用信息,如"目的"虽被大量用户多次使用,但无推荐价值;因此,引入信息熵以规避常用词[22],降低用户对常用词的偏好程度,满足不均匀分布的关键词更能有效度量用户偏好的假设。最后,用户对关键词的偏好评分计算如公式(4)所示:

$$r_{ik} = \begin{cases} 0 & IDF_k < \alpha \\ n_{ik} \times \left(- \sum_{m \in |U_k|} (p_{mk} * \log_2 p_{mk}) \right)^{-\beta} IDF_k & IDF_k \geqslant \alpha \end{cases} \tag{4}$$

其中,r_{ik} 表示用户 i 对关键词 k 的偏好,用作评分,n_{ik} 表示用户 u_i 使用关键词 k 的次数,即词频;U_k 是使用过关键词 k 的用户集合;p_{mk} 是用户 m 使用关键词 k 的次数占关键词 k 被使用总次数的比例;β 是信息熵调节系数,调节信息熵的比重,当 β 越大,则越强调分布不均匀关键词的重要性,反之词频的越重要;IDF_k 是关键词 k 的逆向文档频率,刻画关键词的常用程度,α 是逆向文档频率的过滤阈值,用以过滤不具有推荐价值的常用词。当 IDF 值越小,说明该关键词被大多数用户都使用过,并且分布较为均匀,被识别常用词。基于信息检索理论[23],常用词无法用于识别相关文档,在用户信息推荐中,也无法代表用户偏好,应将其排除在推荐系统之外。

本文将每一个关键词视为描述用户偏好的一个维度,故而所有关键词构成一个 K 维空间,K 是关键词个数,于是用户对所有关键词的评分构成用户在关键词空间中的向量,得到用户偏好的向量空间模型,以刻画用户偏好,表示为 $R_i = \{r_{i1}, r_{i2}, \cdots, r_{iK}\}$。所有用户在关键词空间的偏好向量组成用户 - 关键词评分矩阵:$R = U \times I = \{R_1, R_2, \cdots, R_n\}$,得到 SNCF-RM 的输入。

4 社会网络系数修正方法

社会网络结构提供异于用户偏好的信息。本文利用社会网络结构特征设计社会网络修正系数,从社会网络分析视角提出对用户相似管理的修正方法,整合 UGC 和社会网络关系。

4.1 社会网络修正系数

用户倾向于和相似用户构建社会网络关系,存在用户"同质性"[14]。本文提出三种修正系数将社会网络关系的拓扑特征整合到用户相似矩阵中,以期提高推荐效果和效率。

1)基于社会网络相似度的社会网络修正系数

社会网络相似度假定用户的社会网络关系越紧密,用户的相似度越高[14]。微博中,用户关注行为表示用户偏好被关注用户发布的信息,愿意实时接收被关注用户发布的 UGC。显而易见,即两个用户的共同关注对象越多,则两者的内容相似性越强[14,24],故利用用户的共同关注对象数度量用户的社会网络相似性。用户 i 和用户 j 的社会网络相似度是两者关注对象交集占关注对象并集的比例,计算方法如公式(5)所示。

$$\omega_{ij} = \frac{| Nb(u_i) \cap Nb(u_j) |}{| Nb(u_i) \cup Nb(u_j) |} \tag{5}$$

其中,$Nb(*)$ 是用户的关注用户集。易知,社会网络相似度的取值范围是[0,1],0 表示不存在两个用户共同关注对象,即关注对象完全不同,1 表示两个用户关注对象完全一致。基于社会网络相似度的修正系数矩阵是对称的,即 $\omega_{ij} = \omega_{ji}$。

2)基于社会网络距离的社会网络修正系数

社会网络距离假设距离和相似度紧密相关,用户间距离越远,用户的共同好友数越少,用户相似度越低,故而利用社会网络距离用于修正用户相似性。用户 i 和用户 j 的社会网络距离是用户 i 和用户 j 关注对象并集减去两者关注对象交集所剩下部分占用户 i 和用户 j 关注对象并集的比例,计算如公式(6)

所示。

$$\omega_{ij} = \frac{| Nb(u_i) \cup Nb(u_j) - Nb(u_i) \cap Nb(u_j) |}{| Nb(u_i) \cup Nb(u_j) |} \quad (6)$$

其中，$Nb(*)$是用户的关注用户集。易知，基于社会网络距离的修正系数矩阵是对称矩阵，即 $\omega_{ij} = \omega_{ji}$。

3）基于社会网络度中心性修正系数

社会网络度指用户在微博中的社会关系数目，度中心性是社会网络分析的重要指标，体现用户在社会网络中的重要性。Zeng 和 Wei 通过实证研究证实了用户度中心性显著影响用户之间的相似性[14]。SNCF-RM 采用度中心性作为社会网络修正系数，并与社会网络修正系数进行对比，其计算如公式(7)所示。

$$\omega_{ij} = \frac{DC(u_j)}{DC(u_i) + DC(u_j)} \quad (7)$$

其中，$DC(*)$是用户的度，基于度中心性修正的修正系数矩阵是不对称的，即 $\omega_{ij} \neq \omega_{ji}$。

4.2 用户相似度修正方法

社会网络修正系数从不同视角刻画了用户之间的关系。本文利用社会网络修正系数修正基准协同过滤模型中的用户相似度，将社会网络关系和 UGC 整合到一起，获得修正后的用户相似度，更准确度量消费者之间的相似关系。

设利用皮尔森相关系数[公式(2)]得到的相似矩阵为 $S = \{s_{ij}\}$，s_{ij}表示基于关键词评分得到的用户 i 和用户 j 相似性。在具体计算中，采用实验环境(Matlab 2012b)中提供的内置函数计算两个用户的皮尔森相关系数。然而用户—关键词矩阵的稀疏性导致得不到皮尔森相关系数，则将两个用户的相关系数置为0。

设用户相似度的社会网络修正系数为 ω_{ij}，基于社会网络修正系数对用户相似度的修正方法如公式(8)所示：

$$s_{ij}^{*} = \min(F(s_{ij}, \omega_{ij}), 1) \quad (8)$$

其中，$F(*)$是修订方法，$\min(*)$保证用户的相似度恒小于1。

1）若采用社会网络相似度修订用户相似度，则 $F(s_{ij}, \omega_{ij}) = s_{ij} \times \omega_{ij}$，即度量用户 i 和用户 j 在两种相似度方法下的共同相似度。

2）若采用社会网络距离相似度修订，则 $F(s_{ij}, \omega_{ij}) = s_{ij} / \omega_{ij}$，即用户 i 和用户 j 的距离越进，两者相似度越高。

3）若采用社会网络度中心性距离修订，则 $F(s_{ij}, \omega_{ij}) = s_{ij} \times \omega_{ij}$，即目标用户度越小，则越容易被高度数的用户影响。

修正用户相似矩阵是选择目标用户的近邻集。在实验中，通过对相似度排序，选出较大的 L 个用户作为目标用户 i 的近邻集 N_i。

5 实验与结果分析

本研究基于 C#.NET 平台构建微博非文本内容的预处理过程，构建用户向量模型，得到用户偏好矩阵，以 Matlab 2012b 为计算平台，实现推荐过程。

5.1 数据集及预测与评价方法

本文选择来自新浪微博(http://weibo.com)和来自 Epinion(http://www.epinion.com)的数据集进行实验，利用前者验证所提出框架的有效性。虽然 Epinion 不是微博应用，但具有社会化媒体典型特征，与微博较为相似，都以 UGC 和社会网络关系为核心要素，本文利用该数据集验证 SNCF-RM 能适用于多种社会网络结构环境。

新浪微博目前有5亿注册用户，其中活跃用户数4620万。实验数据收集过程中，选择一家企业用户作为数据收集的种子，截止至2012年10月12日，该企业关注580名用户，拥有1 361 080名粉丝，通过新浪微博开放 API 获取其两级关注用户数(即其关注的用户与其关注的用户关注的用户)，一共得到167 124名用户，平均每人关注287.1位用户。通过对剔除垃圾数据，一级用户(直接被该企业关注用户)剩下487位，社会网络稀疏度为91.63%。针对每名一级用户，抽取其最近100条微博，最后得到47 488条微博(有用户发布微博少于100条)。预处理包括分词以及关键词过滤两个过程。本文采用中国科学院的汉语词法分析系统 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)对文本 UGC 进行分词，其支持中文分词、词性标注等功能，分词准确度达到98.45%，满足实验要求。分词得到的名词被选作为关键词，共382 542个(未去重)，去重并剔除无意义关键词后，得到14 005个关键词，用户关键词矩阵的稀疏度是97.24%。

Epinion 是国外著名的产品点评网站，通过用户信任关系构建社会网络结构，数据集中包含4837名

用户,以及 3479 个被评论产品,用户的评分范围是[1,2,3,4,5]。Epinion 数据集中,用户—项目评分矩阵的稀疏度为 99.2%,用户社会网络关联矩阵的稀疏度为 98.22%。

为验证 SNCF-RM 的有效性,实验数据集随机分为训练数据集和测试数据集两个部分,其中训练集占 70%,测试数据集占 30%。为验证基于社会网络修正系数的 SNCF-RM 的有效性,平均绝对误差(MAE)被用于评价推荐系统的推荐效率,如公式(9)。

$$MAE = \sum_{i=1}^{n} \frac{| r_{k,i} - \hat{r}_{k,i} |}{n} \quad (9)$$

其中,r 是用户对关键词评分的真实值,$\hat{r}$ 是基于协同过滤推荐模型对的用户对关键词的评分的预测值,N 是用户数,n 是每个用户的预测关键词数。

5.2 对比方法

为验证所提出方法的有效性,分别使用如下 5 个推荐模型进行实验,并对比结果。

1)基准协同过滤推荐模型(记为 CF)。即采用 2.1 节中描述的基准方法对用户偏好进行用户偏好预测,实现关键词推荐过程;该方法只使用用户内容发布或点评行为信息进行推荐。

2)基于社会网络相似度的协同过滤推荐模型(记为 SN)。即采用社会网络相似度代替用户相似度矩阵,其他和基准协同过滤推荐模型一致,实现关键词推荐过程;该方法只使用社会网络结构信息进行推荐。

3)基于社会网络相似度修订的 SNCF-RM(记为 SNCF1)。即采用社会网络相似度矩阵修订用户相似度矩阵,其他和基准协同过滤推荐模型一致,实现关键词推荐过程。

4)基于社会网络距离修订的 SNCF-RM(记为 SNCF2)。即采用社会网络距离矩阵修订用户相似度矩阵,其他和基准协同过滤推荐模型一致,实现关键词推荐过程。

5)基于社会网络度中心性修订的 SNCF-RM(记为 SNCF3)。即采用社会网络度中心性修订用户相似度矩阵,其他和基准协同过滤推荐模型一致,实现关键词推荐过程。

5.3 参数选择实验

本文设计两个实验验证基于向量空间模型的用户 - 内容评分方法的有效性,并合理选择公式(4)中的参数取值。实验 1(图 2)在 $\beta = 0$ 的情况下,验证使用 IDF 屏蔽关键词对协同过滤的有效性,其中推荐的近邻集大小 $L = [10,20,\cdots,50]$。当 L 取不同值时,实验结果较为接近,为简洁起见,本文只给出 $L = 10$ 和 $L = 50$ 时的结果,如图 2 所示,上面是对词频进行归一化处理,使用 CF 所得到的结果,下面是未经过归一化得到的结果。实验显示选择参数 IDF 的过滤阈值 α 能过滤常用词,证明了常用词确实影响了协同过滤推荐效果。

IDF 度量关键词被全部用户使用的分布情况,当参数 IDF 过滤阈值 α 越大,常用关键词被过滤的概率越大,推荐过程较少受到常用词的影响;然而当过滤阈值 α 超过一定界限,具有偏好表达能力的关键词也会被过滤,从而影响推荐效果。图 2 中的 IDF 过滤最优值 α 是 2.4,下图中 IDF 过滤条件 α 最优值是 3.2,经过分析,认为经过归一化的结果更能反映实际情况,从而采用 IDF 过滤条件 $\alpha = 2.4$ 作为关键词过滤条件,并且实验测试了不同 L 值的情况,但 α 最优值并不随着 L 值变化而变化,具有鲁棒性。

实验 2 证明整合关键词词频和信息熵的用户偏好评分模型[公式(4)]的有效性,实验固定 IDF 取值。为了规避信息熵的修订系数取不同值使得评分的均值和方差发生改变,本实验将评分数据规范为 z-score 值(以 0 为均值,以 1 为方差)。为了保证参数选择的合理性,实验分别使用 CF 和 SNCF1 进行推荐实验,并且取近邻集大小 $L = \{10,15,\cdots,50\}$。因为多个实验得到的曲线较为接近,为简洁起见,图 3 仅显示 $L = 10$&CF 以及 $L = 50$&SNCF1 实验结果。

实验结果显示出推荐效率和 β 呈 U 型关系(图 3),随着 β 取值递增,MAE 递减,推荐效果上升,但超过一定阈值之后,β 值增加将使得 MAE 值增加,不利于推荐实现。由公式(4)可知,β 越大,关键词的信息熵占的权重越大,用户越偏好分布不均匀的关键词。图 3 显示在一定程度上增加对不均匀分布的关键词的偏好能有效提高推荐效率,因为用户为显示自身知识水平或独特性,会频繁使用部分关键词以提高其辨识度,但用户会减少使用其他用户常用的关键词,以规避“人云亦云”。信息熵虽度量了关键词词频的不均匀性,其并不包含特定用户的偏好信息,盲目 β 值,增加信息熵的权重,用户的偏好水平趋于一致,推荐过程无法体现出用户的个性化偏好,反过来抑制推荐效果。

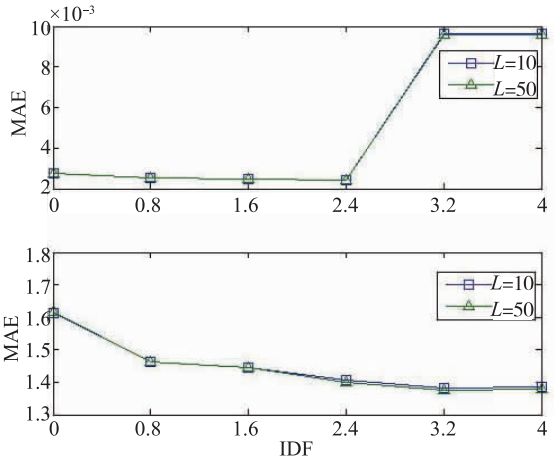


图 2　IDF 过滤

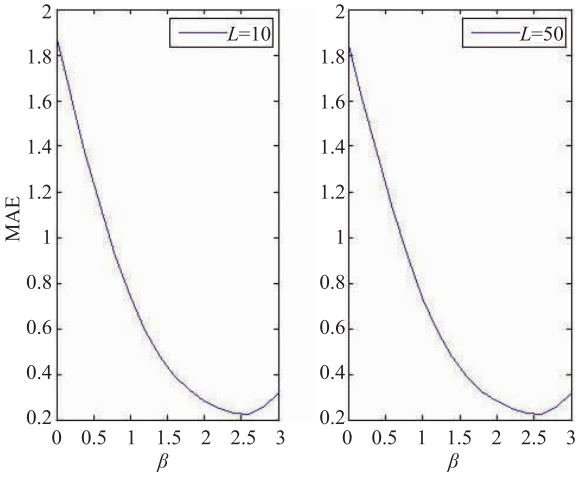


图 3　调节系数 β

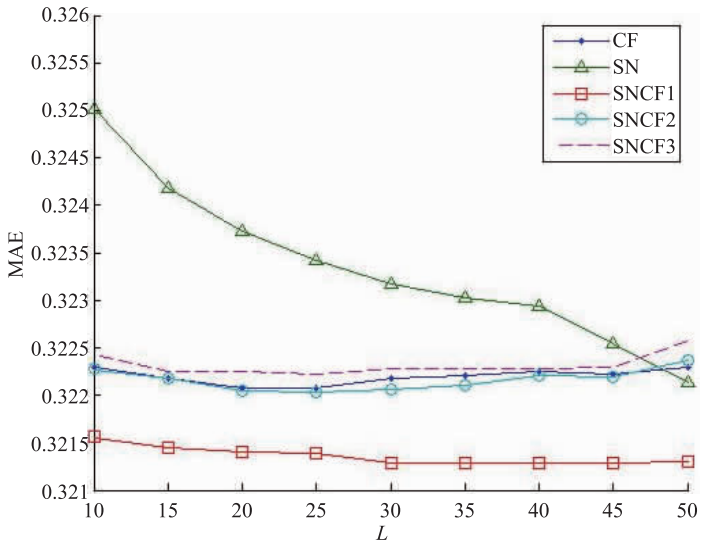


图 4　针对新浪微博数据集的实验

5.4　推荐实验

1)新浪微博数据集

本实验验证基于社会网络特征修订的协同过滤推荐方法的有效性,取 $L=\{10,15,\cdots,50\}$、IDF 屏蔽阈值 $\alpha=2.4$、信息熵调节系数 $\beta=2.6$,并将评分数据规范为 z-score 值(以 0 为均值,以 1 为方差),结果如图 4 所示。

本文采用 t 检验对多个协同过滤模型的推荐结果进行对比分析,结果如表 1 所示,采用 DW 检验也得到相同结果。在显著性水平 $p=0.01$ 下,拒绝 SNCF1 和 CF 推荐效果相同的假设,说明利用社会网络相似度修订的 SNCF-RM 的推荐效率优于基准协同过滤推荐模型,在推荐中考虑用户间社会网络结构是具有理论和实践价值。SNCF2 和 CF 结果差异不显著,但显著高于 SN 方法的推荐结果;SNCF3 劣于 CF,但 MAE 值差距不大;只考虑社会网络结构的推荐模型(SN)推荐效果显著劣于其他模型的推荐效果。本实验进一步研究近邻集大小对推荐结果的影响,近邻集大小 $L=30$ 时推荐效果较好,当 L 取较大值时,近邻集中用户的平均相似度降低,从而降低对用户的评分预测能力;当 L 过小,预测的随机性因素不能被消除,难以有效预测用户对关键词的评分。

2)Epinion 数据集

为了验证 SNCF-RM 能用于不同网络结构环境,本文利用来自在线评论网站 Epinion 的数据集进行推荐实验,实验结果如图 5。在本实验中,不仅仅使用 MAE 作为评价指标,进一步引入信息检索领域的准确率(Precesion)[图 5(b)]和召回率(Recall)[图 5(c)]评价不同模型的推荐效果,取被推荐对象个数 $N=[30,31,\cdots,50]$。

表 1　*t* 检验结果

Model	SN	SNCF1	SNCF2	SNCF3
CF	CF > SN (-3.1924 *)	CF < SNCF1 (19.4589 * *)	CF < SNCF2 (1.9560)	CF > SNCF3 (-4.7220 * *)
SN		SN < SNCF1 (7.6282 * *)	SN < SNCF2 (3.9914 *)	SN < SNCF3 (3.4400 *)
SNCF1			SNCF1 > SNCF2 (-17.1474 * *)	SNCF1 > SNCF3 (-19.9996 * *)
SNCF2				SNCF2 > SNCF3 (-7.5693 * *)

注：* 显著性水平 0.01；* * 显著性水平 0.001。

实验结果显示基于三种社会网络结构修正系数的 SNCF-RM 都取得比基准协同过滤(CF)较好的推荐效率。SNCF1 的推荐效果最好，具有最低的 MAE 值和最高的准确率和召回率。与基准模型(CF)对比，SNCF1 平均提高准确率 12.24%，以及召回率 12.23%；SNCF2 和 SNCF3 和基准模型(CF)对比也都显著提高了推荐效率。其中，与面向新浪微博数据的实验结果不同的是，SNCF2 和 SNCF3 的推荐效率都显著优于基准实验，这是因为 Epinion 中，用户更加关注其他用户评论的产品，用户社会网络带来的相似性更为显著。该实验正面所提出的 SNCF-RM 不仅仅能应用于本文特征下的微博环境，也能应用于其他网络环境，具有较大实践价值。

3)实验结果讨论

针对来自新浪微博和 Epinion 的数据集的推荐实验验证了本文提出的 SNCF-RM 是有效的，和基准协同过滤推荐模型对比，有效提高了推荐效率。社会网络关系和用户间相似性成正相关关系，协同过滤推荐通过用户之间相似度来刻画用户行为的相关性，高相似性的用户对关键词产生相近的偏好。故而，结合不同社会网络指标修正用户评分矩阵会对协同过滤推荐效果产生不同影响，在协同过滤推荐中整合社会网络关系信息能提高协同过滤推荐效率(图 4、图 5)。

首先，基于社会网络相似性修订方法(SNCF1)显著提高了协同过滤推荐效率。社会网络相似性从用户关注行为层面刻画了用户的相关性，同时信息在社会网络中的传播过程使用户接收到来自其关注对象的信息，通过转发或者发布新内容等方式进行呼应，进一步加强了用户间的相似性；

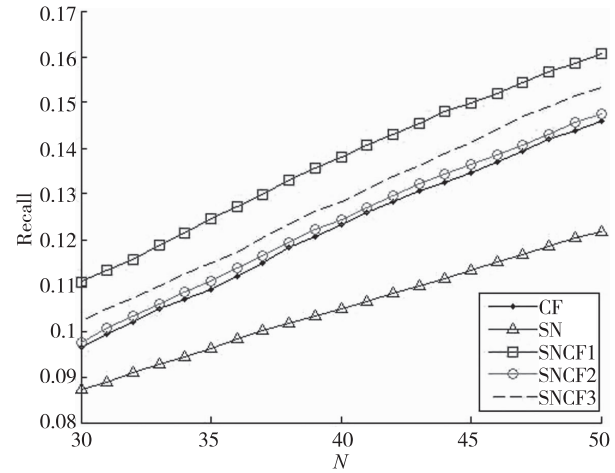
其次，在使用新浪微博数据集的实验中，基于社会网络距离修订方法(SNCF2)取得了和基准协同过滤推荐模型(CF)相近的效率，*t* 检验不能拒绝两者无差异的假设，但在使用 Epinion 的数据集中，SNCF2 优于 CF。社会网络距离也度量了用户的相似性，体现出用户偏好的相关特征；

然后，基于度中心性修订方法(SNCF3)强调高度中心性用户的作用，强调弱关系，在偏好高度中心性内容的社会网络(如 Epinion)中取得效果显著。对于更强调人际交往(强关系)的社会网络(新浪微博)则效果较差。度中心性重视拥有较多社会网络连接的用户。SNCF3 增强了名人效应，使意见领袖发布的内容更容易被推荐给目标用户，增强意见领袖在微博中对用户观点、意见的引导作用，对于企业微博营销实践，如社会化广告等，具有一定指导意义。

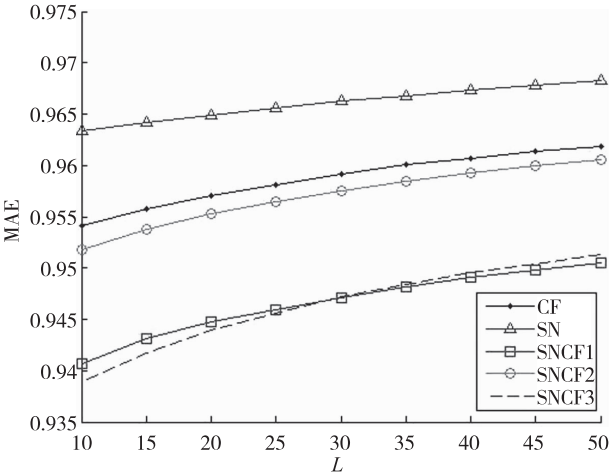
最后，基于近邻集 *L* 的对比显示，近邻集的选择有效影响推荐效率。当近邻集较大时，相似性较低的用户被攘括进来，从而不能准确预测用户偏好；当近邻集较小时，虽然用户相似性很高，但是多种未知变量对推荐的影响难以把握，预测过程中存在大量不确定性。通过扩大近邻集能在一定基数条件下抵消这些变量带来的不确定性，从而选择合适的近邻集大小能有效提高推荐效率。

6　结　论

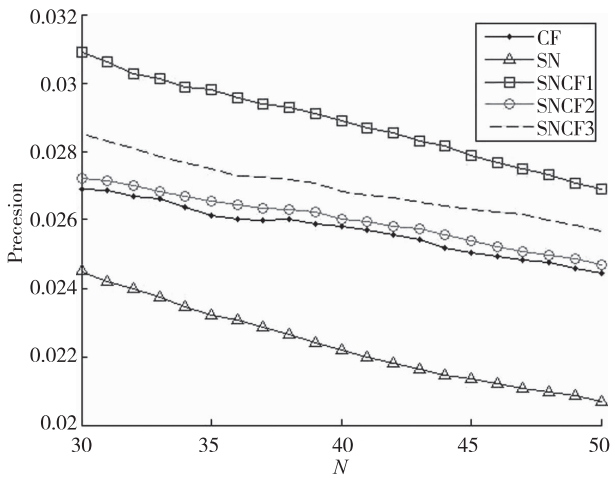
本文针对微博平台社会网络关系复杂、UGC 碎



(a)MAE



(b) 准确率



(c) 召回率

图 5　针对 Epinion 数据集的实验

片化等特征,提出 SNCF-RM,从关键词层面设计微博推荐系统,引入社会网络相似度修订系数修订用户相似度矩阵,有效协同过滤推荐模型中整合社会网络关系。从实验结果来看,SNCF-RM 帮助用户快速有效获取其感兴趣信息,较基准协同过滤推荐模型有更高的推荐精度。本文主要做出如下四方面贡献:①采用协同过滤推荐为背景,验证了用户社会网络关系对用户偏好的影响,并给出在推荐中整合社会网络关系的方法;②从关键词层面解决了推荐的冷启动问题,冷启动是推荐研究重点解决的问题,基于自然语言处理将 UGC 处理为关键词集合能有效代表用户偏好;③从信息检索研究中引入向量空间模型以描述用户偏好,提高用户偏好表达能力;④合理选择被推荐关键词,过滤对推荐无意义的关键词能有效描述微博用户偏好,并提高推荐效率。

研究结果对微博运营和企业微博营销策略有启示意义:①企业微博营销策略应重视用户间社会网络关系对用户偏好的影响,利用社会网络关系提高对用户的识别能力;②企业作为微博中的用户时,应当尽量吸引相似的用户,提高企业的网络密度,即提高与其存在社会网络关系的用户之间的链接数,促使自己的信息能够在小范围内快速传播,并有利于提高推荐效率;③企业应当有效选择潜在关键词作为被推荐的目标关键词,构建和企业产品、品牌相关的关键词库,从而刻画针对企业的用户偏好空间。未来的研究工作将一方面从动态视角处理用户网络关系,而不是作为静态网络,进一步提高推荐效率,另一方面从社会学、心理学出发,研究用户参与微博的不同动机对推荐系统的影响机制,使推荐结构满足不同的用户需求。

参　考　文　献

[1] Java A. Mining Social Media Communities and Content [D]. United States-Maryland: University of Maryland, Baltimore County, 2008.

[2] 冯芷艳,郭迅华,曾大军,等. 大数据背景下商务管理研究若干前沿课题[J]. 管理科学学报, 2013, 16(1): 1-9.

[3] Shang M S, Zhang Z K, Zhou T, et al. Collaborative filtering with diffusion-based similarity on tripartite graphs[J]. Physica A: Statistical Mechanics and its Applications, 2010, 389(6): 1259-1264.

[4] Zhang Z K, Zhou T, Zhang Y C. Personalized recommendation via integrated diffusion on User-Item-Tag tripartite graphs[J]. Physica A: Statistical Mechanics and its Applications,2010,389(1):179-186.

[5] Dao T H, Jeong S R, Ahn H. A novel recommendation model of location-based advertising: context-aware collaborative filtering using ga approach[J]. Expert Systems with Applications,2012, 39(3): 3731-3739.

[6] Lee S K, Cho Y H, Kim S H. Collaborative filtering with

ordinal scale-based implicit ratings for mobile music recommendations[J]. Information Sciences, 2010, 180(11): 2142-2155.

[7] Jannach D, Zanker M, Felfernig A, et al. Recommender systems: An Introduction[M]. Cambridge Cambridge University Press, 2010.

[8] Resnick P, Iacovou N, Suchak M, et al. Grouplens: An Open Architecture for Collaborative Filtering of Netnews[C]. Chapel Hill, North Carolina, United States: ACM, 1994.

[9] Lawrence R D, Almasi G S, Kotlyar V, et al. Personalization of Supermarket Product Recommendations[M]. BerlinSpringer, 2001.

[10] 王丽莎,张绍武,林鸿飞. 基于项目和标签的随机游走个性化信息推荐模型[J]. 情报学报, 2012, 31(3): 289-296.

[11] 朱国玮,周利. 基于遗忘函数和领域最近邻的混合推荐研究[J]. 管理科学学报, 2012, 15(5): 55-64.

[12] Ting I, Chang P S, Wang S. Understanding microblog users for social recommendation based on social networks analysis[J]. Journal of Universal Computer Science, 2012, 18(4): 554-576.

[13] Mcpherson M, Smith-Lovin L, Cook J M. Birds of a feather: homophily in social networks[J]. Annual Review of Sociology, 2001: 415-444.

[14] Zeng X, Wei L. Social ties and user content generation: evidence from flickr[J]. Information Systems Research, 2013, 24(1): 52-70.

[15] 蔡淑琴,胡慕海,叶波,等. 情境化推荐中基于超图模式的用户偏好漂移识别研究[J]. 情报学报, 2011, 30(8): 802-811.

[16] 王茜,杨莉云,杨德礼. 面向用户偏好的属性值评分分布协同过滤算法[J]. 系统工程学报, 2010(4): 561-568..

[17] Xu Y H, Guo X T, Hao J X, et al. Combining social network and semantic concept analysis for personalized academic researcher recommendation[J]. Decision Support Systems, 2012, 54(1): 564-573.

[18] Sinha R, Swearingen K. Comparing recommendations made by online systems and friends[C]. Dublin, Ireland, 2001.

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

[20] Choi S M, Ko S K, Han Y S. A movie recommendation algorithm based on genre correlations[J]. Expert Systems with Applications, 2012, 39(9): 8079-8085.

[21] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Commun. ACM, 1975, 18(11): 613-620.

[22] Aizawa A. An information-theoretic perspective of Tf-Idf measures[J]. Information Processing&Management, 2003, 39(1): 45-65.

[23] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24(5): 513-523.

[24] Yang W, Dia J. Discovering cohesive subgroups from social networks for targeted advertising[J]. Expert Systems with Applications, 2008, 34(3): 2029-2038.

(责任编辑　车　尧)