# 新闻推荐系统研究综述

孟开元,岳宇航,曹庆年

(西安石油大学 计算机学院,陕西 西安 710065)

**摘 要**:通过网络阅读新闻已经成为一种广泛流行的阅读方式,随着网络新闻资源的日益普及,在海量的新闻中,用户很容易被一些偏离自己兴趣爱好的信息所淹没,这种现状促进了新闻推荐系统的发展,其旨在帮助用户在巨大的动态新闻空间发现心仪的新闻,提高用户满意度。分析新闻推荐系统研究过程和现状,对新闻推荐系统的 4 个关键技术进行重点分析,并比较其优缺点。此外,从准确度和非准确度两个方面论述新闻推荐系统的效用评价标准,并分析了目前新闻推荐系统面临的挑战,指出了未来新闻推荐系统的研究重点。

**关键词**:新闻推荐;推荐算法;效用评价;研究综述

**DOI**:10.11907/rjdk.201896 **开放科学(资源服务)标识码(OSID)**:

**中图分类号**:TP301 **文献标识码**:A **文章编号**:1672-7800(2021)001-0249-04

## Research Review of News Recommendation System

MENG Kai-yuan, YUE Yue-hang, CAO Qing-nian

(*School of Computer Science, Xi'an Shiyou University, Xi'an* 710065, *China*)

**Abstract**: Reading news through the Internet has become a widely popular reading method. With the increasing popularity of online news resources, users are easily overwhelmed by information that deviates from their interests and hobbies in reading massive news. This situation has promoted the development of news recommendation system which aims to help users find the interesting news in a huge dynamic news space and improve users' satisfaction. This paper discusses the research process and current status of the news recommendation system, focusing on analyzing the four key technologies of the news recommendation system, and comparing their advantages and disadvantages. In addition, this paper discusses the effectiveness evaluation criteria of the news recommendation system from two aspects of accuracy and inaccuracy, and analyzes the challenges faced by the current news recommendation system, and points out the research focus of the future news recommendation system.

**Key Words**: news recommendation system; recommendation algorithm; effectiveness evaluation; research review

## 0 引言

随着全球化和网络技术的发展,越来越多的人通过网络渠道在线阅读来自全球各地的新闻。然而,新闻域稀疏的用户配置、快速增长的项目数量、加速衰减的项目价值,以及用户偏好的动态转移,使得各类新闻平台的用户越来越难选择自己感兴趣的新闻[1]。因此,如何在瞬时变化的新闻领域,利用一些模型和技术帮助用户找到感兴趣的新闻变得尤为重要。由此,新闻推荐系统应运而生,它不是根据用户显式的查询搜索对信息进行过滤,而是根据用户兴趣主动呈现相关新闻。经过 20 多年的发展,新闻推荐系统已经成为一种帮助用户在信息过载情况下找到自己感兴趣新闻的重要工具。

新闻推荐系统主要通过分析用户的阅读兴趣偏好,帮助用户高效获取自己所需的新闻,被视为解决新闻领域信息爆炸问题的重要手段。与电影推荐等其他领域的推荐系统相比,新闻推荐面临的问题具有独特性,如:新闻制作频率很高、新闻项目相关性变化较快、新闻实时性要求很高等。

**收稿日期**:2020-07-28

**作者简介**:孟开元(1968-),男,硕士,西安石油大学计算机学院副教授、硕士生导师,研究方向为计算机网络与通信;岳宇航(1994-),女,西安石油大学计算机学院硕士研究生,研究方向为推荐系统、计算机接口技术、控制系统;曹庆年(1963-),男,硕士,西安石油大学计算机学院教授、硕士生导师,研究方向为计算机网络与通信。本文通讯作者:岳宇航。

# 1 新闻推荐系统关键技术

## 1.1 新闻推荐算法

### 1.1.1 基于内容的推荐算法

基于内容的推荐算法(Content-Based Filtering,CB),主要用于推荐基于文本类的项目,通常采用浅层模型分析用户的历史阅读记录,从而发现用户的阅读兴趣,进而将与用户阅读兴趣类似且用户评价很高的新闻作为推荐结果。该算法运用过程的核心问题是如何进行项目相似性度量。先构建用户特征并计算项目间的相似度,再将最符合用户兴趣的新闻项目进行推荐。如 Goossen 等[2]结合 TF-IDF 与领域本体的语义进行推荐;Samarinas 等[3]通过引入一种使用单词嵌入来构建用户兴趣模型的方法,实现新闻个性化推荐,一定程度上对传统计算相似度算法进行了优化。然而,对人工特征提取的依赖制约了基于内容的推荐算法发展,很难获取更深层次的新闻特征和用户行为,深度学习在基于内容的推荐算法中的发展弥补了浅层算法的不足,已经成为当前研究热点。

### 1.1.2 协同过滤推荐算法

协同过滤推荐算法(Collaborative Filtering,CF),是新闻推荐系统中应用最广泛的算法[4]。从本质上讲,协同过滤是一种基于用户与项目之间的交互行为数据进行信息过滤的方法,分为基于用户和基于项目的协同过滤两种算法[5]。基于用户的协同过滤是指采用均方差、皮尔逊相关系数、斯皮尔曼相关性等算法计算用户相似度,进而通过基于阈值的方法和 Top-N 推荐,得到 k 个最相似的用户组成目标用户的最近邻集合 K,并将集合 K 中用户感兴趣的且没有接触过的项目推荐给目标用户。基于项目的协同过滤是指通过计算不同用户对不同项目的评分获得项目之间的关系。一般采用余弦向量计算项目相似度。利用带去噪的堆叠自动编码器的 CF 方法、考虑用户行为时间排序协同过滤推荐算法等基于 CF 的改进算法在一定程度上提高了推荐系统的性能[5-6]。

### 1.1.3 基于知识的推荐算法

基于知识的推荐(Knowledge-based Technoques,KB),可看作是一种不依赖于大量项目或用户统计数据,而是直接将用户偏好映射到被推荐新闻项目上的推理技术。基于知识的推荐具有很强的交互性,用户只需要对某个新闻项目有简单的反馈,系统将通过有效的搜索策略进行推荐。

### 1.1.4 混合新闻推荐算法

相比于上述 3 种推荐算法,将基于内容和协同过滤的算法采取加权、切换、混杂、层叠、级联、特征组合和特征补充混合策略而得到的混合推荐算法具有更大的推荐优势。混合算法能够平衡预测精度和其他质量因素,如新颖性或多样性,进而提高系统推荐效率。如:Jonnalagedda 等[7]根据新闻的受欢迎程度与用户配置文件的相关性向用户进行新闻推荐;Hao 等[8]通过混合算法为用户提供了一种能够减少网络浏览中重复单调内容的工具。

## 1.2 新闻推荐算法比较

4 种关键新闻推荐算法的优缺点如表 1 所示。

**Table 1 Comparison of advantuges and disadvantages of news recommendation algorithms**

**表 1 新闻推荐算法优缺点对比**

| 推荐算法 | 优点 | 缺点 |
|---|---|---|
| 基于内容的推荐算法 | 用户独立性高;不存在冷启动问题;推荐结果可解释性高;不存在马太效应 | 无法为新用户推荐;无法挖掘用户的潜在兴趣;缺乏多样性 |
| 协同过滤的推荐算法 | 模型通用性强;可挖掘用户的潜在兴趣;适合复杂的非结构化对象 | 存在数据稀疏性问题;存在冷启动问题;推荐结果可解释性低;存在马太效应 |
| 基于知识的推荐算法 | 不存在冷启动问题;交互性强 | 推荐效果对知识库的依赖性较强,而专门领域的知识和推理规则较难获取[9] |
| 混合推荐算法 | 不存在冷启动问题;不存在马太效应;可挖掘用户的潜在兴趣;可产生意外,实现多样性 | 算法工作量较大,需要大工作量才能得到正确的平衡 |

# 2 新闻推荐系统效用评价

新闻推荐系统的性能评价是为以后更好地完善技术手段,以便得到更有效的推荐系统。而数据集和评价指标是进行新闻推荐系统性能测试的两个关键因素。

## 2.1 常用数据集

目前,新闻推荐系统进行效用评价依赖的常用数据集,主要有加州大学欧文分校推出的 UCI 数据集、由 ComeToMyHead 搜集的 AG 数据库、雅虎推出的“雅虎新闻推荐”数据集以及新闻推荐领域最好的 Adressa 数据集等。如 Del corso 等[10]从 comeToMyHead 中提取新闻数据;Gulla 等[11]对 Adressa 精简新闻数据集进行了介绍,该数据集支持各种类型的新闻推荐。

## 2.2 评价指标

推荐系统通常通过以下 3 种方法之一进行评估:①基于历史数据的离线实验和模拟,Maksai 等[12]进行实验时将数据集分为训练集、验证集和测试集;②实验室研究,李增等[13]通过实验室研究验证推荐结果;③真实网站上的 A/B 测试,Wang 等[14]在在线新闻平台上进行大量实验。本文从准确度和非准确度指标两方面对新闻推荐系统评价指标进行论述。

### 2.2.1 准确度指标

(1)预测准确度指标。预测准确度指推荐系统的预测评级与真实用户评级的接近程度。其中,最典型的评估指标有平均绝对误差(MAE)、均方误差(MSE)、均方根误差(RMSE)以及归一化平均绝对误差(NMAE)。预测准确度评估指标数值越低,则预测准确度越高。

(2)分类准确度指标。分类准确度指推荐系统对一个

项目作出正确或错误决定的频率。评估指标包括准确率、召回率、F1 指标。准确率越高，即推荐系统预测项目中目标项目所占比例越高，但此时召回率越低。因此，在不同情况下需要判断是准确率高还是召回率高才能满足自己的需求。F1 指标即为准确率和召回率的调和平均值，是一个可以反映整体情况的指标。

(3)排序准确度指标。排序准确度是为了评估用户对推荐系统生成的推荐列表排序的满意程度，更适用于评估需向用户呈现排名列表的推荐系统。

2.2.2　非准确度指标

(1)覆盖率。覆盖率(Coverage)指推荐系统能够推荐出来的项目占总项目集合的比例，旨在评估推荐系统挖掘长尾项目的能力。但该定义过于粗略，为了更好地描述覆盖率，故引入信息论中信息熵和经济学中的基尼系数，计算推荐列表中各项目出现次数的分布情况。若分布较平，则覆盖率较高。

(2)新颖性。根据用户历史兴趣进行新闻推荐，其结果往往会缺乏“惊喜感”。21 世纪初，Herlocker 等[15]最先提出新颖性推荐的概念，即向用户推荐不太流行的产品。新颖性可通过新闻项目的流行度或推荐项目与用户的距离进行度量，新颖性越高，准确性指标就会受到一定的挑战，因此现有研究通常对新颖性和准确性指标进行加权测试，以便得到更高的效用评价效果。目前，关于新颖性的研究较少，可作为未来研究重点。

(3)多样性。由于用户的兴趣偏好是广泛的，为了提高用户对推荐结果的满意度，新闻推荐系统应生成多样化的推荐列表，因此多样性也成为预测新闻推荐系统性能的指标之一[16]。同新颖性类似，多样性和准确性之间也需要进行平衡，并且，多样性的程度也应考虑不同用户的偏好广泛程度。

(4)鲁棒性。新闻推荐系统的鲁棒性是衡量系统抗击作弊能力的指标，主要通过比较添加噪声(如对抗训练)后产生的推荐列表和原推荐列表相似度验证系统的鲁棒性。如：将知识图表示方法融入新闻推荐的深度知识感知网络，在实际应用中具有鲁棒性和稳定性[14]。

## 3　新闻推荐领域面临的挑战

本文对新闻推荐领域面临的一些主要挑战进行了分析，这些挑战可作为未来重点研究方向。

### 3.1　数据稀疏性

由于大型新闻推荐系统项目数量巨大，用户之间数据重叠率极低，故存在数据稀疏性问题。尽管通过用户聚类和项目聚类技术推荐[17]、基于排序的地理因子分解[18]、利用 RapidMiner 工具实现的协同过滤推荐[19]等方法可缓解新闻推荐系统的数据稀疏性。但推荐系统数据库中急剧增加的用户数量新闻特征使得推荐质量越来越差，稀疏性问题更加凸显。由此可见，数据稀疏性问题亟待解决。

### 3.2　冷启动问题

冷启动是指当一个用户与新的推荐系统交互时，该系统没有任何可利用的用户兴趣偏好以生成推荐项目，往往产生于协同过滤算法。常见处理方式是在推荐过程中加入关于用户的上下文信息，如用户位置信息、访问时间等。Pereira 等[20]将人口统计信息与协同过滤推荐相结合，有助于缓解用户冷启动问题。Lei 等[21]通过超图学习进行新闻推荐，该算法能够缓解新闻推荐中的冷启动问题，但系统可伸缩性较差。故冷启动问题仍然需要不断探索，以便提高用户对新推荐系统的感知价值。

### 3.3　用户兴趣漂移

用户兴趣漂移即指用户的兴趣偏好随时间推移而发生变化的现象。人们对音乐、电影或书籍的喜好在短时间内通常会有轻微差异，但在新闻领域，人们的阅读偏好会受到外界环境、年龄、文化水平甚至情绪的影响[22]。袁仁进等[23]为缓解新闻推荐系统的用户兴趣漂移，提出了一种面向新闻推荐用户的兴趣模型与更新方法，但还难以解释 F 值呈现先高后低的现象。因此，持续研究用户兴趣偏好实时更新模型、平衡长期偏好和短期偏好对新闻推荐系统的发展也是一项真正的挑战。

### 3.4　可伸缩性问题

可伸缩性能衡量新闻推荐系统扩展过程中系统的计算处理能力。大型新闻网站每天需要处理海量数据，一般通过应用不同类型的集群技术进行聚类以提高系统可伸缩性。现有研究[24]针对新闻推荐系统的可伸缩性问题提出了多种聚类技术；Kuchař 等[25]提出基于关联规则作为分类器的方法可提高系统可伸缩性，但评估结果并不好；Verbitskiy 等[26]使用 Akka 框架实现了基于时间窗口的新闻推荐算法，具有良好的可伸缩性，但该推荐算法点击通过率过低。聚类可以加快计算速率，但它也可能降低系统准确性。因此，如何平衡系统准确性和可伸缩性也是目前一大难点。

## 4　结语

随着网络新闻资源的日益普及，在高度动态的新闻领域中，新闻推荐系统必将是众多学者的研究热点。本文对现有新闻推荐系统相关研究进行了回顾，从新闻推荐系统关键技术、主要评价指标和面临的挑战等方面进行了多角度论述。如何优化算法以提高推荐系统性能？如何应对数据稀疏、冷启动、用户兴趣漂移和可伸缩性等新闻推荐中的挑战？此类问题均将是今后的重点研究方向。

**参考文献：**

[1] DE G. CHAMELEON：a deep learning meta-architecture for news recommender systems[C]. Proceedings of the 12th ACM Conference on Recommender Systems，2018：578-583.

[2] GOOSSEN F，IJNTEMA W，FRASINCAR F，et al. News personalization using the CF-IDF semantic recommender[C]. Proceedings of the

International Conference on Web Intelligence, 2011:25-27.

[3] SAMARINAS C, ZAFEIRIOU S. Personalized high quality news recommendations using word embeddings and text classification models [Z]. EasyChair, 2019.

[4] JANNACH D, ZANKER M, GE M, et al. Recommender systems in computer science and information systems—a landscape of research [C]. International Conference on Electronic Commerce and Web Technologies, 2012:76-87.

[5] CAO S, NAN Y, LIU Z. Online news recommender based on stacked auto-encoder[C]. 2017 IEEE/ACIS 16th International Conference on Computer and Information Science, 2017:721-726.

[6] XIAO Y, AI P, HSU C, et al. Time-ordered collaborative filtering for news recommendation[J]. China Communications, 2015, 12(12):53-62.

[7] JONNALAGEDDA N, GAUCH S, LABILLE K, et al. Incorporating popularity in a personalized news recommender system [J]. PeerJ Computer Science, 2016, 2:63.

[8] HAO W, FANG L, LING G. A hybrid approach for personalized recommendation of news on the Web[J]. Expert Systems with Applications, 2012, 39(5):5806-5814.

[9] MENG X W, CHEN C, ZHANG Y J. A survey of mobile news recommend techniques and applications[J]. Chinese Journal of Computers, 2016, 39(4):685-703.
孟祥武, 陈诚, 张玉洁. 移动新闻推荐技术及其应用研究综述[J]. 计算机学报, 2016, 39(4):685-703.

[10] DEL CORSO G, GULLÍ A, ROMANI F. Ranking a stream of news [C]. Proceedings of The 14th International Conference on World Wide Web, 2005:97-106.

[11] GULLA J A, ZHANG L, PENG L, et al. The adressa dataset for news recommendation[C]. Proceedings of The International Conference on Web Intelligence, 2017:1042-1048.

[12] MAKSAI A, GARCIN F, FALTINGS B. Predicting online performance of news recommender systems through richer evaluation metrics[C]. Proceedings of the 9th ACM Conference on Recommender Systems, 2015:179-186.

[13] LI Z, LIU Y, LI C C. A news recommendation algorithm based on user behavior[J]. Computer Engineering & Science, 2020, 42(3):529-534.
李增, 刘羽, 李诚诚. 基于用户行为的新闻推荐算法的研究[J]. 计算机工程与科学, 2020, 42(3):529-534.

[14] WANG H, ZHANG F, XIE X, et al. DKN: deep knowledge-aware network for news recommendation [C]. Proceedings of The 2018 World Wide Web Conference, 2018:1835-1844.

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

[16] YU B, SHAO J, CHENG Q, et al. Multi-source news recommender system based on convolutional neural networks[C]. Proceedings of the 3rd International Conference on Intelligent Information Processing, 2018:17-23.

[17] GONG S. A collaborative filtering recommendation algorithm based on user clustering and item clustering [J]. Journal of Software, 2010, 5(7):745-752.

[18] LI X, CONG G, LI X, et al. Rank-geofm: a ranking based geographical factorization method for point of interest recommendation [C]. Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, 2015:433-442.

[19] JAIN A F, VISHWAKARMA S K, JAIN P. An efficient collaborative recommender system for removing sparsity problem [M]. ICT Analysis and Applications. 2020:131-141.

[20] PEREIRA A L V, HRUSCHKA E R. Simultaneous co-clustering and learning to address the cold start problem in recommender systems[J]. Knowledge-Based Systems, 2015, 82:11-19.

[21] LEI L, TAO L. News recommendation via hypergraph learning: encapsulation of user behavior and news content[C]. Proceedings of The 6th ACM International Conference on Web Search and Data Mining, 2013:305-314.

[22] Özgöbek Ö, Gulla J A, Erdur R C. A survey on challenges and methods in news recommendation[C]. International Conference on Web Information Systems and Technologies, 2014:278-285.

[23] YUAN J R, CHEN G, LI F. User interest model construction and update for news recommendation[J]. Application research of Computers, 2019, 36(12):3593-3596.
袁仁进, 陈刚, 李锋. 面向新闻推荐的用户兴趣模型构建与更新[J]. 计算机应用研究, 2019, 36(12):3593-3596.

[24] KARIMI, MOZHGAN, JANNACH, et al. News recommender systems-survey and roads ahead[J]. Information Processing & Management, 2018, 54(6): 1203-1227.

[25] KUCHAŘ J, GOLIAN C. News recommender system based on association rules @ CLEF NewsREEL 2017: Clef[C]. Conference & Labs of the Evaluation Forum, 2017:239-254.

[26] VERBITSKIY I, PROBST P, LOMMATZSCH A. Development and evaluation of a highly scalable news recommender system[C]. Toulouse: Working Notes of CLEF 2015 -Conference and Labs of the Evaluation forum, 2015.

(责任编辑:孙 娟)