Computer Science and Application 计算机科学与应用, 2019, 9(1), 38-45
Published Online January 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.91005

Hans汉斯

News Recommendation Combining User Microblog Interests Mining and Collaborative Filtering

Shuai Zhang, Pinghua Chen, Jingyu Chen

School of Computer Science and Technology, Guangdong University of Technology, Guangzhou Guangdong
Email: 413664603@qq.com

Received: Dec. 24th, 2018; accepted: Jan. 7th, 2019; published: Jan. 14th, 2019

Abstract

In order to solve the problems of insufficient diversity and potential interest in content-based news recommendation and the cold start problem of collaborative filtering recommendation methods, a news recommendation algorithm combining microblog user interest mining and collaborative filtering is proposed. By analyzing the data structure of Weibo, UI-FP mining method is used to mine user interest on user Weibo, and this user interest set and user history browsing news record are used as user model interest set as metadata for news recommendation, combined with collaboration. The algorithm implements the fusion of microblog data and news data to solve the aforementioned cold start and potential lack of interest. Experimental results show that the method is in line with expectations.

Keywords

Cold Start, Interest Mining, Microblog, News Recommendation, Collaborative Filtering

融合用户微博兴趣挖掘与协同过滤的新闻推荐

张　帅，陈平华，陈靖宇

广东工业大学计算机学院，广东　广州
Email: 413664603@qq.com

收稿日期：2018年12月24日；录用日期：2019年1月7日；发布日期：2019年1月14日

摘　要

针对基于内容的新闻推荐中存在的多样性不足、潜在兴趣缺失等问题和协同过滤的推荐方法存在的冷启

文章引用: 张帅, 陈平华, 陈靖宇. 融合用户微博兴趣挖掘与协同过滤的新闻推荐[J]. 计算机科学与应用, 2019, 9(1): 38-45. DOI: 10.12677/csa.2019.91005

动问题，提出了一种融合微博用户兴趣挖掘与协同过滤的新闻推荐算法。通过分析了微博的数据结构，使用UI-FP挖掘方法挖掘用户微博上的用户兴趣，将此挖掘兴趣集和用户历史浏览新闻纪录兴趣集作为新闻推荐的元数据，结合协同算法来实现微博数据和新闻数据的融合，从而解决前述的冷启动和潜在兴趣缺失问题。实验结果表明，该方法符合预期。

关键词

冷启动，兴趣挖掘，微博，新闻推荐，协同过滤

1. 引言

互联网最重要的贡献之一是让信息的获取变得更加简单高效，而在个人设备上阅览网络新闻成为人们最重要的信息来源之一，CNNIC 发布的第 35 次《中国互联网络发展状况统计报告》指出到 2014 年 12 月，网络新闻的使用率已达到 80.0%，远超其他网络应用，但与此同时带来的信息爆炸和信息的良莠不齐给用户带来不好的使用体验，因此关于新闻推荐的研究越来越多，而随着推荐技术和算法的发展，对新闻推荐的发展带来重大影响。

目前，国内外的新闻推荐已经有较多的研究。文献[1]使用基于内容的推荐方法，通过提取历史浏览的文本特征构造用户兴趣模型，将候选新闻与用户兴趣模型进行比较得到推荐结果。基于内容的方法可解释性强，推荐理由让用户容易理解，但存在冷启动问题，它在推荐多样性上存在不足，难以发掘用户潜在兴趣[2]。文献[3]使用了协同过滤的新闻推荐方法，通过计算用户行为相似性，向目标用户推荐同类用户关注的新闻。由于协同过滤通常需要数小时积累用户点击才能形成推荐，造成了冷启动问题。文献[4]通过混合协同过滤一定程度上解决了数据稀疏的问题，但需要用户对新闻评分，协同过滤在新闻推荐的个性化方面也表现不足。文献[5]将基于内容的矩阵协同过滤得到用户潜在兴趣矩阵从而进行新闻推荐，但没有考虑上下文等信息。

除了考虑基于内容的推荐、协同过滤推荐等基本技术，结合上下文感知[6] [7] [8]的新闻推荐和基于社会化网络的移动新闻推荐[9]成为近来的研究热点。事实上，许多社交媒体信息都和新闻事件相关，Twitter 上甚至有超过 85%的内容是和新闻有关联的[10]。Abel 等人[10]提出的个性化新闻推荐框架利用微博中的 URL 链接或相似度计算将微博与相关新闻联系起来，然后从新闻中抽取实体、主题等来丰富相关微博的语义信息，并分别建立了 3 种用户偏好档案用于新闻推荐：基于 Hashtag、基于实体和基于主题的用户档案等。然而，上述的结合上下文、社交媒体和传统新闻进行新闻推荐时候都没有考虑用户的潜在兴趣的挖掘，于是本文基于社交媒体的特点，提出了融合用户微博兴趣挖掘与协同过滤的新闻推荐方法。

2. 推荐框架

本文推荐框架主要有二个模块，即用户微博兴趣挖掘模块和推荐模块，总流程架构如图 1 所示。

各模块的功能如下：

1) 兴趣挖掘模块。用户微博中的数据有如文本、图片、视频、标签、关注者和粉丝等各种数据。本模块主要针对文本、标签、关注者、转发和评论数据进行挖掘，用来构建用户兴趣集进行新闻推荐。

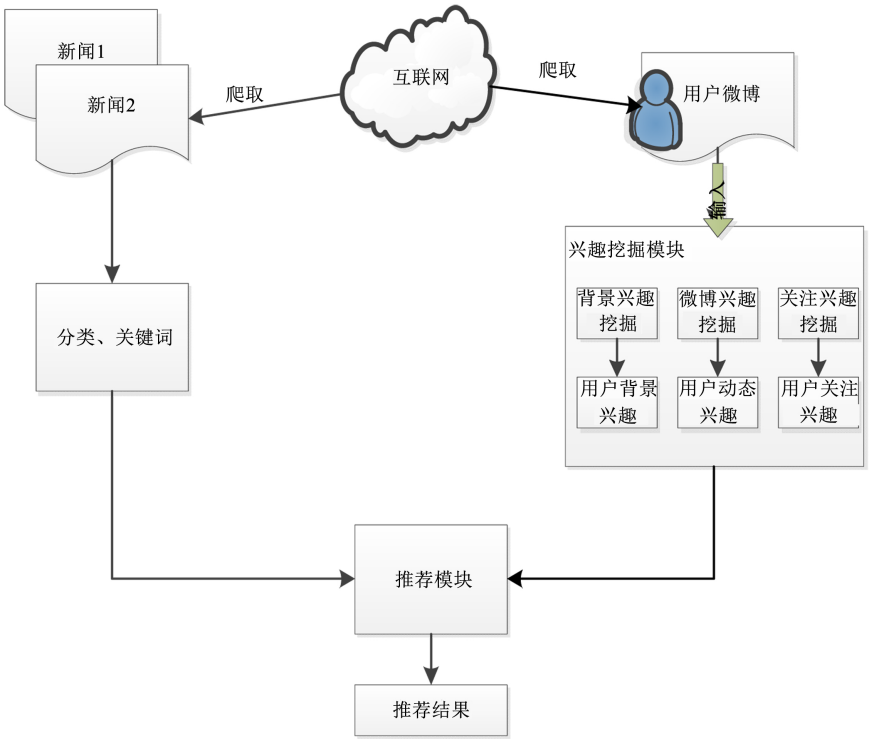

Figure 1. Recommended frame diagram
图 1. 推荐框架图

2) 推荐模块。本模块将经过分类和关键词提取处理后的新闻与用户微博兴趣集计算匹配相似的新闻。通过计算新闻集合用户兴趣集的相似度并结合协同过滤来获得推荐列表，并将推荐列表进行因素调整排序进行推荐，最后验证推荐效果。

3. 兴趣挖掘模块

已有的研究表明，传统的利用用户浏览记录[11]、搜索记录等基于内容方法的新闻推荐无法挖掘出用户的潜在兴趣，不符合个性化的推荐，而微博网络中的用户行为数据和背景信息以及用户的社交关系中蕴含着丰富的用户兴趣，而挖掘用户兴趣用于个性化推荐等也是研究的热点。关于微博兴趣挖掘的研究分为基于背景和基于内容[12]。在本文挖掘模块中，综合考虑了基于用户背景、基于用户文本内容、用户关注者背景等信息。

挖掘模型如图 1 右侧部分所示。在图中用户背景指的是用户基本信息和标签信息等；微博指的是用户发布的微博、转发微博等；用户社交信息指的是用户的关注和粉丝信息，这里主要采用关注者信息。其中用户背景兴趣用 U_B 表示，用户微博动态兴趣用 U_M 表示，用户关注兴趣用 U_{fb} 表示，则有如下定义：

定义 1：用户背景中一般包括简介、职业、标签、性别、地点、年龄等，由于职业、标签、性别、地点、年龄等单词类信息是自然的单词，所以不用分词；而简介信息通常是短文本，需要进行分词，则用户背景的兴趣集是 $U_B=\{kb_1,kb_2,\cdots,kb_i\}$，统计各个词出现次数并归一化作为权重，则 $U_B=\{(kb_1,w_1)(kb_2,w_2),\cdots,(kb_n,w_n)\}$

定义 2：微博发文内容包括原创内容、转发内容和评论内容等，用户的原创内容自然能够代表用户的兴趣，但后两者在一定程度反映用户兴趣，设置评论内容权重 A_1，转发内容 A_2，将微博分词并统计次数并归一化，则有：$U_M=\{(M_1,w_1),(M_2,w_2),\cdots,A_1\{(M_i,w_i),\cdots,(M_j,w_j)\},A2\{(M_k,w_k),\cdots,(M_n,w_n)\}\}$

其中 A_1、A_2 参考文献[13] [14]分别设置为 0.75 和 0.25。

定义 3：用户关注者的信息挖掘主要是被关注者的标签信息和背景信息挖掘，通过爬取用户关注列表并爬取被关注者标签信息、简介信息、基本信息等。微博限制为 20 页关注列表，统计标签词频，简介信息分词并计算权值，方法同定义 1 类似。

$$U_{fb}=\{(t_1,w_1),\ (t_2,w_2),\cdots,(t_n,w_n)\}$$

由于这三类兴趣集都能代表用户兴趣及潜在兴趣，于是综合兴趣集，则引出公式：

$$U_F=\sum_{U\in(U_B,U_M,U_{fb})}U \tag{1}$$

又有研究者 Gu [15]认为，在用户微博中出现过的人名、地域名应该重点考虑，加大该类信息词的权重比例，于是引入：

$$U_p=\{m,p,n,t\} \tag{2}$$

其中，m 表示用户微博发文内容单词与权重，表示为 $\{\langle m_1,w_1\rangle,\langle m_2,w_2\rangle,\cdots\cdots\}$；$p$ 表示为用户微博发文中出现的地名，表示为 $\{\langle p_1,w_1\rangle,\langle p_2,w_2\rangle,\cdots\cdots\}$；$n$ 表示用户微博中出现的人名，表示为 $\{\langle n_1,w_1\rangle,\langle n_2,w_2\rangle,\cdots\cdots\}$；$t$ 表示微博内容分类的类别和权重，表示为 $\{\langle n_1,w_1\rangle,\langle n_2,w_2\rangle,\cdots\cdots\}$。

综合以上考虑，则引出 UI-FP 兴趣挖掘公式：

$$U_{FP}=(1-\alpha)U_F+\alpha U_p,\alpha\in[0,1] \tag{3}$$

其中，α 为调节参数，具体的最优取值由第 4 章的实验结果及分析中得出。

4. 推荐模块

4.1. 新闻分类

新闻分为文本新闻、图片新闻和视频、音频新闻等，其中文本新闻占据大部分。在处理文本新闻时，运用 VSM (Vector Space Model)模型来表示新闻会产生超高维度，在相似度计算时采用 TF-IDF (Term Frequency-Inverse Document Frequency)算法会导致每次计算一对词的相似度时会遍历全部分词会造成高时间复杂度，所以通过将新闻分类可以大大降低计算时间，提高推荐效率。本文采用 textCNN (text Convolutional Neural Networks)模型[16]实现新闻文本的分类。textCNN 分类模型如图 2 所示：

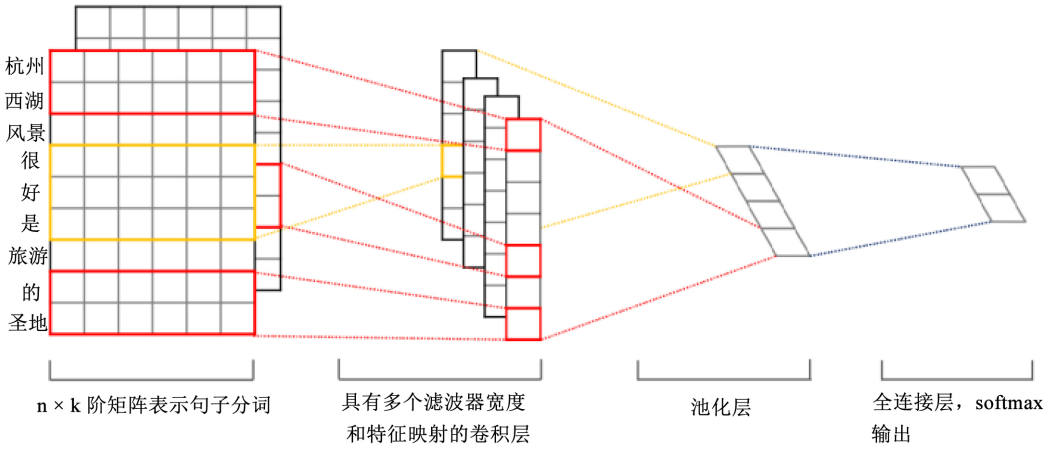

Figure 2. Used for the example sentence textCNN model diagram
图 2. 用于实例句子 textCNN 模型图

分别是嵌入层、卷积层、池化层、全连接层。嵌入层将新闻文本中的句子分割为单词，并词向量化，嵌入层将文本转化为权值矩阵，每个单词转化为词向量矩阵，由此得到一个 $N \times K$ 的矩阵 M，其中每一行代表一个词向量，此矩阵可以是静态的，也可以是动态的；卷积层将嵌入层看作输入层，通过卷积操作将每个词向量矩阵卷积成一维矩阵，提取输入的特征；池化层在卷积层之后，将池化层看作输入层，池化层与卷积层有着相似的特征提取的作用，它将卷积层局部域进行池化操作，将分类特征提取池化；全连接层通过采用 softmax 逻辑回归函数进行分类并将类型 id 输出，将新闻分成按照原国家新闻出版总署制定的新闻出版分类法，分成 14 个一级分类。

4.2. 新闻推荐

由于微博文本是短文本，新闻文本可以长文本也可以是短文本，本文的相似度计算是采用 VSM(向量空间模型)模型下的改进 TF-IDF 算法。VSM 模型通过词项－权值的经典表示算法来描述文本，即每篇微博文本或新闻文本都可以采用一个 $\{K,W\}$ 二元组来表示。K 是词项(文本句子中分词得到)的集合；W 是对应词的权值(如出现次数等)。VSM 常用权值计算方法有评率统计、TF-IDF 算法等。其中 TF-IDF 算法如下：

$$W_{ij} = \left[\frac{num(t_{ij})}{total(D_i)}\right] * \log\left[\frac{N}{n_{ij}}\right] \tag{4}$$

其中，$num(t_{ij})$ 为此单词在这篇文本中出现次数，$total(D_i)$ 为该文本中总单词数，N 为此分类文本集中该单词次数，n_{ij} 为该分类总文本集数量。考虑新闻具有很强的时效性，引入时效参数：

$$D(i) = \left[\frac{-(t_p - t_n)^2}{t_B + 1}\right] \tag{5}$$

其中，t_p 为当前时间，t_n 为新闻时间，t_B 为基准时间戳。则改进后的 TF-IDF 算法为：

$$W_{ij} = \left[\frac{num(t_{ij})}{total(D_i)}\right] * \log\left[\frac{N}{n_{ij}}\right] * D(i) \tag{6}$$

衡量用户挖掘兴趣集，文本 i 之间的相似度算法为：

$$sim(U_{FP,j}) = \sum_{w_i,w_j=1}^{k} W_{Ij} * W_{ij} \tag{7}$$

其中，w_{Ij} 是关键词 j 在用户兴趣中 U_{FP} 的权值，w_{ij} 是关键词 j 在文本 i 中的权值。

本文融合算法中用户浏览新闻的协同过滤算法：

$$UserSim(U,V) = \frac{U_N \cap V_N}{U_N \cup V_N} \tag{8}$$

其中 U_N 为用户 u 的新闻浏览集合，V_N 为用户 v 的新闻浏览集合，取用户 v 的 $V_N - U_N \cap V_N$ 新闻集作为候选推荐新闻，然后找出相似度最高的前 n 名用户，分别找出候选新闻，然后全部加起来选取其中推荐次数最多的前 m 篇新闻。则有：

$$\sum_{I=0}^{n} UserSim(U,V_I) = \sum_{I=0}^{n} V_{IN} - U_N \cap V_{IN} \tag{9}$$

其中 V_I 表示第 I 个用户，V_{IN} 表示第 I 个用户的浏览新闻集。

最后融合兴趣推荐和协同过滤的算法来表示通过挖掘用户 u 对新闻集 I 的感兴趣程度：

$$EXP\left(U,V_I,I_j\right)=\tau\sum_{j=0}^{m}Sim\left(U\left(U\in U_{FP}\right),I_j\right)+\zeta\sum_{I=0}^{n}UserSim\left(U,V_I\right) \tag{10}$$

其中 τ 、ζ 为调节参数，为方便实验，均取值为 1。 $Sim\left(U\left(U\in U_{FP}\right),I_j\right)$ 为基于用户 U 微博兴趣集对新闻 I_j 的相似度计算，式(10)前半部分为基于用户 U 微博兴趣对 m 篇新闻集的相似度集，后半部分按协同过滤得到的 n 个相似用户的推荐新闻集，将俩部分相加得到融合用户微博兴趣的协同过滤新闻推荐候选集。

5. 实验与分析

5.1. 数据集

新闻分类训练和测试数据集：本文的新闻文本分类训练数据是采用清华大学组提供的 THUCNews 新闻文本分类数据集的一个子集。THUCnews 是包含了 74 万篇根据新浪 RSS 订阅频道几年间的历史数据生成。原始数据集中总共有 14 种已分类的新闻文本，每篇新闻有唯一的编码，从 1~700,000+，每种类型的新闻总数不定。

新闻推荐实验中新闻数据集：采用爬虫爬取的新浪网新闻，从 2018 年 7 月 2 日到 2018 年 9 月 15 日共 12,635 篇新闻。其中新闻数据中包含了新闻 ID、分类类型、网址、标题、新闻内容、新闻时间、新闻评论用户 ID 及词数量等数据。

新闻推荐实验中用户微博数据集：新浪新闻数据中的评论用户 ID 就是该用户在微博中的 ID，通过筛选爬取所需的用户微博数据。其中，微博数据中分为微博编号、微博所属用户 ID、内容、时间、地点、背景、标签、关注者信息等共 10 万多条数据。

5.2. 实验

通过对处理好的新闻数据中的新闻评论用户 ID 统计计算用户评论总数，得到关于用户对新闻评论的排序，随机选取评论新闻数量超过 40 篇的 100 名用户，将用户随机分成 10 组，每组用户随机抽取 10 名，10 组用户的抽取存在一定的重复。并将用户新闻数据集中用户评论时间靠后的 15 条评论新闻作为测试集，其余的评论新闻作为训练集并爬取其新浪微博数据集。在通过第 2 节中的算法处理得到用户微博兴趣集。将新闻数据集安装第 3 节中的 VSM 模型处理得到新闻模型数据集。最终按照本文式 3 和式 10 得到实验结果，采用平均值。

评估指标：本文采用的评估指标是在推荐领域广泛使用的评估指标，即准确度(precision)、召回率(recall)和 F1 值。其中，精确度为推荐的新闻中用户阅读(此处用评论代替)的比例，如式(11)给出。召回率为用户所有阅读中被推荐出来的比例，如式(12)。

$$precision=\frac{R_N\left(L\right)}{L} \tag{11}$$

$$recall=\frac{R_N\left(L\right)}{R\left(L\right)} \tag{12}$$

其中， $R_N\left(L\right)$ 代表推荐给用户并被阅读(评论)的数量，L 代表推荐总数， $R\left(L\right)$ 代表用户的阅读(评论)新闻的总数。为了进一步综合衡量准确率和召回率，使用 F1 值。具体计算方法由式(13)给出。

$$\mathrm{F1}=\frac{2\times precision\times recall}{precision+recall} \tag{13}$$

5.3. 实验对比算法

为了验证本文算法的有效性，本文采取了同其他 2 种经典推荐策略进行对比实验，即基于内容的推荐和基于用户的协同过滤推荐。对比算法的简介及算法如下：

1) 基于内容的推荐(CB)：该算法的主要思想是根据用户以往历史浏览新闻纪录，将历史浏览新闻作为用户兴趣集，推荐给用户相似的新闻。

2) 基于用户的推荐(UCF)：通过找到具有相似兴趣的邻居用户，将邻居用户其他的阅览新闻推荐给用户。算法如式(9)。

5.4. 实验结果与分析

在确定用户微博兴趣挖掘算法(式 3)中参数 α 中参数对推荐效果的影响的结果如图 3 所示。α 取值 0.4 时，新闻推荐的 F1 值最大。即说明利用用户微博背景信息、标签、用户关注者信息的挖掘兴趣效果权值大于人名、地名等时推荐效果更好。

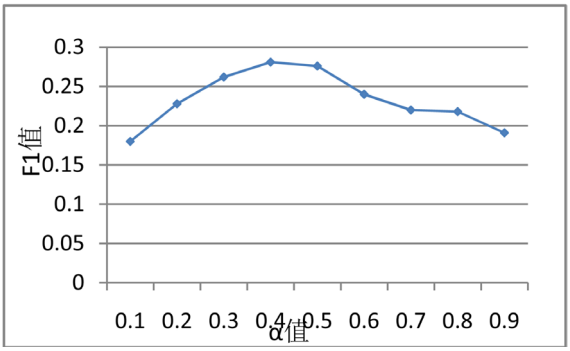

Figure 3. User Weibo mining interest parameters
图 3. 用户微博挖掘兴趣参数

本文的推荐效果实验图如图 4 所示。其中，T@n 为推荐数量，由图 4 可知，在推荐数量较少时，本文推荐 F1 值接近于基于内容的推荐方法，在数量合适时，本文算法高于其他算法的 F1 值，说明本文算法有效提高了推荐的综合指标。对比基于用户协同过滤的算法时，当用户数据少时，本文算法因为结合了挖掘到的用户微博兴趣用于推荐，所以一定程度解决了基于用户协同过滤中的冷启动问题。因此，本文融合用户微博兴趣与协同过滤的新闻推荐算法的实际应用性能要优于传统的基于内容和协同过滤推荐方法。

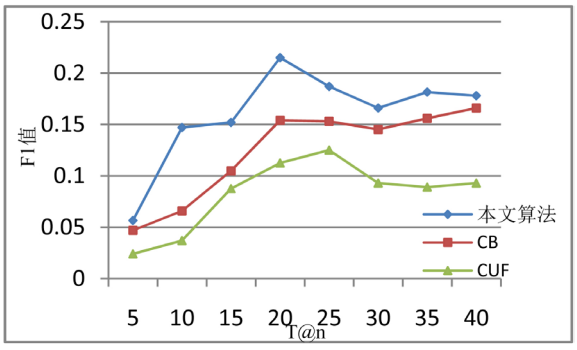

Figure 4. F1 value map under different recommended quantities
图 4. 推荐算法在不同推荐数量下的 F1 值图

6. 总结

本文提出了一种融合用户微博兴趣挖掘和协同过滤的新闻推荐方法，该方法通过将用户微博数据和用户新闻评论记录数据联系起来用于新闻推荐。在挖掘用户微博兴趣中，考虑了用户背景、用户微博发文、用户关注者背景等信息，挖掘了用户兴趣的多样性，在新闻推荐中通过引入带时间权重的算法，使新闻时效性大大增强，再融合协同过滤算法，通过使用对比证明，改进了传统基于内容算法的无法挖掘用户潜在兴趣的问题，也改进了传统协同过滤的冷启动问题，表明了本文方法的有效性。下一步，可以将本文算法融合其他用户社交平台数据做进一步研究，同时结合微博数据中的图片数据进一步挖掘用户潜在兴趣，从而完善挖掘用户的兴趣集更好地用于新闻推荐。

基金项目

国家基金项目(61572144)；广东省科技计划项目(No. 2016B030306002, No. 2015B010110001, 2017B030307002)。

参考文献

[1] Li, L., Chu, W., Langford, J., *et al.* (2010) A Contextual-Bandit Approach to Personalized News Article Recommendation. *Proceedings of the* 19*th International Conference on World Wide Web*, Raleigh, North Carolina, 26-30 April 2010, 661-670. https://doi.org/10.1145/1772690.1772758

[2] Adnan, M.N.M., Chowdury, M.R., Taz, I., *et al.* (2014) Content Based News Recommendation System Based on Fuzzy Logic. *IEEE* 2014 *International Conference on Informatics*, *Electronics & Vision* (*ICIEV*), Dhaka, 23-24 May 2014, 1-6. https://doi.org/10.1109/ICIEV.2014.6850800

[3] Wu, Y., Rui, T. and Ling, L.U. (2016) News Recommendation Method by Fusion of Content-Based Recommendation and Collaborative Filtering. *Journal of Computer Applications*, **2**, 025.

[4] Liu, S., Dong, Y. and Chai, J. (2016) Research of Personalized News Recommendation System Based on Hybrid Collaborative Filtering Algorithm. 2016 2*nd IEEE International Conference on Computer and Communications* (*ICCC*), Chengdu, 14-17 October 2016, 865-869.

[5] 杨武, 唐瑞, 卢玲. 基于内容的推荐与协同过滤融合的新闻推荐方法[J]. 计算机应用, 2016, 36(2): 414-418.

[6] Chen, C., Meng, X., Xu, Z., *et al*. (2017) Location-Aware Personalized News Recommendation with Deep Semantic Analysis. *IEEE Access*, **5**, 1624-1638. https://doi.org/10.1109/ACCESS.2017.2655150

[7] Kumar, V., Khattar, D., Gupta, S., *et al*. (2017) Deep Neural Architecture for News Recommendation. *Working Notes of the* 8*th International Conference of the CLEF Initiative*, Dublin, Ireland, 11-14 September 2017.

[8] Chen, C., Lukasiewicz, T., Meng, X., *et al*. (2017) Location-Aware News Recommendation Using Deep Localized Semantic Analysis. *International Conference on Database Systems for Advanced Applications*, **10177**, 507-524. https://doi.org/10.1007/978-3-319-55753-3_32

[9] Meng, X.W., Chen, C. and Zhang, Y.J. (2016) A Survey of Mobile News Recommend Techniques and Applications. *Chinese Journal of Computers*, **39**, 685-703.

[10] Abel, F., Gao, Q., Houben, G.J., *et al*. (2011) Analyzing User Modeling on Twitter for Personalized News Recommendations. *User Modeling*, *Adaption and Personalization*, **6787**, 1-12.

[11] Zhang, X.Y., Su, Y. and Yan, X.H. (2016) Context-Awareness Recommendation Based on User Browsing Log. *Computer Engineering and Applications*, **52**, 99-104.

[12] Zhong, Z.M., Guan, Y., Hu, Y. and Li, C.H. (2017) Mining User Interests on Microblog Based on Profile and Content. *Journal of Software*, **28**, 278-229.

[13] Hu, Y., Wang, C.J., Wu, J., Xie, J.Y. and Li, H. (2014) Overlapping Community Discovery and Global Representation on Microblog Network. *Journal of Software*, **25**, 2824-2836.

[14] Xu, Z.M., Li, D., Liu, T., Li, S., Wang, G. and Yuan, S.L. (2014) Measuring Similarity between Microblog Users and Its Application. *Chinese Journal of Computers*, **37**, 207-218.

[15] 古万荣, 董守斌, 曾之肇, 等. 基于微博用户模型的个性化闻推荐[J]. 中文信息学报, 2016, 30(1): 93-100.

[16] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, Qatar, 25-29 October 2014, 1746-1751. https://doi.org/10.3115/v1/D14-1181

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org