

一种三部图网络中标签时间加权的推荐方法

顾亦然 陈 敏

(南京邮电大学自动化学院 南京 210003)

摘 要 社会标签可以提供对象高度抽象的内容信息和个性偏好信息,因此标签的使用有助于提高个性推荐的精度。用户的偏好会随时间的变化而变化,网络中的资源也会随着时间推移而增加。如何根据用户兴趣的变化推荐出用户即时感兴趣的网络资源,已成为推荐系统研究的新问题。在用户-标签-对象三部分图网络结构中,结合标签使用频率和用户添加标签的时间,提出了一种利用标签时间加权的资源推荐算法。实验结果表明,利用标签时间加权的算法能有效地提高推荐的精度和多样性。

关键词 个性化推荐,社会标签网络,时间加权,三部分网络

中图法分类号 TP391 **文献标识码** A

One Tag Time-weighted Recommend Approach on Tripartite Graphs Networks

GU Yi-ran CHEN Min

(College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract Social tags can provide highly abstract information about not only item contents but also personalized preferences, hence using labels could improve the accuracy of personalized recommendation. As a result of user preferences changes over time, network resources also will be increased as time goes by. How to recommend the network resources in which users have immediate interest based on the user preferences changes becomes a new research problems in the recommendation system. Combined with using tag frequency and label time on user-object-tag tripartite graphs, we proposed a recommendation algorithm based on tag time-weighted network. Experimental results demonstrate that the usage of tag time-weighted can significantly improve accuracy and diversification of recommendations.

Keywords Personalized recommendation, Social tagging networks, Time-weighted, Tripartite graphs

1 引言

随着互联网技术的迅速发展,接入网络的服务器数量和万维网上的网页数量都呈指数形势增长,信息超载也越来越严重。人们需要从成千上万的电影、书籍和网页中做出选择。个性化推荐,包括个性化搜索被认为是当前解决信息超载最有效的工具。推荐问题的实质就是代替用户预测他从未看到过的对象^[1]。目前应用最为广泛的推荐系统是协同过滤系统,之后则发展出了基于内容的推荐系统和基于网络结构的推荐系统。但是目前的推荐算法仍存在很多的问题,如面临新用户和新产品、打分稀疏性和算法可扩展性等。由于网络资源更新速度很快,资源对象随着时间推移而增加,用户的偏好也会随着时间的变化而变化。如何满足用户对网络资源的即时需求对于推荐系统而言是一个新的挑战。

伴随着 Web2.0 技术的成熟,社会标签系统也在迅猛发展,它以一种新颖的形式创造了更好的平台,让人们更高效地查找网络资源。社会标签系统让用户自由地对所选资源对象添加标签,没有任何要求限制,从而扩充了用户和对象的语义范围,大大提高了推荐算法的精度,这也吸引了很多科学团体

的关注。Golder 等^[2]研究标签的使用形式将标签的功能分成 7 类。刘建国等人^[3]以杂志 PNAS 为实验对象,研究关键词的演化特性。此外,也有学者对社会标签的性能做了研究,如 Cattuto 等^[4]提出了基于记忆的 Yule-Simon 模型来描述标签的衰变形式和其使用频率。张子柯等^[5]提出了超图模型的演化过程,该方法中用户不仅可以给对象添加标签,也可以通过标签检索对象。

本文从复杂网络角度出发,提出了一种利用标签时间加权的资源推荐方法,该方法结合用户使用标签的频率及添加标签的时间,解决了用户对资源的即时需求问题,同时使得推荐算法的精度和推荐的多样性都有显著的提高。

2 理论与方法

本文提出的算法由 3 个数据集组成,定义如下:

1) 数据集组成:用户集 $U = \{U_1, U_2, \dots, U_n\}$;

对象集 $O = \{O_1, O_2, \dots, O_m\}$;

标签集 $T = \{T_1, T_2, \dots, T_r\}$ 。

2) 三部分网络图可以由 3 个邻接矩阵 A, A' 和 A'' 表示^[6]。

到稿日期:2011-09-14 返修日期:2011-11-22

顾亦然(1972-),女,副教授,硕士生导师,CCF 高级会员,主要研究方向为通信网络的性能分析与控制、复杂网络理论在通信系统中的应用等, E-mail:guyr@njupt.edu.cn;陈 敏(1986-),女,硕士生,主要研究方向为复杂网络理论在通信系统中的应用。

A 代表用户-对象间的关系。若用户 U_i 选择过对象 O_j , 则 $a_{ij}=1$, 其他为 0。

A' 代表对象-标签间的关系。若对象 O_j 被赋有标签 T_k , 则 $a'_{jk}=1$, 其他为 0。

A'' 代表用户-标签间的关系。其中 a''_{ik} 表示用户 U_i 使用标签 T_k 的次数。

2.1 基于资源分配的二部图设计

在二部图网络 $G(U, O, E)$ 中, U 和 O 分别为用户集和对象集, E 为连接用户和对象边的集合, 二部图网络结构如图 1 所示。周涛等^[7]提出了一种全新的基于资源分配的算法。假设对象上有一定的初始资源, 每个对象平均分配其资源给所有邻居用户, 然后每个用户再平均分配其资源给所有选择过的对象。定义 $\vec{f}$ 为对象的初始资源向量, $\vec{f}'$ 为最终资源向量, 经过两步再分配后, 可以得到:

$$f'_j = \sum_{i=1}^n \frac{a_{ij}}{k(U_i)} \sum_{s=1}^m \frac{f_s}{k(O_s)}, j=1, 2, \dots, m \quad (1)$$

式中, $k(U_i) = \sum_{j=1}^m a_{ij}$ 为用户 U_i 选择对象的数量, $k(O_s) = \sum_{i=1}^n a_{is}$ 为对象 O_s 的邻居用户的数量。由式(1)得到最终资源向量 $\vec{f}'$ 。把目标用户没有看到过的对象按照矢量 $\vec{f}'$ 中对应元素的大小进行排序——值越大就说明该用户越喜欢。排序靠前的产品, 可以推荐给目标用户。

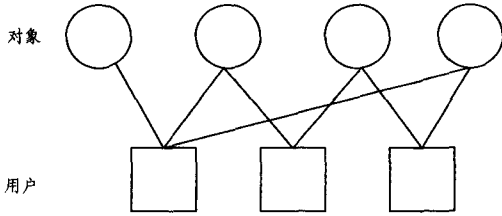


图1 二部图网络结构

2.2 基于标签时间加权的推荐算法

2.2.1 三部图网络结构中时间衰减分析

由于网络资源的不断更新, 用户喜好也在不断变化, 对资源对象的选择是有时间效应的。一般情况下, 用户对最近选择的对象比较感兴趣, 而忽略了很早以前选择的对象。在不同时期, 用户的历史资源选择信息对推荐预测的贡献是不一样的, 在对资源分配时就要考虑时间的影响^[8]。在标签系统中则体现为用户最新添加标签的对象可能是其当前最感兴趣的, 可以通过用户最新使用的标签找到最符合其需求的资源。

举一个简单的例子, 如图 2 所示。目标用户为 U_3 , 对象 O_3 和 O_4 是待推荐对象。假设当前时间 $t=50$, 用户 U_3 在 $t=10$ 时, 选择对象 O_1 并添加标签 T_1 , 在 $t=30$ 时, 选择对象 O_2 并添加标签 T_2 , 对象 O_3 和 O_4 分别赋有标签 (T_2, T_3) 和 (T_1, T_4) 。在一般的资源分配网络中, 对象 O_3 和 O_4 获得的推荐能力是相同的, 用户 U_3 选择他们的机会是一样的。然而, 既然用户 U_3 选择对象 O_1 并添加标签 T_1 的时刻离当前时刻 t 比选择对象 O_2 添加标签 T_2 更近, 那么含有标签 T_2 的对象 O_3 其推荐能力就应该比对象 O_4 大, 用户选择对象 O_3 比选择对象 O_4 的可能性自然也更大。但是由于忽视了用户添加标签的时间衰减影响, 一般的三部图资源分配网络不能得到较好的推荐结果。

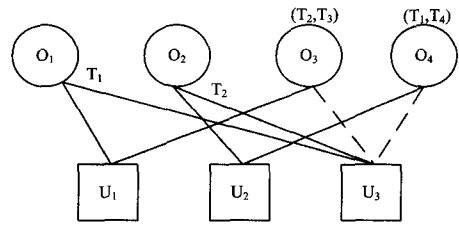


图2 资源分配网络的时间衰减影响

在标签系统中, 用户的评价周期长短不同。对于评价周期较长的用户, 其较早的标签评价对现在的推荐预测影响较小, 而最近添加的标签更能反映出用户的即时喜好, 对推荐预测影响很大。对于评价周期短的用户, 其早期添加的标签对推荐预测影响也比较大。所以定义用户网络活动期为:

$$lifetime(U_i) = t_{in}(U_i) - t_{fi}(U_i) \quad (2)$$

式中, $t_{in}(U_i)$ 表示用户 U_i 最后选择对象并添加标签的时刻, $t_{fi}(U_i)$ 表示用户 U_i 第一次选择对象并添加标签的时刻。

用户不同的网络活动期对资源对象的推荐预测是不同的, 结合用户网络活动期和标签时间衰减的影响, 标签的时间衰减因子定义为:

$$k_{U_i, T_l}(t) = e^{-\frac{(\ln(lifetime(U_i))) \cdot (t - t_{U_i, T_l})}{lifetime(U_i)}} \quad (3)$$

式中, t 为当前时刻, t_{U_i, T_l} 表示用户 U_i 添加标签 T_l 的时间。该式可理解为每个标签在每一次被用户 U_i 使用后所获得的一份推荐能力。

2.2.2 改进的三部图结构算法描述

在标签系统中, 若用户使用某个标签的次数越多, 说明该用户对赋有该标签的对象越感兴趣。将用户 U_i 使用标签的频率与用户添加标签的时间结合起来更能体现出用户的个性偏好。所以本文提出的算法将初始资源定在标签上, 即根据用户使用标签的频率与每次标签由时间加权获得的推荐能力结合起来, 设置初始资源, 最后将标签的初始资源平均分配给添加过该标签的对象, 就得到预测对象的最终资源向量 $\vec{f}''$, 为:

$$f''_j = \sum_{i=1}^r \frac{a'_{ij} a''_{il}}{k'(T_l)} e^{-\frac{(\ln(lifetime(U_i))) \cdot (t - t_{U_i, T_l})}{lifetime(U_i)}} \quad (4)$$

式中, $k'(T_l) = \sum_{j=1}^m a'_{jl}$ 表示标签 T_l 邻居对象的数量。根据式(4)可以得到目标用户 U_i 所有未选择对象的最终资源排序, 将最终资源按照从大到小的顺序推荐前 L 个对象给用户 U_i 。

为了更好地理解该算法, 举个简单的例子, 假设当前时间 $t=100$, 取目标用户 U_1 , 该用户使用过的标签有 T_1, T_2 和 T_3 。 T_1 使用过 3 次, 时间分别在 $t=10, 20$ 和 30 ; T_2 使用过 2 次, 时间分别在 $t=40, 50$; T_3 使用过 1 次, 时间在 $t=60$ 。由式(2)得到用户 U_1 的网络活跃期为 $lifetime(U_1)=70$, 所以, 由式(3)可以得到标签 T_1 的初始资源为:

$$\vec{f}_{T_1} = e^{-\frac{-(\ln 70) \cdot \frac{100-10}{70}}}{3} + e^{-\frac{-(\ln 70) \cdot \frac{100-20}{70}}}{2} + e^{-\frac{-(\ln 70) \cdot \frac{100-30}{70}}}{1} = 0.014 \quad (5)$$

同理, 标签 T_2 和 T_3 的初始资源分别为 $\vec{f}_{T_2}=0.048, \vec{f}_{T_3}=0.088$ 。再按式(4)将 3 个标签的初始资源平均分配给所有含有该标签但没有被用户 U_1 选择过的邻居对象, 就能得到最终资源对象从大到小的排序, 并将前 L 个推荐给用户 U_1 。

3 实验结果及分析

3.1 数据集

本文使用 MovieLens 数据集(www. grouplens. org)来测试算法的性能。MovieLens 是一个评分系统,每个用户可以自由选取电影并给电影打分(1~5 分),2006 年 1 月开始加入标签功能。由于 MovieLens 允许用户自由添加标签,无任何语法限制,导致出现很多奇怪的、长短不同的标签,甚至一些很少用到的符号。为了简化数据,去除 MovieLens 中只出现一次的标签,同时要求去除独立的节点对象,以保证每个用户至少选择一个对象,每个对象至少被两个用户选择过,并且每个对象至少赋有一个标签。数据集的详细信息见表 1。

表 1 数据集的基本信息

数量	类型
957	用户数量
3341	对象数量
3219	标签数量
16430	用户-对象对的数量
32217	总的标签积聚数量

每个数据集被随机分为两个部分:一部分作为训练集,是已知信息;另外一部分隐藏起来作为测试集,以检测算法的准确程度。本文是利用用户使用标签次数结合标签时间进行资源对象推荐的,所以,每个数据集按照用户选择对象添加标签的时间被划分成两个部分。数据以不同的比例分割,如 10% 的测试集对应 90% 的训练集,30% 的测试集对应 70% 的训练集,50% 的测试集对应 50% 的训练集等。

3.2 实验结果

为了测试算法性能,本文从平均排序分、命中率和多样性 3 个指标来分析算法性能。

1)平均排序分^[9]。其用于度量推荐系统的排序准确度,定义为:

$$r_i = L_i / N \quad (6)$$

式中, N 为训练集中用户未选择的对象个数, L_i 为预测集中待预测对象 i 在推荐列表中的位置。排序分越小,说明该系统趋向于把用户喜欢的产品排在前面。将所有用户-对象对的排序分求平均,可以得到平均 $\langle r \rangle$,这个值用来量化推荐算法的精确度, $\langle r \rangle$ 值越小越精确。

将文中提出的利用标签时间加权的推荐算法(Tag Time-Weighted Network, TT-WN)与全局排序算法(Global Ranking Method, GRM)、协同过滤算法(Collaborative Filtering, CF)和基于网络结构的推荐算法(Network-Based Inference, NBI)进行比较。

平均排序分的测试结果如图 3 所示。从图中可以看出,平均排序分大小与数据集中训练集所占比例成反比,本文提出的利用标签时间加权(TT-WN)的推荐方法在精确度上明显好于其他几个方法。

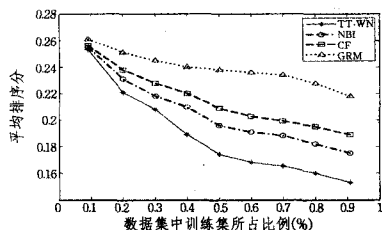


图 3 几种算法对应的平均排序分比较

2)命中率^[7]。该方法依赖于推荐列表的长度,定义为:

$$\text{命中率} = \frac{\text{命中次数}}{\text{列表长度}} \quad (7)$$

式中,若在测试集中出现的用户 U_i 选择了对象 O_j ,并且对象 O_j 也出现在用户 U_i 的推荐列表中,则说明该推荐算法命中 U_i-O_j 组。

表 2 列出了 4 种推荐算法在典型推荐列表长度下的命中率比较。当推荐列表长度 $L=10$ 时,本文提出的推荐方法比 GRM 提高了 76.42%,比 CF 提高了 21.29%,比 NBI 提高了 19.81%。可以看出,本文提出的方法在命中率方面明显好于 GRM、CF 和 NBI 等方法。

表 2 典型推荐列表长度下的命中率

列表长度	GRM(%)	CF(%)	NBI(%)	TT-WN(%)
0	3.52	5.12	4.98	6.21
50	14.31	20.43	23.47	27.63
100	23.44	30.74	34.92	40.33

3)多样性。周涛等^[9]提出利用平均海明距离来度量推荐系统中推荐列表的多样性。用户 U_i 和 U_j 推荐列表的海明距离被定义为:

$$H_{ij} = 1 - Q_{ij} / L \quad (8)$$

式中, L 为推荐列表的长度, Q_{ij} 为系统推荐给用户 U_i 和 U_j 的两个推荐列表中相同对象的个数。推荐列表的多样性定义为 H_{ij} 的平均值 $\langle H \rangle$ 。

表 3 列出了不同推荐算法在不同列表长度下的多样性结果比较。列表长度为 10 时,本文提出的方法多样性比 GRM 提高了 74.07%,比 NBI 提高了 12.30%;在列表长度为 50 和 100 时,本文推荐方法的多样性都比其他方法高。

表 3 不同推荐列表长度下的多样性

列表长度	GRM(%)	NBI(%)	TT-WN(%)
0	0.51317	0.79542	0.89325
50	0.38625	0.65284	0.77364
100	0.31747	0.55743	0.66853

结束语 本文利用标签时间和标签使用频率进行资源推荐,采用 3 个评价指标即平均排序分、命中率和多样性来检测该算法的性能。实验结果表明,相对于 GRM、CF 和 NBI 推荐算法,该算法在准确度、命中率和多样性方面都有明显的优势,提高了推荐算法的性能,同时较好地解决了用户爱好改变对资源的即时需求问题。然而,尽管标签对于资源推荐很重要,但是在标签系统中,标签添加的随意性也带来了标签质量不高的问题,如标签的一词多义和多词一义问题,都会影响资源推荐性能。因此,未来的工作将是解决如何提高标签推荐质量问题。

参考文献

- [1] 刘建国,周涛,汪秉宏. 个性化推荐系统研究进展[J]. 自然科学进展,2009,19(1):1-15
- [2] Golder S A, Huberman B A. Usage patterns of collaborativetagging systems [J]. J. Info. Sci., 2006,32(2):198-208
- [3] Zhang Z K, Lü L Y, Liu J G, et al. Empirical analysis on a keyword-based semantic system [J]. Eur. Phys. J. B 666, 2008:557-561
- [4] Cattuto C, Loreto V, Pietronero L. Semiotic dynamics and collaborative tagging [J]. PNAS, 2007, 104(5):1461-1464

(下转第 129 页)

4.2 汇聚节点软件设计

汇聚节点主要负责建立网络,对节点入网请求进行处理,维护网络路由信息及拓扑结构;将节点采集的数据进行收集并发送给上位机;转发由上位机发送的控制命令。其软件流程图如图 9 所示。

5 实验及结果分析

以某智能家居监控系统为例,对某公寓楼环境参数如温湿度和有害气体进行监控。假设此公寓楼共 5 层,在每层东、西、天花板以及楼梯处各放置 1 个传感器节点,共 20 个传感器节点;节点类型为温度传感器、湿度传感器、CO₂ 气体传感器、SO₂ 传感器,汇聚节点布置在 3 楼中央。下面主要从节点数据测量精度、节点功耗和节点丢包率等方面进行分析。

以温度数据采集为例,将上位机收到的 5 个温度传感器所得到的数据进行处理分析,并与实际用温度计得到的测量值进行对比,其对比图如图 10 所示。

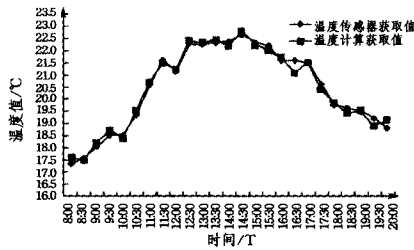


图 10 采集温度与实际温度对比图

从图 10 中可以看出,节点采集的温度值平均误差仅为 0.9%,感知精度较高。

系统运行 40min 后,可以发现,与汇聚节点距离较近的节点既要负责采集数据,又要充当网络的多跳路由节点负责数据包的转发,因此功耗相对较大。节点的功耗与转发数据包的关系如图 11 所示。

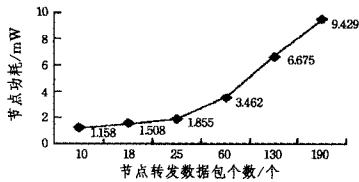


图 11 节点功耗与转发包对应关系

系统运行 15 天后,对实验数据进行分析,计算网络中各类型传感器节点产生的平均数据包和汇聚节点收到的平均数据包并进行对比,判断网络的丢包率,如表 1 所列。

从表 1 中看出,节点的最大丢包率为 4.44%,整个网络的

平均丢包率仅为 2.34%,因此节点具有数据丢包率低的优点。

表 1 网络数据包丢包率统计

传感器	发送包	接收包	丢包率/%
温度	2500	2467	1.32%
湿度	2500	2500	0%
CO ₂	2500	2389	4.44%
SO ₂	2500	2410	3.6%

基于上述实验结果表明,设计的传感器节点不仅能高精度地采集数据,而且满足采集数据的实时性、可靠性的要求。

结束语 针对传统的人工采样环境监测方法具有劳动强度大、实时性低的缺点,本文提出了一种环境监测物联网传感器节点的设计方案,对普通的数据采集节点即传感器节点以及网关节点即汇聚节点进行了软、硬件设计。传感器节点的设计采用 MSP430F1611 作为微控制器,采用 CC2420 作为无线通信模块。汇聚节点采用高端 ARM9 系列的微处理器 S3C2410 作为核心。实验证明,设计的节点具有感知数据精确、功耗低、数据丢包率低、可靠性强等优点。

参考文献

[1] 赵志军,沈强,唐晖,等. 物联网构架和智能信息处理理论与关键技术[J]. 计算机科学,2011,38(8):1-8

[2] Aiger M, et al. Internet of Things in 2020 [EB/OL]. Eposs, May, 2008

[3] 孙利民,李建中,陈渝,等. 无线传感器网络[M]. 北京:清华大学出版社,2005

[4] Tia Gao, Dan G, Matt W, et al. Vital Signs Monitoring and Patient Tracking Over a Wireless Network [C]// Proceedings of the 27 th IEEE EMBS Annual International Conference, September 2005

[5] Pakzad S N, Sukum K, Fenves G L, et al. Multi-iPurpose Wireless Accelerometers for Civil Infrastructure Monitoring [C]// Proceedings of the 5th International Workshop on Structural Health Monitoring (IWSHM 2005). Stanford, CA, September 2005

[6] 何世钧,陈中华,张雨,等. 基于物联网的海洋环境监测系统的研究[J]. 传感器与微系统,2011,30(3):13-15

[7] 杨宏伟,吴挺峰,张维易,等. 基于物联网技术的太湖蓝藻水华预警平台[J]. 计算机应用,2011,31(10):2841-2843

[8] 李长庚,刘威鹏,胡纯意,等. 基于 ARM 和 ZigBee 的 WSN 节点设计与实现[J]. 计算机工程,2010,36(17):135-137

[9] 姜华,袁晓兵,付耀先,等. 无线传感器网络中信道仿真模型的研究[J]. 计算机仿真,2006(11):129-132

(上接第 98 页)

[5] Zhang Z K, Liu C. Hypergraph model of social tagging networks [J]. J. Stat. Mech., 2010, P100005

[6] Zhang Z K, Liu C, Zhang Y C, et al. Solving the coldstart problem in recommender systems with social tags[J]. EPL, 2010, 92 (2): 28002

[7] Zhou T, Ren J, Medo M, et al. Bipartite network projection and

personal recommendation [J]. Phys Rev E, 2007, 76: 046115

[8] Liu J, Deng G S. Link prediction in a use-object network based on time-weighted resource allocation[J]. Physica A, 2009, 388 (17): 3643-3650

[9] Zhou T, Jiang L L, Su R Q, et al. Effect of initial configuration on network-based recommendation [J]. Europhys Lett, 2008, 81: 58004