文章编号:1674-2974(2017)10-0158-07 DOI:10.16339/j.cnki.hdxbzkb.2017.10.021

基于多标签分类的学术文献潜在时间意图识别研究*

沈思[1†],吴玺煜[2]
(1.南京理工大学 经济管理学院,江苏 南京 210094; 2.华南师范大学 计算机学院,广东 广州 510631)

摘 要:为了提高检索结果的时间相关性,将文本特征抽取和多标签分类算法应用于文献检索的潜在时间意图分类研究之中.从检索潜在时间意图分类的角度出发,提出一种基于文本时间信息抽取和Labeled LDA(标签主题模型)的文献潜在时间意图自动分类算法.首先,在获取的文献时间信息基础上,将文献检索潜在时间意图映射至具体时间类别.其次,为了减少时间信息的稀疏性对分类特征学习过程的影响,利用交叉学科中时间短语分布特征优化Labeled LDA分类模型的标签选择过程.最后,将所提算法与其他多标签分类算法进行对比实验,分析和评估文献检索潜在时间意图自动分类的准确率.结果表明,所提算法的AUC的值达到79.6%,较同类基准算法ECC(整体分类链)提高约10.9%,且针对不同学科均取得了较好的分类效果,是一种有效的文献检索潜在时间意图学习方法.

关键词:多标签分类;主题模型;潜在时间意图;文本特征抽取;文本分类
中图分类号:TP391.1 **文献标志码**:A

Research on Identifying Potential Temporal Intentions of Academic Literature Based on Multi-label Classification

SHEN Si[1†], WU Xiyu[2]
(1.School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China;
2. School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: In order to enhance the temporal relevance of retrieval result, the text feature extraction and algorithm of multi-label classification were applied to potential temporal intention classification of literature retrieval. From the perspective of retrieving the classification of potential temporal intentions, an algorithm was proposed to automatically classifiy potential temporal intentions of literature, based on text temporal information extraction and labeled LDA. Firstly, by use of such temporal information, the potential temporal intention of literature retrieval was mapped onto specific temporal categories based on temporal information gained from literature. Secondly, the distribution features of temporal phrases across disciplines were used to optimize the process of label selection of the classification model of labeled LDA in order to reduce the impact of sparsity of temporal information on the learning process of classification features. Finally, the proposed algorithm was compared with other multi-label classification algorithms in specific experiments, and the accuracy of automated classification of potential temporal intentions of literature retrieval was analyzed and evaluated. The result shows that the AUC value of the proposed algorithm rea-

* **收稿日期**:2017-01-30
基金项目:国家自然科学基金资助项目(71503124),National Natural Science Foundation of China(71503124); 江苏省社会科学基金青年资助项目(15TQC03),Jiangsu Province Social Science Young Fund Project(15TQC03).
作者简介:沈思(1983—),女,江苏南京人,南京理工大学讲师,博士
†通讯联系人,E-mail: shensi@njust.edu.cn

ches 94.3%, which increases approximately 4.3%, compared with the algorithm of ECC (Ensembles of Classifler Chains). In addition, the present algorithm has produced favorable classifying effects in different disciplines. Thus, it is an effective learning method for potential temporal intention of literature retrieval.

Key words: multi-label classification; topic model; temporal information need; text feature Extraction; ext classification

目前,检索式的潜在时间意图自动识别研究主要是进行自动抽取能体现潜在时间意图的各类文本特征并应用于现有分类算法中.NTCIR 主办的 TQIC[1] (Temporal Query Intent Classification Task)测评任务要求参加者设计算法分析检索式的潜在时间意图,正确的潜在时间意图识别可以帮助更加容易地找到与未来信息相关的研究方向而不是流行的研究趋势.

在完成对时间信息标注的学术文献基础上[2],针对学术文献检索这一特定应用领域,本文主要解决学术文献潜在时间意图识别的问题.本文通过训练数据获取具有领域特性的时间词汇,并在本领域大量未标记的数据中进行扩展后,与跨学科全局时间词汇相结合作为先验知识,融入产生式分类模型 Labeled LDA 之中,对学术文献的潜在时间意图进行分类.在评价所构建的模型性能时,本文主要选取了由 Read[3] 提出的 ECC 的算法进行比较.ECC 是一种集成框架算法,主要集成了多条随机产生的分类链并通过投票机制来确定最终的分类结果.该算法的优点是通过多条分类链来提高分类的整体准确率,缺点主要是在解释性的分类任务上不能使用.

1　相关研究

典型的针对检索式的潜在时间意图的文本特征选取和分类模型构建方法主要有:Gupta[4] 采用朴素贝叶斯分类算法并采用多项特征对检索式的潜在时间意图进行分类.基于搜狗实验室的查询日志,张晓娟[5]按照检索词和时间词共同出现的频次自动识别检索式的潜在时间意图.结合查询位置、意图和用户的个性化特征,杨丹[6] 提出了 GT-WSearch 个性化 Web 搜索框架,该框架对于改进 Web 搜索结果的质量具有极大的提升.对于识别与事件相关的查询,Kanhabua[7] 进行了相应的探究.对于理解时间查询的意图和完成不同时间检索的应用,比如,时间感知查询自动实现、时间排序、结果呈现的多样化来说,识别与事件相关的查询是第一步的工作.在 AOL 查询日志和 MSN 查询日志两个数据集上,通过考虑隐性和显性的时间信息需求,研究者首先识别了潜在事件.在 TQIC 任务上,Burghartz[8] 完成了相应的探究工作,把特征分成了七个集合,集合包含了 N 元的特征,而被描述的时间触发词被分到了特征集合词汇上,与事件和时间词典相关的特征被单独进行了保存并按照他们各自的特征对时间和词典进行了分类.Zhao[9] 对维基百科页面浏览日志进行了探究,同时从查询中抽取出了两类特征,为一类为内容特征,另一类为时间序列的基本特征,并使用这些特征对歧义或者多种查询意图进行了分类.采用逻辑回归的方法,通过人工标记的方式,Willis[10] 对 TREC(Text Retrieval Conference)数据集中判别是否存在潜在时间意图的 600 项主题进行统计分析,通过内容分析的方法识别与时间敏感相关的潜在 TREC 主题特征.TQIC 测评结果表明[11],在分类算法选择上,其他效果较好的方法还包括支持向量机(SVM)、随机森林(Random Forrest)等分类器以及组合分类器.

上述研究表明,文本潜在时间意图对分类结果性能有很大影响.因此,本研究主要通过利用学术文献的隐含时间意图,在产生式分类模型中加入时间信息和领域特性的先验知识,提高学术文献的主题分类效果.

2　基于多标签的学术文献潜在时间意图自动分类算法

2.1　文本内容特征与时间特征的确定

文本时间特征选择方面,主要基于 Chinese TIMEX2 规范[12]中收录的中文时间词确定本研究的时间触发词.同时,根据本研究关注的研究领域对该规范的时间词进行了调整.一方面,删减了该标准所收录的“春分”、“春节”等在日常时间概念词汇,因为该类词汇在学术文献文本中极少出现.另一方面,追加“未来”、“最近”等综述类文献中频繁出现但 Chinese TIMEX2 却没有列出的时间词作为本文

的时间触发词,并作为一项可用于确定学术文献时间类别的描述特征.

在学术文献中,时间信息主要用于修饰文献的特有表述,例如结合“与有在什么领域……”、“本文拟探讨……”、“本文旨在……”、“作者希望……”、“对……的研究分析表明”等修辞性表述,时间信息可以对研究主题涉及的概念、方法、模型、算法、理论、应用、数据的不同侧面进行描述.因此,通过对时间信息与其描述对象之间的语义关系建模,可以有效区分不同学术文献的潜在时间意图.TempEval2010 测评将该数据集中出现的时间词,按照时间信息与其描述对象的语义关系,划分至表 1 所示的 12 项类别之中.

表 1 基于语义关系的隐含时间意图分类标准

Tab.1 Standard of classification of implied time intention based on semantic relation

隐含时间类别	描 述	举 例
before	描述对象在过去时刻提供的信息	before 1990s
after	描述对象在将来时刻提供的信息	after tomorrow
on_or_before	描述对象在从过去至今提供的信息	since 1990s
on_or_after	描述对象在从至今到将来提供的信息	until 1990s
less_than	两个描述对象存在先后的时间关系,且前一项对象发生时间早于后者	less than 2 hours long
more_than	两个描述对象存在先后的时间关系,且前一项对象发生时间晚于后者	more than 5 minutes
equal_or_less	两个描述对象存在先后的时间关系,且前一项对象发生时间早于后者,或两对象发生时间相同	no more than 10 days
equal_or_more	两个描述对象存在先后的时间关系,且前一项对象发生时间晚于后者,或两对象发生时间相同	at least 10 days
start	描述对象在事件开始时刻提供的信息	the early 1960s
mid	描述对象在事件某一时间点提供的信息	the middle of month
end	描述对象在事件结束时刻提供的信息	the end of year
approx	描述对象在模糊时间点提供的信息	about three year ago

基于隐含时间意图,本文定义了如表 1 所示的类别,并把学术文献标题、摘要和关键词中的时间信息映射到表 1 的不同类别当中.我们定义映射函数:$f_{\mathrm{mod}}:A \rightarrow B$,$A=\{$时间触发词 $\}$,$B=\{$隐含时间类别$\}=\{$before,...,approx$\}$将学术文献标题、摘要和关键词中的时间信息映射到表 1 的不同类别当中.表 2 是映射实例.表 2 是以计算机学科文献为例,描述了部分映射结果.

表 2 基于关键字的学术文献隐含类别确定

Tab.2 Determining the implied classification of academic literature based on keyword

分类结果/查询词	数据挖掘	遗传算法	神经网络	中文信息处理	专家系统
before	历史	过去;历史;长期以来	历史;过去;往年;很久以前	古代;当年;历史;过去	历史;过去;去年;历史
after	将来;未来	未来;将来	未来;将来	未来;	未来;未来十年;
on_or_before	近年来;近十年;最近几年来	近期;近年;近 20 年;至今	近年来;近十年;过去的几十年	近期;近年;十年历史;近几年来;最近几年来	近年;近几年来;至今
on_or_after					今后十年
start	早期;最初	早期;初期;最初	早期;初期;最初	最初	早期;次年夏天开始
mid	目前;当今	目前;当今	目前;当今	目前;当今	目前;当今
...	...	...	...	...	...

在表 2 中,行列交叉的单元格表示待分类的文本时间词,其列标记对应该时间词所描述的查询表示式,其行标记对应该时间词按照表 1 制定的分类标准所映射的时间类别.

2.2 基于 Labeled LDA 的文本分类模型

Labeled LDA[13]将类别标签融入到无监督的主题模型 LDA 中,构造一种有监督的主题模型.该模型对于多标签分类问题的解决证明是非常有用的,在不同的领域具有广泛的应用,比如被应用于

利用微博内容对微博标签分类[14]、利用 RCDC(Research Categorization and Disease Classification category)标签对 NIH(National Cancer Institute)医疗项目分类[15]等领域相关文本的分类任务中.下图 1 给出了 Labeled LDA 的概率图模型表示.

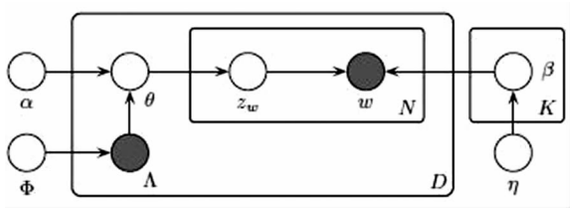

图 1　Labeled-LDA 概率图模型表示
Fig.1　Probabilistic graphical model expression of Labeled-LDA

假设语料中的词项构成词典,且词典长度为 V.一篇文档有 N 个词项组成,标记为

$W=\{w_1,w_2,\cdots,w_n\}$.整个语料由 M 篇文档组成,标记为 $D=\{W_1,W_2,\cdots,W_m\}$,文档集 D 共可得到 T 项主题,即 $t\in 1,\cdots,T$.对每一篇文档 W,定义标签类别向量 $\Lambda^{(w)}=(\Lambda_1{}^{(w)},\Lambda_2{}^{(w)},\cdots,\Lambda_T{}^{(w)})$,且 $\Lambda_t{}^{(w)}=\begin{cases}1,\text{文档 } w \text{ 中第 } t \text{ 项主题对应标签}\\0,\text{其他情况}\end{cases}$

则 Labeled-LDA 对语料 D 中每篇文档 W 的生成过程如下:

(a)对每一个主题 $t\in 1,\cdots,T$,选择超参数 β;

(b)对每一篇文档 W,选择超参数 α,生成文档 W 与标签的映射向量 $\alpha_w=\Lambda^{(w)}\times\alpha$.同时选择 θ,$\theta^{(w)}\sim Dirichlet(\alpha_w=(\alpha_{w1},\alpha_{w2},\cdots,\alpha_{wT}))$;

(c)对文档中 N 个词项中的每个词项 w_n,选择一个主题 z_n,$z_n\sim Multinomial(\theta^{(w)})$,并以 z_n 为条件的概率 $P(w_n\mid z_n,\beta)$ 选出词 w_n.

从上述过程中可以看出,对比于无监督的主题模型 LDA 中任何主题均能被分配到相应的词干上,有监督的 LDA 主题模型则至于某一个主题下的词汇关联.基于吉布斯采样的 Labeled-LDA 训练模型的概率计算式为:

$$P(z_{-i}=j\mid z_{-i},w,d_i)\propto\frac{n_{-i,j}^{(w)}+\beta}{n_{-i,j}^{(*)}+W\beta}\cdot\frac{n_{-i,j}^{(d_i)}+\alpha_w}{n_{-i,j}^{d_i}+K\alpha_w}\tag{1}$$

公式(1)中,$n_{-i,j}^{(w)}$ 表示词 w 分配到主题 j 的数量,$n_{-i,j}^{(*)}$ 表示分配到主题 j 的词的总数,$n_{-i,j}^{(d_i)}$ 表示文本 d_i 中分配到主题 j 的词的数量,$n_{-i,j}^{d_i}$ 表示文本 d_i 中词的数量,α_w 表示考虑超参数 α 情况下文档 W 与标签的映射向量.

基于吉布斯采样的 Labeled-LDA 预测模型的概率计算式为:对于新加入数据集的文档 d',设 $\Lambda_t^{(d')}=1\,\forall\, t\in\{1,\cdots,T\}$,则主题 j 下的后验分布计算式为:

$$P(z_{-i}=j\mid z_{-i},w,d')\propto\frac{n_{-i,j}^{(w)}+\beta}{n_{-i,j}^{(*)}+W\beta}\cdot\frac{n_{-i,j}^{(d')}+\alpha}{n_{-,j}^{d'}+K\alpha_{wj}}\tag{2}$$

式中,$n_{-i,j}^{(w)}$ 和 $n_{-i,j}^{(*)}$ 等词汇-主题共现数量统计由 Labeled-LDA 训练模型得到,仅 $n_{-i,j}^{(d')}$ 和 $n_{-,j}^{d'}$ 等需要根据文本 d' 中被分配到主题 j 的情况进行更新.

2.3　基于 Labeled LDA 的学术文献时间分类的特征权重改进算法

已有研究证明,在 LDA 模型的预测上不能非常好的进行校准.在本研究中,Labeled LDA 的标签预测结果,同样存在上述问题.

针对上述问题,我们提出一种根据不同学科潜在时间意图偏好性的标签主题模型改进特征权重算法.

算法 1:基于学科时间信息的分类特征权重调整算法.

输入:某学科文档集合 S.

输出:基于该学科文档 S 生成的特征权重调整向量 $\overrightarrow{weight}$.

step1:遍历学科文档集合 S,提取各文档中的时间触发词汇,形成候选集 $T_{word}(S)$,并分别计算各时间触发词汇对应的时间关系,找出属于相同类别的触发词集合:$f_{\mathrm{mod}}(t_{word1})=f_{\mathrm{mod}}(t_{word2})=\cdots=f_{\mathrm{mod}}(t_{wordn})$,$t_{wordi}\in T_{word}(S)$.

step2:依据 $T_{word}(S)$ 中的各时间触发词对应的时间关系 $f_{\mathrm{mod}}(t_{wordi})$,将学科文档集合 S 进行子集划分,使得子集 S_j 中各文档的时间信息隶属于同一时间关系,即 $\{t_{word1},t_{word2},\cdots,t_{wordjn}\}\in T_{word}(S_j)$,且 $f_{\mathrm{mod}}(t_{word1})=f_{\mathrm{mod}}(t_{word2})=\cdots=f_{\mathrm{mod}}(t_{wordjn})$.

step3:依次计算隶属于同一时间关系的文档子集 S_j 的时间信息语义倾向性.假设子集 S_j 对应第 i 类时间关系,且子集中不重复的时间词项数为 jn 项.则从第 0 项时间词的词频 $N_{i,0}$ 开始依次统计该子集中各项时间词的词频,并将词频项的平均值作为该时间关系对应的特征权重调整分量 $w_{\mathrm{mod}(i)}$,即

$$w_{\mathrm{mod}(i)}=\frac{(N_{i,0}-N_{avg})^2+(N_{i,1}-N_{avg})^2+\cdots+(N_{i,jn}-N_{avg})^2}{jn\times {N_{avg}}^2} \tag{3}$$

其中，$N_{i,0}+N_{i,1}+\cdots+N_{i,jn}>0$ 且 $N_{\mathrm{avg}}=\mathrm{average}(N_{i,0}+N_{i,1}+\cdots+N_{i,jn})$.

step4：将不同隐含时间类别的分量 $w_{\mathrm{mod}(i)}$ 组成特征权重调整向量 $\overrightarrow{weight}$，即 $\overrightarrow{weight}=\{w_{before},w_{after},\cdots,w_{\mathrm{mod}(i)},\cdots,w_{mid}\}$，再逐一对不同学科文献的 Labeled LDA 隐含时间意图的分类结果进行调整，详见算法 2.

算法 2：潜在时间意图标签分类权重调整算法.

输入：和某检索主题相关的文档集合 D，以及文档集合 D 中各文档所属学科类别.

输出：改进后的对于学术文献的预测标签集合 $p(L_t^{(d')}\mid d')$.

1)将文档集合 D 划分为训练集 D^+ 和测试集 D^-，并基于训练集 D^+ 学习 Labeled LDA 模型.

2)利用已学习的 Labeled LDA 模型，生成文档集合中的测试集 D^- 中各文档标签分配结果.对其中每一个新加入的文档 d'，对应的多重 l 标签分配结果可表示为 $p(\Lambda_t^{(d')}\mid d')$，$\Lambda^{(d')}=(\Lambda_1{}^{(d')},\Lambda_2{}^{(d')},\cdots,\Lambda_T{}^{(d')})$.

对新加入数据集的文档 d'，对基于公式(2)计算所有的标签 $\Lambda_t^{(d')}=1\,\forall\, t\in\{1,\cdots,T\}$ 在该文档中的概率，按照概率从高至低进行排序.

3)对文档集合 D' 按照文档所属学科类别进行划分，并以划分结果中的学科文档集为输入，根据算法 1 依次生成各学科的特征权重调整向量 $\overrightarrow{weight}$.

4)对每一个新加入的文档 d'，首先查找该文档对应学科.再根据对应学科的权重调整向量 $\overrightarrow{weight}$，调整 $p(\Lambda_t^{(d')}\mid d')$ 标签分配概率值.首先计算各项标签 Λ_t 对应的时间关系 $f_{\mathrm{mod}}(\Lambda_t)$，随后查找权重向量 $\overrightarrow{weight}$ 中第 k 个表示该时间关系的分量，用公式 $p(L_t^{(d')}\mid d')=p(\Lambda_t^{(d')}\mid d')\cdot\overrightarrow{weight_k}$ 将该标签分配概率值进行更新.

5)重复 4)，直至测试集中所有文档均得到修正后的标签分配结果.

3 实验与结果分析

3.1 实验数据描述

统计结果表明，人文社会科学领域的文献内容中时间词存在比例远高于自然科学领域文献集合.因此，我们以 cnki 数据库为来源，收集了 11 个学科的学术文献题录文本，具体学科和对应文献数量以及各学科包括时间词文献所占比例如表 3 所示.

表 3 各个学科包含时间词文献分布表

Tab.3 Distribution of literature of temporal words of different disciplines

学科类别	文献数	包括时间词的文献数	内容时间词总数	显性时间数量	显性时间百分比/%	隐性时间数量	隐性时间百分比/%	相关时间数量	相关时间百分比/%
计算机科学	144 288	21 679	29 989	5747	19.16	17 500	58.35	6 742	22.48
经济学	84 261	59 073	148 824	47 795	32.12	41 986	28.21	59 043	39.67
历史学	46 897	39 951	169 110	32 253	19.07	53 721	31.77	83 136	49.16
图书馆、情报与文献学	109 276	59 958	98 023	32 259	32.91	20 128	20.53	45 636	46.56
新闻学与传播学	53 719	35 482	94 357	26 519	28.10	24 247	25.70	43 591	46.20
法学	43 642	27 806	120 957	13 791	11.40	7 046	5.83	100 120	82.77
中国文学	69 768	52 477	100 098	38 885	38.85	30 372	30.34	30 841	30.81
管理学	48 425	28 287	77 598	16 712	21.54	16 629	21.43	44 257	57.03
哲学	41 510	27 971	53 558	15 516	28.97	13 676	25.53	24 366	45.49
政治学	21 262	15 868	60 678	11 530	19.00	9 594	15.81	39 554	65.19
社会学	34 264	23 171	87 768	19 356	22.05	9 369	10.67	59 043	67.27

3.2 实验结果以及分析

3.2.1 分类过程与参数选择

在参数选择方面，对文本进行 TFIDF，PMI 和信息熵等不同特征选择时，本文保留排名前 50%的特征作为输入文本并去除噪音.在训练集和测试集构建时采用 9∶1 比例，把 36,409 个标签的文献随机地分成了训练和测试两种数据集合，并参考已有研究选择 AUC(area under the receiver operating characteristic curve)得分作为 Labeled-LDA 分类任务的测评指标.在分类过程方面，采用抽样的方式对

测试集的样本进行标签分类,再计算标签分类结果的 AUC 值.针对包含时间触发词的标签集合 $Set_{labeled}$ 作为正样本采样结果,采用 n 次迭代且根据公式(3)计算 AUC 值

$$AUC=\frac{n'+0.5n''}{n} \tag{4}$$

其中,n' 为从 $Set_{labeled}$ 中取出的正样本的概率大于负样本的概率的次数,n'' 为概率相等的次数,n 为总迭代次数.

3.2.2　分类性能评价

本实验在通过 TF-IDF,PMI 和熵获取的单个内容特征的基础上,基于比较原有的标签主题模型和本研究改进的模型方法,对比了本文方法在不同学科之间的分类性能表现情况.整个查询分类的结果具体如表 4 所示.如该表所示,在文本特征选择方面,PMI 值相较于其他文本特征的选择方法 AUC 的值最低为 0.739,其性能最差.同时,信息熵的选取特征的方法取得了最好的效果,AUC 的值最好达到了 0.795.从表中还可以看出,即使选择效果最差的 PMI 值作为文本特征,本文方法也比原始的 Labeled LDA 算法在 AUC 值有所提升,从 0.739 提升到了 0.754.

表 4　不同文本特征下的整体性能

Tab.4　Whole performance of different text features

AUC 值	TFIDF	PMI	信息熵
原始 Labeled LDA 算法	0.755	0.739	0.791
本文特征权重调整算法	0.761	0.754	0.795

表 5　不同特征在单一学科上的性能

Tab.5　Performance of different features in single discipline

学科	TFIDF		PMI		熵	
	原有的主题模型性能（AUC 值）	本文改进的方法（AUC 值）	原有的主题模型性能（AUC 值）	本文改进的方法（AUC 值）	原有的主题模型性能（AUC 值）	本文改进的方法（AUC 值）
经济学	0.784	0.801	0.815	0.812	0.843	0.837
管理学	0.778	0.792	0.704	0.715	0.805	0.828
图书馆、情报与文献学	0.781	0.794	0.750	0.771	0.747	0.754
计算机科学	0.801	0.831	0.874	0.900	0.922	0.939
新闻学与传播学	0.717	0.722	0.716	0.710	0.789	0.795
政治学	0.692	0.686	0.787	0.792	0.734	0.764
社会学	0.721	0.702	0.740	0.719	0.749	0.774
历史学	0.746	0.760	0.682	0.698	0.820	0.820
中国文学	0.737	0.750	0.686	0.771	0.778	0.790
法学	0.706	0.693	0.769	0.763	0.798	0.750
哲学	0.826	0.831	0.746	0.773	0.735	0.739
平均值	0.7536	0.761	0.752	0.766	0.794	0.796

表 6　分类性能对比的结果

Tab.6　Result of comparing the performance of classification

学科	ECC 分类算法(AUC 值)	本文改进的方法(AUC 值)		
		TFIDF	PMI	信息熵
经济学	0.725	0.801	0.812	0.837
管理学	0.682	0.792	0.715	0.828
图书馆、情报与文献学	0.726	0.794	0.771	0.754
计算机科学	0.750	0.831	0.900	0.939
新闻学与传播学	0.632	0.722	0.710	0.795
政治学	0.714	0.686	0.792	0.764
社会学	0.651	0.702	0.719	0.774
历史学	0.728	0.760	0.698	0.820
中国文学	0.721	0.750	0.771	0.790
法学	0.636	0.693	0.763	0.750
哲学	0.592	0.831	0.773	0.739
平均值	0.687	0.761	0.766	0.796

表 5 则对比了不同学科下本文的特征权重调整算法的分类性能,从表中可以看出,即使学术文献的不同学科对时间触发词分布有所影响,信息熵的选取特征的方法依然取得了最好的效果,其在 11 个学科中的 AUC 平均值达到了 0.796.同时,就单一学科文本分类结果而言,本文提出的方法在计算机学科上 AUC 达到了最高值,为 0.939.

表 6 则对比了本文的特征权重调整算法和同类基于权重调整的多标签分类算法 ECC 的性能对比.从表中可以看出,本文方法在相同的文本特征选择基础上,分类结果的 AUC 值均高于 ECC 方法.同样以信息熵作为文本特征时效果最好,本方法的 AUC 值平均值高过 ECC 算法达到了 10.9%.

4 结 语

综上所示,本研究以学术文献中隐含时间意图为分类对象,在 Labeled-LDA 的标签语义关系的分类基础上,提出了一种潜在时间意图标签分类权重调整算法.根据不同的文本特征选择方式,以及在不同学科上的分类实验表明,本文提出的方法能够区分不同文献、不同学科在隐含时间意图之下的时间关系偏好性,从而更好地优化学术文献的隐含时间意图分类结果.因此,本文的方法可用于更好地从语义知识层面来挖掘学术文献的隐含的时间信息,帮助分析以时间触发词作为文本标签时研究主题之间的时间关联性.

参考文献

[1] JOHO H, JATOWT A, BLANCO R. NTCIR temporalia: a test collection for temporal information access research[C]// Proceedings of the 23rd International Conference on World Wide Web. Seoul, Republic of Korea, 2014: 845－850.

[2] 沈思,苏新宁,谢靖,等. 基于清华汉语树库的时间表达式抽取模型构建研究[J]. 图书情报工作,2012,56(18):127－132.

[3] READ J, PFAHRINGER B, HOLMES G, *et al*. Classifier chains for multi-label classification[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg, 2009:254－269.

[4] GUPTA D, BERBERICH K. Temporal query classification at different granularities[C]// String Processing and Information Retrieval. London, UK, 2015:156－164.

[5] 张晓娟,陆伟,周红霞. 用户查询中潜在时间意图分析及其检索建模[J]. 现代图书情报技术,2011 ,30(11): 38－43.

[6] 杨丹,申德荣,陈默. 基于地理-时间意图和偏好的个性化 Web 搜索框架 GT-WSearch[J]. 计算机科学,2015,42(7):240－244.

[7] KANHABUA N, NGOC NGUYEN T, NEJDL W. Learning to detect event-related queries for web search[C]//Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, 2015: 1339－1344.

[8] BURGHARTZ R, BERBERICH K. MPI-INF at the NTCIR-11 temporal query classification task[C]// Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo, Japan, 2014:443－450.

[9] ZHAO Y, HAUFF C. Temporal query intent disambiguation using time-series data[C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2016: 1017－1020.

[10] WILLIS C, SHERMAN G, EFRON M. What makes a query temporally sensitive? [C]// Proceedings of the 39th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval. Beijing, China, 2016: 1065－1068

[11] JOHO H, JATOWT A, BLANCO R, *et al*. Building test collections for evaluating temporal IR[C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2016: 677－680.

[12] ACE (Automatic Content Extraction) Chinese Annotation Guidelines for TIMEX2 (Summary)[EB/OL].[2016－12－19].http://www.ldc.upenn.edu/Projects/ACE/docs/Chinese-TIMEX2-Guideline-Summary_v1.

[13] RAMAGE, D, HALL, D, NALLAPATI, R, *et al*. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora[C] //Proceedings of the 2009 Conference on Empirical Methods in Natural Language. Cambridge, Massachusetts USA, 2009: 248－256.

[14] KIRCZ G. Rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval [J]. Journal of Documentation, 1991, 47(4): 354－372.

[15] PARK J, BLUME-KOHOUT M, KRESTE R, *et al*. Analyzing NIH funding patterns over time with statistical[C] // Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, Arizona USA, 2016:698－704.