基于维基百科层次分类框架的主题推荐系统的研究*

谢 科，刘奕群，岑荣伟，马少平，茹立云，杨 磊

智能技术与系统国家重点实验室，清华信息科学与技术国家实验室（筹），清华大学 计算机系，北京 100084

E-mail: oeddyo@gmail.com

摘 要：在用户使用互联网的过程中，并不一定经常持有明确的目的性，比如浏览新闻网站时用户可能会被各种不同主题的新闻链接所吸引。但是总体来说，特定用户的兴趣在一段时期内来讲，是趋于固定的。如果能在用户点击日志中，识别其可能感兴趣的主题，同时预测其感兴趣的其他主题或条目，可以帮助用户“探索”他们可能感兴趣的内容。随着推荐系统领域的发展，电影推荐以及音乐推荐已趋近成熟，但此类研究多是建立在已有数据集上，从来没有过对用户点击行为进行分析并做出浏览推荐的研究。本文作者从某浏览器点击记录中，提取出部分用户的浏览记录，分析并整理为实验所用的数据集。同时，我们提出了一种简单有效的框架，即通过对用户点击链接文本的分析，利用汉语维基百科建立索引并分析用户的兴趣，采用协同过滤算法预测用户可能感兴趣的其他主题。通过实验，我们的推荐框架可以较好地描述用户兴趣，即使用简单的协同过滤算法也可以达到良好效果。同时我们的算法运算速度快，可对用户兴趣发掘、文本分类及协同过滤相关的研究领域产生一定的指导意义。

关键词：兴趣发掘；浏览推荐；推荐系统

Research of a Topic Recommendation System Based on Wikipedia Category Hierarchy

Xie Ke, Liu Yiqun, Cen Rongwei, Ma Shaoping, Ru Liyun, Yang Lei

State Key Laboratory of Intelligent Technology and Systems,Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084

E-mail: oeddyo@gmail.com

Abstract: When users are surfing the Internet, it is not necessary that they hold specific purposes. For example, users would be attracted by different news with diverse topics. But generally, a specific user has stable interest in terms of a short period of time. If we can specify their interests from their browsing history or clicking history, and recognize the topics they like, we can help them discover topics they might like. With the rapid growth of Recommendation System, computer can recommend movies and music with high precision, but researches in these fields are based on existing data-sets. To our best knowledge, there are no previous researches on recommending items for users based on users' click history. In this paper, we propose a new simple but effective framework, upon which we can analyze users' interests by anchor-text they clicked. We use Chinese-edition of Wikipedia to build an index, and "search for" users' interests on the index. At the same time, we performed an experiment by using KNN recommendation algorithm to predict users' likes and dislikes. According to our experiments' results, our framework can produce high quality recommendations. At the same time, the speed for generating the recommendations is high. We believe our research can give insights for researchers in the field of discovering users' interests, text categorization and collaborative filtering.

Keywords: user's interest discovery; browsing recommendation; recommendation system

1 前言

随着搜索引擎的快速发展，用户可以非常方便快捷地找到自己所需要的信息，从而满足自身的信息需求。然而，搜索引擎仅能够解决用户明确持有信息需求的应用场景，很多情况下，用户并不清楚自己需要什么信息。比如，在电影推荐的场景中，某用户可能知道自己喜欢观赏动作片

* 本文承自然科学基金(60736044，60903107）和高等学校博士学科点专项科研基金（20090002120005）的资助。

类型，但是在没有看过功夫片之前，这个用户可能并不知道功夫片的存在，因此也不可能通过搜索引擎找到潜在的可能喜欢的功夫片。而推荐引擎的出现，则可以在很大程度上帮助用户“探索”他们可能感兴趣的内容。同时，推荐引擎可以帮助企业更精准地定位用户需求，从而在帮助用户实现需求的同时创造更多价值[2]。如果我们能够较为准确地通过用户的行为定位其兴趣，那么就有可能帮助其发掘出这些用户还有可能感兴趣的其他内容。

在本文中，我们提出了利用维基百科分类树，进行用户兴趣发掘的框架，并利用 K 邻近算法进行推荐实验。通过实验证实我们的推荐框架具有较好的准确性和实用性。

本文的后续内容组织如下：第二部分是相关工作概述，第三部分简单描述我们使用的用户点击数据，第四部分详细介绍我们的兴趣发掘框架，第五部分描述我们在此框架下进行实验的一些结果。最后，在第六部分我们提出结论以及改进方向。

2 相关工作概述

在早期的推荐系统研究中，研究者们通常采用一个预定好的数据集，比如来自明尼苏达大学的 MovieLens 数据集，并在数据集的基础上测试推荐算法和评价算法效果。这些数据集中，一般包含特定数量的用户，以及特定数量的待推荐物品(item)，同时有每个用户对这些物品的实际评分。在推荐算法的评价指标中，最常采用的是 RMSE，即 Root Mean Square Deviation，来衡量推荐系统预测用户的打分值与用户真实打分值之前的差异[3]。

然而，推荐系统算法的有效性和进行实验的数据有较大的关联。首先，数据噪声会较为严重地影响到推荐效果。其次，数据可能非常稀疏，因为用户通常不会非常明确地向数据收集系统表达自己的喜好，因此，在通用的数据集上进行的实验有利于横向比较。但是，在电影推荐上有效的算法框架并不一定适合其他领域。比如说，在浏览推荐的任务中，研究者是无法获得用户的评分数据，只能通过其他方法来预测用户兴趣。浏览数据的稀缺也是浏览推荐任务鲜有研究者触及的原因之一。我们的研究得到某浏览器公司的支持，并得到了较大规模的用户浏览数据。

在相关研究中，Xiaobin Fu 等人[4]利用关联规则，在用户浏览数据上进行推荐，但并不考虑用户点击的 anchor 文本信息。这样的推荐效果虽然关联度非常高，但是很明显的局限性是，它可能给用户推荐出并不是非常有价值的内容。例如，在他们展示的 demo 中，用户点击加州大学伯克利分校的网页后，为用户推荐出密苏里大学，俄亥俄大学的页面。这样的结果非常合理，但是对于用户来说，意义可能并不大，因为结果过于集中和专一。

利用英文维基百科构造内容分类和文本挖掘的工作也有研究者涉及，如 Fabian M. Suchane 等人[5]曾用 Wikipedia 建立起一个语义知识库。Wikipedia 贡献者们用他们的集体智慧，为每一个页面指定了分类，同时构建起一个完整的分类树，在我们的研究中，我们利用这棵分类树，来分析用户点击所代表的用户兴趣。

3 浏览数据

在我们的实验中采用的用户浏览数据，全部来源于某浏览器的真实用户浏览数据。数据中记录了使用该浏览器的用户的编号，点击链接文本，链接所对应的 URL。

表 1 用户浏览日志格式

Field	含义
Machine Number	用户所对应的机器号，用于区分用户
Anchor Text	用户所点击的链接文本
URL	点击链接所对应的 URL

4 推荐框架

在我们的推荐框架中，主要包含两个部分，即兴趣发掘和协同过滤。在兴趣发掘的部分，我们对维基百科的所有页面建立倒排索引，利用维基百科的分类树，通过将链接文本作为 query 的方法，来确定一个特定的链接文本所属的页面。由于维基百科的每一个页面均有对应的人工标注的分类，我们利用查询到的页面的分类，回溯维基百科的分类树，并在用户兴趣向量上为对应分类树上每个结点加上对应分值，最后，用用户兴趣向量作为输入，进行协同过滤。本节后续部分将详细介绍我们使用的方法。

4.1 兴趣发掘

4.1.1 页面索引

中文维基百科共有页面 353,055 个，我们将其所有页面用开源搜索引擎 Lucene 建立倒排索引，分别索引标题字段、正文及分类字段。由于贡献者们通常活跃于互联网，因此在维基百科里的条目通常可以非常快的更新速度，这也是我们选用索引维基百科的原因之一。

4.1.2 维基百科分类

在维基百科页面中，每一个页面均有一到数个人工标注的分类。由于维基百科的分类会有志愿者严格控制，分类质量较高。图 1 为部分分类树的截图。

2 应用科学: 3
  -3 健康科学: 1
    4 兽医学: 4 / 医学: 106 / 卫生保健: 8 / 药学: 29 / 护理学: 4
  3 军事科学: 0
  3 工業設計: 9
    4 消費品: 18
  3 工程技术: 42
    4 交通运输: 9 / 化学工程: 56 / 土木工程: 26 / 工程技术模板: 0 / 巨型工程: 5 / 度量儀器: 34 / 控制论: 18 / 机械工程: 69 / 材料科学: 57 / 照明: 25 / 環境工程: 程: 2 / 设计: 49 / 鐵路車輛工學: 3 / 電機工程: 26 / 项目管理: 25
  3 应用物理: 6
    4 地球物理学: 11 / 大气科学: 10 / 材料科学: 57 / 物理化学: 24 / 生物物理学: 7
  3 建筑: 35
    4 地板: 8 / 屋顶: 15 / 窗: 7 / 建筑业: 1 / 建筑学: 21 / 建筑师: 10 / 建築列表: 3 / 建築物: 58 / 東亞建築: 0 / 綠色建筑: 2 / 裝飾: 3 / 金字塔: 18
  3 建筑学: 21
    4 室内設計: 11 / 建筑史: 32 / 建筑技术: 7 / 建筑构造: 7 / 建筑经济: 5 / 建筑装饰: 15 / 建筑设计: 24 / 建築則例: 0 / 建築學家: 4 / 建築風格: 9 / 环境艺术: 3 / 建
  3 應用數學: 15
    4 博弈论: 36 / 密码学: 44 / 数值分析: 17 / 数字信号处理: 40 / 数理经济学: 6 / 混沌理论: 4 / 生物數學: 1 / 算法: 48 / 统计学: 66 / 計量經濟學: 9 / 运筹学: 7 / 小
  3 核科学: 0
    4 核武器: 21
  3 测绘学: 27
    4 地图学: 10 / 坐标参考系: 3 / 大地测量学: 5 / 误差理论: 6
  3 資訊科學: 28
    4 人工智能: 64 / 信息技术: 31 / 信息检索: 5 / 图表: 27 / 圖書資訊科學: 126 / 大眾媒體: 26 / 情报检索: 2 / 搜索: 7 / 生物信息學: 16 / 知识表示: 11 / 系統理論 科技教育: 5
  3 都市設計: 6
    4 城市中轴线: 2
  3 量度: 17
    4 度量儀器: 34 / 度量指标: 9 / 度量衡: 47 / 比率: 13 / 测绘学: 27 / 测试: 8 / 评价: 7 / 调查: 10
2 心理学: 152
  3 人類成長: 18
    4 人際關係: 48 / 兒童: 36 / 婦產醫學: 4 / 年齡: 6 / 成人禮: 9 / 死亡: 61 / 老年: 7 / 育兒: 7 / 青少年: 1 / 青春期: 3
  3 人類行為: 4
    4 人權互動: 8 / 人際關係: 48 / 歧視與差別待遇: 1 / 虐待: 5 / 選舉學: 1
  3 價值觀: 11
    4 功利主義: 6
  3 夢: 10
  3 市场心理学: 20
  3 应用心理学: 7
  3 心理学小作品: 0
  3 心理学爱好者: 17
  3 心理学分支: 11
    4 發展心理學: 15
  3 心理学家: 20
    4 俄罗斯心理学家: 3 / 各国心理学家: 0 / 奧地利心理學家: 5 / 德國心理學家: 3 / 心理學家小作品: 0 / 精神病學家: 1 / 美国心理学家: 56 / 英国心理学家: 8

图 1 维基百科分类树部分截图

对于每一个页面，其分类均为分类树的叶子结点上的分类。因此，我们需要回溯以便可以获得更加泛化的分类。

比如，检索用户的点击记录“中韩乒乓球大对决”，我们将其对应到页面“乒乓球”，如图 2

所示。而其在分类树中的结构为：

体育<-球类运动<-乒乓球

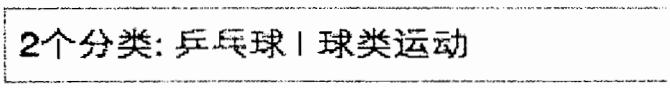

图2 页面“乒乓球”的分类

对于数据集中每一个用户 u，我们对其每次点击记录进行检索，并按照公式(1)，计算单次点击用户对某一分类 i 的喜好值

$$Score(i)=1-\left(\frac{1}{k^n}\right) \qquad \text{公式(1)}$$

其中，n 为 i 这个分类在分类树中的深度，而 k 是常数。在试验中我们发现，$k=2.1$ 时有较好的实验结果。由公式(1)可以知道，越是一般化的分类，n 值越小，每次获得的分数越低。这样的原因是，泛化的分类通常可能多次获得评分，比如无论链接是“乒乓球”还是“足球”都会给“体育”这个分类加上对应分值，因此需要给泛化的分类降权。

4.1.3 兴趣向量

我们定义一个 N 维向量 Tr，并定义 $Tr(i)$ 为 Tr 中第 i 维，N 为分类树中结点数。

$$Tr=[T(1),...,T(N)]$$

同时，我们用 $Tr(i)$ 表示向量 Tr 的第 i 维。

对于用户 u 所有点击的文本 Pu，我们将 $p \in Pu$ 作为 query 在维基百科索引中进行检索，将对应的结果页面上的所有分类在分类树上进行回溯，并记回溯路径上的所有分类结点为集合 C。那么对于 p 生成的 C，我们可以生成一个兴趣向量

$$Tr(p)=[T(1),...,T(N)]$$

其中，若 c 属于 C，则

$$T(c)=Score(c)$$

否则

$$T(c)=0$$

那么对于用户 u，他的兴趣向量为

$$Tu=\sum_{p\in Pu} Tr(p)$$

由于在 Tu 中，各维的值范围不确定，若用户只点击某一主题的链接，那么对应维度的值将会过大。由此，我们对 Tu 进行归一化。用 $\max(Tu)$ 表示取 Tu 所有维度中的最大值，那么则有用户兴趣向量为：

$$Tu(i)_{norm}=\frac{5\times Tu(i)}{\max(Tu)}$$

这样便将用户的兴趣向量每一维的值限制在 $[0,5]$ 的范围内，可以方便我们实验结果的评测。

4.2 推荐算法

我们在数据集上测试了 K 最近邻协同过滤算法。由于本文的主旨并不在于比较不同的推荐算法，而是提出主题推荐的方法和框架，因此我们仅使用 K 最近邻协同过滤算法进行实验验证。

4.2.1 K 最近邻协同过滤算法

对于特定用户$u \in U$，我们在用户集合U中，寻找K个与之最“相似”的用户[6]。在我们的算法中，我们定义用户i和j之间的相似度为 Cosine-based Similarity，即

$$sim(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|_2 \times \|\vec{j}\|_2} \qquad \text{公式(2)}$$

其中$\vec{i}$,$\vec{j}$为用户i与用户j的兴趣向量。在计算每个用户与其他所有用户的相似度之后，我们取与该相似度最高的K个用户，记为集合Nb。

我们预测用户u对某一个主题t的评分为

$$rating(u,t) = Ru + \frac{\sum_{v \in Nb} (Rv(t) - Rv) \times sim(u,v)}{\sum_{v \in Nb} |sim(u,v)|}$$

其中，Nb为用户u的K近邻集合，Ru为用户u的评分平均值，$Rv(t)$为集合Nb中的用户v对主题t的评分值，而Rv为用户v的评分平均值，$sim(u,v)$即按公式(2)计算的两个用户间的相似度。

5 实验设计及结果

我们对浏览器日志中 2011 年 04 月 02 号的日志 4,300,000 条点击数据进行清洗处理。处理过程主要是滤去色情内容、广告内容以及删除掉所有单天点击超过 1,000 个链接的用户点击记录。最后日志记录共有 2,256,075 条点击记录。在清理后，数据集中共有用户数为 46,197，待推荐主题共 32,129。矩阵稀疏程度为 0.1519%。

5.1 评价指标

为了方便横向比较，我们采用常用的评价指标 RMSE，即 Root Mean Square Deviation 作为评价指标。RMSE 的定义为:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - R_i)^2}{n}}$$

n为预测评分的数量，Pi为我们对某用户u对主题i喜好程度的预测分数，R_i为用户u对主题i表示的真实预测分数。由此，RMSE 值越小，说明推荐越精确。若 RMSE 为 0，则表示算法可以精确到每次预测的分数均与用户的真实喜好程度一致。

我们将评分数据按$P = 7:3$的比例，分为训练集和测试及，然后挑选出产生最佳结果的产数。最后再在该参数下试验P的不同取值对实验结果的影响。

5.2 实验结果与讨论

我们在不同参数下重复进行实验，得到了一系列 RMSE 值，如图 3 所示。可以看到，考虑到我们数据集的稀疏性，RMSE 值稳定维持在 1.1 左右，这表示我们对用户对某主题评分的预测是比较准确的。同时，在$K = 2$时我们得到最好的 RMSE 值，这也验证了数据的稀疏性，因为K在取较大值时，加入了一些相似度并不高的用户，产生了噪音，影响了算法对其评分的计算准确度。可以预见，如果持续增加用户数据，我们可以达到更好的实验效果。而图 4 中的结果也验证了这个假设。

在固定 K=2 以后，我们用不同的训练集大小进行训练，结果如图 4 所示，可见数据越稠密，推荐的效果就越好。

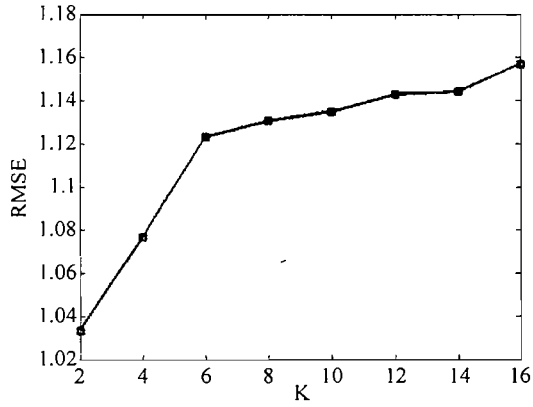

图3 RMSE 随 K 值的变化图

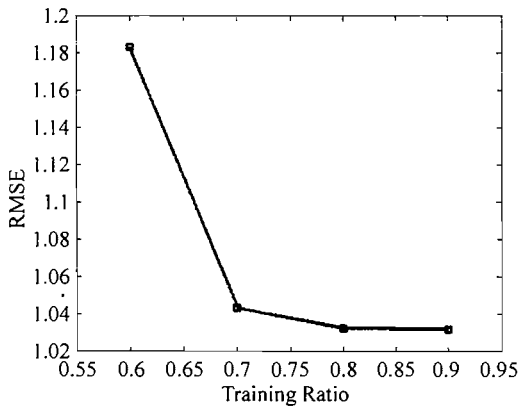

图4 RMSE 随训练集大小的变化

6 结语

本文提出了用维基百科分类进行用户兴趣分析及主题推荐的框架。由于维基百科分类规则明确，页面分类质量较高，因此我们可以从用户的点击较准确地获得其对应的主题。由第 5 节中的实验结果可以看出，本框架可以较好地刻画出用户兴趣，且简单易行。在此基础上，简单的协同过滤算法便可以产生较好的实验效果。同时，该框架不需要大量计算，仅需一次索引检索，并回溯遍历分类树即可得出用户的兴趣，适用于在线计算。

在此基础之上，我们将会从以下几个方向继续改进算法：

1. 进一步完善其在稀疏数据集上的表现。
2. 加入新词发现，主体词提取等技术，进一步增加获得分类的准确度。
3. 将页面进行主题聚类，并用推荐的主题进行页面推荐，而非仅推荐主题。
4. 融合多种推荐算法，以期获得更佳的推荐效果[7]。

参 考 文 献

[1] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering[C]. IEEE Internet Computing, Jan/Feb.: 76-80, 2003.

[2] Herlocker, J. L. and Konstan, J. A. and Terveen, L. G. and Riedl, J. T. Evaluating collaborative filtering recommender systems[C]. ACM Transactions on Information Systems, 5-53, 2004.

[3] Fu, X. and Budzik, J. and Hammond, K. J. Mining navigation history for recommendation[C]. Proceedings of the 5th international conference on Intelligent user interfaces. 106-112, 2000.

[4] Suchanek, F. M. and Kasneci, G. and Weikum, G. Yago: a core of semantic knowledge[c]. Proceedings of the 16th international conference on World Wide Web[C]. 697-706, 2007.

[5] Balabanovi, M. and Shoham, Y. Fab: content-based, collaborative recommendation[C]. Communications of the ACM, 5-53, 2004.

[6] Bell, R. M. and Koren, Y. Lessons from the Netflix prize challenge[C]. ACM SIGKDD Explorations Newsletter. 75-79, 2007.