SNS 背景下基于 Tag 和 Rating 相似度融合的协同过滤①

王卫平，张丽君

(中国科学技术大学 管理学院，合肥 230026)

摘 要：SNS 即社会性网络服务的出现为 Tag 技术的应用提供契机，以基于 SNS 的网站为背景，将 Tag 信息作为补充信息融入协同过滤推荐系统，提出了 SNS 背景下基于 Tag 和 Rating 相似度融合的协同过滤，以降低数据稀疏性对推荐精度的影响。首先分别计算基于 Tag 信息和 Rating 评分信息的用户相似度，然后将这两种相似度进行融合得到综合相似度，最后据此进行协同过滤推荐。实验结果表明本文提出的算法能提高推荐的精度。

关键词：SNS；Tag；协同过滤；相似度融合；稀疏性

Collaborative Filtering Based on Similarity Fusion of Tag and Rating Under the Background of SNS

WANG Wei-Ping, ZHANG Li-Jun

(School of Managenment, University of Science and Technology of China, Hefei 230026, China)

Abstract: The emergence of social networking services (SNS) provides an opportunity for the application of tag. In this paper, we choose the website based on SNS as the background, and then integrate tag information into collaborative filtering recommendation system. So we proposed collaborative filtering recommendation system based on similarity fusion of tag and rating under the background of SNS. It can help us to reduce the influence of data sparsity on the recommendation accuracy. First, we calculate the user similarity based on Tag information and Rating information respectively, and then get the integrated similarity by the fusion of these two similarities. Finally, Collaborative Filtering can be executed based on this integrated similarity. Experimental results show that the proposed algorithm can improve the accuracy of the recommended.

Key words: SNS; Tag; collaborative filtering; similarity fusion; sparsity

1 研究背景

近年来，“社会性网络服务”一词流行开来，所谓社会性网络服务(Social Networking Services, SNS)，专指旨在帮助人们建立社会性网络的互联网应用服务[1]。其理论基础是源于哈佛大学心理学教授 Stanley Milgram 在 1967 年创立的六度分割理论，即任何一个人最多通过六个人就可以认识一位陌生人。基于六度分割理论，每个人的社交圈都在不断扩大，最后成为一个大型的网络。

基于社会性网络服务的网站也如雨后春笋般出现，如 Douban.com，del.icio.us，flicker 等。国内自 2005 年开始，校园社交网站如开心网、人人网、聚友网、海内网等也陆续出现。除了此类新兴的基于社会性网络服务的网站之外，另一类就是老网站开发新功能[2]，如 amazon.com 的书评系统，douban.com 的影评系统，以及一些购物网站的商品评价等。人们在这些网站上，可以自由地发表评论，从而找到与自己趣味相投的朋友，也可以找到自己喜欢的对象，还可以建立相应的论坛以便讨论交流。从 2004 年 12 月-2006 年 12 月的两年中，网站的数量几乎增加了一倍。同一时期，所有网站的成员总数增长了 4 倍[3]。所以我们有理由相信，社会性网络服务今后必将成为一种发展趋势。

Tag 的兴起正是源于社会性网络服务的应用—Del.icio.us 书签服务[4]。Tag 即社会性标签，其本质是一种分类系统，它允许用户自由地使用一些词或短语来标注自己喜爱的资源。Tag 不同于一般的关键词，

① 收稿时间:2011-02-21;收到修改稿时间:2011-03-17

它允许用户使用并未在文章中出现过的词或短语进行标注。Tag 也不同于一般目录结构的分类系统，分类是系统预先设定好的，而 Tag 是用户自己动态添加的；分类可以是多层树状结构，而 Tag 只有一层，不同 Tag 之间只有平行关系，而无父-子节点关系；用户可以同时为一个对象设定多个 Tag，而在目录结构的分类中一个对象一次只能存放在一个分类目录下[4]。

另外，电子商务的发展为我们提供丰富产品的同时，也使我们置身于信息海洋中，面临信息过载的问题。于是，推荐系统应运而生，其中以协同过滤推荐系统最为成功。但是其推荐质量却受到数据稀疏性问题的影响[5]。即随着电子商务网站规模的扩大，用户数目和产品数目都呈现指数级增长，但是由于各种原因用户评分数据极端稀疏，经两个用户共同评分过的项目更是少之又少，这给相似度计算带来困难，进而影响推荐系统的推荐质量。

传统协同过滤推荐系统就是在经两个用户共同评分的项目基础上利用 Pearson 相关系数来计算用户之间的相似度，但这样得到的用户相似度受数据稀疏性的影响较大从而影响了推荐的精度。所以本文在传统协同过滤的基础上，利用 Tag 信息为我们提供额外的有价值的补充信息[6]。即先分别基于 Tag 信息和 Rating 信息来计算用户之间的相似度，再将分别计算的相似度融合，并以这一综合相似度为基础进行协同过滤。实验证明本文提出的方法提高了推荐的质量。

2 基本定义

2.1 用户-项目评分矩阵 $R_{m\times n}$

在协同过滤推荐系统中，用户-项目评分矩阵有着举足轻重的作用，它是以一个 m×n 的矩阵来存储数据。其中 m 行代表 m 个用户，n 列代表 n 个项目，第 i 行第 j 列的元素 r_{ij} 代表用户 U_i 对项目 I_j 的评分。也就是说 r_{ij} 就是一个评分信息，即文章后面所指的 Rating。

2.2 标签 Tag

在一些社会性网络中，允许用户自由地使用一些词或短语来标注自己喜爱的资源。这些词或短语就是我们所说的标签 Tag。

2.3 用户-Tag 相关度矩阵 $Rel_{m\times s}$

首先给出用户与 Tag 之间的相关度的定义，是指用户与 Tag 之间联系的紧密程度。这是本文提出的一个新概念，而相关度矩阵就是用来存储这些具体的用户-Tag 相关度数值的 m×n 的矩阵，也是一种数据存储形式。其中 m 行代表 m 个用户， s 列代表 s 个 Tag，第 i 行第 k 列的元素 rel_{ik} 代表用户 U_i 与 Tag t_k 之间的相关度。

2.4 用户 U 和 V 的相似度 simrat（U，V）

用户间的相似度是表示两个用户之间喜好的相似程度。而 simrat（U，V）表示基于 Rating 评分信息，利用 Pearson 相关系数计算所得的用来衡量用户 U 与 V 之间相似度的系数，这个值越大说明 U 和 V 的相似度越大。

2.5 用户 U 和 V 的相似度 simtag（U，V）

与 simrat（U，V）相类似，都是用来衡量用户 U 与 V 之间相似度的系数，不同的是 simtag（U，V）是基于 Tag 信息计算所得的用户相似度系数。同样的，值越大表示相似度越大。

3 基于Tag和Rating相似度融合的协同过滤推荐

鉴于用户-项目评分矩阵 $R_{m\times n}$ 十分稀疏，故经两个用户共同评分的项目数则更是少的可怜，这就导致传统的协同过滤推荐系统在计算用户相似度时许多用户之间的相似度都为零，从而影响了推荐系统的推荐质量。于是本文以传统推荐系统为基础，再利用基于 SNS 网站收集来的 Tag 信息计算用户之间的相似度，并将这两种方法计算所得的相似度进行融合得到一个综合相似度，进而据此进行协同过滤推荐。其具体的算法步骤如下：

(1) 经 Tag 预处理，得到热门 Tag 集合 Tags；

(2) 计算基于 Tag 信息的用户相似度 simtag(U，V)；

(3) 计算基于 Rating 评分信息的用户相似度 simtag(U，V)；

(4) 相似度融合，得到综合的用户相似度 sim(U，V)；

(5) 寻找目标用户的最近邻居集合 neighbors(U_c)；

(6) 预测目标用户对目标项目的评分 P(U_c，I_c)。

3.1 Tag 预处理

Tag 是用户动态添加的，由用户自由选择而没有任何限制，这就使得社会标签面临着很多问题，例如

一词多义、多词同义，无意义词汇，语法错误，拼写错误等等[7]。这些问题使得 Tag 信息中包含了大量的噪音数据，从而严重影响了应用 Tag 信息计算的精度。为此，在应用 Tag 信息之前，需要先进行预处理，去除其中噪音数据的干扰。所以对于收集的 Tag 信息首先除去其中错误的词语。另外，Xin Li, Lei Guo, Yihong(Eric) Zhao 等人[8]通过所有人使用 Tag 的分布发现热门 Tag 被大多数用户使用。徐雁斐，张亮，刘炜等人[9]的研究则证明了这些少量频繁使用的热门 Tag 集合是稳定的。鉴于此，本文从收集来的 Tag 中提取少量热门 Tag 形成热门 Tag 集合 Tags 进行分析，以降低噪音数据对分析结果的影响。

3.2 计算基于 Tag 信息的用户相似度 simtag（U，V）

为方便后面的描述，在这里先作如下定义：

1）$n(I,t)$：表示以 Tag t 标注项目 I 的用户数，即项目 I 被 Tag t 标注的次数。

2）markeditem(U)：表示所有被用户 U 贴过标签的项目集合。

3）相关度 rel：用来度量某一关系紧密程度的数值。

3.2.1 计算项目 I 与 Tag t 之间的相关度 rel(I，t)[10]

$$rel(I,t)=TIF(I,t)\times IDF(t)$$

其中 TIF(I，t)表示贴在项目 I 上的 Tag t 在所有贴在项目 I 上的 Tag 中所占的比例；IDF(t)表示 Tag t 的稀疏程度，t 越稀疏则 IDF(t)的值越大，与倒文档频率相似。即：

$$TIF(I,t)=\frac{n(I,t)}{\sum_{t_k\in Tags}n(I,t_k)}$$

$$IDF(t)=\ln\frac{\sum_{I_j\in Items}\sum_{t_k\in Tags}n(I_j,t_k)}{\sum_{I_j\in Items}n(I_j,t)}$$

3.2.2 计算用户 U 与 Tag t 之间的相关度 rel(U，t)

$$rel(U,t)=\sum_{I_j\in markeditem(U)}rel(I_j,t)$$

前面已经计算出了各项目 I 与 Tag 之间的相关度，这里将所有用户 U 评价过的项目 I_j 与 Tag t 之间的相关度加总求和作为 U 与 Tag t 之间的相关度。

3.2.3 计算基于 Tag 信息的用户相似度 simtag（U，V）

根据前面 3.2.2 计算所得的各用户与 Tag 之间的相关度建立用户-Tag 相关度矩阵 $Rel_{m\times s}$，并据此计算各用户之间的相似度：

$$simtag(U,V)=\frac{\sum_{t_k\in Tags}rel(U,t_k)\times rel(V,t_k)}{\sqrt{\sum_{t_k\in Tags}rel(U,t_k)^2\times\sum_{t_k\in Tags}rel(V,t_k)^2}}$$

3.3 计算基于 Rating 评分信息的用户相似度 simrat（U，V）

利用 Pearson 系数法计算基于 Rating 评分信息的用户相似度：

$$simrat(U,V)=\frac{\sum_{I\in I_{UV}}(r_{U,I}-\overline{r_U})(r_{V,I}-\overline{r_V})}{\sqrt{\sum_{I\in I_{UV}}(r_{U,I}-\overline{r_U})^2\sum_{I\in I_{UV}}(r_{V,I}-\overline{r_V})^2}}$$

其中，I_{UV} 表示用户 U 和 V 共同评价过的项目集合，$\overline{r_U}$、$\overline{r_V}$ 分别表示用户 U 和 V 的平均评分。

3.4 相似度融合

鉴于数据稀疏性的影响，使得分别计算的用户间相似度 simtag（U，V）和 simrat（U，V）有很多都为 0，而基于 Rating 评分信息计算所得的用户相似度 simrat（U，V）经受了诸多研究学者多年的实践验证，所以本文就以 simrat（U，V）为基础，用 simtag（U，V）作为补充信息替换 simrat（U，V）为 0 的用户间相似度，从而降低数据稀疏性的影响。其数学表达式如下：

$$sim(U,V)=\begin{cases}simrat(U,V) & simrat\neq 0\\ simtag(U,V) & \text{otherwise}\end{cases}$$

3.5 寻找目标用户的最近邻居集合 neighbor（U_c）

将目标用户 U_c 与各用户 U 之间的相似度按照降序排列，根据设定的阈值 k 选取前 k 个相似度最大的用户 U 组成邻居集合 neighbors（U_c）。

3.6 评分预测 P（U_c，I_c）

目标用户 U_c 对目标项目 I_c 的评分可以用下面的公式计算：

$$P(U_c,I_c)=\overline{r_{U_c}}+\frac{\sum_{U\in neighbors(U_c)}Sim(U_c,U)(r_{U,I_c}-\overline{r_U})}{\sum_{U\in neighbors(U_c)}Sim(U_c,U)}$$

4 实验结论

4.1 实验数据

本实验的数据来源于 GroupLens 研究小组在网站 http://www.grouplens.org 上提供的 MovieLens 数据集。我们使用的是包含有 Tag 信息的数据集，该数据集包含了 71567 个用户对 10681 部电影的 10000054 个评分

记录和 95580 个标签信息，每个用户至上评价了 20 部电影，评分的取值范围为 1-5。

出于减小计算规模和降低噪音数据干扰的考虑，在实验中我们随机抽取了 500 个用户和 800 部电影以及相关的 117 个热门 Tag 作为实验样本，并按照 9:1 的比例分别构造训练数据集和测试数据集。

4.2 评价标准

关于推荐系统性能的评价标准很多，其中平均绝对误差（MAE）被广泛用于协同过滤推荐系统的推荐质量评价。同时，考虑到实验中的训练样本和测试样本都是用户对电影的评分，利用 MAE 来计算预测评分与真实评分之间的偏差简单易行。因此本文采用 MAE 来评价基于Tag和Rating相似度融合的协同过滤推荐系统的性能，MAE 的值越小说明推荐系统的性能越好。

$$MAE=\frac{\sum_{i=1}^{n}|p_i-q_i|}{n}$$

其中，n 表示预测的评分数目，p_i 表示预测的评分值，q_i 表示实际的评分值。

4.3 实验结论

为了衡量本文提出的 SNS 背景下基于 Tag 和 Rating 相似度融合的协同过滤推荐系统的性能，我们以传统的协同过滤推荐系统作为比较对象，实验结果如图 1 所示。分析该实验结果可知，与传统的协同过滤推荐系统（传统 CF）相比，本文提出的协同过滤推荐系统（本文 CF）其平均绝对误差（MAE）较小，能提高推荐的精度。

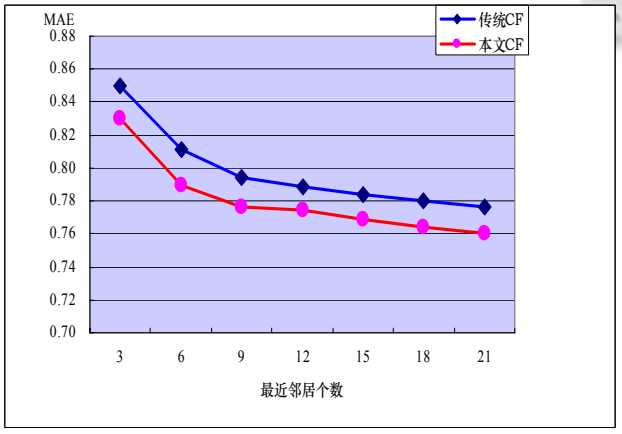

图 1 实验结果

5 未来展望

虽然本文提出的协同过滤推荐方法在平均绝对误差上较传统的协同过滤推荐较小，提高了推荐的精度，但是仅仅提高了 2.5%左右。另外我们对于 Tag 的处理也不够精确，例如对于同义词、多义词等未作处理。今后的研究方向是对 Tag 做更精确的处理，首先消除一词多义、多词同义的现象，再对 Tag 作聚类处理，通过计算用户与类标签的相关度进而计算用户之间的相似度。

参考文献

1 雷环,彭舰.SNS 中结合声誉与主观逻辑的信任网络分析.计算机应用研究,2010,27(6):2321−2323.

2 任建华,汪赫瑜.协同过滤技术在社会性网络服务的应用.中国新通信,2007,17.

3 Golbeck J. The dynamics of Web-based social networks：membership, relationships and change. First Monday, 2007, 12(11).

4 林森.基于 Tag 技术的知识个性化推荐及系统[硕士学位论文].武汉:华中科技大学,2006.

5 Wang J, de Vries AP, Reinders MJT. Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion. Annual ACM Conference on Research and Development in Information Retrieval, 2006.501−508.

6 Tso-Sutter KHL, Marinho LB, Schimidt-Thieme L. Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms. Proc. of the 2008 ACM symposium on Applied Computing.

7 Liang HZ, Xu Y, Li YF, Nayak R. Collaborative Filtering Recommender Systems based on Popular Tags. Proceedings of the Fourteenth Australasian Document Computing Symposium, 2009-12-04.

8 Li X, Guo L, Zhao YH. Tag-based social interest discovery. Proc. of the 17th International Conference on World Wide Web, 2008.

9 徐雁斐,张亮,刘炜.基于协同标记的个性化推荐.计算机应用与软件,2008,25(1):9−13.

10 杨丹,曹俊.基于 Web2.0 的社会性标签推荐系统.重庆工学院学报,2008,22(7).