doi:10.3969/j.issn.1000-436x.2014.02.003

基于用户相似度的协同过滤推荐算法

荣辉桂[1]，火生旭[1]，胡春华[2]，莫进侠[1]
（1. 湖南大学 信息科学与工程学院，湖南 长沙 410082；2. 湖南商学院 计算机与信息工程学院，湖南 长沙 410205）

摘 要：协同过滤推荐算法通过研究用户的喜好，实现从海量数据资源中为用户推荐其感兴趣的内容，在电子商务中得到了广泛的应用。然而，当此类算法应用到社交网络时，传统的评价指标与相似度计算的重点发生了变化，从而出现推荐算法效率偏低，推荐准确度下降问题，导致社交网络中用户交友推荐满意度偏低。针对这一问题，引入用户相似度概念，定义社交网络中属性相似度，相似度构成与计算方法，提出一种改进的协同过滤推荐算法，并给出推荐质量与用户满意度评价方法。实验结果表明：改进算法能有效改善社交网络中的推荐准确性并提高推荐效率，全面提高用户满意度。

关键词：协同过滤；用户相似度；属性相似度；互动相似度；用户满意度
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2014)02-0016-09

User similarity-based collaborative filtering recommendation algorithm

RONG Hui-gui[1], HUO Sheng-xu[1] , HU Chun-hua[2], MO Jin-xia[1]
(1. School of Information Science and Engineering, Hunan University, Changsha 410082, China;
2. School of Computer and Information Engineering, Hunan University Commerce, Changsha 410205, China)

Abstract: Collaborative filtering recommendation algorithms widely used in e-commerce, recommend interesting content for users from massive data resources by studying their preferences and interests. The focus of similarity and evaluation have been changed when applied to social networks, however, they cause low efficiency and accuracy of the recommendation algorithms. User similarity was introduced for redefining the attribute similarity and similarity composition as well as the method of similarity calculating, then a new collaborative filtering recommendation algorithm based on user attributes was designed and some methods for user satisfaction and quality of recommendations were presented. The experimental result shows that the new algorithm can effectively improve the accuracy, quality and user satisfaction of recommendation system in social networks.

Key words: collaborative filtering; user similarity; attribute similarity; interactive similarity; user satisfaction

1 引言

随着互联网的发展，数据资源每天以几何数量级增加，为解决用户复杂的需求和庞大数据之间的矛盾，个性化推荐系统应运而生，其应用日益广泛[1]。个性化推荐技术通过研究用户的喜好和兴趣，为用户推荐其所需的各种资源，最初应用于电子商务个性化服务中[2]。随着社交网络的兴起，个性化推荐技术也在社交网络中得到了广泛的应用。与传统的基于内容过滤的直接分析内容进行推荐不同，协同

收稿日期：2013-09-28；修回日期：2014-01-05
基金项目：国家自然科学基金资助项目(61273232, 61304184)；湖南大学“青年教师成长计划”基金资助项目(531107021115)；教育部新世纪优秀人才支持计划基金资助项目(NCET-13-0785)；教育部人文社会科学研究青年基金资助项目(10YJC630080)；湖南省自然科学基金资助项目(11JJ2033)；湖南省教育厅重点科研基金资助项目(11A062)
Foundation Items: The National Natural Science Foundation of China (61273232, 61304184); The Development Plan for Young People of Hunan University (531107021115); The Program for New Century Excellent Talents in University(NCET-13-0785); The Science Foundation of Ministry of Education of China (10YJC630080); The Natural Science Foundation of Hunan Province (11JJ2033); The Major Scientific Research Fund of Hunan Provincial Education Department of China(11A062)

过滤分析用户的兴趣，在用户群中找出与目标用户相似的用户，综合这些相似用户对不同项目的评分，产生目标用户对这些项目喜好程度的预测，从而产生推荐[2]。

目前，主流协同过滤推荐算法分为 2 类：基于用户的协同过滤推荐算法[1]和基于项目的协同过滤推荐算法[3,4]。基于用户的协同过滤推荐算法根据用户对项目的评分矩阵，计算用户之间的相似度，找出目标用户的最邻近邻居集合，最后，对最近邻居集合进行加权，从而产生目标用户的推荐集。此类算法能够有效地使用其他相似用户的反馈信息，为用户产生推荐。但是由于用户涉及的信息量相当有限，用户对项目的评分相对稀少，造成评分矩阵相对稀疏，数据冷启动问题严重，难以找到相似用户集，在这种情况下，仅使用少量评价数据不可能产生精确推荐，大大降低了推荐系统有效性。基于项目的协同过滤推荐算法根据对用户已评分项目相似项目的评分进行预测，从某种程度上减少了评分矩阵稀疏性和冷启动问题对推荐质量的影响。虽然项目间相似性相对稳定，但用户的喜好和兴趣是不断变化的，推荐集覆盖率较低，此类算法也没有提出有效解决这一问题的方法，用户对推荐的满意度较低。

虽然协同过滤推荐算法在信息过滤方面呈现出了极大的优势，但随着电子商务和社交网络的快速发展和相互间的不断融合，算法在不同领域中的应用也凸显出一些问题：①冷启动问题；②稀疏性问题；③最初评价问题。社交网络包含用户的基本资料信息的同时，也包含大量用户交互、互动行为信息，如何有效利用这 2 类信息为用户产生推荐，也成为个性化推荐研究的一个重要议题。

为解决这些问题，文献[6]中，在弱关系的微博类社交网络中，采用基于用户聚类的方法，提出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，有效缓解了矩阵稀疏性和冷启动问题。文献[7]里提出了一种递归预测算法，该算法让那些最近邻的用户加入到预测处理中，即使他们没有对给定的项目进行评分。对所需评分值不明确的用户，预测它的递归，整合到预测过程中。此方法用另一种方式缓解了矩阵稀疏对推荐质量的影响，提供了推荐精度。Alfred 等在文献[8]中提出了要利用社交网络中隐式的用户间互动数据为用户产生推荐。在下面情形下：信息接收方可能会拒绝用户初始阶段发送的互动信息，一些用户接收到大量并不期望接受的信息，降低了用户满意度，文献中提出的方法解决了这一问题。Brŏzovský L 和 Petřiček 在文献[9]中，通过估算目标用户与用户间的吸引度，针对几种协同过滤推荐算法在社交网络中的应用进行了评价，分析了几种算法的优缺点。文献[10]中深入分析了交友约会、招聘等网站的特性，利用用户个人信息，对个人信息进行分类取值，找到用户的喜好，为用户产生推荐。这种推荐方式在用户个人信息不足或者很少的情况下，不能为目标用户产生满意的推荐。Tingting Wang 等在文献[11]中预测新用户的行为，文献中对用户进行分类，把用户行为分为浏览、点击、发信 3 类，对 3 类行为给予不同权值；利用用户信息计算新用户相似度，根据与新用户相似用户的行为分析，预测新用户的行为；但是没有考虑初始的用户行为，在用户行为较少的情况下，不能产生良好的推荐。文献[12]中，在交友约会网站中，利用社会图谱，根据用户的个人信息、喜好及对推荐集用户的匹配要求，对用户进行聚类，利用 SimRank 对产生的推荐进行排序，为用户产生推荐，相比其他协同过滤推荐算法，这种方法取得了较好的推荐结果和用户满意度。文献[13]中，利用用户个人信息和互动信息，估算出满足用户喜好的配对模型，根据这一模型，使用 Gale-Shapley 算法预测满足用户喜好模型的用户，为用户产生推荐。RichiNayak 等在文献[14]中利用用户在过去联系过的用户间的社会关系，来预测新用户之间的社会关系，计算相似度，为用户产生推荐，这种方法未能解决矩阵稀疏问题。

基于搜索匹配和用户属性的推荐系统已经广泛应用于社交网络中，但此类推荐系统有很大的限制，一些用户得到了很多并不期望的推荐，一些用户仅仅得到很少的推荐。基于内容的推荐算法可以根据用户的类别、标签等信息缓解数据冷启动问题，但往往推荐精度不够高。社交网络中，庞大的用户之间存在某种关系，把这种关系划分为显式关系和隐式关系。显式关系是指用户间明确的建立并确认了关系，隐式关系是用户间尚未确立关系。在显式信息不足的情况下，有效地使用隐式信息可提高推荐系统的精确度。这些方法从一定程度上减少了矩阵稀疏性和冷启动问题对推荐算法的影响，但没有从根本上解决协同过滤推荐算法在社交网络

中的实际应用。

前述研究表明，社交网络中用户属性和互动信息仍未能充分利用，推荐效率与准确度偏低，可见现有推荐算法难以满足日益复杂的社交网络的推荐需求。针对这一问题，本文引入用户相似度概念，重新定义社交网络中相似度属性，相似度构成及其计算方法，提出一种改进的协同过滤推荐算法，并给出推荐质量与用户满意度评价方法。

2 用户相似度定义及描述

传统的相似度有皮尔逊相关系数法、向量余弦法、调整的向量余弦法、约束的皮尔逊相关系数法、斯皮尔曼相关系数法等，在不同的应用领域中，选取不同的相似度计算方法。由于社交网络的特殊场景，本文重新定义了相似度及其计算方法。算法中的相似度由 2 部分构成：一部分是由用户属性决定的用户属性相似度，通过计算用户间的距离 $D_{\mathrm{A-B}}$ 度量，距离值越小，用户间的属性相似程度越高；另一部分由用户间的互动信息决定互动相似度，其计算与目标用户相似发件人和收件人的用户数度量，值越大，用户间的互动相似程度越高。最后将 2 部分相似度进行线性拟合，计算得出用户间总相似度。

2.1 用户属性相似度及计算

社交网络中用户属性包括用户个人信息和其他选填项，另外在交友网站中，用户需要填写自己理想对象的匹配条件，以便得到更好的推荐。用户属性分为 2 类：一类是数值型属性（如年龄、身高、收入等）；一类是名称型的属性（如体型、教育水平、婚否等）。对于数值型属性，计算不同用户之间数值型属性的绝对差值 $\|d\|=\left|Attr_{\mathrm{A}}-Attr_{\mathrm{B}}\right|$。不同属性绝对差值的最小和最大差距为 $[\xi_1,\xi_n]$，将这个区间平均划分成 $n-1$ 个等距的小区间：$\{[\xi_1,\xi_2],[\xi_2,\xi_3]\cdots[\xi_{n-1},\xi_n];\xi\in[0,+\infty]\}$，当用户间的数值型属性的绝对差值落在其中的某个小区间，对每个小区间依次给定数值型属性距离{0，1，2，…，$n-1$，n}（这里只划定 3 个区间），针对不同的区间，得到用户间的数值型属性距离 D_{Num}；对于每个名称型属性，根据每个名称型属性先前设定的取值数 N，确定编码的位数 $n=\mathrm{lb}N$，然后对不同的取值进行格雷编码并依次串连起来，计算不同用户间格雷编码之间的海明距离，得到不同用户间的名称型属性距离 D_H。

若定义用户 A 与 B 间的距离来度量用户间的属性相似度，每一个属性的权重为 ω_i，则所有属性权重值满足 $\sum_{i=1}^{n}\omega_i=1$。

1) 对于数值型的属性距离 D_{Num}，根据前面的说明，定义不同的取值区间：

若 $\xi\in[\xi_1,\xi_2]$，则 $d_{Num}=0$；

若 $\xi\in[\xi_2,\xi_3]$，则 $d_{Num}=1$；

…

若 $\xi\in[\xi_{n-1},\xi_n]$，则 $d_{Num}=n-1$。

针对数值属性，用户间的距离计算为

$$D_{Num}=\sum_{i=1}^{n}\omega_i d_i \tag{1}$$

一般情况下，只划分为 3 个区间。

若 $0\le\left|Attr_{\mathrm{A}}-Attr_{\mathrm{B}}\right|\le 5$，则 $d_{num}=0$；

若 $5<\left|Attr_{\mathrm{A}}-Attr_{\mathrm{B}}\right|\le 10$，则 $d_{num}=1$；

若 $\left|Attr_{\mathrm{A}}-Attr_{\mathrm{B}}\right|>10$，则 $d_{num}=2$。

2) 对于名称型的属性距离 d_{Num}，对不同的取值进行编码。

因名称型属性的取值范围较为单一，可采用二进制编码来表示，比如名称型属性中体型可描述为：瘦、匀称、胖，对应的二进制编码可描述为：00、01、11，其他属性可以此类推。最终将用户的全部名称型属性编码串联起来，形成一个二进制串 B_{Nom}；采用计算二进制串 B_{Nom} 的海明距离来度量用户间名称型属性的距离；其权重是 n 个名称型属性的权重的平均值：$\overline{\omega_{Nom}}=\dfrac{\sum_{i=1}^{n}\omega_i}{n}$。

则

$$D_H=\overline{\omega_{Nom}}D_{HM}(D_{Nom\mathrm{A}},D_{Nom\mathrm{B}}) \tag{2}$$

3) 最终得到 2 个用户 A 与 B 的信息属性距离为

$$D_{\mathrm{A-B}}=D_{Num}+D_H \tag{3}$$

即 $D_{\mathrm{A-B}}=\sum_{i=1}^{n}\omega_i d_i+\overline{\omega_{Nom}}D_{HM}$。

$D_{\mathrm{A-B}}$ 越小，相似度越大，而 $D_{\mathrm{A-B}}$ 越大，相似度越小）。计算示例：若 A={23，183 cm, 0101000000}，B={26, 176 cm, 1100010100}，则用户 A 与 B 间的距离 $D_{\mathrm{A-B}}$=1(0+1+4)=5。

2.2 用户互动相似度及计算

社交网络中用户行为存在多种情况，如用户间的信息浏览、信息互发、收取信息与拒绝信息等。

为适应社交网络的特殊场景，算法中重点考虑积极、成功的互动（如果用户 U 给用户 V 发送了信息，同时，U 也收到了 V 的回复）。因此，互动相似度可定义为：如果发信人 S_1 和 S_2 都给收信人 R_1 和 R_2 发送了信息，则 R_1 和 R_2 是相似的发信人，相似度为相同的收件人数量；同理，如果收件人 R_1 和 R_2 收到发信人 S_1 和 S_2 发送的信息，则 R_1 和 R_2 为相似的收信人，相似度为收件人 R_1 和 R_2 相同发件人的数量。互动相似度可用数学定义如下。

相似发信人：$U_1 \overset{S}{\sim} U_2 : \exists u(U_1 \to u \cap U_2 \to u)$

相似收信人：$U_1 \overset{R}{\sim} U_2 : \exists u(u \to U_1 \cap u \to U_2)$

其中，$U_1 \to U_2$ 表示 U_1 向 U_2 发送了消息。

1) 发信相似度 S_S：

$Sim_S=Num$ (用户 U_1 与 U_2 同时向相同用户 u 发送信息的用户数)

2) 收信相似度 S_R

$Sim_R=Num$ (用户 U_1 与 U_2 同时向相同用户 u 发送信息的用户数)

3) 用户 A 与 B 的互动相似度 S_I

根据用户间的互动信息，相似的发信人和相似的收信人，计算用户间互动相似度

$$Sim_I = Sim_S + Sim_R \tag{4}$$

其中，若用户 U_1 与 U_2，相似的发信人又是相似的收信人，则直接将这些用户 C_{S-R} 加入到推荐集中。

3　引入用户相似度的协同过滤推荐算法

用户相似度是将用户属性相似度和互动相似度 2 部分相似度进行线性拟合并计算得到。社交网络中，大量用户只填写必须的信息，用户信息缺失相对严重，用户间产生互动信息相对较少。因此，为产生较好的推荐集，算法应结合实际情况，2 部分相似度权重的定义应该有所不同。

3.1　用户相似度计算

根据前述说明，社交网络中的用户信息由用户属性和用户互动（行为）信息构成。在社交网络的不同的应用场景下，用户属性相似度（用户间距离）$D_{\mathrm{A-B}}$ 和用户互动相似度 Sim_I 对于总体相似度的影响不同，所对应的权重 α 与 β 的取值不同，可根据实际应用进行设置。对于计算出的 2 个子相似度进行线性拟合，计算得出用户间的相似度 $Sim_{\mathrm{A-B}}$。若 $Sim_{\mathrm{A-B}}$ 值越小，说明用户间的相似程度越高；若 $Sim_{\mathrm{A-B}}$ 值越大，则说明用户间的相似程度越低。

用户 A 与 B 的总相似度

$$Sim_{\mathrm{A-B}} = \alpha D_{\mathrm{A\sim B}} + \beta \frac{1}{Sim_I} \tag{5}$$

即

$$Sim_{\mathrm{A-B}} = \alpha(\sum\nolimits_{i=1}^{n} \omega_i d_i + \overline{\omega_{Nom}} D_{HM}) + \beta \frac{1}{Sim_I}$$

其中，α 与 β 是 2 个子相似度在用户间相似度中的权重，满足 $\alpha+\beta=1$。$Sim_{\mathrm{A-B}}$ 越小，用户 A 与 B 之间的相似度越大。

3.2　基于用户相似度的协同过滤推荐算法

综合前面的论述，算法 1 给出了为目标用户 U_0 产生推荐集合的过程。

算法 1　A&I_CF(U_0, U, int N)

//算法为目标用户 U_0 产生其推荐集 C；

//算法最后输出目标用户 U_0 的推荐集 C。

输入：目标用户 U_0，备选用户集 U，产生推荐个数 N。

Begin：

1) 相似度计算

① 用户属性相似度计算。

根据 $D_{Num} = \sum_{i=1}^{n} \omega_i d_i$ 计算用户 U_0 的数值型属性距离 D_{Num}。

对于用户 U_0 名称型属性，对属性取值进行格雷编码，将用户名称型取值格雷码串连，计算出海明距离 D_H。

根据 $D_{\mathrm{A-B}} = D_{Num} + D_H$ 计算用户 U_0 与其他用户间的距离，用来度量用户 U_0 与其他用户间属性相似度，即用户间的距离 $D_{\mathrm{A-B}}$。

② 用户互动相似度计算。

找到与用户 U_0 相似的发信人用户，并统计其数量。

找到与用户 U_0 相似的收件人用户，并统计其数量。

根据 $Sim_I = Sim_S + Sim_R$ 计算出 U_0 互动相似度 Sim_I。

③ 根据 $Sim_{\mathrm{A-B}} = \alpha D_{\mathrm{A-B}} + \beta \frac{1}{Sim_I}$ 计算得出用户 U_0 与其他用户间的总相似度 $Sim_{\mathrm{A-B}}$。

2) 产生推荐集

① 确定候选集 C。根据用户间的互动信息，找出和目标用户 U 相似用户 $\{U_1, U_2, \cdots, U_n\}$ 产生互动的用户集 $\{C_{U_1}, C_{U_2}, \cdots, C_{U_n}\}$。

② 产生推荐集。上一步产生的候选集求并集

$C = C_{U_1} \cup C_{U_2} \cdots C_{U_n}$，将候选集 C 里的用户按照相似度进行排序，得出最后的推荐集合 $R = R_{C\_Ranked} \cup C_{S-R}$，按照 Top-$N$ 排序算法为用户产生推荐。

输出：目标用户 U_0 的推荐集 $C = C_{U_1} \cup C_{U_2} \cdots C_{U_n}$。

End

基于用户相似度的系统过滤推荐算法通过计算用户相似度，计算得到用户相似度值越小，表明用户间相似程度越高，按照相似度降序对用户排序，产生推荐候选集，再使用 Top-N 方法取到候选集排在前 N 位的用户推荐给目标用户。

3.3 算法复杂度分析

算法复杂度是衡量算法效率的标准，通常可分为时间复杂度和空间复杂度。随科技的发展，算法执行所需的存储空间对于算法的影响逐渐弱化。

通过对上述用户相似度算法分析，算法执行过程只需要存储用户属性信息、交互信息、推荐集信息，存储空间的占用较小；且随用户的增加，存储空间线性增加，数量级上没有变化；此外，当前硬件发展使得较小的代价即可获得较大的存储容量，因此，该算法中时间复杂度成为衡量算法效率的重点，本文聚焦于此算法的时间复杂度分析。

该算法执行的时间开销集中在相似度计算公式 $Sim_{\mathrm{A-B}} = \alpha(\sum_{i=1}^{n} \omega_i d_i + \overline{\omega_{Nom}} D_{HM}) + \beta \frac{1}{Sim_I}$ 中，可见算法时间复杂度由相加的 2 部分构成。若用户属性中有 M 个数值型属性和 N 个名称型属性，用户集数量级为 n。则有以下分析：

for(in i=0; i<n; i++){ //执行 n+1 次

for(int j=0; j<n; j++){ //执行 n (n+1)次

//用户属性相似度计算

a:数值型 //执行 $M \times n$ (n+1)次

b:名称型 //执行 lb $N \times n$(n+1)次

a+b //执行 n (n+1)次

//用户交互相似度计算

c:发信相似度 //执行 n (n+1)次

d:收信相似度 //执行 n (n+1)次

c+d //执行 n (n+1)次

//相似度线性拟合

$\alpha(a+b)+\beta(c+d)$ //执行 n (n+1)次

}

//按照用户间的相似度对用户进行排序

Rank_Users By Similarity //执行 n+1 次

}

//为用户产生推荐

Generate Recommendations //执行 N_0 次

算法执行总次数：$f(n)=n+1+n(n+1)+M \times n(n+1)+\mathrm{lb}N \times n(n+1)+n(n+1)+n(n+1)+n(n+1)+n(n+1)+n+1+N_0$。

$f(n)=(M+\mathrm{lb}\ N+6) \times n^2+(M+\mathrm{lb}\ N) \times n+(N_0+2)$

其中，M、N、α、β、N_0 均为常量，利用时间复杂度计算原则，忽略常量、低次幂和最高次幂的系数，计算得出算法的时间复杂度：

$$T(n)=T(n^2)+T(n)+T(1)$$

$$T(n)=O(n^2)$$

从上述分析可知：本文中算法时间复杂度在 $O(n^2)$内，在不增加额外存储空间的前提下，其时间复杂度与文中引用 2 个经典推荐算法处于同一数量级，未增加过多的时间开销。

4 仿真实验与性能分析

4.1 仿真实验环境设置

实验运行在 Apache Mahout 开源项目基础之上，该开源平台的主要目标是创建一些可伸缩的机器学习算法，包含聚类、协同过滤、分词分类、集群等算法应用。利用 Apache 提供的工具，通过 Taste 库建立一个推荐引擎，Taste 是基于用户和基于项目的推荐，并且提供了许多推荐选项，以及用户自定义的界面[15]。

为检验本文提出算法的有效性，实验环境贴近真实随机、复杂网络的社交网。以反映算法在真实环境中的有效性为目标，设置了如下的实验环境和实验过程。

1) 利用随机方法生成用户属性及其取值，形成用户属性相似度，实验数据来自真实的社交网络。

2) 获取训练集，训练集搜集用户在过去一个月的数据。测试集的数据源于训练集中最后一周的用户数据，并且测试集的用户是训练集中活跃用户，保证数据的真实有效性。

3) 对用户属性和互动信息进行统计分析，得出用户间互动信息的分布，互动信息的汇总为算法度量提供数据。

4)利用统计得出的数据和评价标准度量算法有效性。

可见，算法的实验数据使用社交网络中具有代表性网站用户的真实历史数据，最大限度的模拟和描述真实社交网络应用情景。尽管实际社交网络应

用场景比实验描述的环境要复杂，但以上实验设置基本描述了当前社交网络实体的主要情况，基本符合社交网络的真实情况。本文中提出的算法在时间复杂度上与传统经典算法相比较，处于同一数量级，故可认为三者的执行效率是基本一致；且实验环境中算法是线下执行并生成推荐，初步可不考虑执行时间对算法有效性的影响。

以下通过仿真实验就本文提出的算法在社交网络中的应用进行验证，通过对协同过滤推荐算法（collaborative filtering recommendation algorithm）、基于互动的推荐算法（interaction-based, recommendation algorithm based interaction）、基于用户相似度的协同过滤推荐算法（A&I-based, recommendation algorithm based on attribute and interaction）3 种算法的基线成功率、成功率、召回率、覆盖率等评价指标对算法进行比较与分析，得到较真实的度量评价，为社交网络中推荐算法选择与应用打下基础。

4.2　测试数据获取

1) 获取用户数值型及名称型属性

由于目前社交网络中用户个人信息的获取比较困难，但信息相对简单，故实验时选取随机函数产生确定的用户属性取值，也能反映用户属性的平均状况，最大限度地逼近真实的使用环境。实验利用随机函数产生的用户属性值计算用户属性相似度，最终和用户行为相似度拟合为用户相似度。

建立 3 个数据字典$\{U_{Num},U_{Nom},\cdots,User\}$，分别表示用户数值型属性、名称型属性、用户属性。针对前两者，对每个子集建立属性取值字典：①数值型属性$V_{Num_i}=v_{Num_1},v_{Num_2},\cdots,v_{Num_j}$；②名称型属性$V_{Nom_i}=v_{Nom_1},v_{Nom_2},\cdots,v_{Nom_j}$。随机生成用户属性取值后，形成用户属性取值：$User_i=\{V_1,V_2,\cdots,V_j\}$，进而得到用户属性相似度。

2) 获取用户交互数据

测试数据选用目前流行的在线交友网站的真实历史数据，该数据集包含不记名用户的用户属性和用户互动信息，训练集搜集过去 4 周所有的互动信息，并剔除那些受欢迎用户的互动。测试集搜集的互动信息是紧接着训练集后的 6 天的数据，测试集里的用户是训练集中活跃的用户，即在训练期内与其他用户产生了互动信息。最后，产生了训练集 700 000 条互动信息，测试集 120 000 条互动信息；训练集大概有 60 000 个发信人用户，110 000 收信人用户；测试集有 25 000 个发信人用户，47 000 个收信人用户。

为构建候选集，需要用到初始的训练集，然后使用训练集的一个子集来评价推荐算法的质量。选用的训练集包含 4 周内的 1.3 亿条互动信息，从其中获取的测试集大约包含 300 000 条用户间的互动信息。对 4 周的训练集进行遴选，第 4 周的数据质量高一些，因为随着时间的推移，用户间互动信息的数量会增加。

由于受欢迎的用户对其他用户的回复往往是消极的，为使度量更合理、有效，在测试集中剔除那些受欢迎的用户。这里，定义在过去的一个月收到多余 50 条互动信息的用户为受欢迎的用户：$Pop_{\text{user}}:(R_{Msg})_{30d}\geq 30$。通过统计得到下面用户行为分析结果如表 1 所示，作为后续计算度量标准的源数据。

表 1　　实验数据用户分布

数据类型	发信人	收信人	S+	R+	R+S+
相似用户集	7 789 252	5 785 978	616 568	224 066	9 680
孤立用户	61 024	114 044	35 808	50 939	4 916
相似度为平均值	127	507	18	47	2
相似度为 1	980	957	3 218	3 059	2 931
相似度为 2	1 032	1 046	2 837	2 753	992
相似度为 3	971	1 081	2 513	2 550	404
相似度为 3~50	27 748	32 426	27 347	32 854	993

4.3　实验结果及性能分析

为更好地评价推荐算法的质量和用户满意度，引入几个度量因子：推荐准确度、覆盖率、基线成功率、成功率、召回率。

P 是所有用户集合的训练集，在集合 P 的元素中选取一个可能存在成功互动的子集，集合 C 为生成的推荐候选集。这样就隐含了一个发信人的集合 S 和 2 个收信人的集合 R 和 Q；R 是可能收到发信人集合 S 中用户发送信息收信人集合，Q 是在测试阶段，实际收到集合 S 中用户发送信息的收信人集合。$M(C)$是在测试期内实际上发生互动的集合，$nm(C)$是候选集中互动集 $M(C)$中的互动数量，$nm(C,+)$是互动集 $m(C)$中成功互动的数量，$n(S)$是 S 集合中用户的个数，$n(S,R)$是由 S 中用户和 R 中用户互动的数量，$ns(S,R,+)$是 S 和 R 之间成功互动的

数量, 同样的 $ns(S,Q,+)$是 S 和 Q 之间成功互动的数量（+表示一个积极、成功的互动）。

1) 精确度

$$P=\frac{nm(C,+)}{nm(C)} \tag{6}$$

生成的候选集中积极、成功的互动数量占总互动的比重称作推荐的准确度。$C,+$表示积极、成功的互动，C 表示所有互动。

2) 覆盖率

$$Cov=\frac{n(M)}{n(N)} \tag{7}$$

其中，集合 N 是用户集合，$n(N)$是用户数量；集合 M 是用户 N 中收到推荐的用户集合, $n(M)$是用户集合 N 中收到推荐的用户数量。

3) 基线成功率

$$BSR(S,Q)=\frac{ns(S,Q,+)}{n(S,Q)} \tag{8}$$

在测试期内产生实际的互动中，发信人集合 S 中用户发送的积极、成功互动数量占总互动数量的比重。

4) 成功率

$$SR=\frac{nm(C,+)}{nm(C)} \tag{9}$$

成功率是候选集中成功互动的数量占总互动数量的比重。

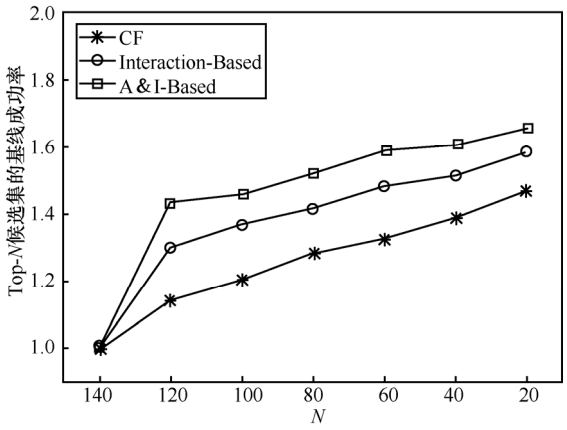
5) 召回率

$$Rc=\frac{nm(C,+)}{nm(S,Q,+)} \tag{10}$$

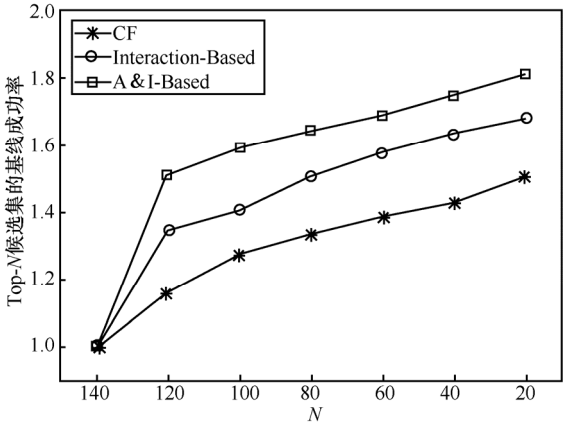
候选集中积极、成功互动数量占在测试期内实际上收到互动信息数量的比重称作召回率。

在实验中, 针对相似度计算中 2 个权重 α 与 β, 需满足条件：$\alpha+\beta=1$，根据实际测试数据及线性规划的最小二乘拟合法，不断去调整 α 与 β 的值, 以获得最佳的实验结果。针对{α=0.4, β=0.6; α=0.5, β=0.5; α=0.6, β=0.4}3 组不同取值组合，考虑用户间所有互动信息的情况下, 得到 3 种算法 BSR 在不同的 α 与 β 取值下的取值表现。

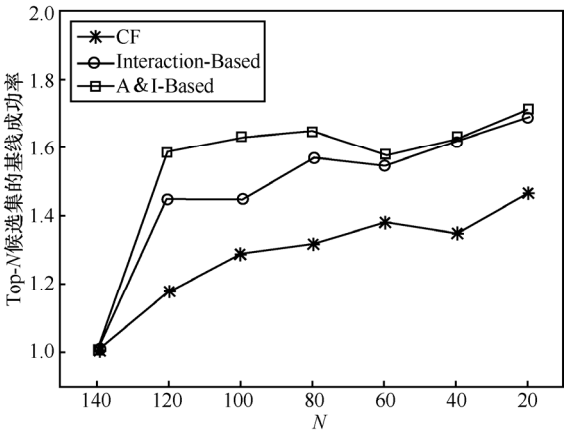
通过图 1 中 3 种算法的推基线成功率（BSR）曲线比较显示, 随 α 变大, BSR 递增, 当 α>0.6 时, 相比 α=0.5 时，BSR 出现下降趋势，通过多次实验发现，现在 α=0.57 与 β=0.43 时，算法取得较好的推荐效果，有着较高的基线成功率和覆盖率。



(a) α=0.4, β=0.6



(b) α=0.5, β=0.5



(c) α=0.6, β=0.4

图 1 不同 α、β 取值下 Top-N 候选集的基线成功率

图 2 和图 3 实验结果数据展示了在 α=0.57 与 β=0.43 时, 分别在考虑所有互动和只考虑积极互动情况下，得出 BSR 结果的曲线比较。

1) 考虑用户间所有互动，包括积极、消极、成功互动，在这种情况下，3 种算法的基线成功率比较如图 2 所示。

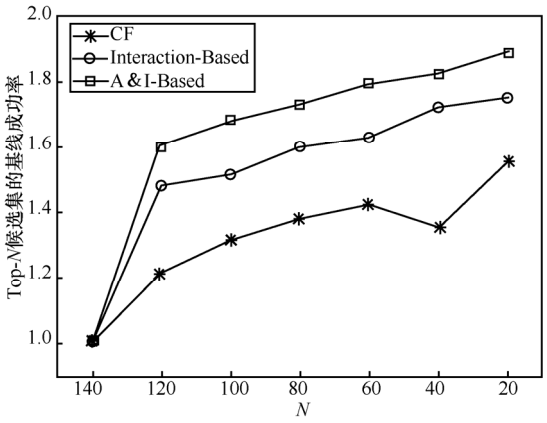


图 2　全部互动下算法基线成功率比较

图 3 中 3 种算法 BSR 曲线表明, 在考虑用户所有互动情况下，基于用户相似度的协同过滤推荐算法的基线成功率明显高于另外 2 种算法。

2) 在所有的互动中，剔除用户间消极、不成功的互动，只考虑用户间积极、成功的互动，在这种情况下，3 种算法的基线成功率比较如图 3 所示。

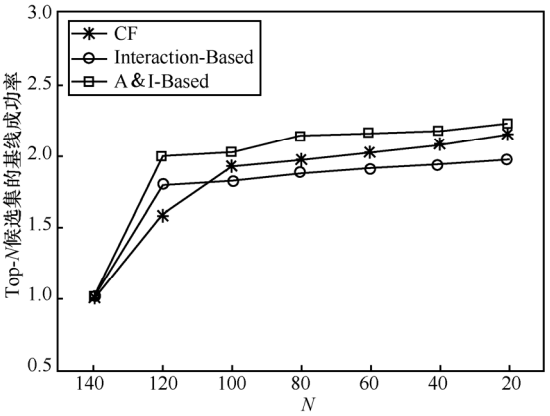


图 3　积极互动下基线成功率比较

图 3 表明，只考虑积极成功互动情况下，基于用户相似度的协同过滤推荐算法的 BSR 同样优于另外 2 种算法。实验表明，无论是在考虑所有的互动下，还是只考虑积极互动情况下，A&I-Based 推荐算法的基线成功率都优于另外 2 种算法。

为更全面度量基于用户相似度的系统过滤推荐算法的推荐质量，图 4 展示 3 种算法的基线成功率、成功率、召回率、覆盖率 4 个度量值，全方位立体度量算法质量。

图 4 呈现了协同推荐算法、基于用户互动信息推荐算法、基于用户资料和互动信息推荐算法的度量。在基线成功率、成功率、召回率、覆盖率 4 个度量中，A&I-Based 推荐算法的成功率、推荐质量、覆盖率都优于另外 2 种算法。

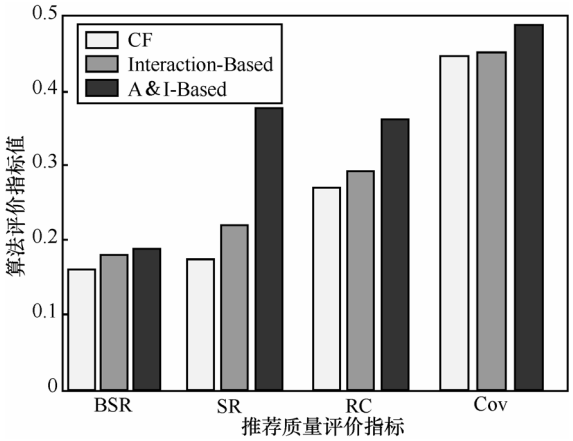


图 4　3 种算法的推荐质量度量

根据文中对算法复杂度的分析，算法在 $\odot(n^2)$ 时间内即可执行完算法，为用户产生推荐，与另外 2 种算法复杂度处于同一数量级。在不增加算法额外存储空间的情况下，实验证明基于用户属性和互动信息的推荐算法精确度、基线成功率、覆盖率都优于基于全部互动的推荐方式（包括积极、成功和消极、不成功的互动）。实验还表明，按用户相似度的 Top-N 排序算法在社交网站中得到推荐集有较好的推荐质量。

5　结束语

本文在定义用户相似度构成与计算方法基础之上，提出一种基于用户属性和用户互动信息的协同过滤推荐算法，并应用到社交网络中的智能推荐过程；通过与 2 类经典协同过滤推荐算法进行比较，仿真实验结果表明本文提出的算法有以下优点。

1) 相似度 $Sim_{\mathrm{A-B}}=\alpha(\sum_{i=1}^{n}\omega_i d_i+\overline{\omega_{Nom}}D_{HM})+\beta\frac{1}{Sim_f}$ 在 $\odot(n^2)$ 时间内即可执行完算法, 为用户产生推荐；其中，α=0.57 与 β=0.43 取值时，算法取得较好的推荐效果，有着较高的基线成功率和覆盖率。

2) 在考虑用户所有互动情况下, 基于用户相似度的协同过滤推荐算法的基线成功率明显高于另外 2 种算法。

3) 在考虑积极、成功的互动信息的情况下，基于用户相似度的协同过滤推荐算法的精确度，基线成功率，覆盖率都优于基于全部互动的推荐方式（包括积极、成功和消极、不成功的互动）；且按

用户相似度的 Top-*N* 排序算法在社交网站中得到推荐集有较好的推荐质量。

参考文献：

[1] ZHAO Z D, SHANG M S. User-based collaborative-filtering recommendation algorithms on hadoop[A]. Knowledge Discovery and Data Mining, WKDD'10 Third International Conference on IEEE[C]. 2010. 478-481.

[2] 吴颜, 沈洁, 顾天竺等. 协同过滤推荐系统中数据稀疏问题的解决[J]. 计算机应用研究, 2007, 24(6):94-97.
WU Y, SHEN J, GU T Z, *et al*. Algorithm for sparse problem in collaborative filtering[J]. Application Research of Computers, 2007, 24(6): 94-97.

[3] 罗奇, 余英, 赵呈领等. 自适应推荐算法在电子超市个性化服务系统中的应用研究[J]. 通信学报, 2006,(11): 183-186.
LUO Q, YU Y, ZHAO C L, *et al*. Research on personalized service system in E-supermarket by using adaptive recommendation algorithm[J]. Journal on Communications,2006(11):183-186.

[4] 邓爱林, 朱扬勇, 施伯乐. 基于项目评分预测的协同过滤推荐算法[J]. 软件学报, 2003, 14(9) :1621-1628.
DENG A L, ZHU Y Y, SHI B L. A collaborative filtering recommendation algorithm based on item rating prediction[J]. Journal of Software, 2003, 14(9):1621-1628.

[5] 张中峰, 李秋丹. 社交网站中潜在好友推荐模型研究[J]. 情报学报, 2011, 30(12):1319-1325.
ZHANG Z F, LI Q D. Latent friend recommendation in social network services[J]. Journal of The China Society For Scientific and Technical Information, 2011, 30(12):1319-1325.

[6] 陈可寒, 韩盼盼, 吴健. 基于用户聚类的异构社交网络推荐算法[J]. 计算机学报, 2013, 36(2):349-359.
CHEN K H, HAN P P, WU J. User clustering based social network recommendation[J]. Chinese Journal of Computers, 2011, 36(2): 349-359.

[7] ZHANG J Y, PEARL P. A recursive prediction algorithm for collaborative filtering recommender systems[A]. Proceedings of the 2007 ACM Conference on Recommender Systems[C]. ACM, 2007.57-64.

[8] KRZYWICKI A, WOBCKE W, CAI X. Interaction-based collaborative filtering methods for recommendation in online dating[A]. Web Information Systems Engineering-WISE 2010[C]. Springer Berlin Heidelberg, 2010.342-356.

[9] BRŎZOVSKÝ L, PETŘIČEK V. Recommender system for online dating service[D]. Charles University in Prague, 2007.

[10] PIZZATO L, REJ T, CHUANG T. RECON: a reciprocal recommender for online dating[A]. Proceedings of the fourth ACM conference on Recommender systems ACM[C]. 2010.207-214.

[11] WANG T T, LIU H Y, HE J, *et al*. Predicting New User's Behavior in Online Dating Systems[M]. Advanced Data Mining and Applications. Springer Berlin Heidelberg, 2011.266-277.

[12] CHEN L, NAYAK R, XU Y. A recommendation method for online dating networks based on social relations and demographic information[A]. Advances in Social Networks Analysis and Mining (ASONAM) International Conference on IEEE[C]. 2011.407-411.

[13] HITSCH G J, HORTACSU A, ARIELY D. Matching and sorting in online dating[J]. The American Economic Review, 2010,100(1):130-163.

[14] NAYAK R, ZHANG M, CHEN L. A social matching system for an online dating network: a preliminary study[A]. Data Mining Workshops (ICDMW) IEEE International Conference on[C]. 2010. 352-357.

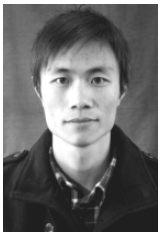
[15] OWEN S, NAIL R, DUNNING T, *et al*. Mahout in Action[M]. Manning, 2011.

作者简介：

荣辉桂（1975-），男，湖南株州人，博士，湖南大学讲师、硕士生导师，主要研究方向为大数据、云计算、电子商务。

火生旭（1986-），男，青海西宁人，湖南大学硕士生，主要研究方向为智能推荐、电子商务、移动互联网。

胡春华（1973-），男，湖南新化人，博士，湖南商学院教授，主要研究方向为云计算、服务计算、电子商务。



莫进侠（1987-），男，湖南邵阳人，湖南大学硕士生，主要研究方向为云计算、移动互联网等。