# 基于用户上下文序列的个性化新闻推荐方法研究

温宇俊[1],袁晖[2]
(1. 中国传媒大学 理工学部,北京 100024;2. 中国传媒大学 新媒体研究院,北京 100024)

**摘要:** 面对互联网上丰富的新闻时,用户更希望阅读到自己感兴趣的那一部分新闻内容,避免太多资讯带来的信息过载问题,所以比搜索引擎等更加智能的个性化推荐技术成了新闻受众目前最迫切的需求之一。新闻推荐往往基于用户阅读记录实现,而当用户的阅读记录过少时,通常采用了流行度等简单方法实现,这类做法通常忽略了用户阅读的新闻序列之间其实还是有一定的关联,本文针对用户阅读过程中冷启动问题,从用户上下文序列的角度探讨其在新闻推荐中的影响。

**关键词:** 新闻推荐;记录序列;冷启动
**中图分类号:** TP391.1 **文献标识码:** A **文章编号:** 1673-4793(2018)04-0048-05
DOI:10.16196/j.cnki.issn.1673-4793.2018.04.008

# Research on News Recommendation Method Based on User Behavior Sequence

WEN Yu-jun[1],YUAN Hui[2]
(1. Faculty of Science and Technology,Communication University of China,Beijing 100024,China;
2. New Media Institute,Communication University of China,Beijing 100024,China)

**Abstract:** When facing a lot of news on the Internet,users may want to read some news that they are interested in,which can avoid information overload caused by too much information. So personalized recommendation technology,which is more intelligent than search engines,has become one of the most popular technologies for the news audience. The news recommendation often requires the user to read the record. Currently,when users have little reading history,system recommend news often by popularity factor or other simple method. In fact,it often ignores the connection between the series of news reading record. To solve the user cold-start problem,this paper discusses its influence on news recommendation from the perspective of user context sequence.

**Key words:** news recommendation; record sequence; cold-start

自大数据技术蓬勃发展以来,个性化新闻推荐系统也逐渐出现在互联网的应用服务中,它目的是为了帮助新闻阅读者在无法准确描述需求时,寻找有用的资讯信息;它通过研究发现用户浏览行为的历史数据中的潜在价值,从而对新闻行业中的未知应用领域做出有效的预测与评估,因其具有行业独立性,目前个性化新闻推荐已成为一个重要的研究方向。用户在浏览新闻的过程中会受到短期兴趣的影响,往往是由于用户对个别的关键词产生了阅读的兴趣,继而开始追踪此类新闻的连续报道,从这个

**收稿日期:** 2018-05-12
**基金项目:** 国家科技支撑计划(2014BAK10B01);中央高校基本科研业务费专项资金(3132016XNG1605)
**作者简介:** 温宇俊(1981-),男(汉族),广东湛江人,中国传媒大学博士研究生、副教授. E-mail: wenyujun@cuc.edu.cn

角度来说，用户容易受到先前所阅读新闻的影响，因此上下文信息对用户的短期兴趣建模而言是不可或缺的。但如果纯粹从内容的角度来考虑，又容易引起推荐的多样性不足，用户往往希望在阅读过程中系统能够推荐一些新鲜的，甚至可能不在自己兴趣范围之内的新闻内容，从而能够扩展自己的阅读范围，本文针对这样的问题，从用户上下文序列的角度探讨其在新闻推荐中用户的阅读趋势的影响。

## 1　研究现状

中文新闻个性化推荐技术，也称中文新闻个性化过滤技术，在目前已成为推荐系统中的成熟分支之一。现有的新闻推荐体系从算法结构类型上仍然可以按照传统的推荐系统分类体系分为三种类型：基于内容的个性化新闻推荐、基于协同过滤技术的个性化新闻推荐以及基于混合模型的个性化新闻推荐。

基于内容的新闻推荐[1]属于推荐技术在新闻领域的很重要的一种类型[2-3]，最初的新闻推荐技术也是从内容来考虑模型的构建。算法在构建用户兴趣模型以后，会查找新出现的新闻与用户兴趣模型具有一定相似性的内容，继而按照某种排序方式进行推荐。主要方法有通过关键词的角度考虑构建兴趣模型，如 NewsDude[4] 等，也有从主题模型[5]的角度考虑用户兴趣等，基于内容的个性化新闻推荐算法主要特点为：1. 不需要大量的用户历史数据，以用户阅读过的历史新闻为出发点即可挖掘其兴趣爱好；2. 对新用户新闻内容推荐与用户的潜在兴趣发现方面有较大的局限性，推荐的多样性稍显不足，往往会较为集中的聚焦在用户兴趣相关的主题及分类中。

基于协同过滤的推荐技术[6]在推荐领域适用范围较大，效果也较好，尤其在商品推荐方面一贯以来都非常的行之有效。早期的 Google News 将协同过滤思想引入到新闻推荐领域[7]，主要解决了大数据下的新闻推荐中的用户聚类的问题。基于协同过滤的个性化新闻推荐方法主要有以下特点：1. 主要通过用户之间、新闻之间的协同性进行推荐，推荐结果的多样性、新颖度方面普遍好于基于内容的个性化新闻推荐方法。2. 协同过滤在发现用户的潜在兴趣方面，即使是复杂结构的新闻类型（如视频新闻、图片新闻等）也都可以适用，表现也都较为出色；3. 协同过滤目前主要需要解决的是普遍存在的冷启动问题，用户数据往往会因为稀疏度非常高，计算相似用户难度过大。

基于混合模型的新闻推荐技术并无特定的模式，一般的实现方法都是将基于内容的推荐方法与基于协同过滤的推荐方法通过不同的方式构造混合模式，从而形成新的推荐框架。如今日头条[8]算法模型建立在它的投票机制上，以人群为单位，针对某篇新闻，用户对其进行投票，最后统计该新闻所属的用户群体中的得票率，根据得票率进行用户群推荐，这属于在流行度基础上的一种混合推荐算法。针对不同的特定场景，使用混合推荐模型往往会针对性的解决部分难以处理的问题，获得更好的效果。

## 2　基于上下文序列的推荐算法

当用户的新闻阅读历史较少的时候，我们往往会采用基于内容的方法构造用户兴趣模型，而这样构造的兴趣偏好往往会比较集中，因此这样做出的推荐结果多样性以及新颖度是较为不足的。针对用户冷启动问题，我们从用户上下文序列的角度探讨其在新闻推荐中用户的阅读趋势的影响。

### 2.1　连续词袋模型

Mikolov[9]在关于词向量的论文中基于此思想提出了一个统计语言模型 CBOW（连续词袋模型），可以对文本语言进行词预测，通过前面的 c 个词或者是前后连续的 c 个词出现的前提下计算某个词出现的概率，大致思想如下：1. 文本训练集切分，对所有的文本进行分词，以词为计算的最小单元；2. 设定滑动窗口，将滑动窗口内的词相关的词向量（初始化为 1）送进该神经网络进行训练。3. 输出为一棵霍夫曼树的内部节点模型参数及所有的词向量。其训练过程采用了一个三层神经网络进行实现，如图 1 所示。

我们对比该过程看待新闻推荐系统中的用户点击记录，很容易可以联想到用户在某一段时间内的连续阅读操作集合，相当于训练语料中的独立文本，该连续操作序列相当于文本中的词序列，操作之间

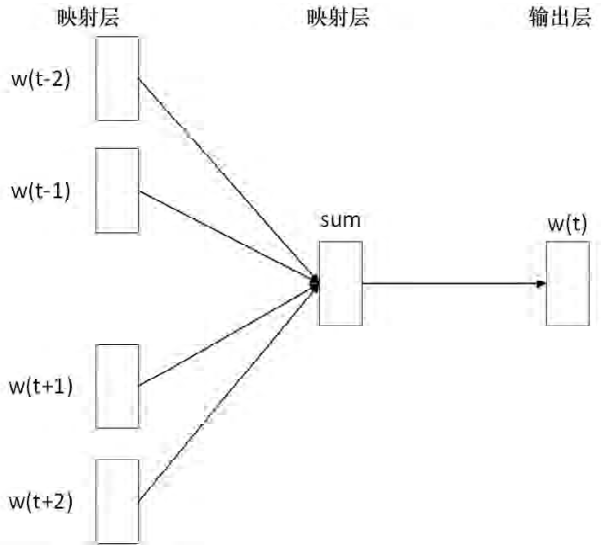


图 1 CBOW 模型

的共现序列可以作为样本，所以，我们将用“词”和“项”互换，这样我们就可以对新闻项进行类似的计算，并使用该模型进行推荐。

## 2.2 行为序列选择

我们给定所有的新闻集合为 D，当用户 u 在浏览该集合中的新闻的过程中，我们可以通过记录用户浏览日志的方式，跟踪该用户阅读过的新闻文章序列。我们用 D 表示所有的文章序列，定义序列 $Ru = <d1,\cdots,dm>$，是一个有序的新闻列表，其中 $di \in D$ 是被用户已经阅读过的新闻。我们用 Dt 表示少于一个时间阈值 $\triangle t$ 内用户阅读的所有新闻文章序列，$\triangle t$ 实际上为用户阅读前后两篇新闻之间的时间间隔，如图 2 所示。

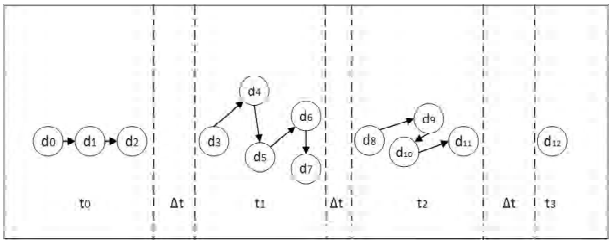


图 2 用户阅读序列

其中 $\{t_0,\cdots\cdots,t_n\}$ 是一个不定长的时间片段，$\triangle t$ 是两次浏览序列产生之间的时间间隔，该时间间隔也为一个不定长度的值，因为人阅读新闻的时间点无法固定，且不具备规律性。如果两个序列有相似的上下文，那么一个用户想要阅读的下一篇文章应该是类似的，在本文中我们通过向量的方式对其进行表示，从而在向量空间层面挖掘出下一篇新闻。

由于新闻阅读序列的特殊性，并不像商品推荐中容易丢失行为序列性，它如同文本内容中的词一样，存在一定的一定程度的前后因果性，虽然这种因果性看起来不强，但我们通过序列选择仍然能保存下来这部分关联信息。

## 2.3 新闻推荐

与传统新闻推荐算法的目的一致，我们基于连续词袋模型主要用于解决已知用户阅读新闻序列的前提下，预测下一步可能会阅读的新闻；但本文提出的新闻推荐方法也有和传统新闻推荐算法不一样的地方，因为推荐即我们主要面向于解决新闻系统的长尾效应[10]，一方面可以为用户提供更多样化的推荐列表，另一方面也为新闻的传播路径提供更多的可能性，推荐流程图如图 3 所示。

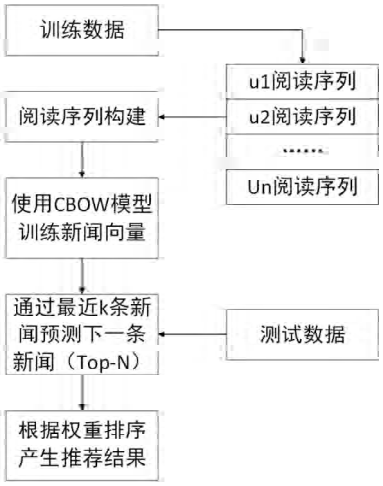


图 3 推荐流程图

CBOW 模型的每个非节点的模型参数及叶子结点的向量与词向量模型一样，是训练过程中产生的副产品，我们想获取新闻推荐结果，需要将用户最近浏览的新闻序列作为神经网络的输入，待推荐新闻中的新闻项作为最终的叶子结点，计算分类路径，并获取到达此节点的概率 $P(d_i \mid context(d_i))$，最后对概率最高的做 Top－N 推荐。

$$P(d_i \mid context(d_i)) = \prod_{j=1}^{l^d} p(c_j^d \mid x_d, \theta_{j-1}^d) \quad (1)$$

其中 $c_j^d = \{0,1\}$ 为所构造的 CBOW 模型的霍夫曼树上的第 j 个节点的分类标记，$\theta_{j-1}^d$ 为节点上 LR 回归的参数。

## 3　实验及结果分析

### 3.1　实验数据

我们采集了某新闻网站的用户行为数据，通过清洗后选择部分用户作为实验验证数据，用户数 1000 名，时间跨度为 1 个月，总用户记录 60000，新闻数量为 4000 篇，对该数据集按照阅读时间分为 4 周，每周按时间取前 80% 的新闻作为训练集，同时去除部分未在训练集中出现过的新闻，主要观察阅读次数少于 20 条新闻的 300 位用户的推荐结果，以此验证算法面向用户冷启动的效率。

### 3.2　参数设置

主要需要确定的 2 个参数需要确定，1) 有序列时间间隔 △t; 2) 此外梯度上升速率设置为 η = 0.005。

我们将用户浏览新闻的时间间隔做了统计，分别统计用户两条新闻产生的时间间隔为 10 分钟内、10 – 20 分钟、20 – 30 分钟、30 – 40 分钟、40 – 50 分钟、50 – 60 分钟、60 – 90 分钟、90 – 120 分钟、120 – 180 分钟、180 – 240 分钟、240 – 300 分钟、300 – 360 分钟以及 360 分钟以外的平均情况，由训练集中用户的新闻阅读时间间隔分布统计情况我们可以看到，用户浏览两条新闻之间的时间间隔大部分都在 10 分钟以内，我们猜测若用户两条浏览记录的间隔时间过大，表明这是两段阅读行为，因此我们选择最合理阅读分段时间间隔 △t 为 10 分钟，意味着用户在阅读完一篇新闻后若 10 分钟内没有新的阅读行为产生，那么就可以认为先前的新闻集合可以作为一个浏览序列。

### 3.3　结果分析

为了评测模型的有效性，我们主要考虑采用同样是基于上下文序列的马尔科夫模型的协同推荐算法( MR) 进行实验比较。

实验主要对准确率和召回率的 F1 值、多样性进行实验结果评测以及综合比较分析，对于推荐列表数量我们仍然采用 Top – N 推荐，N 分别为 10、20、30。若部分用户在测试集中实际阅读记录不满被推荐新闻数目时，按推荐未命中处理。

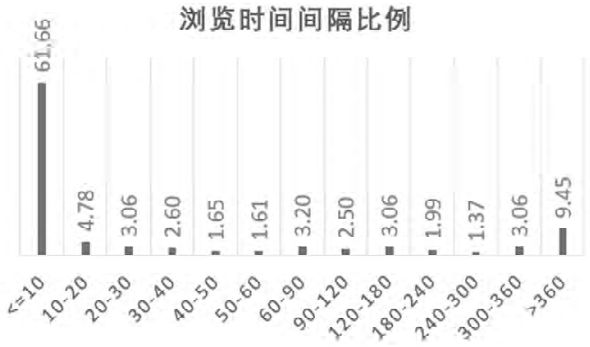


图 4　时间序列间隔

( 1) 准确度 F1 值

表 1　F1 值

| | Top@ 10 | Top@ 20 | Top@ 30 |
|---|---|---|---|
| MR | 0.046 | 0.103 | 0.181 |
| CR | 0.112 | 0.134 | 0.216 |

从准确度上来说，基于马尔科夫模型上下文推荐算法与 CR 算法在推荐列表内容少的时候效果差不多，但在推荐列表内容逐渐增多的情况下，CR 算法效果更佳，准确度的 F1 值提高了约 3% 左右，主要原因是对于多层的训练过程中产生的语义联想的能力要强，因此协同性更好。

( 2) 多样性

表 2　多样性

| | Top@ 10 | Top@ 20 | Top@ 30 |
|---|---|---|---|
| MR | 0.231 | 0.341 | 0.421 |
| CR | 0.402 | 0.438 | 0.474 |

从多样性来看，CR 算法要比基于马尔科夫的新闻上下文推荐算法 MR 要好，多样性提高了 5 % 左右。主要原因是在推荐过程中因为融入了序列选择的概念，给用户进行推荐过程中可选择范围扩大。

## 4　结论

本文主要面对用户的少量数据，从阅读序列的上下文角度，引入连续词袋模型进行模型训练，基于当前的上下文把最符合用户阅读趋势的下一条新闻推荐给用户，实验表明，在用户冷启动方面与经典的基于马尔科夫模型的新闻推荐算法相比，主要在推荐的多样性上具有更好的效率。

## 参考文献


[1] Pazzani, Michael J, Billsus. Content - based recommendation systems [C]. Adaptive Web Springer - Verlag, 2007, 325 - 341.

[2] Ahn J W, Brusilovsky P, Grady J. Open user profiles for adaptive news systems: help or harm [C]. International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May, DBLP, 2007, 11 - 20.

[3] Mooney R J, Roy L. Content - based book recommending using learning for text categorization [C]. ACM Conference on Digital Libraries, ACM, 2000, 195 - 204.

[4] Etzioni O, Ller J, Rg P. Proceedings of the third annual conference on Autonomous Agents [C]. Conference on Autonomous Agents, ACM, 1999.

[5] Kuang W, Luo N. User interests mining based on Topic Map [C]. International Conference on Fuzzy Systems & Knowledge Discovery, IEEE, 2010, 2399 - 2402.

[6] Su X, Khoshgoftaar T M. A survey of collaborative filtering techniques [M]. Hindawi Publishing Corp, 2009.

[7] Das A S, Datar M, Garg A. Google news personalization: scalable online collaborative filtering [C]. International Conference on World Wide Web, ACM, 2007, 271 - 280.

[8] 今日头条推荐算法原理全文详解 [OL]. http://36kr.com/p/5114077.html.

[9] Mikolov T, Chen K, Corrado G. Efficient estimation of word representations in vector space [J]. arXiv Preprint arXiv, 2013, 1301 - 3781.

[10] Zhang L, Qin T, Teng P. Using Key Users of Social Networks to Solve Cold Start Problem in Collaborative Recommendation Systems [J]. Information Technology Journal, 2013, 12(22): 7004 - 7008.


（责任编辑：宋金宝）