基于动态参考书目推荐的英语创意写作辅助教学系统的设计与实现

王梦雪 李 俊 贾清源 费 腾
(武汉大学资源与环境科学学院 湖北 武汉 430079)

摘 要 近年来,国内不少高校英语专业开设创意写作(creative writing)课程,旨在提高学生的英语写作水平,培养学生的创想和才思。该课程中,学生将利用大量的英文文献,提高英语运用能力。然而,基于教师水平、阅读量和经验的参差不齐,参考文献的推荐不可能贴合所有学生的写作需求。因此,设计一个服务于英语创意写作课程的动态参考书目推荐辅助教学系统。该系统基于内容的推荐算法和基于物品的协同过滤算法进行推荐:分析学生写作样本中词频特性和易读性等写作风格的特征;根据用户行为挖掘文献之间的关联性。测试结果表明,该系统能够很好地满足用户需求,并且快速准确地为用户推荐参考书目。

关键词 文本推荐 在线学习 相似度计算 协同过滤

中图分类号 TP311 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.02.017

DESIGN AND IMPLEMENTATION OF AN ASSISTANT TEACHING SYSTEM FOR ENGLISH CREATIVE WRITING BASED ON DYNAMIC REFERENCE RECOMMENDATION

Wang Mengxue Li Jun Jia Qingyuan Fei Teng
(*School of Resource and Environmental Sciences*, *Wuhan University*, *Wuhan 430079*, *Hubei*, *China*)

Abstract In recent years, English creative writing courses have been taught in the class of many Chinese universities. This course aims at promoting students' English writing ability and encouraging for imagination and inspiration. In this course, students use a large number of English documents to improve their English application ability. However, due to the uneven level of teachers, reading volume and experience, the recommendation of references cannot meet the writing needs of all students. In view of this, we designed an assistant teaching system for the class of English creative writing based on dynamic reference recommendation. Content-based recommendation algorithm and item-based collaborative filtering algorithm were used in the system. We analyzed the characteristics of writing styles such as word frequency and readability in students' writing samples, and found the relevance between documents according to user behavior. Test results show that the system can meet the needs of students well and provide fast and precise recommendations.

Keywords Text recommendation Online study Similarity calculation Collaborative filtering

0 引 言

创意写作是在英美高校非常普遍的一门课程,它以一种具有想象力的、独特的又赋有诗意的方式表达作者的思想情感。近年来,国内已有几所大学开设创意写作课程。在创意写作学习过程中,学生需要阅读大量优秀的英文作品,因此,合适的参考书目尤为重要。如何满足不同用户的需求,在海量参考文献中为每个用户提供精准的、个性化的参考书目,并通过在线推荐系统进行实时推荐,是本文研究的目的所在。

推荐系统是能够为用户提供所需产品信息建议的软件工具和技术手段[1]。目前各平台采用的推荐系统算法主要是基于内容的推荐算法和基于协同过滤的推荐算法[2]。基于内容的推荐算法的主要思想是为用户推荐与他们所喜欢的产品内容相似度最高的产品[3],

收稿日期:2018-03-30。湖北省高等学校省级教学研究项目(2017012);武汉大学教学研究项目(2017JG003)。**王梦雪**,本科生,主研领域:推荐算法。**李俊**,本科生。**贾清源**,本科生。**费腾**,副教授。

对于文本相似性,可以通过提取文本特征来度量,主流的方法是利用 TF-IDF 词频统计算法提取词频特征[4]。除此之外,本文提出用易读性作为文本特征的另一个指标,其大小用 Flesch 易读性公式[5]衡量,Microsoft Word 就是应用 Flesch 公式来计算文本易读性的[6]。基于协同过滤的推荐算法是使用最广泛的推荐技术,其中基于物品的协同过滤被认为是相对稳定的算法[7-8]。通过计算待推荐产品与用户已评分过的产品间的相关性对产品进行评分预测,从而将预测评分高的产品加入推荐列表。然而,无论是基于内容的推荐算法还是协同过滤,都有自身的优点和缺陷,针对这一点,许多学者提出同时使用这两种方法以解决冷启动问题,提高精度[9-10]。

本文结合基于内容和基于产品的协同过滤推荐算法设计并实现了一个基于动态参考书目推荐的英语创意写作辅助教学系统。首先利用基于内容的推荐实时向用户推送相似文体和文风的参考文章,并通过多用户协同过滤的推荐,不断提高系统推荐的准确率。该系统不仅能应用于在线创意写作平台,还能应用于新闻、微博、商品信息、旅游文记、论文期刊等其他个性化文档推荐的项目中。

1　整体框架

图 1 展示了构建英语创意写作动态参考书目推荐在线系统的研究框架。该系统分三个模块进行构建:底层数据库模块、中层推荐算法模块和顶层的用户模块。数据库模块存储有文本特征数据和用户信息数据;推荐算法模块进行基于内容和基于协同过滤的混合推荐;用户模块用于前端交互,主要涉及账号密码、用户文章、推荐文章等的输入或输出和其他交互操作。

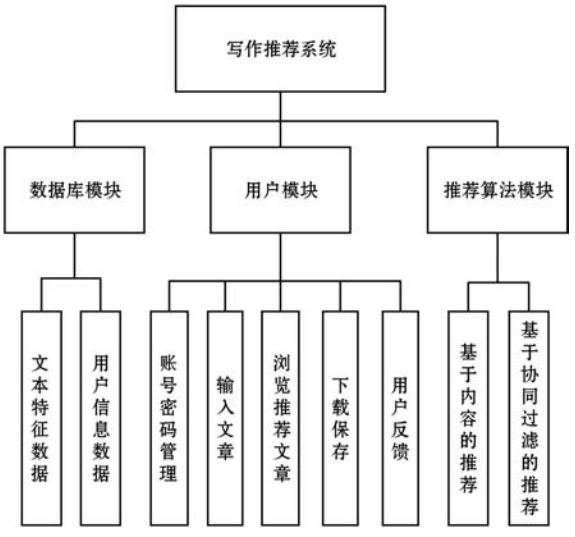

图 1　创意写作动态参考书目推荐系统的研究框架

该推荐系统的运行流程如图 2 所示。在系统投入使用之前要对文库文章进行预处理,分析所有文章的词频特征和易读性特征,将特征值存储在底层数据库中。推荐过程分为两部分:一是基于文本的内容推荐,用户输入文章片段后,在线分析该文本的词频和易读性特征并与文库中文章的特征值比较,计算二者的相似度并将结果排序,输出相似度高的文章列表;二是基于物品的协同过滤推荐,用户查看推荐的文章后,构成浏览记录,对用户的浏览记录进行分析处理,计算文章之间的支持度和置信度,基于此,判断某些文章的关联度并对关联度进行排序,输出与用户浏览记录关联度高的文章列表,作为对基于文本内容推荐的补充。

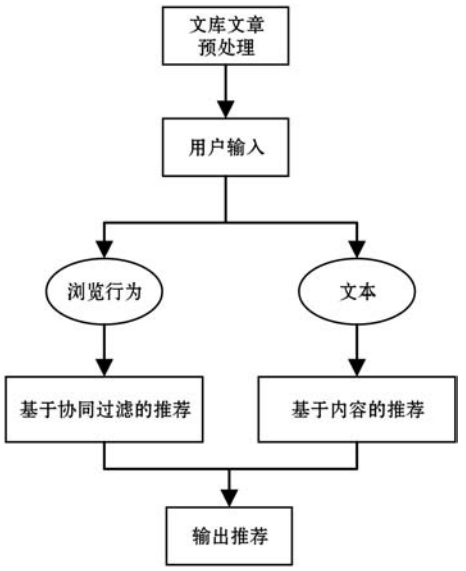

图 2　创意写作动态参考书目推荐系统的运行流程

2　基于内容的推荐

基于内容的推荐需要计算用户内容和产品内容之间的相似度,在大多数情况下需要对描述内容的信息进行分析,其中对用户兴趣的描述来自用户自己提供的信息[11]。本文在基于内容的推荐中,从特征词频相似性方面判断用户提供的片段与文库文章是否相似,并辅佐以易读性差异计算进一步衡量相似度。本文先用 TF-IDF 算法计算输入片段的词频特征,然后计算输入片段和文库文章排名较高的前 100 个词的 TF-IDF 值的余弦相似度,再计算输入片段的易读性,将它与文库文章的易读性进行差值计算,最后对二者的计算结果加权,得到最终的结果。

2.1　相似度计算

根据空间向量模型,一个文件空间中的文件可以看作一组特征值的集合,统计每个特征词的出现频率。将词频信息表示成向量模式,该向量就是文本的特征

向量,进而可以利用向量间的余弦相似度计算或者 Jaccard 公式得到文本相似度。

选取特征词最常用的方法是 TF-IDF 算法。TF-IDF 通过统计文件中每个单词在该文件的出现频率和在所有文件中的出现频率,给该文件中每个单词赋权值,TF 指词频,IDF 指逆向文件频率[12]。TF-IDF 衡量的是给定单词与一篇特定文件的相关性,若一个单词的 TF-IDF 值高,那么该单词在一个特定文件中出现频率高而在该文件集中出现频率相对低,说明该单词具有很好的类别区分能力[13],将它们作为标识该文件的特征词。这样做的目的是找到衡量文章内容相似性的可靠依据,一篇英语文章中无意义的介词出现频率一般会比具有实际含义的动词或名词高。如果单纯按一篇文章的最高词频计算,那么所有文章的特征词都会充斥着大量的介词、冠词、连词,甚至是无意义的名词和动词。而 TF-IDF 算法的作用则能降低停止词的权值,提高实义词的权值,筛选出一篇文章中独有且出现次数多的单词,提高相似性计算的准确率。

TF-IDF 算法计算过程如下:设一个文件集中有 N 个文本文件,f_{ij}为标识为 i 的单词在文件 j 中的出现频次,那么词频 TF_{ij}定义如下:

$$TF_{ij}=\frac{f_{ij}}{\max\limits_{k} f_{kj}} \tag{1}$$

TF_{ij}是f_{ij}标准化得到的结果,标准化过程是f_{ij}除以一个文本文件中所有单词的最大频率值。所以,文件 j 中出现次数最多的单词的 TF 值为 1,其他单词的 TF 值都小于 1。

设单词 i 在 n_i 个文件中出现过,那么 IDF_i 定义如下:

$$IDF_i=\log_2\frac{N}{n_i} \tag{2}$$

若一个文件 j 有 m 个不同的项,那么该文件的内容可以表示为 m 维向量:

$$\boldsymbol{d}_j=(w_{1j},w_{2j},\cdots,w_{mj}) \tag{3}$$

式中:

$$w_{mj}=\frac{f_{ij}}{\max\limits_{k} f_{kj}}\log_2\frac{N}{n_i} \tag{4}$$

$\boldsymbol{d}_j$ 即为文件的特征向量,用这个值来计算文件之间的相似度。在基于内容的推荐系统中,用 $Content\text{-}BasedProfile(c)$表示用户特征,用 $Content(s)$表示产品特征[14],有如下所示计算该相似性的函数:

$$u(c,s)=score(ContentBasedProfile(c),Content(s)) \tag{5}$$

其中 $score$ 的计算方式有很多,本文使用夹角余弦相似法,因为该方法计算简便,且能够得到较为精确的结果。该方法是用向量空间中两个向量夹角的余弦值衡量两个对象之间的相似度,计算方法如下:

$$u(c,s)=\cos(w_c,w_s)=\frac{\sum\limits_{I=1}^{K} w_{i,c}w_{i,s}}{\sqrt{\sum\limits_{I=1}^{K} w_{i,c}^2}\sqrt{\sum\limits_{I=1}^{K} w_{i,s}^2}} \tag{6}$$

两个特征向量的夹角余弦值越大,向量之间的夹角就越小,说明两个文本文件越相似。

本文先统计文章的单词词频,取频次最高的若干个单词,然后用 TF-IDF 算法从中筛选出 100 个能标识该文章的特征单词。将这些单词的 TF-IDF 值作为文章的特征向量,计算出文库文章和输入语句特征向量的夹角余弦值,得到的结果即为二者的词频相似度,作为评价文档相似性的一个指标。

2.2 易读性差异计算

易读性用来衡量文章难度,本文将它作为另一个文本特征,使用 Flesch 公式计算文本的易读性。该公式用单词音节数衡量单词难度,用文本的平均句长衡量句子的难度。

Flesch 易读性公式形式如下:

$$Reading\ Ease(RE)=206.835-0.846wl-1.015sl \tag{7}$$

式中:wl 为每 100 个单词的平均音节数;sl 为句子的平均单词数;RE 代表易读性指数,范围为 0 ~ 100。RE 值越大,文本越容易,RE 值在 0 ~ 30 被认为很难,是美国大学生水平,60 ~ 70 被定义为标准难度,相当于初中生水平。

本文对文库文章和输入语句的单词平均音节数和句子平均单词数进行统计,用 Flesch 公式计算出用户输入和文库文章的易读性差值,作为评价文档相似性的另一个指标。

2.3 基于文本内容的推荐

通过对文本信息的分析计算,得到词频相似度和文章易读性两个指标。本文在决定最终的计算公式时,采用熵权法[15]确定这两个指标的权重系数。

设词频特征相似度结果的权重系数为 a,易读性差异计算结果的权重系数为 b,可得到用户输入片段和文库文章的相似度的计算公式:

$$sim(c,s)=a\times u(c,s)+b\times|RE_c-RE_s|=a\times\frac{\sum\limits_{I=1}^{K} w_{i,c}w_{i,s}}{\sqrt{\sum\limits_{I=1}^{K} w_{i,c}^2}\sqrt{\sum\limits_{I=1}^{K} w_{i,s}^2}}+b\times|RE_c-RE_s| \tag{8}$$

式中:$sim(c,s)$表示用户输入片段与文库文章的相似

度,将相似度结果从大到小排序,优先推荐相似度高的文章。

3　基于协同过滤的推荐

在基于物品的协同过滤推荐中,分析每次推荐后产生的用户喜好数据,如果多个用户同时看了某些文章,可以判断这些文章存在着隐含的联系。据此将与用户阅读过的文章关联性强的文章推荐给该用户,帮助其当前的写作,作为对基于内容推荐结果的补充。基于物品的协同过滤的理论之一是数据挖掘中的关联规则,用支持度(support)和置信度(confidence)来反映两个物品之间的关联度,支持度表示两个物品同时出现的概率。

本文采用隐式评分[16],即不需要用户显式输入评分数值,仅通过用户在浏览推荐结果片段后是否点击“阅读全文”来判断用户是否对该文章感兴趣。若判断为是,则将该文章加入到该用户的阅读列表,文库列表中的每一篇文章和其他文章的关联性都要在阅读列表中进行统计计算。对于任意两篇属于文库列表的文章 A 和 B,它们之间的支持度为:

$$s_{\text{support}} = P(A \cup B) = \frac{\text{同时出现 } A \text{ 和 } B \text{ 的文章个数}}{\text{文章个数}} \quad (9)$$

A 对 B 的置信度表示如果用户阅读过 A,他也会喜欢 B 的概率,公式为:

$$c_{\text{onfidence}}(A \rightarrow B) = P(B|A) = \frac{\text{同时出现 } A \text{ 和 } B \text{ 的文章个数}}{\text{出现 } A \text{ 的文章个数}} \quad (10)$$

给支持度和置信度设置阈值,若 A 对 B 的支持度和 A 对 B 的置信度分别大于这两个阈值,则判断 B 是 A 的强相关性文章,将 B 添加到 A 的相关文章列表中。遍历文库中的所有文章,为每篇文章都建立对应的相关文章列表。若用户阅读文章 A,则从列表中筛选出用户没有读过的文章 B,按照关联度从大到小排序,将排序结果推送给用户。

4　测试分析及改进

本文设计了一种个性化在线推荐系统平台,系统界面如图 3 所示。用户在左侧文本框输入写作片段,系统根据目前的内容在左下角实时呈现 5 篇推荐书目的标题,当输入文字较多后推荐列表会趋于稳定。用户点击后以片段方式呈现在右侧文本框供用户试阅,当用户对此文章感兴趣可以点击“full text”阅读全文。此时,系统会在右下角会列出与该文章相关度最高的 5 篇文章,并将浏览行为记录下来,用于计算更新各文章的相关文章列表。

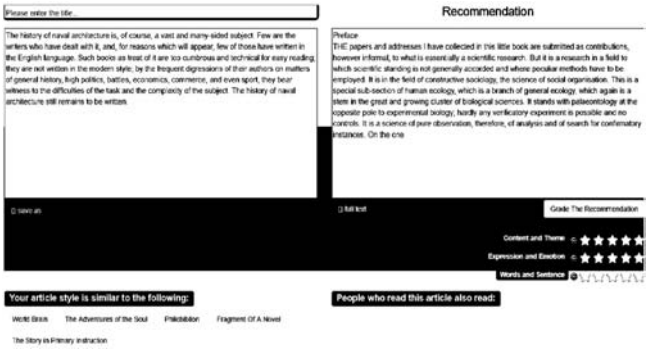

图 3　创意写作动态参考书目推荐系统界面

本文从各网站采集了 580 短篇英文小说作为实验测试数据,测试平台中设计了评分系统,选择 100 名学生作为用户分别从内容和主题、表达和情感、用词和句长三个方面进行评价,每个方面评分为 1 ~ 5 分。用户综合推荐的 5 篇文章对结果进行评分,对评分结果求平均值并进行归一化,得到用户对推荐结果的满意程度,结果如表 1 所示。

表 1　用户满意程度及评分标准

评分标准	满意程度
内容和主题	75.90%
表达和情感	77.44%
用词和句长	84.44%

由用户的反馈可以看出,本系统能够较好地满足用户的需求。虽然在情感表达和主题内容上稍有欠缺,但是在词汇和句子难度方面系统能提供较为精准的推荐。

本系统与已有的教学资源推荐系统相比[17-18],优势在于使用基于内容和基于物品的混合式推荐系统,发挥了两种推荐方法各自的优点。在基于内容的推荐中,分别从文章相似度和易读性两个角度出发进行相似计算,从而提高了推荐的准确性。

5　结　语

本文提出了一个基于动态参考书目推荐的英语创意写作辅助教学系统。该系统能根据用户在线输入的英文写作内容,提取多维写作风格特征,进行实时动态相关参考读物推荐。此外,作为一个多用户系统,还基于协同过滤算法,将其他用户的接受推荐行为也纳入推荐考虑范围,利用用户贡献内容(UGC)对系统的贡献,对其基于写作风格的推荐进行补充和修正。该系统在使用中,反应迅速、推荐准确,深受测试用户的好

评。可以作为创意写作课程教学与课后辅导信息化的有利工具。

在之后的改进中可以考虑更多方面,如利用自然语言处理进行情感分析,通过文章主题分类提高推荐效果,使推荐系统更符合用户的预期。

参 考 文 献

[1] 尹书华,傅城州. 基于百科大数据的旅游景点推荐系统应用研究[J]. 旅游论坛,2017,10(3):107 - 115.

[2] 刘红霞. 基于协同过滤技术的推荐系统综述[J]. 信息安全与技术,2016,7(3):24 - 26,48.

[3] 单京晶. 基于内容的个性化推荐系统研究[D]. 长春:东北师范大学,2015.

[4] 周源,刘怀兰,杜朋朋,等. 基于改进 TF-IDF 特征提取的文本分类模型研究[J]. 情报科学,2017,35(5):111 - 118.

[5] Börstler J, Caspersen M E, Nordström M. Beauty and the Beast: on the readability of object-oriented example programs [J]. Software Quality Journal, 2016, 24(2): 231 - 246.

[6] 陈爱文. 文本难度对高中生英语阅读理解影响的研究[D]. 漳州:闽南师范大学,2016.

[7] 杨强,杨有,余春君. 协同过滤推荐系统研究综述[J]. 现代计算机(专业版),2015(13):3 - 6.

[8] 刘辉,郭梦梦,潘伟强. 个性化推荐系统综述[J]. 常州大学学报(自然科学版),2017,29(3):51 - 59.

[9] Balabanović M, Shoham Y. Fab: content-based, collaborative recommendation [J]. Communications of the ACM, 1997, 40(3): 66 - 72.

[10] Thorat P B, Goudar R M, Barve S. Survey on collaborative filtering, content-based filtering and hybrid recommendation system[J]. International Journal of Computer Applications, 2015, 110(4):31 - 36.

[11] Soares M, Viana P. Tuning metadata for better movie content-based recommendation systems [J]. Multimedia Tools and Applications, 2015, 74(17): 7015 - 7036.

[12] 武永亮,赵书良,李长镜,等. 基于 TF-IDF 和余弦相似度的文本分类方法[J]. 中文信息学报,2017,31(5):138 - 145.

[13] 王嘉旸,杨丽萍,闫天伟. 基于向量空间模型的文本相似度计算方法[J]. 科技广场,2017(2):9 - 13.

[14] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. IEEE transactions on knowledge and data engineering, 2005, 17(6): 734 - 749.

[15] 陆添超,康凯. 熵值法和层次分析法在权重确定中的应用[J]. 电脑编程技巧与维护,2009(22):19 - 20,53.

[16] 陈华月. 基于加权关联规则和浏览行为的个性化推荐[D]. 重庆:重庆大学,2005.

[17] 汤娟梅,唐岭. 个性化英语阅读文章推荐系统的设计[C]//第十四届全国计算机辅助教育学会年会论文集. 华南师范大学,2010:573 - 577.

[18] 杨淑枝. 基于协同过滤的协作学习活动推荐系统[C]//第十四届全国计算机辅助教育学会年会论文集. 华南师范大学,2010:627 - 633.

(上接第 18 页)

[6] 胡文,宰祥顺. 基于 BP 神经网络与隐马尔科夫模型的推荐算法[J]. 哈尔滨商业大学学报(自然科学版),2017,33(5):551 - 555.

[7] 伊华伟,张付志,巢进波. 基于模糊核聚类和支持向量机的鲁棒协同推荐算法[J]. 电子与信息学报,2017,39(8):1942 - 1949.

[8] Wang H, Wang N, Yeung D Y. Collaborative Deep Learning for Recommender Systems[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2015:1235 - 1244.

[9] Elkahky A M, Song Y, He X. A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems[C]//Proceedings of the 24th International Conference on World Wide Web. ACM,2015:278 - 288.

[10] 杜丹琪,周凤. 基于 TimeRBM 和项目属性聚类的混合协同过滤算法[J]. 计算机应用研究,2018,35(2):349 - 353.

[11] 刘静,武文琪,李骁,等. 基于用户兴趣和项目属性的协同过滤算法[J]. 计算机应用与软件,2017,34(5):33 - 37.

[12] 宫志晨,李学俊,张晶晶,等. 基于多向测度和属性相似度的混合协同过滤[J]. 计算机应用与软件,2015,32(6):62 - 65.

[13] 胡健,覃慧,梁雪雷. 基于用户量化属性的多维相似度的协同过滤推荐算法[J]. 江西理工大学学报,2017,38(3):86 - 91.

[14] 王三虎,王丰锦. 融合用户评分和属性相似度的协同过滤推荐算法[J]. 计算机应用与软件,2017,34(4):305 - 308,321.

[15] Kim B M, Li Q, Park C S, et al. A new approach for combining content-based and collaborative filters[J]. Journal of Intelligent Information Systems, 2006, 27(1):79 - 91.

[16] Jiang X Y, Li S. BAS: beetle antennae search algorithm for optimization problems[EB]. arXiv preprint arXiv:1710.0724, 2017.

[17] Jiang X Y, Li S. Beetle antennae search without parameter tuning(BAS-WPT) for multi-objective optimization[EB]. arXiv preprint arXiv:1711.02395,2017.

[18] 梁丽君, 李业刚, 张娜娜,等. 融合用户特征优化聚类的协同过滤算法[J/OL]. 智能系统学报,2018: 1 - 7[2018 - 05 - 11]. http://ns.nki.net/cms/detail/23.1538.tp.20180423.1526.014.html.