基于主题的个性化新闻推荐系统的设计与实现

刘金亮，卢美莲
（北京邮电大学网络技术研究院宽带网中心，北京 100876）

摘要: 随着互联网的飞速发展所带来的"信息过载"问题使准确的新闻推荐技术变得越来越重要。本文提出了一个基于内容与协同过滤混合的新闻推荐方法，将概率主题模型引入到新闻推荐中，挖掘用户的主题兴趣，并在此基础上完成了新闻推荐系统的设计与实现。文章介绍了个性化推荐和主题模型相关技术，详细描述了新闻推荐方法和系统设计，并对系统进行了实现。

关键词：计算机应用；个性化推荐；主题模型；基于内容；协同过滤
中图分类号：TP391

The Design and Implementation of Personalized News Recommendation System Based on News Topic

Liu Jinliang, Lu Meilian
(State Key Lab of Networking & Switching Technology,Beijing University of Posts&Telecommunications, Beijing 100876)

Abstract: With the rapid development of the Internet brought about by the "information overload" make accurate news recommend technology becomes more and more important. In this paper, we proposed a recommended methond based on news topic and collaborative filtering. This method use topic models into the text recommend, in-depth excavation user's interest. Besides, we completed a personalized news recommendationg system design and implementation. We introduce the personalized recommendation and topic model technologies, a detailed description of the news recommended method and system design.A news recommendation system based on this design will be illustrated.

Key words: Computer application;Personalized Recommendation; Topic Models; Content-Based; Collaborative Filtering

0 引言

随着互联网的快速发展和广泛应用，网络资源逐渐成为人们获知信息的重要渠道，全球范围内每天都会有数以亿计的网络信息涌现，面对如此海量的信息资讯，用户仅仅依靠自身搜索往往不能够有效地获取高质量的适合自己的有用信息[1]。

专业的新闻资讯类网站的数量数以万计，非专业的个人维护的资讯类网站的数量也很多。大量的新闻资讯网站为人们提供了充足的信息，与此同时，也为用户带来了新的问题与挑战。用户往往不会只喜好或需要一个单一分类的新闻信息，而是一般都同时关注数个类别的新闻，这就需要用户同时关注数个网站的更新情况，而如此做法，无论是时间上，还是精力上，都会给用户带来负担。

个性化新闻推荐系统将个性化推荐应用于新闻信息的推荐。它可以根据用户的兴趣特点和行为，帮助用户从互联网上的海量信息中轻松获得自己感兴趣的新闻资讯信息，并发掘用

作者简介：刘金亮，(1988-)，男，硕士研究生，移动互联网应用
通信联系人：卢美莲，（1967-），女，副教授，移动互联网应用，下一代网络技术. E-mail: mllu@bupt.edu.cn

户可能感兴趣的内容，并且不需要用户花费时间去寻找，可以节约大量的时间，以此实现新闻网站及网站用户的利益双赢[2,3]。

本文设计并实现了一个基于新闻主题的个性化新闻推荐系统，混合基于内容和协同过滤的推荐方法，应用主题模型表达用户对新闻的兴趣，最终针对每个用户的兴趣对其进行个性化的新闻推荐。

1 个性化新闻推荐

个性化推荐技术可以为不同的客户提供有针对性的服务，以满足其特定的需求。个性化新闻推荐则是个性化推荐在新闻处理领域中的一个延伸应用。通过分析用户的历史行为为各个用户构建用户兴趣模型，并以此为依据预测用户可能感兴趣的新闻。其主要包含两种方法：协同过滤推荐和基于内容推荐。

1.1 协同过滤推荐

协同过滤，就是利用某个兴趣相投、拥有共同经验的群体的喜好来向使用者对其感兴趣的新闻进行推荐。协同过滤利用了用户的历史行为（偏好、习惯等）将用户聚类成簇，这种推荐通过计算相似用户，假设被其他相似用户喜好的新闻当前用户也感兴趣[4]。

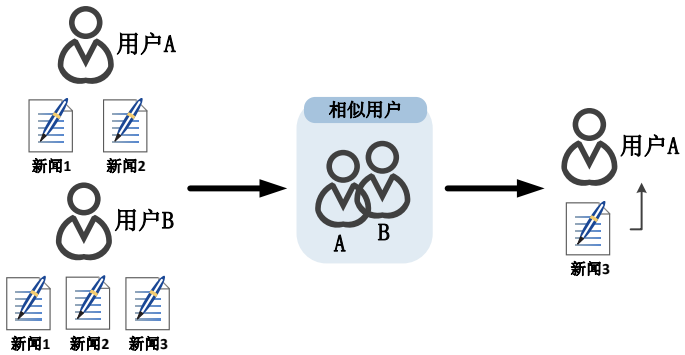

图 1 协同过滤推荐

Fig. 1 Collaborative Filtering Recommendation

如上图所示，协同过滤的推荐方法通常包括两个步骤：

1） 根据用户行为数据找到和目标用户兴趣相似的用户集合；

2） 找到这个集合中的用户喜欢的，且目标用户没有浏览过的新闻推荐给目标用户。

基于内容推荐的主要思想是，根据新闻内容，结合用户的历史行为数据对用户进行兴趣建模，并通过新闻匹配计算新闻与目标用户兴趣间的相似度，确定与用户兴趣相似的新闻，将相似度排名最高的新闻推荐给目标用户。

对于以文本为特征的新闻来说，较多采用基于内容推荐方法。基于内容的个性化推荐方法的基本过程如下图 2 所示，一般分为三步：

1） 新闻特征表示：为每个新闻抽取出一些特征（也就是新闻的内容）来表示此新闻，即新闻的文本表示；

2） 用户兴趣模型：利用用户过去喜欢的新闻的特征数据，来构建学习出此用户的喜好特征，得到用户的兴趣；

3） 新闻推荐生成：通过比较上一步得到的用户兴趣与候选新闻的特征，为此用户推荐一组相关性最大的新闻，即推荐结果输出。

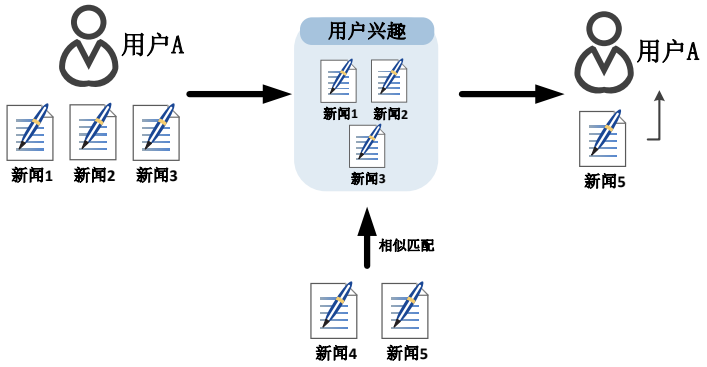

图 2　基于内容推荐

Fig. 2　Content-based Recommendation

2　主题模型

2.1　主题模型简介

主题模型是一种生成性的概率模型，一般基于如下观点构建：文本是主题上的概率分布；而主题则是词语上的概率分布。主题模型即在文本和词语之间加入了潜在语义分析中的语义维度，主题即文本和词语语义的高度抽象和压缩表示。主题模型从一个文本集合中挖掘指定个数的潜在主题模型，通过这些主题模型表示一篇文本，从而达到特征降维的目的，在同一个主题中的词语通常比较相关或相近[2]。

图 3 是在多篇新闻文本集上通过主题模型训练得到的一部分主题. 每个主题中的词语按照在该主题中的概率降序排列. 其中主题 1 表示“医疗”相关的概念，主题 2 表示了“农村”相关的概念等。

Topic 1: 0.02045		**Topic 2: 0.02157**		**Topic 3: 0.05312**	
医院	0. 210183	农村	0. 121180	法院	0. 093612
医生	0. 054501	农民	0. 094800	律师	0. 036817
患者	0. 033472	农业	0. 030296	证据	0. 036817
病人	0. 020039	地方	0. 024510	案件	0. 036532
医疗	0. 015635	问题	0. 024340	法律	0. 030348
抗生素	0. 014094	城乡	0. 023489	事实	0. 027019
细菌	0. 013433	粮食	0. 018383	司法	0. 025972
专家	0. 012112	条件	0. 016851	检察院	0. 024165
药物	0. 011892	目标	0. 016170	被告人	0. 021501
症状	0. 011782	城镇	0. 015489	行为	0. 021311
Topic 4: 0.01762		**Topic 5: 0.31948**		**Topic 6: 0.06814**	
环境	0. 034730	银行	0. 121689	社会	0. 206417
环保	0. 030581	资金	0. 063161	政治	0. 044171
空气质量	0. 023051	贷款	0. 043086	利益	0. 043690
数据	0. 021668	业务	0. 023937	道德	0. 022808
核电站	0. 019517	金额	0. 021930	公众	0. 020594
空气	0. 018441	债券	0. 021003	公民	0. 020210
日本	0. 017519	现金	0. 020540	权利	0. 019247
浓度	0. 017212	风险	0. 019768	政府	0. 019055
大气	0. 016905	利息	0. 017606	意识	0. 018285
污染物	0. 016290	金融	0. 016988	法律	0. 018189

图 3　新闻文本集在主题模型的训练结果

Fig. 3　Training Results of Topic Models in News’ Dataset

主题模型的起源是隐性语义索引(Latent Semantic Index , LSI)[5]，隐性语义索引并不是概率模型，因此也算不上一个主题模型，但是其基本思想为主题模型的发展奠定了基础。在 LSI 的基础上，Hofmann[6]提出了概率隐性语义索引(probabilistic Latent Semantic Indexing ，

pLSI)，该模型被看成是一个真正意义上的主题模型。而 Blei 等人[7]提出的 LDA (Latent Dirichlet Allocation)又在 pLSI 的基础上进行了扩展得到一个更为完全的概率生成模型。近年来，与特定的任务相结合，出现了越来越多的基于 LDA 的概率模型。

2.2 LDA 主题模型

LDA 模型是当前最具有代表性，也是最流行的一种概率主题模型，在文本挖掘、知识发现、话题跟踪以及多文本摘要等领域得到了广泛的良好应用[8]。LDA 是一种非监督机器学习技术，可以用来识别大规模文本集或语料库中潜藏的主题信息。它是一个生成性的三层贝叶斯模型，将词语和文本通过潜在的主题相关联。类似于许多概率模型，LDA 模型也基于词袋(Bag of Words)假设，即在模型中不考虑词语的顺序而只考虑它们的出现次数。

LDA 生成的概率模型图如图 4 所示：图中的空心点表示隐含变量，实心点表示可观察到的变量值，矩形表示重复步骤过程。外层矩阵表示从 Dirichlet 分布中为文本集中的每个文本 m 反复抽取主题分布 θ_m；内层矩形表示从主题分布中反复抽样产生文本 m 的词 $\{w_1,w_2,...,w_n\}$，θ_m 表示第 m 篇文本的主题概率分布，φ_k 表示主题中的词语概率分布；K 代表主题数目，M 代表文本数目，N_m 表示第 m 篇文本的词语个数，$w_{m,n}$ 和 $z_{m,n}$ 分别表示第 m 篇文本中第 n 个词语及其所属主题。α 和 β 是主题模型的两个参数，α 反应了文本集合中主题的相对强弱，β 则反映了主题中词语的概率密度[9]。

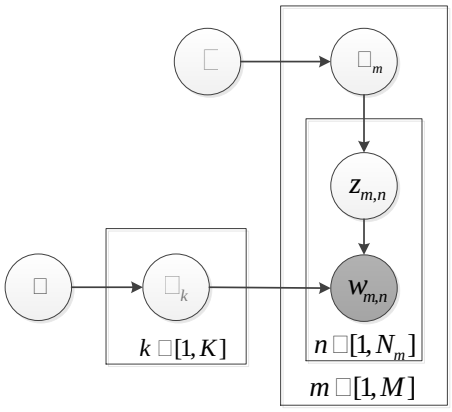

图 4 LDA 图模型

Fig. 4 Graph Model of LDA

3 推荐系统的设计与实现

3.1 推荐方法设计

本文设计的推荐系统采混合的推荐方法，在基于内容推荐方法的基础上加入协同过滤方法，运用 LDA 主题模型抽象新闻内容作为新闻的特征表示，分析用户行为构建用户兴趣模型，找到用户的主题兴趣偏好与相似兴趣用户组，最终通过计算新闻与用户模型间的相似度为用户推荐新闻。

3.1.1 新闻特征表示

本文所述的基于主题的个性化新闻推荐方法将新闻表示为一个二维的模型，包括新闻的主题特征向量及用户阅读列表，表示为 $F_n=\{T_n,K_n\}$。

新闻模型表示的主题特征向量 T_n 利用 LDA 主题模型计算得出，为一组向量表示：

$T_n=\{<t_1,w_{n1}>,<t_2,w_{n2}>,<t_3,w_{n3}>,\ldots,<t_m,w_{nm}>\}$，向量中得每一维$<t_i,w_{ni}>(i\le m)$表示新闻在编号为$i$主题上的权值，m 为主题个数。

K_n是对该新闻有过行为的一组用户，表示为$K_n=\{user1,user2\ldots\}$

3.1.2 用户兴趣模型

推荐系统的用户兴趣模型表示为$F_u=\{T_u,K_u\}$，其由两部分组成：主题兴趣偏好和相似用户列表。

用户的主题兴趣偏好由一组权值向量表示：$T_u=\{<t_1,w_{u1}>,<t_2,w_{u2}>,<t_3,w_{u3}>,\ldots,<t_m,w_{um}>\}$，$<t_i,w_{ui}>(i\ \ m)$代表用户对第 i 个主题$t_i$的兴趣偏好度为$w_{ui}$，m 为主题个数。

T_u的计算与更新基于用户的阅读历史行为和新闻的主题向量：$T_u=\frac{1}{A}\sum_{j=1}^{A}x_j?\ T_n\ \{w_{u1},w_{u2},w_{u3},\ldots,w_{um}\}$，其中 A 为用户历史行为个数，$x_j$为其中某一个行为的行为权值（区分浏览、评论、转发等多种行为），T_n表示用户某个历史行为对应的新闻的主题特征向量。

用户的相似用户列表K_u则通过分析用户的历史行为数据，找到与该用户浏览行为相似的一组用户，表示为$K_u=\{user1,user4\ldots\}$。

3.1.3 新闻推荐生成

本文基于主题特征的推荐方法中新闻推荐生成的过程如下：

1) 计算新闻与用户兴趣模型的匹配度值

对于每一个新闻模型表示$F_n=\{T_n,K_n\}$和每个用户兴趣模型$F_u=\{T_u,K_u\}$，计算它们间的匹配度：

a) 根据用户的主题偏好向量T_u与新闻主题特征向量T_n进行余弦相似度的计算：

$$Sim(T_u,T_n)=\frac{T_u'\ T_n}{\|T_u\|\bullet\|T_n\|}$$

b) 对用户的相似用户列表K_u与已对新闻产生过行为的用户列表进行 Jaccard 相似度计算：$Sim(K_u,K_n)=\left|\frac{K_u\cap K_n}{K_u\cup K_n}\right|$，即基于协同过滤方法比较用户兴趣相似的用户组$K_u$与新闻的阅读用户列表$K_n$的重合程度。

c) 计算新闻与用户兴趣的匹配度：$Sim(F_u,F_n)=\frac{m?\ Sim(T_u,T_n)\ \ n\ \ Sim(K_u,K_n)}{\sqrt{m^2+n^2}}$，$m$、$n$均为设定的用于调整主题特征相似度和用户列表相似度的比例系数。如果$Sim(F_u,F_n)$大于设定阈值，则将新闻加入到用户的新闻推荐列表中。

2) 推荐列表处理：新闻的流行度和新颖度是随着时间变化的，这不同于其他系统的物品推荐（商品和电影），因此，推荐系统在新闻推荐时在计算与用户兴趣匹配度的同时需要综合考虑内容重复冗余、时间性等因素，对用户的新闻推荐列表进行多步的处理，提升新闻的推荐准确率及新颖性。

3.2 推荐系统实现

基于主题的新闻推荐系统外围架构如图 5 所示，新闻推荐系统运行于新闻网站后台，推荐系统通过读取后台数据库中的用户行为和新闻内容，分析用户的行为日志，基于推荐算法，给用户生成推荐列表，最终展示到网站的界面上供用户浏览，用户可使用移动客户端或 Web 客户端查看新闻推荐内容。

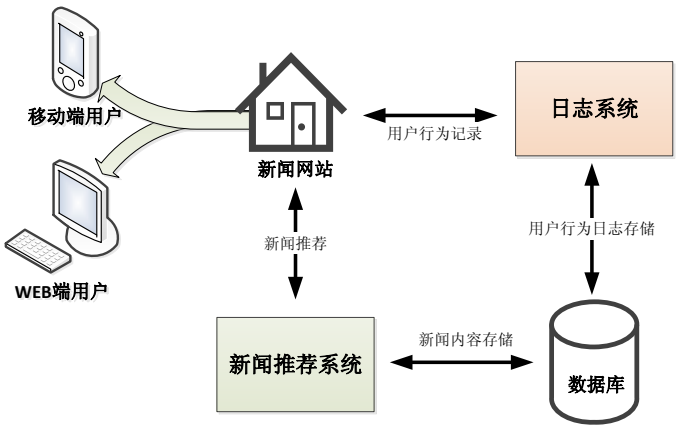

图 5 推荐系统

Fig. 5 Recommendation System

基于本文所述的推荐方法和推荐系统架构，我们对基于主题的个性化新闻推荐系统进行了实现，系统分为四个模块：预处理模块、主题模型模块、用户模型构建模块和新闻推荐处理模块。推荐系统的功能模块结构如图 6 所示：

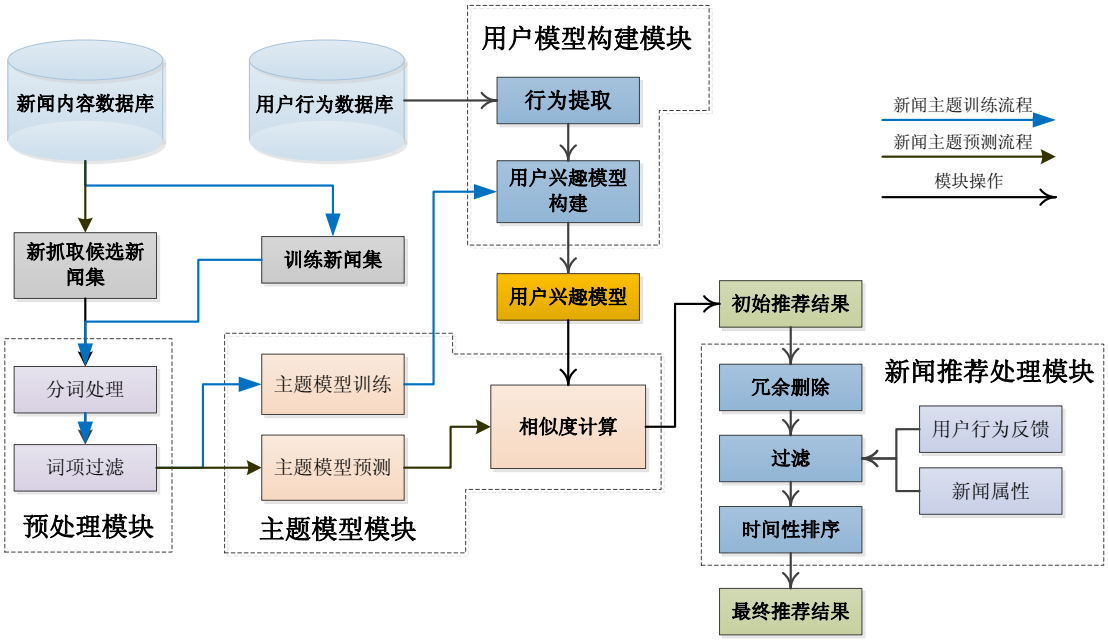

图 6 推荐系统功能模块图

Fig. 6 Function Modules of Recommendation System

预处理模块：对中文新闻的词语进行与处理。从新闻内容数据库中读取新闻文本内容，并对其进行分词过滤和词性标注：去除标记词、虚词与标点符号，保留能够代表新闻内容的名词。

主题模型模块：模块包括主题模型训练和主题模型预测。主题模型训练通过对训练新闻

集合进行计算得到新闻的主题特征向量，用于构建用户模型和对新进入系统的新闻的主题向量预测。主题模型预测则根据训练新闻集的主题向量预测得到新进入系统新闻的主题特征向量，向用户进行推荐计算。

用户模型构建模块：负责用户兴趣模型的构建。读取用户行为数据库，对于每个用户根据其阅读历史行为结合训练新闻集的主题特征向量以及其他用户的历史行为构建个性化的用户兴趣模型。

新闻推荐处理模块：对准备推荐的初始新闻推荐列表中的每个新闻文本进行相应处理和排序。对每个用户的初始新闻推荐列表进行删除重复新闻、过滤用户已有过行为的新闻、并将剩余新闻按其时间性进行排序，最终生成每个用户的最终新闻推荐列表，并据此向用户推荐其感兴趣的新闻文本。

4 结束语

本文介绍了一种基于主题的个性化新闻系统的设计实现方案，利用 LDA 主题模型挖掘新闻内容的主题，将新闻内容表示在一个低维的融入语义信息的主题空间上，有效的挖掘用户兴趣和新闻内容间的联系，构建准确的、个性化的用户兴趣模型，并结合协同过滤方法为用户进行新闻推荐。由于互联网中信息的极速膨胀及个性化推荐的迅速发展，对新闻推荐系统的研究实现越来越多，该方案为个性化新闻系统提供了一个良好的范例。后续的工作将针对用户模型和新闻推荐生成进行改进和优化，从而使得其在新闻推荐的准确性和多样性进一步提高。

[参考文献] (References)

[1] 曾春,邢晓春. 个性化服务技术综述[J]. 软件学报,2002,13(10):1952-1962.
[2] 陈宏,陈伟. 基于多主题追踪的网络新闻推荐[J]. 计算机应用,2011,31(9):2426-2428.
[3] 唐朝. 资源自适应个性化新闻推荐系统的研究与实现[D]. 杭州：浙江大学,2010.
[4] 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.
[5] 徐戈,王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报,2011,34(8):1423-1435.
[6] Deerwester S., Dumais,S.T.,Furnas, G.W., Landauer, T. K., & Harshman, R.Indexing by Latent Semantic Analysis.Journal of the American Society for Information Science,1999,41:391-407
[7] Hofmann, T. Probabilistic latent semantic analysis.In Proceedings of 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval,1999:50-57.
[8] Blei D,Andrew Y NG,Jordan M.Latent dirichlet allocation[J].Journal of Machine Learning Reserach,2003,3:993-1022
[9] 孙玉婷. 基于概率主题模型的中文话题检测与追踪研究[D]. 武汉：华中科技大学,2010