基于主题模型的英语写作批阅系统个性化推荐模块设计与实现

葛 昊，叶 艳，包西林，吴 敏

（中国科学技术大学 现代教育技术中心，合肥 230026）

摘要：伴随着自然语言处理技术的发展，计算机代替人工评阅作文已成为可能，计算机评阅的便利，带来了巨大的用户量及作文量，而针对不同用户准确的推荐各自感兴趣的范文可以更加丰富及开阔自己的写作能力。本文提出了一种新的推荐算法，将概率主题模型引入到个性化推荐中，并在实验室的写作批阅平台（简称 CMET）上进行了实现。全文介绍了协同过滤推荐和主题模型的相关技术，对传统的协同过滤算法进行了改进，在此基础上，提出了新的推荐算法——基于主题模型的协同过滤（BaseLDACF），并将该推荐算法进行了实现。

关键词：主题模型；计算机应用；个性化推荐；协同过滤

中图分类号：TP391 **文献标志码**：A **文章编号**：1671－1807(2013)06－0151－05

人类步入信息时代以来，计算机技术日新月异，信息技术从各个领域影响着人类的生活。与此同时，在教育领域，随着技术的进步，计算机辅助教学和 e-Learning 也越来越流行。在这样大的环境下，CMET 顺应而生。随着系统用户量及作文量的积累，海量的作文，仅仅依靠用户自身搜索不能够全面有效地获取适合的有用范文。

写作批阅平台的作文量的积累，为用户提供了充足的范文库，但也为用户带来了“信息过载”的问题。用户从充足的范文库中获取自己感兴趣或者对自身有用的范文变的越发困难。CMET 运用了本文的个性化推荐技术，从传统的单向主动模式转变为双向主动模式，以用户为中心，主动分析用户的需求，并尽可能的满足不同用户的需求，最终将范文推送给用户，即“用户需要什么，我就提供什么”，节约了用户大量的时间去寻找范文。

本文提出并实现了 BaseLDACF，针对不同用户进行个性化的范文推荐。

1 相关推荐技术

推荐技术大致分为基于内容（BaseContent）、基于用户的协同过滤（BaseUserCF）和基于物品的协同过滤（BaseItemCF）三种[1]。BaseLDACF 大体就是将 BaseContent 和协同过滤算法相融合。通过分析用户的历史行为为各个用户构建用户兴趣模型，并以此为依据，采用 BaseLDACF 算法预测用户可能感兴趣的范文。

1.1 BaseContent 推荐

基于内容的推荐是信息过滤技术的延续与发展，它是建立在物品的内容信息上作出推荐的，而不需要依据用户对物品的评价意见。其主要思想是，根据物品内容，结合用户的历史行为对用户进行兴趣建模，并通过物品匹配计算物品与目标用户兴趣间的相似度，将相似度排名最高的物品推荐给目标用户。

1.2 BaseUserCF 推荐

BaseUserCF 算法的诞生标志着推荐系统的诞生，它一般采用最近邻技术，利用用户的历史喜好信息计算用户之间的距离，然后利用目标用户的最近邻居用户推荐。该算法是基于这样的假设：为一用户找到他真正感兴趣的内容好方法是首先找到与此用户有相似兴趣的其他用户，然后将它们感兴趣的内容推荐给此用户。

从上面的描述及图 1 可知，BaseUserCF 算法主

收稿日期：2013－03－29

作者简介：葛昊（1989—），男，河南范县人，中国科学技术大学，硕士研究生，研究方向：教育软件工程；叶艳（1978—），女，安徽颍上人，中国科学技术大学，工程师，硕士，研究方向：教育软件工程；包西林（1989—），男，安徽池州人，中国科学技术大学，硕士研究生，研究方向：教育软件工程；吴敏（1962—），男，安徽蚌埠人，中国科学技术大学，教授，硕士，研究方向：教育软件工程。

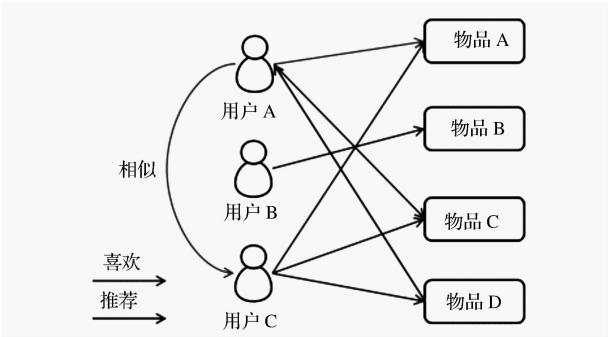

图 1　基于用户协同过滤推荐

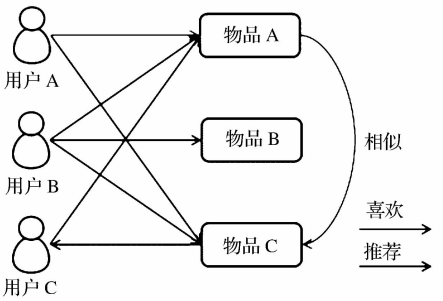

图 2　基于物品协同过滤推荐

要包括两个步骤：

1）找到和目标用户兴趣相似的用户集合；

2）找到这个集合中的用户喜欢的，且目标用户没有操作过的物品推荐给目标用户。

我们可以通过如下的余弦相似度计算出步骤 1）所需的用户之间的相似度：

$$w_{uv}=\frac{|N(u)\cap N(v)|}{\sqrt{|N(u)\parallel N(v)|}} \tag{1}$$

u 和 v 表示用户，$N(k)$，$k\in\{u,v\}$为用户 k 曾经有过正反馈的物品集合。

有了与目标用户相似的用户集合后，我们便可以通过如下公式度量用户 u 对物品 i 的感兴趣程度：

$$p(u,i)=\sum_{v\in S(u,K)\cap N(i)} w_{uv}r_{vi} \tag{2}$$

其中，$S(u,K)$包含和用户 u 兴趣最接近的 K 个用户，$N(i)$是对物品 i 有过行为的用户集合，w_{uv}是用户 u 和用户 v 的兴趣相似度，r_{vi}代表用户 v 对物品 i 的兴趣。

1.3　BaseItemCF 推荐

BaseItemCF 算法给用户推荐那些和他们之前喜欢的物品相似的物品。不过，BaseItemCF 算法并不利用物品的内容属性计算物品之间的相似度，它主要通过分析用户的行为记录计算物品之间的相似度。该算法认为，物品 A 和物品 B 具有很大的相似度是因为喜欢物品 A 的用户大都也喜欢物品 B[3]。如图 2 所示。

BaseItemCF 算法主要分为两步：

1）计算物品之间的相似度；

2）根据物品的相似度和用户的历史行为给用户生成推荐列表。

为了避免推荐出热门的物品，有如下的公式：

$$w_{ij}=\frac{|N(i)\cap N(j)|}{\sqrt{|N(i)\parallel N(j)|}} \tag{3}$$

$|N(i)|$是喜欢物品 i 的用户数，$|N(i)\cap N(j)|$是同时喜欢物品 i 和物品 j 的用户数。公式(3)惩罚了物品 j 的权重，因此减轻了热门物品会和很多物品相似的可能性[4]。

2　主题模型

主题模型是用来判断两个文档是否相似。其目的是尽可能地判断出两个文档在语义方面的相关性，而不是像传统的相似性判断算法，如 TF-IDF 等，是通过查看两个文档共同出现的单词的多少来判断的。但是传统的方法没有考虑到文字背后的语义关联，可能在两个文档共同出现的单词很少甚至没有，但两个文档是相似的。

以两个句子为例，第一个是“最近流感比较厉害”，第二个是“板蓝根可能要涨价”。

如果由人来判断，我们一看就知道，这两个句子之间重复的词语虽然没有，但仍然是很相关的。如果按传统的方法——只取决于字面上的词语重复，来判断这两个句子的相关性，二者肯定不相似，因此在判断文档相关性的时候需要考虑到文档的语义，而语义挖掘的利器是主题模型。

2.1　主题模型是什么

主题模型，顾名思义，就是对文字中隐含主题的一种建模方法。在主题模型中，主题就是一个概念、一个方面，它表现为一系列相关的词语的条件概率。形象来说，一个主题就好像是一个桶，里面装了出现概率较高的词语，正是这些词语共同定义了这个主题[5]。

主题模型作为一种生成模型，对一篇文档产生的过程进行了建模。其基本思想是，一个人在写一篇文档的时候，会首先想这篇文章要讨论哪些话题，然后思考这些话题应该用什么词描述，从而最终用词写成一篇文章。可知，要生成一篇文档，它里面的每个词语出现的概率为：

$$p(\text{词语}|\text{文档})=\sum_{\text{主题}} p(\text{词语}|\text{主题})\times p(\text{主题}|\text{文档}) \tag{4}$$

式(4)中 p(词语|文档)表示每篇文档中每个词

语出现的概率；p(词语|主题)表示每个主题中每个词语出现的概率；p(主题|文档)表示每篇文档中各个主题出现的概率。三者分别以“词语－文档”、“词语－主题”和“主题－文档”三个矩阵进行表示[6]。

给定一系列文档，通过对文档进行分词，计算各个文档中每个词语的词频就可以得到“词语－文档”矩阵 C。主题模型就是通过矩阵 C 进行训练，学习出“词语－主题”和“主题－文档”矩阵。主题模型训练学习的方法代表性的有基于 LDA 的概率模型。

2.2 LDA 模型

LDA 模型是一个产生式 3 层概率模型，有 3 种元素，即文档、主题和词语。每一篇文档都会表现为词的集合，这称为词袋模型。每个词在一篇文章中属于一个主题。LDA 通俗的来说：假想一个人，他现在要写 m 篇文章，一共涉及了 K 个主题，每个主题下的词分布为一个从参数为 $\vec{\beta}$ 的 Dirichlet 先验分布中抽样出来的多项式分布。对于每篇文章，他首先会从一个泊松分布中抽样一个值作为文章长度，再从一个参数为 $\vec{\alpha}$ 的 Dirichlet 先验分布中抽样出一个多项式分布作为该文章里面出现每个主题下词的概率；当他想写某篇文章中的第 n 个词的时候，首先从该文章中出现每个主题下词的多项式分布中抽样一个主题，然后再在这个主题对应的词的多项式分布中抽样一个词作为他要写的词。不断重复这个随机生成过程，直到他把 m 篇文章全部写完。概率图模型表示如下图 3 所示。

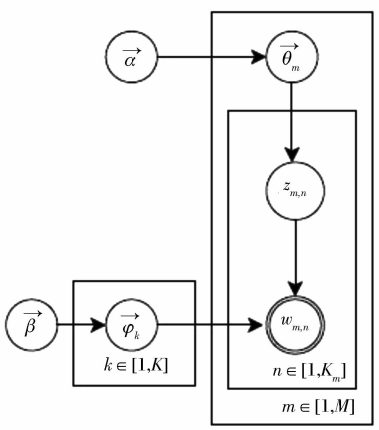

图 3 LDA 概率图模型

本文采用了 Gibbs Sample 算法学习 LDA，该算法的运行方式是每次选取概率向量的一个维度，给定其他维度的变量值抽样当前维度的值，不断迭代，直到收敛输出待估计的参数。得到模型参数 $\vec{\theta}_m$ 和 $\vec{\varphi}_k$ 的计算公式如下：

$$\varphi_{ki}=\frac{n_k^{(1)}+\beta_i}{\sum_{i=1}^{v} n_k^{(i)}+\beta_i} \tag{5}$$

$$\theta_{mk}=\frac{n_m^{(k)}+\alpha_k}{\sum_{k=1}^{k} n_m^{(k)}+\alpha_k} \tag{6}$$

上式中，φ_{ki} 是当前词 i 在主题 k 上的概率，θ_{mk} 是文档 m 包含主题 k 的概率；α_k 是文档 m 下主题 k 的多项分布的 Dirichlet 先验参数，β_i 是主题 k 下词 i 的多项分布的 Dirichlet 先验参数；$n_k^{(i)}$ 是词 i 在主题 k 上出现的个数，$n_m^{(k)}$ 是文档 d 中出现主题 k 的个数。

2.3 LDA 的改进

综上所述，可以看出将一个词分配到一个主题中受到词在主题中的概率和词所在的文档拥有的主题概率二者影响，且由于高频词在主题中和文档中占有的比例都较大，导致主题的分配向高频词的主题倾斜。因此，本文对传统的 LDA 模型进行了改进，为文档单词赋予了权重，在 LDA 的初始化阶段，通过加上各个单词的权重 weight 代替加 1 进行改进。伪代码如下：

```
foreach document i in range(0,|D|):
foreach word j in range(0,|D(i)|):
z[i][j] = rand() % K
NZD[z[i][j], D[i]] += weight[i][j]
NWZ[w[i][j], z[i][j]] += weight[i][j]
NZ[z[i][j]] += weight[i][j]
```

D 为文档集合，z[i][j]是文档随机被赋予的主题，NWZ(w,z)记录词 w 被赋予主题 z 的次数，NZD(z,d)记录文档 d 中被赋予主题 z 的词的数量。

计算单词的权重采用了类似于 TFIDF 的主要思想，即权重 weight = TFC * IDF，这里的 TFC 表示词条 t 在文档 d 的个数除以文档 d 中居中的词条个数，IDF 则是如果包含词条 t 的文档越少，IDF 越大。

3 基于主题模型的协同过滤

3.1 改进 BaseUserCF 算法

传统的 BaseUserCF 算法并没有考虑到物品的热门程度，我们知道，热门的物品相对于冷门的物品对计算用户之间相似度的影响较之弱，就是说两个用户对冷门物品采取过同样的行为更能说明他们兴趣的相似度。因此，可将式(1)做以下改进：

$$w_{uv}=\frac{\sum_{i\in N(u)\cap N(v)}\frac{1}{\log 1+|N(i)|}}{\sqrt{|N(u)\|N(v)|}} \tag{7}$$

公式(7)通过 $\frac{1}{\log 1+|N(i)|}$ 惩罚了用户 u 和用户 v 共同兴趣列表中热门物品对他们相似度的影响。

而公式(2)仅仅通过将用户兴趣相似度 w_{uv} 与用

户 v 对物品 i 的兴趣 r_{vi} 相乘求和来度量用户 u 对物品 i 的感兴趣程度，并未考虑到用户 u 对物品 i 的兴趣值 r_{ui} 与 r_{vi} 之间的差别所带来的影响。如：假设用户对物品的最高兴趣值为 5，当用户 v 对物品 i 的兴趣值 $r_{vi}=4$ 且用户 v 对所有物品的平均兴趣值为 3，而用户 u 对物品 i 的兴趣值 r_{ui} 只有 2 并且用户 u 对所有物品的平均兴趣值为 4，此时，若通过式(2)简单的将 w_{uv} 和 r_{vi} 相乘求和，求出的用户 u 对物品 i 的感兴趣程度便会出现较大误差，为了消减这一问题所带来的误差，本文引入了修正因子 ModRate，该因子弥补了兴趣值差异所带来的问题，式(8)便是 ModRate 的公式：

$$\text{ModRate}\begin{cases} \text{fk}-[\text{fk}],\text{fk}=\dfrac{r_{uix}/\bar{r}_u}{r_{vi}/\bar{r}_v} \text{ 且 } \text{fk}>1 \\ \text{fk}-[\text{fk}],\text{fk}=\dfrac{r_{vi}/\bar{r}_v}{r_{uix}/\bar{r}_u} \text{ 且 } \text{fk}>1 \end{cases} \tag{8}$$

其中 r_{uix} 是用户 u 对 i 的预测兴趣值，$\bar{r}_k$ 是用户 k 平均兴趣值的表示。

将 ModRate 引入式(2)，即得最终的 BaseUserCF 改进式(9)：

$$p(u,i)=\sum_{v\in S(u,K)\cap N(i)} w_{uv}(r_{vi}+\text{ModRate}) \tag{9}$$

3.2　BaseLDACF 算法

对改进的 BaseUserCF 算法，我们可从式(9)中充分理解，不难发现，式(9)在计算用户 u 对物品 i 的感兴趣程度时，以用户 u、v 之间的相似度 w_{uv} 作为桥梁，而物品 i 与用户 u 之间相似度重要性却被忽视了。举个例子来说，用户 u 对物品 m，n，k 有过操作，通过与 u 相似的用户集而得到的物品 i，j 等，且通过式(9)计算得出的推荐 i 的可能性小于推荐 j，但 i 与 u 操作过的物品的相似度远高于 j 与 u 操作过的物品的相似度，依常理判断知，此时用户希望得到的推荐是 i 甚过于 j，由此，便是式(9)所会带来的问题。

为了解决上述问题，本文通过将主题模型融入到改进的 BaseUserCF 中，为式(9)添加新的变量 sim，该变量是通过主题模型所得到的主题－文档概率集，使用 JS 散度(Jensen-Shannon Divergence)来计算待推荐物品与目标用户操作过的物品集的相似度集合，取出最小的相似值作为 sim 值。通过上述分析，BaseLDACF 算法可描述如下式(10)：

$$p(u,i)=\sum_{v\in S(u,K)\cap N(i)}(w_{uv}+\text{sim})*(r_{vi}+\text{ModRate}) \tag{10}$$

4　推荐模块的实现

在 CMET 系统中推荐模块分为两种推荐引擎，一个是给用户登录系统后在首页给出推荐的，另一个是给用户写完一篇作文后作出推荐的。为方便描述，在此，我们将这两种推荐引擎分别命名为 RecEngine1 和 RecEngine2。

RecEngine1 使用了本文提出的 BaseLDACF 推荐算法，RecEngine2 使用了本文改进的 LDA 主题模型进行内容上的相似性推荐。推荐模块通过读取后台数据库中的用户行为和作文内容，基于推荐算法，分析用户的行为日志，给用户生成推荐列表，最终展示到网站的界面上。推荐模块加入到 CMET 系统中，如下简图 4：

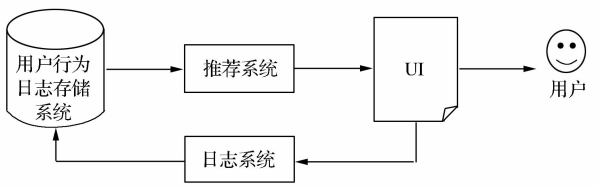

图 4　推荐模块

推荐引擎的架构图如图 5 所示。由图 5 可看出，推荐引擎架构主要包括 4 个部分：

1) 部分 A 负责从数据库或者缓存中拿到用户行为数据，通过分析不同行为，生成当前用户的特征向量。该模块的输出是用户特征向量。

2) 部分 B 负责将用户的特征向量通过特征－范文相关矩阵转化为初始推荐范文列表。

3) 部分 C 负责计算出作文间的内容相似性，此部分在不同引擎中，功能有所不同：

在 RecEngine1 中，将待测范文与用户操作过的范文输入部分 C 中，得到待测范文与用户的相似值 sim 输出给部分 B；

在 RecEngine2 中，直接是通过部分 C 计算新写作文与系统中所有范文的内容相似性，将计算结果排序输出给部分 B，此时 B 将直接根据 C 所得结果转化为初始推荐范文列表。

4) 部分 D 负责对初始的推荐列表进行过滤、排名等处理，从而生成最终的推荐列表。

根据图示 4、5 即在 CMET 系统中实现了推荐模块。

5　结束语

全文通过不断的改善传统的协同过滤算法，将主题模型与协同过滤算法相混合，提出了一种全新的推荐算法 BaseLDACF，利用 LDA 主题模型挖掘作文

的主题，引入了作文与用户的相似值 sim，结合 BaseUserCF 的公式，对用户进行了更准确的作文推荐。本文对推荐引擎 RecEngine1 推荐准确性做了大幅提升，但对 RecEngine2 在写完一篇作文后进行的推荐仅仅只依据内容上的相似度，推荐准确性及用户的满意度都差强人意，后续的工作将针对 RecEngine2 进行改进，可以通过将作文的层次结构作为推荐的影响因素、将 BaseItemCF 算法融入到 RecEngine2 中等举措，从而使推荐模块更加完善。

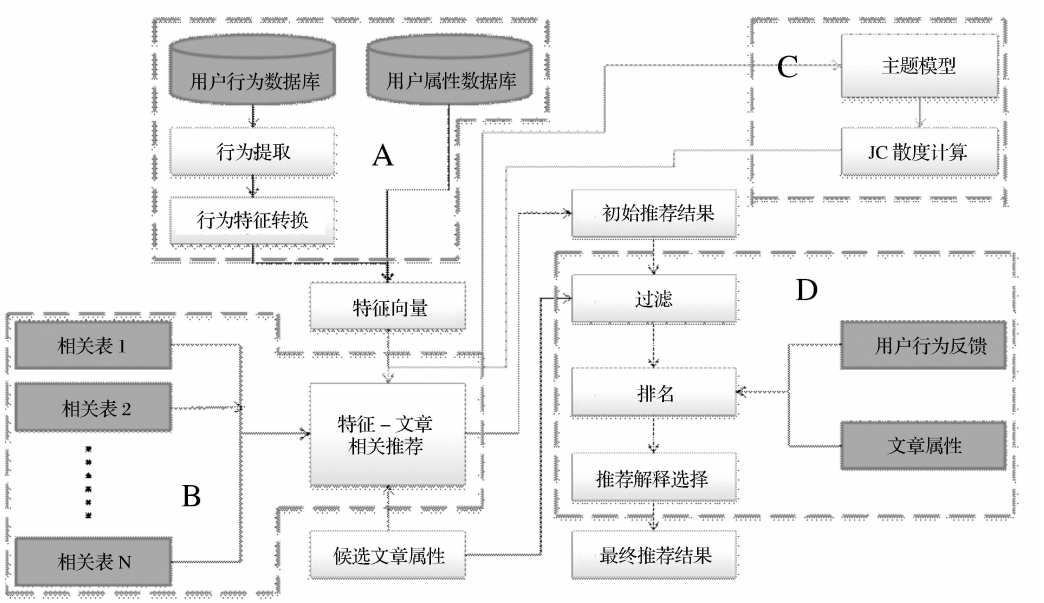

图 5　推荐引擎架构图

参考文献

[1] 刘建国，周涛，汪秉宏．个性化推荐系统的研究进展[J]．自然科学进展，2009，19(1)：1－15.

[2] J BREESE，D HEEHERMAN，C KADIE. Empirical analysis of predictive algorithms for collaborative filtering[C]//Proc of the 14th Conf on Uncertainty in Artificial Intelligence (UAI98). San Francisco，CA：Morgan Kaufmann. 1998：43－52.

[3] 邓爱林，朱扬勇，施伯乐. 基于项目评分预测的协同过滤推荐算法[J]. 软件学报，2003，14(9)：1621－1628

[4] 赵亮，胡乃静，张守志. 个性化推荐算法设计[J]. 计算机研究与发展，2002，39(8)：986－991.

[5] 徐戈，王厚峰．自然语言处理中主题模型的发展[J]．计算机学报，2011，34(8)：1423－1435.

[6] 孙玉婷．基于概率主题模型的中文话题检测与追踪研究[D]．武汉：华中科技大学，2010.

[7] BLEI D，ANDREW Y NG，JORDAN M. Latent dirichlet allocation[J]. Journal of Machine Learning Reserach，2003(3)：993－1022.

[8] MICHAL ROSEN-ZVI，TOM GRIFFITHS，MARK STEYVERS，PADHRAIC SMYTH. The author-topic model for authors and documents[G]. In UAI.

The Design and Implementation of Personal Recommendation Module in CMET Based on Topic Model

GE Hao，YE Yan，BAO Xi-lin，WU Min

(Center of Modern Educational Technology，University of Science and Technology of China，Hefei 230026，China)

Abstract：With the development of Natural Language Processing，it's achievable to review the composition by computer instead of human. The convenience of the technique also bring about large number of users and compositions，it's useful for improving the users' writing ability by recommending compositions which the user is interested in accurately. In this article，a new recommendation algorithm which uses probabilistic topic model in personal recommendation is put forward and implemented in the CMET. The article introduces the Collaborative Filtering Recommendation，the Correlated Topic Model and the improvement on the traditional Collaborative Filtering Recommendation Algorithm. Based on the improved Algorithm，the article puts forward a new recommendation algorithm：Collaborative Filtering based on LDA and implements the algorithm.

Key words：topic model；computer application；personal recommendation；collaborative filtering