基于用户关注度的个性化新闻推荐系统

彭菲菲，钱 旭
（中国矿业大学 机电与信息工程学院，北京 100083）

摘 要：为满足用户需求，以用户为中心，解决用户关注度不断变化、数据稀疏性、优化时间和空间效率等问题，提出基于用户关注度的个性化新闻推荐系统。推荐系统引入个人兴趣和场景兴趣来描述用户关注度，使用雅克比度量用户相似性，对相似度加权求和预测用户关注度，从而提供给用户经过排序的新闻推荐列表。实验结果表明，推荐系统有效地提高了推荐精准度和覆盖度，改善了系统可扩展性和自动更新能力，具有良好的推荐效果。

关键词：个性化推荐；协作型过滤；用户关注度；推荐算法
中图分类号：TP311 **文献标志码**：A **文章编号**：1001-3695(2012)03-1005-03
doi:10.3969/j.issn.1001-3695.2012.03.056

Personalized recommendation system for news based on user concern

PENG Fei-fei, QIAN Xu
(*School of Mechanical Electronic & Information Engineering, China University of Mining & Technology, Beijing* 100083, *China*)

Abstract: There were many problems, such as the shift of user concern, data sparse, and system scalability and automatic update capabilities. This paper gave a new personalized recommendation system for news based on user requirements and user-centered. It described user concern using personal preference and situational interest, used Jacobi to measure user similarity and forecasted user concern with similarity-weighted, and then provided ordered news recommendation list for every user. Experimental results show that the recommendation system can effectively resolve above problems, and it is higher accuracy and coverage and better recommendation results.

Key words: personalized recommendation; collaborative filtering; user concern; recommendation algorithm

面对海量变化迅速的网络新闻，用户面临的选择越来越多，在这样的环境下，为了能够更好地为用户推荐比较符合用户兴趣的新闻列表，个性化新闻推荐系统成为网络新闻检索领域的一项重要研究内容。协作型过滤是至今为止应用最成功的个性化推荐技术。具体做法是先利用用户的历史信息计算用户之间的相似性，然后利用与目标用户相似性较高的邻居对其他信息的评价来预测目标用户对特定信息的喜好程度，根据这一喜好程度为目标用户推荐一个经过排名的推荐列表[1]。协同过滤推荐算法具有推荐新信息的能力，不需要考虑信息的内容[2]。但依然存在用户关注度不断变化、数据稀疏性等问题。

本文为了满足用户的需求，使用协同过滤算法，引入个人兴趣和场景兴趣，用户是否选择了新闻推荐列表中新闻，满足用户关注度的自适应性，从而提高个性化新闻推荐系统的可扩展性和自动更新能力。

1 基于用户关注度的个性化新闻推荐系统

个性化新闻推荐系统主要包括创建用户关注度字典、个性化推荐、系统增量更新三个方面的内容，系统架构如图 1 所示。它通过合理的方法利用用户的相关信息，为不同的用户提供不同的新闻推荐排序，从而实现新闻推荐的个性化。为了提高系统运行时的空间与时间效率，当计算用户相似度时，只存储前 K 个最相似的用户，而且只基于这 K 个相似用户来计算预测分值。

由图 1 可以看出，推荐过程主要组成部分为创建用户关注度字典、寻找相似用户、计算预测关注度、产生推荐。

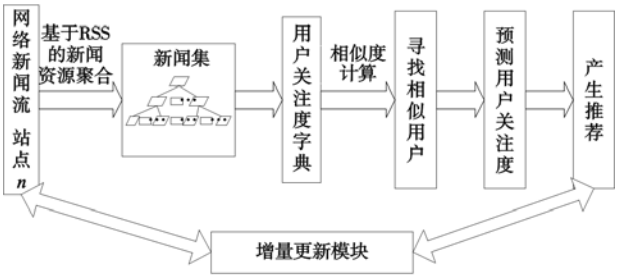

图1 推荐系统架构

1.1 创建用户关注度字典

为了描述用户关注度，通过分析用户行为来构建用户关注度字典。用户行为包括个人兴趣和场景兴趣。影响个人兴趣的因素包括用户查看新闻标题时是否点击查看详文，在新闻详文页面上的停留时间，对新闻的评分，是否收藏，是否分享与转载，是否选择了所推荐的内容。这些因素均能反映出用户对该条新闻的关注程度。如果用户查看新闻标题继而查看了详文，表明用户关注这则新闻。如果用户对这则新闻越关注，停留时间越长，阅读速度越慢。此外，考虑到用户浏览新闻内容是否是客观地关注，所以需要对新闻评分。随着关注度的提高，用户会收藏内容，接着分享与转载。用户关注度是不同等级的，分为不喜欢、喜欢、比较喜欢、非常喜欢。场景兴趣是指当突发性新闻事件发生时，对个人兴趣的影响。

收稿日期：2011-08-26；**修回日期**：2011-09-28

作者简介：彭菲菲(1983-)，女，山东章丘人，博士研究生，主要研究方向为机器学习、人工智能(feifei_gaoren@163.com)；钱旭(1962-)，男，教授，博士，主要研究方向为信息融合技术、知识工程.

基于上面的分析,定义:u 为其中某个用户,n 为其中某条新闻。当用户对新闻标题关注,则打开新闻标题查看详文,如下:

$$\text{link}_{un}=\begin{cases}1 & \text{打开}\\ 0 & \text{未打开}\end{cases} \tag{1}$$

其中:$\text{link}_{un}=1$,表示用户 u 打开新闻 n 的标题,查看新闻 n 的详文。当用户 u 查看了新闻 n 的详文后,发现不是期望的内容,使用式(2)表示:

$$\text{score}_{un}=\begin{cases}1 & \text{喜欢}\\ 0 & \text{不喜欢}\end{cases} \tag{2}$$

其中:$\text{score}_{un}=1$,表示用户 u 对新闻 n 感兴趣,可以进一步关注。

当用户查看了长度不同的新闻详文时,所用时间却相同。通过计算用户 u 的平均阅读速度来衡量用户的关注度,如下:

$$v_u=\frac{1}{m}\sum_{i=1}^{m}\frac{\text{newslen}_{ui}}{t_{ui}} \tag{3}$$

其中:m 表示新闻量;newslen_{ui} 表示用户 u 所阅读的第 i 条新闻的长度;t_{ui} 表示用户 u 所阅读的第 i 条新闻详文时所用的时间;v_u 表示用户 u 的平均阅读速度。

在花费了用户一部分时间之后,如果用户对所读的新闻较关注或非常关注,分别使用式(4)(5)来表示:

$$\text{collect}_{un}=\begin{cases}1 & \text{收藏}\\ 0 & \text{未收藏}\end{cases} \tag{4}$$

其中:$\text{collect}_{un}=1$ 表示用户 u 对阅读的新闻 n 较关注。

$$\text{share}_{un}=\begin{cases}1 & \text{分享或转载}\\ 0 & \text{未分享或转载}\end{cases} \tag{5}$$

其中:$\text{share}_{un}=1$ 表示用户 u 对阅读的新闻 n 非常关注。

结合上面提出的制约条件,给出用户的关注程度:

$$\text{attention}_{un}=\begin{cases}3 & \text{link}_{un}=1\wedge\text{score}_{un}=1\wedge v_u<v_{un}<\text{share}_{un}=1 \quad \text{非常关注}\\ 2 & \text{link}_{un}=1\wedge\text{score}_{un}=1\wedge v_u<v_{un}\wedge\text{collect}_{un}=1 \quad \text{较关注}\\ 1 & \text{link}_{un}=1\wedge\text{score}_{un}=1\wedge v_{un}<v_u \quad \text{关注}\\ 0 & \text{otherwise} \quad \text{不关注}\end{cases} \tag{6}$$

其中:attention_{un} 表示用户 u 对新闻 n 的关注度。根据用户的关注程度,按照式(7)构建用户关注度的嵌套字典。

$$\text{concern}_u=\{u:\{n:\text{attention}_{un}\}\} \tag{7}$$

所有的用户均具有好奇心,均希望了解最新发生的突发性的新闻事件 SN,所以当突发性新闻事件 SN 发生时,系统将 SN 推荐给所有用户,并设定为非常关注,即 $\text{attention}_{SN}=3$。

假设有 K 个不同的注册用户、N 条新闻,算法为 K 个用户分别构建用户关注度字典,算法描述如下:

a)初始化所有参数;

b)如果用户 i 打开并查看了新闻 j,记录用户的阅读时间和新闻的长度,利用公式计算平均速度;

c)如果新闻是突发性新闻,则执行 d),否则执行 e);

d)将 3 存入 attention_{ij} 中;

e)如果用户 i 打开并查看了新闻 j,对新闻 j 的感兴趣,阅读速度小于平均速度,且点击分享或转载了新闻 j,则执行 f),否则执行 g);

f)将 3 存入 attention_{ij} 中;

g)如果用户 i 打开并查看了新闻 j,对新闻 j 的感兴趣,阅读速度小于平均速度,且收藏了新闻 j,则执行 h),否则执行 i);

h)将 2 存入 attention_{ij} 中;

i)如果用户 i 打开并查看了新闻 j,对新闻 j 的感兴趣,阅读速度小于平均速度,则执行 j),否则执行 k);

j)将 1 存入 attention_{ij} 中;

k)将 0 存入 attention_{ij} 中;

l)更新 attention_{ij} 的记录。

1.2 计算及存储用户相似度

推荐算法的基础在于度量任意两个用户之间相似度的能力。两个用户的关注度相同,即两个用户关注度字典的交集;两个用户的关注度总值,即两个用户关注度的并集。采用雅克比相似性度量方法,计算用户相似度,得到用户相似度矩阵。相似度为

$$\text{similarityValues}[i][j]=\frac{(\text{double})\,\text{agreeConcern}}{(\text{double})\,\text{totalConcern}} \tag{8}$$

其中:变量 i 和 j 分别指两个用户;agreeConcern 指用户共同的关注度集合;totalConcern 指用户关注度并集;similarityValues[i][j]是两个用户的相似度。

如果用户 1 与用户 2 相似,则用户 2 与用户 1 必定相似,所以相似度具有对称性,可以选用稀疏矩阵或其他存储结构。为了优化相似度的稀疏性,选用下三角矩阵作为用户相似度的存储矩阵。此外,为了去除自身相似性的比较,在遍历用户时无须全部遍历,第一层循环为式(9),第二层循环则为式(10),即第二层循环从第一层循环中的用户在整个列表中位置开始。

$$\text{for}(\text{int } i=0;i<\text{number\_users};i++) \tag{9}$$

其中:变量 i 是指某个用户编号,number_users 是用户总数。

$$\text{for}(\text{int } j=i+1;j<\text{number\_users};j++) \tag{10}$$

其中:变量 j 是指某个用户编号。

1.3 预测用户关注度

基于 1.2 节提出的用户相似度来计算每个用户与所有条目的预测关注度。预测关注度值是加权相似度之和与相似度总和的比值,如式(11)所示。

$$P\_\text{concern}_i=\frac{\sum_{i=1}^{n}\text{similarityValues}_i*\text{concernNeighbor}_i}{\sum_{i=1}^{n}\text{similarityValues}_i} \tag{11}$$

1.4 产生推荐

将使用式(11)得到的预测关注度进行排序,为了保证系统的时间和空间效率与推荐的高效性,选取前 N 条新闻作为推荐。

2 实验结果分析

2.1 实验数据

应用 Java、JSP、AJAX 技术构建实验仿真模型。使用可定制的 RSS 订阅资源,RSS 的基础技术是 XML,遵从 W3C 的标准规范。目前 RSS 标准已经得到了广泛应用,下面给出了结构化信息体的主要核心内容,如下所示:

〈outline title = "新闻中心 RSS" text = "新闻中心 RSS"〉//主类

〈outline text = "新闻要闻" title = "新闻要闻" type = "rss" xmlUrl = http://rss. sina. com. cn/news/marquee/ddt. xml htmlUrl = " www. sina. com. cn"/〉//主类的一个子类

〈outline text = "国内要闻" title = "国内要闻" type = "rss" xmlUrl = http://rss. sina. com. cn/news/china/focus15. xm htmlUrl = " www. sina. com. cn"/〉//主类的一个子类

〈outline text = "国际要闻" title = "国际要闻" type = "rss" xmlUrl = http://rss. sina. com. cn/news/world/focus15. xml htmlUrl = " www. sina. com. cn"/〉//主类的一个子类

在了解了 RSS 资源的结构之后,使用 Java 编写信息资源聚合器,将 RSS 资源下载到本地。使用 RSS 资源聚合器从热点网站(如新浪、搜狐、网易)获取可定制的体育类新闻,包括 19 类新闻,共 5 082 条新闻,作为实验使用的数据集。

2.2　评测标准

采用统计度量方法中平均绝对偏差 MAE(mean absolute error)作为度量推荐系统好坏的标准[10]。MAE 通过预测的用户关注度与实际用户关注度之间的偏差,衡量推荐系统的推荐质量。MAE 值越小,推荐质量越高。对于用户关注度数据,预测的关注度集合表示为$\{p_1,p_2,\cdots,p_n\}$,实际用户关注度集合为$\{q_1,q_2,\cdots,q_n\}$,则平均绝对偏差 MAE 定义如式(12)所示。

$$\mathrm{MAE}=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N} \tag{12}$$

选用覆盖度 coverage、推荐准确度 precision 和召回率 recall 作为推荐系统的评判标准,并且与传统的新闻推荐系统进行对比实验。Coverage 是一项被广泛使用的用以评价推荐系统推荐覆盖度的评判标准,指的是推荐系统为用户推荐的新闻集对用户关注度的覆盖范围[10]。其计算为

$$\mathrm{coverage}=\sum_{u\in U}(RT_u\cap RS_u)/\sum_{u\in U}(RS_u) \tag{13}$$

其中:u 是某个用户;U 是用户集;RT_u 是推荐系统为 u 推荐的新闻集;RS_u 是 u 在测试数据集上选择阅读的新闻集。Coverage 越高,表示推荐系统对用户关注度的覆盖能力越强。

准确度和召回率分别定义为

$$\mathrm{precision}=\frac{\text{推荐给用户并被用户阅读的新闻数目}}{\text{推荐给用户的新闻总数目}}$$

$$\mathrm{recall}=\frac{\text{推荐给用户并被用户阅读的新闻数目}}{\text{所有阅读的新闻总数目}}$$

2.3　实验结果及分析

在 50 名学生的帮助下进行了实验,以张三和李四作为基准,得到用户关注度的分布情况,如图 2 所示。其中,点代表用户,两个点的距离代表了两个用户间相似情况。用经过加权的用户关注度来预测用户关注度,为不同用户提供不同的推荐。

通过平均绝对偏差 MAE 的计算,推荐效果分析如图 3 所示,即在不同的推荐系统上的 MAE 随着近邻用户数目的变化情况。基于用户关注度的个性化新闻推荐系统平均绝对偏差 MAE 的值相对小,则其具有较好的推荐效果。

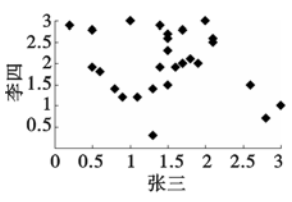

图2　用户关注度分布

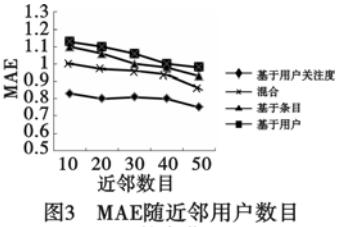

图3　MAE随近邻用户数目的变化

随着推荐数目的增加,推荐系统的覆盖度逐渐取得最大值。由于不同的推荐系统的性能不同,当推荐数目较少时,性能差的只能覆盖小范围,覆盖度低;但是,随着推荐数目的增加,各个推荐系统在不同阶段取得一致,如图 4 所示。

为用户推荐新闻数目的准确度随推荐新闻数目的变化而变化,不同的推荐系统达到最佳准确度时,新闻数目一般在 46 ~65 条。不同性能的推荐系统准确度最佳时,为用户推荐的新闻数目也不同。当推荐的数目过多时,相似度越来越来低,关注度越来越低,所以准确率开始下降,如图 5 所示。

推荐系统与推荐新闻的数目成正比,随着为用户推荐的新闻数目增加,不同的推荐系统召回率逐渐提高。推荐给用户的新闻数目增加,用户阅读新闻的可能性增加,被用户阅读的新闻越多,如图 6 所示。

由图 4 ~6 可知,基于用户关注度的个性化新闻推荐系统与其他推荐系统相比,其具有较高的精准度,具有较好的推荐能力。

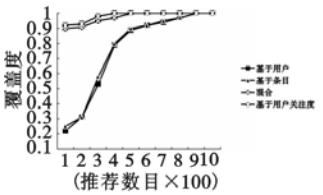

图4　不同推荐系统的覆盖度比较

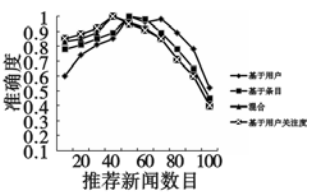

图5　不同推荐系统的准确度比较

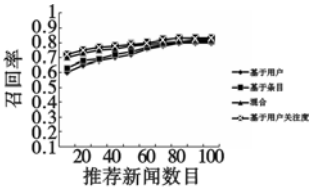

图6　不同推荐系统的召回率比较

3　结束语

本文详细分析用户阅读新闻时的行为特点、个人兴趣和场景兴趣。影响个人兴趣的因素包括用户查看新闻标题时是否点击查看详文,在新闻详文页面上的停留时间,对新闻的评分,是否收藏,是否分享与转载,是否选择了所推荐的内容。这些因素均能反映出用户对该条新闻的关注程度;场景兴趣体现的是用户对突发性的新闻事件的反应。本文使用用户关注度来描述用户行为特点,给出构建用户关注度字典的步骤和实现方法;接着计算用户间的相似度,经过比较,本文选用雅克比相似度度量方法;进一步基于相似度预测用户关注度,从而产生新闻推荐列表,经过排序后推荐给用户。本文提出的推荐系统可有效地解决用户关注度随时可变的问题,由于传统的推荐系统均具有数据稀疏性使得存储空间更为浪费,本文提出了相应的解决方法,同时优化了时间和空间效率。实验表明,本文提出的推荐系统可提高推荐精准度和覆盖度,改善了系统可扩展性和自动更新能力,具有良好的推荐效果可以提供更能满足用户需求的新闻推荐列表。

参考文献:

[1] LINDEN G, SMITH B, YORK J. Amazon. com remmendations item-to-item collaborative filtering[J]. IEEE Trans on Internet Computing, 2003, 7(1): 76-80.

[2] 曾春, 邢春晓, 周立柱. 个性化服务技术综述[J]. 软件学报, 2002, 13(10): 1952-1961.

[3] 刘建国, 周涛, 汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展, 2009, 19(1): 1-15.

[4] RENDA E M, STRACCIA U. A personalized collaborative digital library environment: a model and an application[J]. Information Processing and Management, 2005, 41(1): 5-21.

[5] JEH G, WIDOM J. Scaling personalized Web search[C]//Proc of the 12th International Conference on World Wide Web. New York: ACM Press, 2003.

[6] 周军锋, 汤显, 郭景峰. 一种优化的协同过滤推荐算法[J]. 计算机研究与发展, 2004, 41(10): 1842-1847.

[7] 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): 1622-1623.

[8] 张忠平, 郭献丽. 一种优化的基于项目评分预测的协同过滤推荐算法[J]. 计算机应用研究, 2008, 25(9): 2658-2683.

[9] MARMANIS H, BABENKO D. Algorithms of the intelligent Web[M]. [S. l.]: Manning Publications, 2009: 70-115.

[10] HERLOCKER J, KONSTAN J, TERVEEN L, *et al*. Evaluating collaborative filtering recommender systems[J]. ACM Trans on Information Systems, 2004, 22(1): 5-53.