

文章编号:1001-9081(2011)09-2426-03

doi:10.3724/SP.J.1087.2011.02426

基于多主题追踪的网络新闻推荐

陈 宏¹, 陈 伟²

(1. 浙江科技学院 校园网管理中心, 杭州 310023; 2. 浙江大学 计算机科学与技术学院, 杭州 310027)

(chenhong@zust.edu.cn)

摘要: 针对网络新闻推荐准确率偏低的问题, 提出一种基于多主题追踪的网络新闻推荐算法。基于多主题追踪的推荐算法采用多个用户模型表示用户对不同主题的兴趣, 并动态更新用户模型以动态反映用户的兴趣变化。实现了网络新闻推荐系统的核心推荐算法, 并在标准路透社新闻数据集(RCV1)上验证了算法的有效性, 有效提升了新闻推荐的准确率。

关键词: 新闻推荐; 多主题; 用户模型

中图分类号: TP311.13 **文献标志码:**A

Web news recommendation based on multiple topic tracking

CHEN Hong¹, CHEN Wei²

(1. Information Center, Zhejiang University of Science and Technology, Hangzhou Zhejiang 310023, China;

2. College of Computer Science and Technology, Zhejiang University, Hangzhou Zhejiang 310027, China)

Abstract: A Web news recommendation method based on multiple topic tracking was proposed to improve the precision of recommendation. The proposed algorithm used multiple user profiles to represent user's interests in different topics, and dynamically updated user's profile to reflect the changing of user's interests. The central recommendation algorithm was implemented, and experiments on Reuters Corpus Volume 1 were carried out. The experimental results show that the proposed algorithms can effectively improve the precision of recommendation.

Key words: news recommendation; multiple topic; user profile

0 引言

随着互联网的迅速发展, 上网浏览新闻已经成为许多网民的习惯。中国互联网络信息中心2010年发布的第25次《中国互联网络发展状况统计报告》的统计数据表明: 网络新闻是网民最常使用的网络应用之一, 使用率仅次于网络音乐, 达到了80.1%, 其传播的深度和速度远远领先于传统媒体。然而海量的网络新闻也给人们带来了信息过载问题。推荐系统为用户提供自适应的内容, 能够较好地解决信息过载问题, 已经成为信息检索领域的重要研究方向, 并在商业系统中取得了巨大的成功。

现有的推荐系统方法主要分为三大类: 基于内容的、协同过滤的及混合的推荐方法^[1]。基于内容的方法向用户推荐与过去喜好的内容相似的内容。协同过滤方法则向用户推荐与该用户具有相似兴趣的其他用户过去的偏好信息。而混合型推荐方法则结合了基于内容的和协同过滤方法的优点, 并避免了两个方法存在的问题。网络新闻推荐系统是当前最为流行的推荐系统之一。网络新闻内容通常是高度动态的, 大多数新闻在线的时间比较短。该特性使得纯粹的协同过滤方法在新闻推荐系统中并不是很适用^[2]。大多数先前的工作都是基于内容的方法来作新闻的推荐^[3-5], 但是现有的方法都较少考虑网络新闻及用户兴趣的动态变化特性。

本文提出了一种基于多主题追踪^[4]的网络新闻推荐算法, 该算法采用多个用户模型表示用户的兴趣, 并动态跟踪用户的兴趣变化。在路透社新闻数据集(Reuters Corpus Volume

1, RCV1)上验证了提出的算法, 有效地提高了推荐的准确率。

1 基于多主题追踪的网络新闻推荐

日常生活中, 人们通常对多个主题的新闻感兴趣, 持续关注若干个进展中的新闻事件。如某从事投资的网民在世界杯期间, 除了关注财经类的新闻外, 还可能持续关注世界杯相关的新闻。因此不难理解, 可以把用户阅读的新闻按其所属的主题分成若干个类, 每一类表示用户的一个兴趣偏好。这样对于待推荐给用户的新闻, 只要比较该新闻同这些类别之间的相似程度, 即可判断用户对该新闻是否感兴趣。

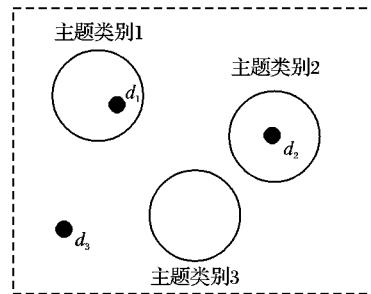


图1 追踪用户对多个主题的兴趣

图1给出了追踪用户对于多个主题的兴趣的基本思路。假设用户已经阅读了一些新闻, 这些新闻属于3个互不相同的类别, 即图中的3个主题类别。对于3篇待推荐给用户的新闻 d_1 、 d_2 和 d_3 , 计算它们与3个主题类别之间的相似度。从图中可知, d_3 与用户已经阅读的新闻的主题类别相似度较低, 而 d_1 与 d_2 则与已阅读的新闻有较大的相似度。这样, 可以优

收稿日期:2011-01-04;修回日期:2011-01-24。 基金项目:浙江省教育厅科研项目(Y200908583)。

作者简介:陈宏(1967-),男,浙江杭州人,助理研究员,硕士,主要研究方向:管理工程; 陈伟(1983-),男,浙江义乌人,博士,主要研究方向:信息检索、数据挖掘。

先将新闻 d_1 与 d_2 推荐给用户。

用户的兴趣总是在不断变化着的,比如世界杯开赛后,用户会突然关注世界杯相关的新闻报道。因此,必须有效地捕捉用户动态更新的兴趣,依据用户持续的反馈,不断地更新用户兴趣模型,为用户提供全方位的新闻推荐服务。

总结起来,基于多主题追踪的网络新闻推荐需要综合考虑多主题的用户模型的表示、用户模型的更新以及推荐列表的产生等问题。

1.1 多主题用户模型

传统的新闻推荐算法如 Rocchio 通常将表示用户特性的信息都包含在一个用户模型中^[4]。由于用户的兴趣通常涉及较多的方面,为用户建立单一的用户模型会降低模型的代表性,并导致用户模型过于概化,无法很好地体现用户对多个主题兴趣的情况,进而影响推荐系统的推荐效果。

这里给出多主题用户模型的定义。多主题用户模型是多个主题的用户偏好向量集合,定义为 $P = \{p_1, p_2, \dots, p_i, \dots, p_m\}$ 。其中每个用户偏好向量 $p_i = [w_{i,1}, w_{i,2}, \dots, w_{i,j}, \dots, w_{i,n}]$ 表示用户对某一主题的偏好程度,其中 $w_{i,j}$ 表示的是与某个主题 i 相关的特征 j 的权重, n 为包含的特征的总数。

1.2 多主题用户模型的更新

当用户阅读了推荐给他的某篇新闻后,需要对用户模型进行更新。假设给用户推荐了一篇新闻报道 d (将 d 表示成一个向量 $d = [d_1, d_2, \dots, d_j, \dots, d_n]$, 其中 d_j 表示的是特征 j 在某篇新闻报道中的 TF-IDF 值^[6]),计算 d 和所有用户偏好向量 p_i 之间的余弦相似度,如下:

$$\cos(d, p_i) = \frac{d \cdot p_i}{|d| \cdot |p_i|} = \frac{\sum_{j=1}^m d_j \cdot w_{i,j}}{\sqrt{\sum_{j=1}^m d_j^2} \cdot \sqrt{\sum_{j=1}^m w_{i,j}^2}}$$

若相似度值小于指定阈值 t_{min} ,且用户阅读了该新闻,自动以新闻报道 d 创建一个新的用户偏好向量 p_{m+1} ,并将其添加到多主题用户模型 P 中。若存在相似度值大于指定阈值 t_{min} 的 p_i (取相似度最大的 p_i),且用户阅读了该新闻,则更新与 d 最相似的用户偏好向量 p_i , $p_i = p_i + d$ 。若用户未阅读该新闻,同样更新用户偏好向量 p_i , $p_i = p_i - \gamma \cdot d$ (γ 用于控制 d 对 p_i 的影响度,实验中置为 0.5)。

对于每一个用户模型 p_i ,系统通过它推荐的所有信息来计算它代表用户使用兴趣的有效性。

$$effect(p_i) = \frac{p_i \text{ 推荐且用户收听过的新闻总数}}{p_i \text{ 推荐的新闻总数}}$$

同时,我们认为随着时间的推移, p_i 的代表性会慢慢衰减,即用户模型具有一定的时新性。因此,本文引入了一个衰减函数如下:

$$e^{-\beta\rho} = 1/2$$

其中 ρ 表示用户模型 p_i 的半衰期。

维护多个用户模型可以提高表示用户兴趣特征的精确度,但用户模型个数的增加势必会增大系统的负担。因此,考虑将精确度低的用户模型和用户长时间不用的用户模型优先去掉。本文采用用户模型的精确度与衰减函数的乘积来表示这一特性。即,当系统中的用户模型个数超过预定个数 M 时,将 $effect(p_i) \cdot e^{-\beta\rho}$ 值最小的用户模型 p_i 从系统中删除。

1.3 推荐列表的产生

综合考虑新闻与用户模型之间的相似度,以及用户模型本身的精确度和衰减程度,根据三者的乘积最终将获得一个

用于新闻排序的值($score$)。在最终提交给用户的新闻推荐列表中,新闻报道将根据 $score$ 的值降序排列。 $score$ 的计算公式如下:

$$score(d) = \cos(d, p_i) \times effect(p_i) \times e^{-\beta\rho}$$

2 实验和结果分析

2.1 数据集

本文在 RCV1 上验证算法的有效性。RCV1 包括从 1996 年 8 月到 1997 年 8 月期间超过 800 000 篇经过人工分类的新闻报道,并已用 XML 标签对数据进行过标注。数据按时间(单位:d)排列,并已编号。除 RCV1 外,实验中还用到了文本检索会议(The Text REtrieval Conference, TREC11)所使用的数据集,它提供了人工归纳总结的 50 个 RCV1 数据覆盖的主题以及与每一个主题相对应的训练新闻列表和测试新闻列表^[7]。

2.2 评价标准

推荐新闻的准确性是新闻推荐系统的重要指标。因此,需要对推荐结果的准确度,也就是新闻推荐系统的推荐性能进行评价。目前衡量推荐系统准确度的两种评价指标为:查准率($precision$)和查全率($recall$)^[6]

对于某一新闻数据集合,设有用户 I , I 喜欢的相关新闻集合为 R ,用 $|R|$ 表示该集合中的新闻数目。假设待评价的新闻推荐系统由 I 提供推荐服务,推荐结果集合为 A ,用 $|A|$ 表示该集合中的新闻数目。另外,设 $|R_a|$ 表示集合 R 和 A 的交集中的新闻数目。则查准率和查全率的定义如下:

查准率 指系统推荐出的用户感兴趣的新闻与推荐的新闻总数 $|A|$ 的比值,即: $precision = |R_a| / |A|$ 。

查全率 指系统推荐出的用户感兴趣的新闻与用户喜欢的新闻总数 $|R|$ 的比值,即: $recall = |R_a| / |R|$ 。

由于使用的新闻数据集包含的新闻数据量非常庞大(RCV1 包括超过 800 000 篇的新闻),其中与 TREC 11 给出的与 50 个主题相关的新闻也达 60 000 多篇。由于用户阅读的新闻数据量有限,因此这里只考查推荐的查准率。

TREC 11 还使用了统一的 T11SU 指标来评价推荐系统。T11SU 评价标准定义如下:

$$T11NU = (2 \times | \text{推荐的用户感兴趣的新闻} | - | \text{推荐的用户不感兴趣的新闻} |) / (2 \times | \text{用户所有感兴趣的新闻} |)$$

$$T11SU = \text{Max}(T11NU, 0.5) - 1$$

2.3 实验结果

实验中比较了两组策略的推荐效果,第一组直接利用训练数据获得用户的偏好向量集合,并基于此作推荐;第二组策略在第一组策略的基础上,利用前文提出的算法加上对用户偏好向量集合的动态更新。图 2 给出了策略一和策略二分别推荐 100、300、400 和 1 000 篇新闻时查准率。从图中可以看出,策略二的效果较策略一有明显的提升,当推荐列表大小为 400 和 1 000 时,提升的幅度达到了 25%。策略二由于加入了动态更新的因素,推荐的效果随着推荐数的增多而有所提升。而策略一中由于训练数据有限,在推荐列表大小为 1 000 时,推荐的效果反而下降了。

此外,还统计了针对每个月的数据推荐 10、30、40、100 和 1 000 篇新闻的查准率。如图 3 和图 4 所示。

从图 3 中可以看出,查准率在不同月份之间差别较大,这

主要跟数据集本身有紧密的联系: RCV1 中的新闻涉及范围比较大, 将其按月分开后, 与 TREC 11 提供的 50 个主题相关的新闻的分布也会随之不稳定。此外, 不难发现当推荐列表增大到 1000 时, 推荐的水平明显低于其他情况。从图 4 中可以看出, 策略二的推荐查准率较策略一稳定, 且推荐列表大小为 1000 时, 其查准率水平与其他情况的差距, 相比策略一有所减小。

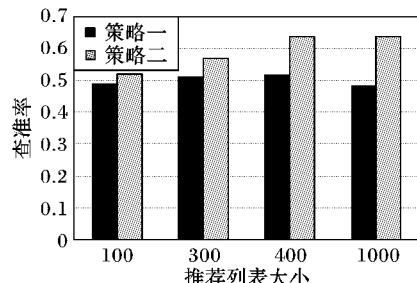


图 2 策略一与策略二推荐 100、300、400 和 1000 篇新闻的查准率

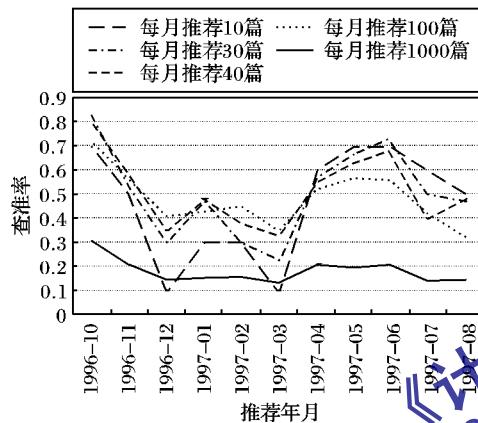


图 3 策略一每月推荐 10、30、40、100 和 1000 篇新闻的查准率

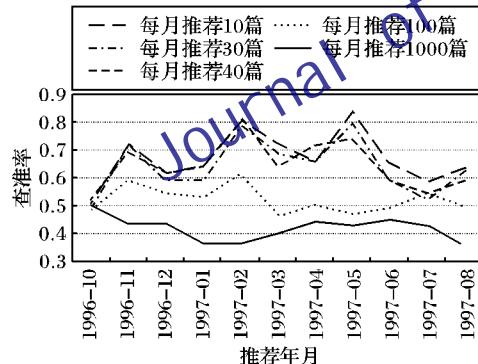


图 4 策略二每月推荐 10、30、40、100 和 1000 篇新闻的查准率

此外, 将策略二的查准率转化为 $T11SU$ 后, 其总体水平

亦优于 TERC11 的公布的数据。表 1 和表 2 分别是策略二一次性推荐和分月推荐时的 $T11SU$ 值。

表 1 一次性推荐 $T11SU$ 的值

推荐列表大小	$T11SU$	推荐列表大小	$T11SU$
100	0.12	400	0.825
300	0.42	1000	0.828

表 2 分月推荐 $T11SU$ 的值

推荐列表大小	$T11SU$	推荐列表大小	$T11SU$
10	0.90545	40	0.92727
30	0.91034	100	0.76909

3 结语

本文研究了基于内容的网络新闻推荐技术。与传统的新闻推荐系统仅维护单一的用户模型不同, 本文在新闻推荐中引入了多主题追踪技术, 结合代表用户多种兴趣特征的多主题兴趣模型, 在提高新闻推荐的准确度方面作了研究。最后在标准数据集 RCV1 上验证了基于多主题追踪的新闻推荐算法的有效性。

参考文献:

- [1] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734 - 749.
- [2] BILLUS D, PAZZANI M. Adaptive news access[C]// The Adaptive Web, LNCS 4321. Berlin: Springer-Verlag, 2007: 550 - 570.
- [3] AHN J, BRUSILOVSKY P, GRADY J, et al. Open user profiles for adaptive news systems: help or harm? [C]// Proceedings of the 16th International Conference on World Wide Web. New York: ACM, 2007: 11 - 20.
- [4] KON P, CARDENAS A, BUTTLER D, et al. Tracking multiple topics for finding interesting articles[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2007: 560 - 569.
- [5] ARDISSONO L, CONSOLE L, TORRE I. An adaptive system for the personalized access to news [J]. AI Communications, 2001, 14(3): 129 - 147.
- [6] BAEZA-YATES R, RIBEIRO-NETO B. Modern information retrieval [M]. New York: Morgan Kaufmann, 2005.
- [7] VOORHEES E. Overview of TREC 2002[C]// TREC 2002: Proceedings of the 11th Text Retrieval Conference. Washington: NIST Special Publication, 2002: 1 - 15.
- [8] ZHANG Y-C, MEDO M, REN J, et al. Recommendation model based on opinion diffusion [J]. Europhysics Letters, 2007, 80(6): 68003.
- [9] ZHOU T, JIANG L-L, SU R-Q, et al. Effect of initial configuration on network-based recommendation [J]. Europhysics Letters, 2008, 81(5): 58004.
- [10] PALLA G, DERENYI I, FARKAS I, et al. Uncovering the overlapping community structures of complex networks in nature and society [J]. Nature, 2005, 435(7043): 814 - 818.
- [11] 杨博, 刘大有, LIU JIMING, 等. 复杂网络聚类方法[J]. 软件学报, 2009, 20(1): 54 - 66.

(上接第 2411 页)

- [5] LIU R R, JIA C X, ZHOU T, et al. Personal recommendation via modified collaborative filtering [J]. Physica A, 2009, 388(4): 462 - 468.
- [6] CHANG Y I, SHEN J H, CHEN T I. A data mining-based method for the incremental update of supporting personalized information filtering [J]. Journal of Information Science and Engineering, 2008, 24 (1): 129 - 142.
- [7] SARWAR B, KARYPIS G, KONSTAN J, et al. Item-based collaborative filtering recommendation algorithms [C]// WWW'01: Proceedings of the 10th International World Wide Web Conference. New York: ACM, 2001: 285 - 295.

- [8] ZHANG Y-C, MEDO M, REN J, et al. Recommendation model based on opinion diffusion [J]. Europhysics Letters, 2007, 80(6): 68003.
- [9] ZHOU T, JIANG L-L, SU R-Q, et al. Effect of initial configuration on network-based recommendation [J]. Europhysics Letters, 2008, 81(5): 58004.
- [10] PALLA G, DERENYI I, FARKAS I, et al. Uncovering the overlapping community structures of complex networks in nature and society [J]. Nature, 2005, 435(7043): 814 - 818.
- [11] 杨博, 刘大有, LIU JIMING, 等. 复杂网络聚类方法[J]. 软件学报, 2009, 20(1): 54 - 66.