顾及事件地理位置的新闻推荐方法研究

袁仁进 陈 刚

(信息工程大学地理空间信息学院 郑州 450052)

摘 要 为研究新闻事件发生地对新闻推荐系统性能的影响,提出了一种顾及事件地理位置的新闻推荐算法。首先,设计了提取新闻事件发生地的相关算法;其次,结合向量空间模型、TF-IDF 算法和 word2vec 工具构建了新闻特征向量;接着,着重讨论了用户兴趣模型的构建问题;最后,运用余弦相似度方法计算用户兴趣模型与候选新闻集之间的相似性,从而完成推荐。实验结果表明,设计的新闻事件发生地抽取算法的性能较好,准确率达到 93.6%,以此为基础构建的新闻推荐算法与协同过滤推荐算法相比仅考虑新闻内容的推荐算法在 F 值上有所提高。

关键词 推荐系统,地理位置,用户兴趣模型,信息抽取,向量空间模型

中图法分类号 TP391 **文献标识码** A

Research on News Recommendation Methods Considering Geographical Location of News

YUAN Ren-jin CHEN Gang

(Institute of Geospatial Information, Information Engineering University, Zhengzhou 450052, China)

Abstract In order to research the impact of news event place on the recommendation performance of news recommendation system, a News recommendation algorithm Considering Geographical Position(NCGP) method is proposed. Firstly, an algorithm was designed to extract the place of news event. Secondly, the vector space model, TF-IDF algorithm and word2vec tool were used to construct the news feature vector. Then, constructing the user interest model was discussed deeply. Finally, the cosine similarity method was used to calculate the similarity between the user interest model and the candidate news set to complete the recommendation. The experimental results show that the performance of the proposed news event place extraction algorithm is better, and the precision can reach 93.6%, besides, the F-value of NCGP is improved compared with the collaborative filtering recommendation algorithm and the recommendation algorithm that only considers news content.

Keywords Recommendation system, Geographical location, User interest model, Information extraction, Vector space model

新闻推荐系统经过一定时间的发展,目前已经有了一些经典算法,主要包括协同过滤新闻推荐[1-2]、基于内容的新闻推荐[3]、基于知识的新闻推荐[4-5]以及混合新闻推荐[6-7]。随着移动通信技术的发展以及智能手机、平板等移动终端越来越便捷化,移动新闻推荐技术也逐步发展起来。孟祥武等[8]对移动新闻推荐技术的现状及应用做了综述性总结,同时一些学者结合用户地理位置在不同的推荐算法下对新闻推荐方法进行了深入研究[9-11]。移动新闻推荐与用户所处地理位置或用户的移动轨迹有关,通过位置传感器感知用户的地理位置,然后根据位置信息或轨迹进行新闻推荐,但未考虑新闻事件发生地对新闻推荐系统的影响。在实际情况中,新闻中的事件地理位置(即事件发生地)也会对用户的阅读兴趣产生较大影响。其主要包括两方面:1)当某些地区发生重大事件(如体育盛事、重大灾害、军事情况等)时,用户的关注点将被吸引到这些特定地区,此时应当为用户持续推荐这些地区的用户感兴趣的新闻而非其他地区相似内容的新闻;2)当用户去旅游、出差或者去某地区从事相关活动之前,一般想要详细了解该地区的人文和社会环境信息,从而尽可能全面地掌握该地区的情况,此时应当根据用户的浏览记录提取出用户感兴趣的地区,然后给用户推荐该地区的用户感兴趣的新闻。可见,在新闻推荐系统中考虑新闻地理位置时能够提高推荐区域的准确性,然而目前鲜有学者研究新闻事件地理的位置对新闻推荐的影响,大都集中在移动新闻推荐的研究中。

针对上述问题,为探讨新闻事件地理位置对推荐效果的影响,本文将新闻事件的地理位置作为一个重要需求引入到新闻推荐中,提出了一种顾及事件地理位置的新闻个性化推荐算法(News recommendation algorithm Considering Geographical Position,NCGP),从而更好地为用户提供服务,提升用户满意度。

1 算法思想和流程框架

本文根据用户的阅读记录,综合考虑新闻事件地理位置和新闻内容来构建用户兴趣模型,从而实现对用户的个性化新闻推荐。

定义 1(新闻特征向量) 由于新闻所包含的内容属于文本类型,因此使用一个多维向量($d=(w_1,w_2,\cdots,w_q)$)来表示

袁仁进(1994-),男,硕士生,主要研究方向为个性化推荐、战场环境数据分析,E-mail:1104372469@qq.com;**陈 刚**(1971-),男,教授,博士生导师,主要研究方向为虚拟地理环境、战场环境仿真与分析、个性化推荐,E-mail:chenggang_vga@sina.com(通信作者)。

新闻的内容，将新闻文本向量化的结果称为新闻特征向量。

定义 2（“用户-地点-新闻”三层层次结构的用户兴趣模型（User-Location-News User Interest Model，ULN-UIM））　对于某一用户，针对有地理位置的新闻集，将新闻事件地理位置作为一个重要参考点，构建 ULN-UIM。

定义 3（用户-新闻类别-新闻）　三层层次结构的用户兴趣模型（User-Categories-News User Interest Model，UCN-UIM）。对于某一用户，针对无地理位置的新闻集，根据新闻内容进行聚类，构建 UCN-UIM。

本文的算法流程框架如图 1 所示。

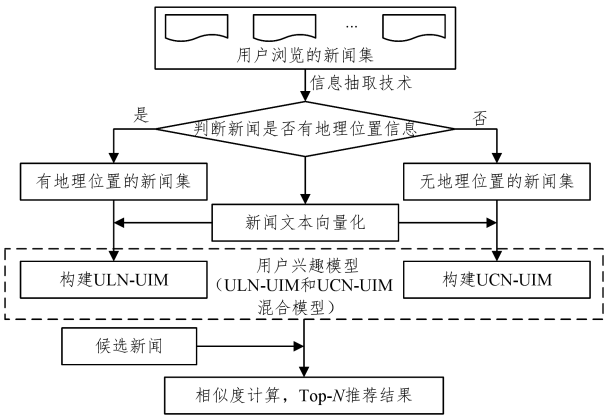

图 1　NCGP 算法流程图

算法思想主要包括 4 个方面：1）新闻事件地理位置信息抽取；2）新闻文本向量化；3）用户兴趣模型构建；4）相似度计算。

1）新闻事件地理位置信息抽取（见第 2 节）：一般来讲，用户浏览的新闻并非都带有地理位置信息，因此对于用户阅读过的新闻记录，需通过信息抽取技术将新闻集分成有地理位置的新闻集与无地理位置的新闻集两部分。

2）新闻文本向量化（见第 3 节）：新闻所包含的内容属于文本类型，若要构建本文的用户兴趣模型，首先应将新闻文本向量化，得到新闻特征向量。

3）用户兴趣模型构建（见第 4 节）：构建用户兴趣模型是推荐系统中最重要的一步。在前面两步的基础上，对用户浏览记录中有地理位置的新闻集构建 ULN-UIM；对用户浏览记录中无地理位置的新闻集构建 UCN-UIM；最后根据两种新闻集的数量权重构建 ULN-UIM 和 UCN-UIM 的混合模型，该混合模型为该用户最终的用户兴趣模型。

4）相似度计算：相似度计算方法很多，最常用的主要有 Pearson 相似性和余弦相似性。余弦相似性通过计算两个向量之间的夹角余弦来衡量两者之间的相关性。由于该方法计算简单，并且本文中用户兴趣模型和新闻最终都采用向量表示，因此本文采用余弦相似性方法（如式（1）所示）来计算用户兴趣模型和候选新闻集中每个新闻特征向量之间的相似性，并选取相似度靠前的 N 条新闻推荐给用户。

$$sim(i,j)=\cos(i,j)=\frac{i\cdot j}{|i|*|j|} \tag{1}$$

2　新闻事件地理位置信息抽取

2.1　新闻事件发生地抽取算法的原理

对新闻进行命名实体识别（Named Entity Recognition，NER）是确定新闻事件发生地的前提。NER 的任务是从文本中识别出诸如人名、组织名、日期、时间、地点、特定的数字形式等内容，并添加相应的标注信息[12]。近年来对 NER 的研究呈现出活跃的态势，研究方法主要包括基于规则的方法和基于统计的方法[13]。在中文命名实体识别领域，张华平等设计的 NLPIR 汉语分词系统处于领先地位，在封闭测试中对地名的识别 F 值能达到 94.53%[14]。大多数情况下，一条新闻包含多个地点，如何从多个地点中判别出事件发生地是目前的一个难题。鞠久朋等[15]采用条件随机场（CRF）和规则相结合的方法构建出 GSNER 系统，并在北大语料中进行了测试，结果判断出事件发生地的准确率高达 92.86%，该方法的准确率较高，能识别出机构名；但模型构建复杂，需要提前标记大量词语标注进行训练，只能识别国内地名，且新闻事件发生地呈多级显示而不利于集中用户的感兴趣区域。

鉴于 NLPIR 汉语分词系统的高准确率，本文设计了一种中文新闻事件发生地抽取算法，具体表现为 NLPIR 汉语分词系统、地名数据库以及基于规则模型的方法相结合。

第 1 步　使用 NLPIR 汉语分词系统对新闻标题以及新闻内容进行分词，得出标题词语集合与内容词语集合。

第 2 步　利用词语相似度将第一步中的标题词语集合与内容词语集合分别与地名数据库相匹配，识别出标题和内容中的地名信息，地名数据库包括以下几个表：“国家-首都”表、“省-省会”表、“省-地级市”表、“地级市-县区”表。

第 3 步　采用基于规则模型的方法最大程度地确定新闻事件发生地。该规则模型主要考虑两部分：1）地名判定标准；2）新闻定位尺度。

1）地名判定标准

不同于一般的文学作品，新闻写作具有一定的规则，一般根据时间、地点、人物、事件等层次进行描述。因此，针对中文新闻事件发生地的提取，本文总结出以下几条地名判定标准：

①一般情况下，新闻内容中的行政区划分级别比新闻标题中的详细；

②若标题中只有一个地名，则该地名为事件发生地的概率很大；

③若标题中有多个地名，则选取内容地名中与标题相关的地名集合作为候选集，候选集中地名靠前的为事件发生地的概率最大；

④若标题无地名，则内容中地名靠前的为事件发生地的概率最大。

2）新闻定位尺度

提取新闻事件发生地的目的是确定用户感兴趣的区域，因此对新闻事件发生地的粒度划分十分重要。新闻中的地点与行政区域密不可分，以中国为例，行政区划由大到小分为“省-市-县-镇-村”等多级架构，新闻中的地点一般以行政区划中的某一等级命名，因此新闻之间的事件发生地存在复杂的层次包含关系，导致确定新闻事件发生地时存在以下 3 个问题：

①新闻事件发生地过于详细（县-镇-村等行政区），容易导致地点分散，难以形成用户的重点关注区域；

②新闻事件发生地过于广阔（国-省等），容易导致囊括所有新闻，造成用户重点关注区域被淹没；

③同样内容的新闻在不同的新闻发布机构中可能被定义成不同等级的行政区。

因此,本文对新闻事件发生地做以下规定:统一以市级行政区为标准,若地点为市级以下则以该地点所属的市级城市代替,若地点为省级则以省会城市代替,若地点为国家则以首都代替。

2.2 新闻事件发生地抽取算法的设计

新闻事件发生地抽取算法如算法 1 所示。

算法 1 新闻事件发生地抽取算法

输入:新闻标题 news_title,新闻内容 news_content

输出:新闻事件发生地 pos

Begin:

step1 使用 NLPIR 汉语分词系统提取出 news_title 中的地名集合 s1 和 news_content 中的地名集合 s2.

step2 if s1 和 s2 均为空集:

则 pos 返回值为 NULL;

step3 if s1 非空:

遍历 s1,首先将县级地名转换成对应的地级市地名,接着将市级地名保存到列表 s1_city 中,将省级地名保存到列表 s1_pro 中;

if s1_city 非空:

则 pos 返回值为 s1_city[0];

else if s1_pro 非空且 s2 为空集:

则 pos 返回值为 s1_pro[0]对应的省会城市;

step4 if s2 非空且 s1_city 为空集:

遍历 s2,首先将县级地名转换成对应的地级市地名,接着将市级地名保存到列表 s2_city 中,将省级地名保存到列表 s2_pro 中,将国级地名保存到列表 s2_cou;

if s2_city 非空:

if s1_pro 非空:

将 s2_city 中不属于 s1_pro 中的城市剔除,并将 s2_city 剩余城市保存到列表 s2_city_copy 中;

if s2_city_copy 非空:

则 pos 返回值为 s2_city_copy[0],否则将 s1_pro[0]的省会返回给 pos;

else:则 pos 返回值为 s2_city[0];

else:

if s2_pro 非空:

则 pos 返回值为 s2_pro[0]的省会;

else if s2_cou 非空:

则 pos 返回值为 s2_cou [0]的首都;

else:

则 pos 返回值为 NULL;

return pos.

End

3 新闻文本向量化

目前主流的文本特征表示模型主要包括 4 种:布尔模型、概率检索模型、语言模型和向量空间模型(Vector Space Model,VSM)[16]。考虑到新闻数据的高维性以及为便于新闻聚类从而构建用户兴趣模型,本文将采用向量空间模型来表示新闻特征向量。

对于新闻集 $D=\{d_1,d_2,\cdots,d_n\}$,其 VSM 表示为:

$$M=\begin{bmatrix} w_{11} & \cdots & w_{1q} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nq} \end{bmatrix} \tag{2}$$

其中,$[w_{i1},w_{i2},\cdots,w_{iq}]$表示新闻 d_i 的新闻特征向量。本文构建 VSM 的关键有两个方面:1)新闻关键词提取;2)关键词向量化。

1)新闻关键词提取:为避免无关冗余内容对新闻文本向量化产生影响,将提取新闻关键词来代表该新闻内容。最常用和有效的计算方法为 TF-IDF 算法,该方法是信息检索领域的成熟技术,本文不详细展开。

为使关键词权重值处于[0,1]区间内,使用余弦归一化的方式对权重进行归一化处理,权重计算公式为:

$$w_{(i,j)}=\frac{\text{TF-IDF}(i,j)}{\sqrt{\sum_{i=1}^{T}\text{TF-IDF}(i,j)^2}} \tag{3}$$

新闻 d_i 中的关键词 kw_i 及其权重 wei_i 如式(4)所示:

$$kw_{d_i}=\{(kw_1,wei_1),\cdots,(kw_h,wei_h)\} \tag{4}$$

2)关键词向量化:目前将词语转换成向量的有效工具为 word2vec,该工具通过 CBoW 模型和 Skip-gram 模型实现词语向量化。常用的 word2vec 工具为 Python Gensim 主题模型中的 word2vec,本文将以此为基础实现新闻文本向量化,其步骤如下。

步骤 1 对于新闻 d_i,首先使用 Gensim 主题模型中的 word2vec 工具将 kw_{d_i} 中的关键词 $kw_i(i\in[1,h])$转化为向量 $kwvec_i$,维度为 q;

步骤 2 对新闻关键词加权求和得出新闻 d_i 的特征向量:

$$[w_{i1},w_{i2},\cdots,w_{iq}]=\sum_{i=1}^{h}wei_i * kwvec_i \tag{5}$$

4 用户兴趣模型的构建

4.1 构建 ULN-UIM

对有地理位置的新闻集构建 ULN-UIM,ULN-UIM 由三层层次结构组成:用户-地点-新闻。如图 2 所示,第一层节点为用户,第二层节点为用户感兴趣的地点,第三层节点为新闻。

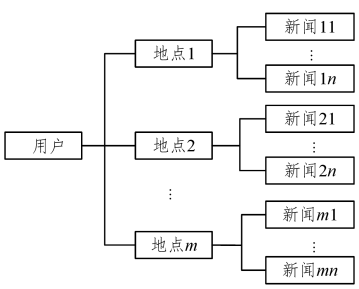

图 2 ULN-UIM 结构图

若用户浏览的新闻聚集于 m 个不同的地点,则 ULN-UIM 可用如下公式表示:

$$ULN\text{-}UIM=\{(L_1,k_1,n_1),\cdots,(L_m,k_m,n_m)\} \tag{6}$$

其中,L_i 表示地点 i 的特征向量,k_i 表示地点 i 的权重,n_i 表示地点 i 包含的用户浏览过的新闻数量。

地点 i 的特征向量根据该地点所包含的所有已浏览过的新闻特征向量根据兴趣度加权平均求出,即地点 i 的特征向量 L_i 的计算公式为:

$$L_i=\frac{\sum_{e_j\in E_j} e_j \cdot I_j}{\sum_{e_j\in E_j} \cdot I_j} \tag{7}$$

其中,E_j 表示地点 i 中用户浏览过的新闻集合,e_j 表示新闻特征向量,I_j 表示为该类别中第 j 个新闻的用户兴趣度,用户浏览过某新闻即表示用户对该新闻感兴趣,因此将 I_j 设为 1,

则式(7)可化简为：

$$L_i=\frac{\sum_{e_j\in E_j} e_j}{n_i} \tag{8}$$

k_i 的值根据地点 i 中用户浏览过的新闻数量占该用户感兴趣的所有地点中的新闻数量总和的权重来计算。

$$k_i=\frac{n_i}{\sum_{i=1}^{m} n_i} \tag{9}$$

在计算时，ULN-UIM 表示为：

$$V_{\text{ULN-UIM}}=(k_i * L_1,\cdots,k_i * L_m)^{\mathrm{T}} \tag{10}$$

4.2 构建 UCN-UIM

对无地理位置的新闻集构建 UCN-UIM，首先需对新闻内容进行聚类。目前，在数据挖掘领域的聚类算法主要包括基于模型的算法、基于网格的算法、基于密度的算法、基于距离的算法 4 种[17]。本研究的数据为新闻数据，具有海量、高维等特点，同时采用了向量空间模型来表示新闻的文本特征，因此基于以上考虑，本文采用基于距离的算法作为新闻聚类方法。Steinbach 等[18]和 Krishna 团队[19]研究得出 K-means 聚类算法，其改进算法——Bisecting K-means 聚类算法，其收敛速度更快、聚类效果更优。综上原因，本文将在向量空间模型的基础上，采用 Bisecting K-means 聚类算法来实现新闻的分类，具体原理请参考文献[18]。

UCN-UIM 采用三层层次结构表示：用户-新闻类别-新闻。如图 3 所示，第一层节点为用户，第二层节点为新闻类别，第三层节点为新闻。

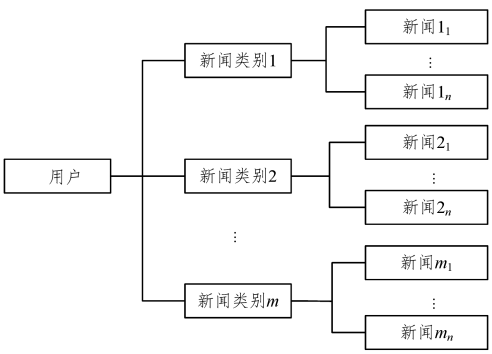

图 3　UCN-UIM 结构图

若新闻集聚类成 m 个不同的新闻类别，则 UCN-UIM 可用如下公式表示：

$$UCN\text{-}UIM=\{(T_1,u_1,n_1),\cdots,(T_m,u_m,n_m)\} \tag{11}$$

其中，T_i 表示新闻类别 i 的特征向量，u_i 表示新闻类别 i 的权重，n_i 表示新闻类别 i 包含的新闻数量。

UCN-UIM 中 T_i 和 u_i 的计算与 ULN-UIM 类似：

$$T_i=\frac{\sum_{e_j\in EE_j} e_j}{n_i} \tag{12}$$

其中，EE_j 表示新闻类别 i 中的新闻集合。

$$u_i=\frac{n_i}{\sum_{i=1}^{m} n_i} \tag{13}$$

在计算时，UCN-UIM 表示为：

$$V_{\text{UCN-UIM}}=(u_i * T_1,\cdots,u_i * T_m)^{\mathrm{T}} \tag{14}$$

4.3 构建用户兴趣模型

一般来说，用户阅读新闻的记录既包含有地理位置的新闻，也包含无地理位置的新闻，因此应当构建 ULN-UIM 和 UCN-UIM 混合模型来表征该用户的用户兴趣模型。用户兴趣模型 V_h 如式(15)所示：

$$V_h=p * V_{\text{ULN-UIM}}+(1-p) * V_{\text{UCN-UIM}} \tag{15}$$

其中，p 表示权重调控因子，其计算公式为：

$$p=\frac{\text{有地理位置的新闻集的新闻数量}}{\text{用户阅读的新闻总数}} \tag{16}$$

最后，使用式(1)中的余弦相似度计算出候选新闻 d_i(V_{d_i} 为 d_i 的特征向量)与用户之间的相似性，并将 Top-N 个新闻推荐给用户，计算公式如式(17)所示：

$$sim(user,V_{d_i})=cos(V_h,V_{d_i}) \tag{17}$$

5　实验及对比分析

5.1 数据准备

本文使用 DataCastle 提供的用户浏览新闻记录作为实验数据集，该数据集从国内财新网随机采集，共包括 10000 名用户在 2014 年 3 月期间浏览的 116225 条记录。每条浏览记录包括用户编号、新闻编号、浏览时间、新闻标题以及新闻文本内容等，假设用户浏览了某新闻代表该用户对该新闻感兴趣。

5.2 评估指标

本文实验采用的评估指标包括准确率和 F 值，F 值由准确率和召回率组合计算。准确率和召回率由混淆矩阵表示，如表 1 所列。

表 1　混淆矩阵

	被推荐	未被推荐
喜欢	True Positive(TP)	False Negative(FN)
不喜欢	False Positive(FP)	True Negative(TN)

准确率 P、召回率 R 的计算公式为：

$$P=\frac{TP}{TP+FP} \tag{18}$$

$$R=\frac{TP}{TP+FN} \tag{19}$$

F 值使用准确率 P、召回率 R 共同计算：

$$F=\frac{2PR}{P+R} \tag{20}$$

5.3 实验结果与分析

5.3.1 新闻事件发生地提取实验

从新闻数据集中随机不重复选取 10 组数据，每组数据包含 200 条新闻，利用第 2 节提出的新闻事件发生地提取算法提取出新闻的事件发生地，同时将人为提取出的新闻事件发生地作为测试集。10 组实验结果的准确率如图 4 所示。

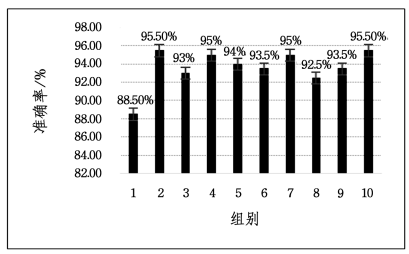

图 4　新闻事件发生地提取算法的准确率

将 10 组数据求平均值，得出新闻事件发生地提取算法的准确率较高为 93.6%，因此可将该算法用于后续推荐系统工作中以获取用户的感兴趣区域。对出现提取错误的新闻进行分析，得出以下 4 种类型的新闻会使提取算法出现差错，如表 2 所列。

表 2 事件发生地提取出错的 4 种新闻类型

类型	新闻标题	新闻内容	真实地点	算法预测地点	分析
1.有机构名，无地名	马航召开新闻发布会通报失联航班最新情况	3 月 9 日下午三点，马航在首都国际机场旁边的国都大饭店召开发布会通报失联航班最新情况	北京	无地点	地名库中无“首都国际机场”、“国都大饭店”等机构名，无法确定其位置。
2.地名缩写	事涉乌市城市改造杨刚案发地产腐败	在中央党校研究生班在职学习三年后，1991 年，杨刚到农八师石河子市担任了 10 个月的副师长，…	乌鲁木齐	石河子	“乌市”代表乌鲁木齐，但地名缩写使得 NLPIR 汉语分词系统无法正确分词，地名库也无法匹配。
3.地名重叠	广东“观音开库”十万人烧香	2 月 25 日，广东佛山西樵山宝峰寺迎来十万游客登山向南海观音“借库”。ReeseCaviezel/东方 IC2/6	佛山	广州	NLPIR 汉语分词系统在分词时出现问题，将“广东佛山西樵山宝峰寺”分成了“广东”、“佛”、“山西”，使得真实地名佛山被略过。
4.国外音译地名	保加利亚苏联红军纪念碑遭“恶搞”	从上至下依次为保加利亚索非亚一座描述苏联红军战士的青铜纪念碑在 2013 年 8 月 21 日、2011 年 6 月 17 日及 2012 年 3 月 15 日的照片…	索菲亚	华沙(保加利亚首都)	NLPIR 汉语分词系统在分词时将“索菲亚”识别成了音译名字，因此无法将其判定为地名。

经总结，该提取算法仍存在以下两点缺陷：

1)无法识别出“首都国际机场”“国都大饭店”等带有地理位置信息的机构名；

2)NLPIR 汉语分词系统存在分词误差，某些特殊情况下(如表 2 中的类型 2－类型 4 情况)会出现地名分词错误，导致事件发生地判断错误。

5.3.2 NCGP 推荐算法实验

在实验数据预处理阶段，将用户浏览记录少于 30 条的用户剔除，共得到 417 名用户的 32770 条浏览记录。这些数据随机分成 5 组，每组数据包含 200 名用户，允许每组数据之间有重复，并将每个用户的后 10 条记录作为测试集，其余数据作为训练集。为保证实验结果的可信度，将 5 组实验结果取平均值作为最终的实验结果。实验中，采用 NLPIR 汉语分词系统对新闻标题和内容进行分词，根据实际新闻内容采用改进的哈尔滨工业大学信息检索中心的停用词表去除停用词。

为验证本文所提 NCGP 算法的推荐性能，实验中将基于用户的协同过滤推荐算法(User-Based CF)[18]、基于 UCN-UIM 的推荐算法(一种基于内容的推荐方法)作为对比算法，来探讨顾及新闻事件地理位置进行个性化推荐能否提升用户满意度。由于推荐结果个数会对评估指标产生影响，因此在实验中考虑了 5 种推荐结果个数(10，15，20，25，30)；在基于 UCN-UIM 的推荐算法和 NCGP 算法中，涉及到 Bisecting K-means 聚类算法，由于聚类簇数 M 会对推荐性能产生影响，因此对聚类簇数 M 选取 5 种聚类情况(10，15，20，25，30)；在 User-Based CF 算法中，将用户浏览过的新闻评分记为 1，未浏览的新闻评分记为 0，最近邻用户个数会对推荐结果产生影响，因此考虑的最近邻用户个数包括(5，10，15，20)4 种。将每种算法中不同情况的实验结果进行比较，选取最优的结果作为每种算法的实验数据。3 种算法在不同推荐结果数上的 F 值对比如表 3 所列。

表 3 不同算法的 F 值对比情况

算法	$N=10$	$N=15$	$N=20$	$N=25$	$N=30$
User-Based CF	$F=0.168$	$F=0.177$	$F=0.171$	$F=0.163$	$F=0.152$
基于 UCN-UIM	$F=0.195$	$F=0.214$	$F=0.220$	$F=0.191$	$F=0.183$
NCGP	$F=0.217$	$F=0.235$	$F=0.241$	$F=0.224$	$F=0.209$

图 5 展示了表 3 中这 3 种算法在 F 值上的变化趋势，F 值都表现为先高后低。F 值是综合准确率和召回率的一种评价指标，其变化趋势与算法中准确率和召回率的变化速率有关，具体内涵还需深入思考。在数据方面，F 值由高到低依次为 NCGP 算法、基于 UCN-UIM 的推荐算法和基于用户的协同过滤推荐算法，当推荐结果数在 15～25 之间时推荐效果更好，在此期间 NCGP 算法的 F 值比基于 UCN-UIM 的推荐算法平均提升了 2.5%，比基于用户的协同过滤推荐算法平均提升了 6.3%。

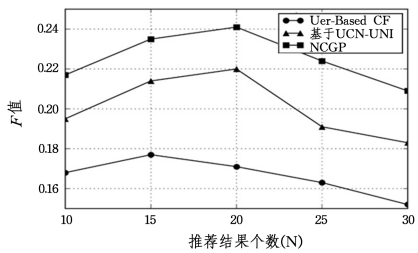

图 5 不同算法的 F 值比较

综上可知，考虑新闻事件地理位置信息后的推荐效果优于仅考虑新闻文本特性以及传统的协同过滤算法，但与基于 UCN-UIM 的推荐算法相比性能提升不高，其原因可能为新闻事件发生地抽取算法仍存在一定的误差，同时推荐系统还受其他因素(如时间、用户特性、用户地点等)的影响。

结束语 本文的主要贡献如下：1)设计了新闻事件发生地提取算法，并测试得出该算法的准确率为 93.6%，具有较高的准确性；2)考虑了新闻事件发生地，构建了 ULN-UIM 和 UCN-UIM 混合模型作为用户兴趣模型。

本文针对新闻事件发生地会对新闻推荐系统的推荐性能产生影响的情况，顾及新闻事件地理位置，提出了 NCGP 推荐算法。实验将本文提出的 NCGP 推荐算法与传统的基于用户协同过滤推荐算法、基于 UCN-UIM 的推荐算法相比较，结果表明在 F 值上本文提出的方法更优，提升了推荐性能。总体来看，本文提出的 NCGP 推荐算法为构建顾及新闻事件地理位置的新闻推荐系统提供了一种思路，但实际上推荐结果还受用户上下文情境的影响，因此下一步应当增加用户所处情境的因素来进一步研究该推荐问题。

参 考 文 献

[1] RESNICK P，IACOVOU N，SUCHAK M，et al. GroupLens：an open architecture for collaborative filtering of netnews[C]// ACM Conference on Computer Supported Cooperative Work. ACM，1994：175-186.

[2] DAS A S，DATAR M，GARG A，et al. Google news personaliza-

tion: scalable online collaborative filtering[C]// International Conference on World Wide Web. ACM, 2007: 271-280.

[3] BILLSUS D, PAZZANI M J, CHEN J. A learning agent for wireless news access[C]// Proceedings of the 5th International Conference on Intelligent user Interfaces. 2000: 33-36.

[4] IVÁN C, CASTELLS P. Ontology-Based Personalised and Context-Aware Recommendations of News Items[C]// Ieee/wic/acm International Conference on Web Intelligence and Intelligent Agent Technology. IEEE Computer Society, 2008: 562-565.

[5] IJNTEMA W, GOOSSEN F, FRASINCAR F, et al. Ontology-based news recommendation[C]// Edbt/icdt Workshops. ACM, 2010: 16.

[6] CANTADOR I, BELLOGÍN A, CASTELLS P. A multilayer ontology-based hybrid recommendation model[J]. Ai Communications, 2008, 21(2-3): 203-210.

[7] 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.

[8] 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.

[9] 陶永才，李俊艳，石磊，等. 基于地理位置的个性化新闻混合推荐研究[J]. 小型微型计算机系统，2016，37(5)：943-947.

[10] SON J W, KIM A Y, PARK S B. A location-based news article recommendation with explicit localized semantic analysis[C]// International ACM SIGIR Conference on Reserach and Development in Information Retrieval. ACM, 2013: 293-302.

[11] YOON H G, SONG H J, PARK S B, et al. A personalized news recommendation using user location and news contents[J]. Applied Mathematics & Information Sciences, 2015, 9(2): 439-449.

[12] 张晓艳，王挺，陈火旺. 命名实体识别研究[J]. 计算机科学，2005，32(4)：44-48.

[13] 郭喜跃，何婷婷. 信息抽取研究综述[J]. 计算机科学，2015，42(2)：14-17.

[14] 俞鸿魁，张华平，刘群，等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报，2006，27(2)：87-94.

[15] 鞠久朋，张伟伟，宁建军，等. CRF 与规则相结合的地理空间命名实体识别[J]. 计算机工程，2011，37(7)：210-212.

[16] 姚清耘. 基于向量空间模型的中文文本聚类方法的研究[D]. 上海：上海交通大学，2008：27.

[17] 李佳珊. 个性化新闻推荐引擎中新闻分组聚类技术的研究与实现[D]. 北京：北京邮电大学，2013：20-29.

[18] STEINBACH M, KARYPIS G, KUMAR V. A comparison of document clustering techniques[C]// World Text Mining Conference. 2000.

[19] KRISHNA B S V, PROFESSOR S, ENGINEERING M C O, et al. Comparative study of K-means and Bisecting k-means techniques in wordnet based on document clustering[J]. Human Movement, 2012, 13(2): 127-131.

[20] 项亮. 推荐系统实践[M]. 北京：人民邮电出版社，2012：44-59.

（上接第 461 页）

[4] ZHANG J M, SHEN Y X. Review on spectral methods for clustering[C]// Control Conference. IEEE, 2015: 3791-3796.

[5] CHE W F, FENG G C. Spectral clustering: A semi-supervised approach[J]. Neuro Computing, 2012, 77(1): 119-228.

[6] ZHAO Y C, ZHANG S C. Generalized Dimension-Reduction Framework for Recent-Biased Time Series Analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(2): 231-244.

[7] LANGONE R, MALL R, ALZATE C, et al. Kernel Spectral Clustering and Applications[M]// Unsupervised Learning Algorithms. Springer International Publishing, 2016.

[8] 李瑞琳，赵永华，黄小磊. 一种基于 MPI 的稀疏化局部尺度并行谱聚类算法的研究与实现[J]. 计算机工程与科学，2016，38(5)：839-847.

[9] LIU G, LIN Z, YAN S, et al. Robust recovery of subspace structures by low-rank representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 171-184.

[10] ELHAMIFAR E, VIDAL R. Sparse subspace clustering [C]// CVPR. 2009: 2790-2797.

[11] LU C Y, MIN H, ZHAO Z Q, et al. Robust and efficient subspace segmentation via least squares regression[C]// ECCV. 2012: 347-360.

[12] 邹小林，冯国灿. 基于正则割(Ncut)的多阈值图像分割方法[J]. 计算机工程与应用，2012，48(19)：174-178.

[13] WANG S, SISKIND J M. Image Segmentation with Ratio Cut [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2003, 25(6): 675-690.

[14] SRINIVASARAO P, SURESH K, RAVI K B. Image Segmentation using Clustering Algorithms[J]. International Journal of Computer Applications, 2015, 120: 36-38.

[15] 刘萍，黄纯万. 基于 SimRank 的作者相似度计算[J]. 情报理论与实践，2015，38(6)：109-114.

[16] ZHENG W, ZOU L, CHEN L, et al. Efficient SimRank-Based Similarity Join[J]. Acm Transactions on Database Systems, 2017, 42(3): 16.

[17] CHEN W F, FENG G C. Spectral clustering with discriminate cuts[J]. Knowledge-Based Systems, 2012, 28(7): 27-37.

[18] BOOBALAN M P, LOPEZ D, GAO X Z. Graph clustering using k-Neighbourhood Attribute Structural similarity [J]. Applied Soft Computing, 2016, 47: 216-223.

[19] ALZATE C, SUYKENS J A. Hierarchical kernel spectral clustering[J]. Neural Networks, 2012, 35(2): 21-30.

[20] 刘敏，韩宾，郭有倩. 一种改进的基于 K-means 的信息聚类算法研究[J]. 信息通信，2015(9)：35-36.

[21] FANG R, POUYANFAR S, YANG Y, et al. Computational Health Informatics in the Big Data Age: A Survey[J]. ACM Computing Surveys, 2016, 49(1): 12.

[22] ZHU X F, LI X L, ZHANG S C. Block-Row Sparse Multiview Multilabel Learning for Image Classification[J]. IEEE Transactions on Cybernetics, 2016, 46(2): 450-461.

[23] 李翠平. 一种基于 SimRank 的结点相似度计算方法：CN104933312 A[P]. 2015.

[24] GAO Y, WANG M, TAO D C, et al. 3-D object retrieval and recognition with hypergraph analysis [J]. IEEE Transactions on Image Processing a Publication of the IEEE Signal Processing Society, 2012, 21(9): 4290-4303.