科技文献关键词自动标注算法研究

倪 娜[1] 刘 凯[1,2] 李耀东[1]

(中国科学院自动化研究所复杂系统智能控制与管理国家重点实验室(筹) 北京 100190)[1]

(中国民航信息网络股份有限公司研发中心分销产品研发部 北京 100029)[2]

摘 要 未标注或遗失关键词给科技文献的分类和导航工作带来一定困难,针对这一问题,提出了基于文献摘要内容的关键词自动标注算法。该算法使用标注过关键词的文献摘要作为训练文本,分别采用语言模型、Latent Dirichlet Allocation(LDA)模型、Probabilistic Author-Topic 模型及语言模型+LDA 模型的组合模型对训练集中的摘要文本和关键词建模,建立关键词和组成摘要文本特征词之间的关系,然后利用这些模型在未标注关键词的科技文献摘要上进行关键词的预测。在中英文数据上的实验结果表明,自动标注的关键词能较好地反映科技文献的内容;在所有模型中,语言模型+LDA 组合模型的效果最佳。

关键词 语言模型,标签预测,Latent Dirichlet Allocation,Probabilistic Author-Topic Model

中图法分类号 TP391 **文献标识码** A

Study of Automatic Keywords Labeling for Scientific Literature

NI Na[1] LIU Kai[1,2] LI Yao-dong[1]

(State Key Laboratory of Intelligent Control and Management of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)[1]

(Dept. of Distribution System, R&D Center, TravelSky Technology Limited, Beijing 100029, China)[2]

Abstract Keywords of scientific literatures provided by authors are helpful for readers. But there are also some scientific literatures that are not labeled with keywords due to all sorts of reasons. So this paper proposed a new abstract-based automatic keywords prediction algorithm for scientific literatures without keywords. The abstracts of scientific literatures, which had been given keywords by authors, were used as the training data set. Four text modeling methods: language model(LM), latent dirichlet allocation(LDA), probabilistic author-topic model, and a combination of LM and LDA were employed to model the abstracts and the keywords in training set to build the relations between keywords and terms of abstracts. Then the trained models were used to predict keywords for the abstracts of scientific literatures without keywords. The experimental results on both Chinese data sets and English data sets show that the keywords predicted by the proposed algorithms can reflect the content of scientific literature well. Among all of the models, the combination of LM and LDA is best.

Keywords Language model, Tag prediction, Latent dirichlet allocation, Probabilistic author-topic model

1 引言

科技文献关键词是由作者给出的,是对文献内容、采用的方法及相关研究领域的一个精炼的描述。准确的关键词标注,可以使读者在不阅读文章内容的情况下对文章涉及的领域和关键技术点有一个概括的了解。但是,科技文献关键词的标注格式通常不统一,发表时间较长的文献常常出现关键词缺失的情况,给文献的检索和分类查看带来一定的困难。因此,根据现有科技文献的信息(如已标注的关键词、摘要)对新资源及关键词缺失的资源进行关键词推荐和标注,有利于对文献进行归类、整理,便于文献的检索等。

现有的方法大都是根据文献内容,采用基于统计或规则的方法进行关键词抽取。索红光等[1]提出了一种基于词汇链的关键词抽取方法;李素建等[2]采用最大熵模型研究关键词的自动标引;张雪英等[3]提出了一种基于 N-gram 权重计算和关键词筛选算法的中文文本关键词自动抽取方法;罗准辰等[4]设计了一种基于分离模型的关键词提取方法,亦即将关键词提取分为关键词提取和关键词串提取两个问题进行处理。上述方法在关键词抽取领域取得了较好的效果,但它们抽取的关键词完全依赖于在文本中出现的词且所需的语料通常较长。通过对科技文献资源的分析发现,作者给文献添加的关键词有时在文献本身出现的次数很少甚至根本不出现;

到稿日期:2011-10-20 返修日期:2012-03-03 本文受 973 国家重点基础研究发展计划(2007CB311007),国家自然科学基金(61072084)资助。

倪 娜(1984—),女,博士生,主要研究方向为智能信息处理、数据挖掘,E-mail:na.ni@ia.ac.cn;刘 凯(1985—),硕士;李耀东(1977—),男,博士,副研究员,主要研究方向为智能信息处理、系统复杂性。

另外，多数科技论文库都直接在网页上展现摘要，而全文的获取则较为繁琐。因此，现有的基于文献全文的方法并不适用于科技文献关键词的标引。

从问题的实质来看，自动为科技文献标注关键词与社会化标签推荐(Social Tag Recommendation)的研究内容非常类似。近年来，社会化标注系统发展迅速，涌现出大量相关网站，如 Flickr[1)]，delicious[2)] 等。在这些网站上，用户可以根据自己的兴趣和需要对相关资源添加标签。因此，很多学者开始研究社会化标签推荐技术，向用户推荐其感兴趣的内容，同时帮助用户更好地对资源进行标注，如 Heymann 等[5]提出了一种基于规则的社会标签预测方法；Krestel 等[6]对标签空间采用 LDA 模型进行建模，对资源进行标注以利于资源的检索；Sun 等采用语言模型进行网络文本资源的标签推荐[7]；Yin 等采用概率模型进行个性化标签预测[8]；Shepitsen 等采用层次聚类对社会标签系统进行个性化的内容推荐[9]。

但是二者也存在着一些显著区别：1)推荐的目的不同。上述社会标签推荐的重点在于个性化的标签推荐，而科技文献关键词推荐的主要目的是向未标注或标注不完全的文献推荐标签，这一过程中没有用户的参与，因而没有个性化的需求。2)推荐的要求不同。科技文献的关键词必须反映文献的主要内容、研究领域和采用的方法等，社会标签则常常是用户依据自己的偏好，对资源添加有利于浏览和分类的信息。因此，社会化标签推荐的方法对基于内容的关键词标注有很大的指导意义，但又不能直接应用于其中。

本文采用科技文献摘要作为分析数据，借鉴社会化标签推荐的部分思路，提出了一种基于概率话题模型的科技文献关键词标注算法。该算法首先在训练文本集合中，采用 LDA、语言模型和概率作者话题模型建立关键词与组成文本的特征词之间的关系，然后预测未标注的文本摘要的可能的关键词信息。实验结果表明，本文所采用的算法能够在文献信息较少的情况下，达到较好的预测效果；同时，本方法易实施，可应用于不同语言的科技文献。

2 基于摘要内容的关键词预测方法

2.1 问题描述

本文所述的关键词预测指根据已标注关键词的科技文献摘要，建立关键词和摘要之间的关系模型，借助该模型，预测未标注关键词摘要的关键词。

关键词预测的训练集由已标注关键词的文献摘要构成，其数学描述如下：

$$Training\ Set: C_{tr}=(t_1,d_1),(t_2,d_2),\cdots,(t_N,d_N)$$

式中，$t\in K$ 表示关键词，K 为训练文本中所有关键词的集合，$d\in D$ 表示摘要文本，摘要文本 d 由特征词(term)组成，用 w 表示。训练过程即根据上述已知信息，建立关键词与组成文本的特征词或潜在话题之间的概率关系。

本文关键词标注过程为：给定一篇未标注的科技文献摘要 d_u，从候选关键词集中选出合适的关键词推荐给这篇文献，即对于所有 $t\in K$，计算 $p(t|d_u)$，则为该篇文献标注的 n 个关键词为：

$$REC(t)=\arg\max_{t\in K}^{n} p(t|d_u) \tag{1}$$

2.2 关键词标注算法

2.2.1 基于语言模型的算法

语言模型在自然语言处理相关领域，如语音识别、机器翻译、信息检索等有广泛的应用。Ponte 和 Croft[10] 在 1998 年最早将统计语言模型应用到信息检索领域，近年来其已经成为该领域一种常用的文本建模方法。采用 Zhai 和 Lafferty[11] 提出的 Dirichlet 平滑方法，根据语言模型对文本建模可得：

$$p(w_i|d)=\frac{N_d}{N_d+\mu}\cdot\frac{tf(w_i,d)}{N_d}+\left(1-\frac{N_d}{N_d+\mu}\right)\cdot\frac{tf(w_i,C_{tr})}{N_C} \tag{2}$$

式中，N_d 为文本 d 的长度(token 数目)，N_C 为整个训练文本集合词数，$tf(w_i,d)$，$tf(w_i,C_{tr})$ 为特征词 w_i 在文本 d 及整个训练文本集合中的词频，μ 为 Dirichlet 平滑因子，通常为文本集合中所有文本的平均长度。

这里将语言模型引入进行关键词的预测。根据 2.1 节中的描述，对于一篇未标注的摘要文本 d_u 进行关键词预测即为估计概率 $p(t|d_u)$ 大小的过程。由贝叶斯理论可得：

$$p(t|d_u)=p(d_u|t)p(t)/p(d_u)\propto p(d_u|t)p(t) \tag{3}$$

采用一元语言模型，摘要文本 d_u 可看作独立的特征词组成的序列 $\{w_1,w_2,\cdots,w_m\}$，由查询似然模型[10]可得：

$$p(d_u|t)=\prod_{i=1}^{m}p(w_i|t) \tag{4}$$

在训练集 C_{tr} 上，采用式(2)的方法得到特征词和关键词的关系如下：

$$p(w_i|t)=\frac{N_t}{N_t+\mu}\cdot\frac{tf(w_i,t)}{N_t}+\left(1-\frac{N_t}{N_t+\mu}\right)\cdot\frac{tf(w_i,C_{tr})}{N_c} \tag{5}$$

式中，N_t 为关键词 t 在训练数据中出现的次数，$tf(w_i,t)$ 为特征词 w_i 在关键词 t 标注过的文本 d^t 中出现的次数的和：

$$tf(w_i,t)=\sum_{d^t}tf(w_i,d^t)$$

对关键词 t 在训练文本集中出现的概率 $p(t)$，采用极大似然估计计算如下：

$$p(t)=\frac{tf(t,C_{tr})}{\sum_{t\in T}tf(t,C_{tr})} \tag{6}$$

综合上述计算结果，由式(1)可得为未标注关键词文献 d_u 推荐的关键词列表。

2.2.2 基于 LDA 的算法

Latent Dirichlet Allocation(LDA)，是由 Blei 等[12]提出的一种产生式模型，近年来广泛应用于信息检索和机器学习等领域[13,14]，它在文本分类、协作过滤等方面都表现出了良好的性能。在 LDA 模型中，引入了话题这一潜在变量。其基本的观点是文本表示为若干个潜在话题的概率组合，而话题则是组成文本的特征词的概率分布，则给定文本 d 情况下，w_i 的概率计算如下：

1) http://www.flickr.com/

2) http://del.icio.us/

$$p(w_i|d)=\sum_{j=1}^{T}p(w_i|z=j)p(z=j|d)=\sum_{j=1}^{T}\phi_i^{(j)}\cdot\theta_j^{(d)} \quad (7)$$

式中，$z=j$ 为文本集合的第 j 个话题。采用 Gibbs 采样[13,15]方法估计模型参数，可得到 $\phi_i^{(j)}$，$\theta_j^{(d)}$ 的计算公式：

$$\phi_i^{(j)}=\frac{C_{ij}^{WT}+\beta}{\sum_{k=1}^{W}C_{kj}^{WT}+W\beta},\quad \theta_j^{(d)}=\frac{C_{dj}^{DT}+\alpha}{\sum_{k=1}^{T}C_{dk}^{DT}+T\alpha}$$

式中，C_{ij}^{WT} 为 w_i 用于描述话题 j 的次数，C_{dj}^{DT} 为文本 d 中的词用于描述话题 j 的总次数，W 为特征空间的维数，T 为潜在话题的数目。

对未标注文本 d_u 进行关键词预测时，根据 LDA 模型的思想，本文在 d_u 和关键词 t 之间引入中间变量话题 z，则文本对关键词的预测概率为：

$$p(t|d_u)=\sum_z p(t|z)p(z|d_u) \quad (8)$$

通过对训练文本采用 LDA 模型建模，可得到关键词 t 和话题 z 之间的关系 $p(t|z)$。然后在测试集中，通过已经建好的模型，仍然采用 Gibbs 采样，推断新文本中的话题分布 $p(z|d_u)$。

在训练集合中，如果直接采用式(7)所示的模型对文本进行建模，则无法获得关键词 t 和话题 z 的关系。因此，本文采用伪文本的形式表示训练语料，即将所有同一关键词标注的摘要文本合并成一篇文本，此时的关键词 t 等价于原 LDA 模型中的文本 d。即合并后的数据集中文本数目等于原数据集中关键词的数目。通过对训练文本进行 Gibbs 采样，可得 $p(t|z)$的值。

给定新摘要文本 d_u，在已经建立的模型基础上进行 Gibbs 采样，可得到更新后的模型参数，则 $p(z|d_u)$的值为：

$$p(z|d_u)=\frac{C_{\inf dj}^{DT}+\alpha}{\sum_{k=1}^{T}C_{\inf dk}^{DT}+T\alpha} \quad (9)$$

式中，$C_{\inf dj}^{DT}$ 为 C_{dj}^{DT} 在训练集合上进行 Gibbs 采样更新后的值。通过式(8)和式(9)的计算，可得最后推断的结果。

2.2.3 基于 PAT 的算法

Probabilistic Author-Topic Model(PAT)，是由 Setyvers 等在 LDA 模型的基础上，针对科技文献中作者和话题的关系提出的[16]。该模型认为，每个作者可表示成话题上的概率分布，而每个话题则为特征词上的概率分布。当生成一个文本时，首先为文本中的每个特征词随机选择一个作者，该作者从他在话题上的多项式分布中选择一个话题，然后从该话题在特征词上的多项式分布中选择一个特征词。对文本中的每一个特征词重复这一过程，即实现了整个文本的生成。采用 Gibbs 采样方法进行参数估计，其模型的数学表达为：

$$p(w_i|z=j)=\frac{C_{ij}^{WT}+\beta}{\sum_{l=1}^{W}C_{lj}^{AT}+W\beta}$$

$$p(z=j|a)=\frac{C_{kj}^{AT}+\alpha}{\sum_{l=1}^{T}C_{kj}^{AT}+T\alpha}$$

本文研究的是关键词和组成摘要文本的特征词之间的关系。对于一篇科技文献来说，我们认为关键词和特征词的关系类似于 PAT 模型中作者和特征词的关系。因此，可将 PAT 模型中的作者替换成关键词，对科技文献文本进行建模，进而实现对未知关键词科技文献的关键词预测。

类似于 LDA 模型，仍然采用式(8)的方法对未标注文本的关键词进行推断。不同的是，这里采用 PAT 模型直接建立关键词和组成文本的特征词之间的关系，应用 Gibbs 采样对参数进行推断，可得：

$$p(t|z=j)=\frac{C_{kj}^{AT}+\alpha}{\sum_{l=1}^{K}C_{lj}^{AT}+K\alpha} \quad (10)$$

式中，K 为关键词的数目。同样，在测试文本上，可用式(9)的方式，对通过训练集建立的文本模型进行更新，从而联合式(9)和式(10)，得到对标注文本的关键词推荐结果。

2.2.4 结合 LDA 和 LM 的方法

在 Wei 等[14]的文章中，将 LDA 模型和语言模型做了线性组合，在 TREC 等数据集上的实验表明，将二者结合能够提高单独采用任意一种方法进行信息检索的性能。因此本文也尝试将二者结合，进行未标注文本的关键词预测。如式(11)所示，引入组合系数 λ，将语言模型和 LDA 模型的建模结果进行线性组合，再将该模型代入式(4)的查询似然模型，采用式(3)的方法进行关键词的预测。

$$p(w_i|t)=\lambda p_{LM}(w_i|t)+(1-\lambda)p_{LDA}(w_i|t) \quad (11)$$

式中，组合系数 λ 的值在 0 到 1 之间变化，不同取值对预测效果的影响将在实验部分进行介绍。

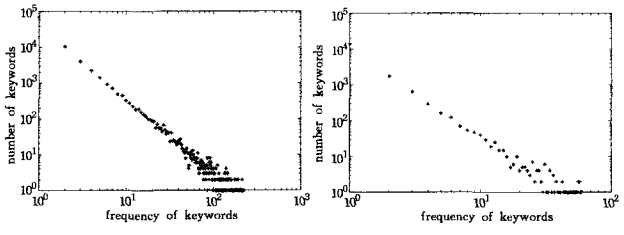
3 实验与评价

为验证方法的有效性，本文分别在中文和英文两个数据集上进行了实验。

3.1 实验语料与实验设置

3.1.1 中文科技文献

本文从万方数据资源[3]选择了 8 个与自动化和计算机技术相关的期刊，如《软件学报》、《计算机学报》、《自动化学报》等，下载了从 2000 年－2010 年全部文献的摘要和关键词信息。去掉摘要字数小于 50 字的文献后，得到了大小为 63041 的数据集。数据集中文献的关键词分布如图 1(a)所示。



(a) 万方数据　　(b) Elsevier 数据

图 1　语料关键词分布

对摘要文档采用 ICTCLAS[4]分词，去除停用词，去掉在整个语料集合中词频小于 10 的特征词，共得到 13524 个 unique word。同样，也去掉了在语料中出现少于 10 次的关键词。将语料按照 9:1 的比例进行分割，其中 1/10 用作测试集，余下的作为训练集。得到的语料统计数据如表 1 所列。

3) http://www.wanfangdata.com.cn/

4) http://ictclas.org/

表 1　预处理后的中英文数据统计信息

语料	关键词数	特征词数	训练集	测试集	文献平均关键词数
万方	2935	13524	46698	5188	1.3920
Elsevier	1614	6186	4398	488	3.0203

3.1.2　英文科技文献

本文收集了 Elsevier Science Direct[5] 中与计算机科学方向相关的几个期刊(如 Advances in Engineering Software、Expert Systems with Applications 等)从 2000 年至今的全部文章。共得到摘要 5873 篇,关键词 29486 个。关键词的分布类似于万方数据,如图 1(b)所示。

根据空格对摘要文本进行断词,采用 Porter Stemmer[6] 抽取所有特征词的词干,并去除了停用词及在语料中出现次数少于 3 次的词(英文语料规模较小)。对于关键词,也进行了词干抽取,并去掉了在语料中出现少于 3 次的词。同样按照 9:1的比例分割了训练集和测试集,得到的数据集的统计信息如表 1 所列。

3.1.3　模型训练

对于 LDA 和 PAT 模型,采用 Matlab Topic Modeling Toolbox 1.4[7] 进行训练。实验中的参数,如话题数目、组合系数等的取值对于本研究仍是一个开放性的问题。本文的目的不在研究参数的选择,因此,仅对参数的不同取值对实验结果的影响做了讨论。

3.1.4　评价准则

本文采用常用的准确率、召回率和 F1 值评价各种方法的预测效果。此外,还采用了 Top-k accuracy 和 Exact-k accuracy[17] 两种方法进一步比较方法的优劣,这两种评价标准的定义如下:

Top-k accuracy:在前 k 个预测结果中,文本被至少一次正确标注的百分比;

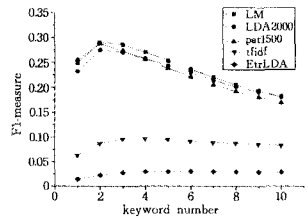
Exact-k accuracy:预测 k 个关键词时,文本被第 k 个预测结果正确标注的百分比。

3.2　关键词预测实验结果

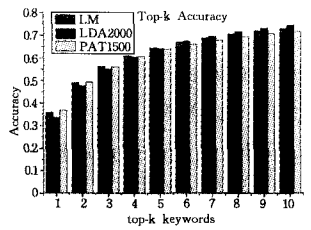
3.2.1　万方数据实验结果

在中文语料——万方数据集分别采用 LM、LDA、PAT 3 种算法进行了未标注文本的关键词预测。结果如图 2 所示,图 2(a)为在预测关键词数目为 1～10 时 3 种算法的 F1 值。其中 LDA2000 表示在 $T=2000$ 时 LDA 算法的结果,PAT1500 表示 $T=1500$ 时 PAT 模型的 F1 值曲线。由于 F1 值平衡了召回率和准确率,表示的是方法的综合性能,因此这里不再展示算法的准确率和召回率。由图可知当预测关键词数目为 2 时,3 种算法都取得了最好的效果。图 2(b)和图 2(c)为 3 种方法的 Top-k accuracy 和 Exact-k accuracy 结果。通过对实验结果的分析可以发现:从整体来说,语言模型的算法得到的结果最好,PAT 在预测的关键词较少时表现较好,而 LDA 则在预测的关键词较多时表现更佳。

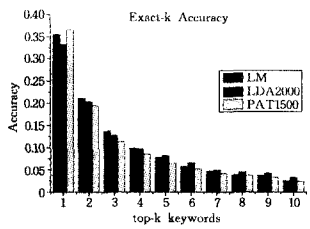
为了验证所提方法的性能,本文将实验结果与现有关键词抽取技术进行了对比。鉴于科技文献摘要通常较短小,现有的一些关键词抽取技术很难在其中挖掘出有意义的词串,因此,本文只采用了常用的基于统计的文本分析技术:TFIDF 算法、LDA 算法等,在预处理(分词、取出停用词)后的训练语料上进行了关键词抽取的实验。抽取的前 10 个关键词的 F1 值如图 2(a)所示,其中曲线 TFIDF 表示采用 TFIDF 算法进行关键词抽取的结果,Etr-LDA 表示采用 LDA 算法直接从测试摘要文本中抽取关键词的结果。实验结果表明,本文提出的方法远远优于直接从文献中抽取关键词的方法。



(a) 预测前 k 个关键词时,本文采用的 3 种算法及 TFIDF 算法的 F1 值



(b) 预测前 k 个关键词时,算法的 Top-k accurary



(c) 预测前 k 个关键词时,算法的 Exact-k accurary

图 2　万方数据集实验结果

3.2.2　Elsevier 数据上的实验结果

分别将本文采用的 LM、LDA 和 PAT 3 种方法在Elsevier数据集上进行了实验。3 种算法得到的预测结果的 F1 值如图 3 所示。由图 3 可见,Elsevier 数据集上的实验结果比万方数据上的结果稍差一些,这是由于该数据集的数据量与万方数据相比较小;同时在这个数据集上,LDA 方法的结果好于 PAT,总体实验结果仍然是 LM 模型最好。3 种算法的 Top-k accuracy 和 Exact-k accuracy 结果和万方数据类似,由于篇幅所限,在这里不进行单独展示。

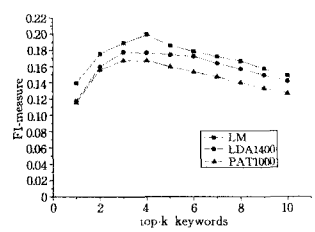


图 3　Elsevier 数据集上 3 种方法的 F1 值对比

从上述实验结果中可以看出,在两个数据集上,语言模型均表现出了较好的性能。但是,采用 PAT 模型进行标注的结果并没有达到很好的效果。分析可能的原因如下:概率话题模型假设文本和特征词之间存在一个潜在的话题层,通过这

5) http://www.sciencedirect.com/

6) http://tartarus.org/~martin/PorterStemmer/

7) http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

一话题层建立特征词和关键词之间的关系模型，但可能由于科技文献摘要文本通常比较短小，特征词在文本中的分布也不同于一般文本，属于相同主题的特征词并不一定经常共现，因此话题模型挖掘出的话题信息相比语言模型，在某种程度上不能较好地反应科技文献摘要文本的结构。

3.2.3 话题数目对实验结果的影响

对于概率话题模型，所选取的话题数目对实验结果影响较大。本文不研究如何对话题数目进行取值，仅通过选取不同数目的话题，考察其对预测结果的影响。在 Elsevier 数据集上，当话题数目在 200～2000 中取值时，LDA 模型和 PAT 模型预测结果的 F1 值变化如图 4 所示。

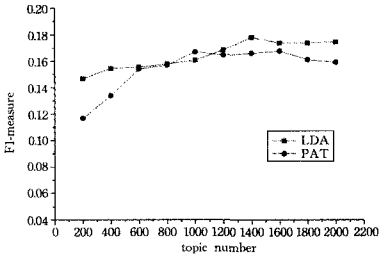


图 4 Elsevier 数据集上话题数目对 LDA 方法和 PAT 方法 F1 值的影响

3.2.4 话题模型和 LDA 模型结合的实验结果

如 3.2.3 节实验结果所示，当 $T=1400$ 时，LDA 模型在 Elsevier 数据集上取得了最好的预测结果。因此，如式(11)所示的方法，这里将语言模型与 $T=1400$ 时的 LDA 模型相结合，λ 取 0 到 1 时的预测结果如表 2 所列。当 $\lambda=0$ 和 1 时，方法即为 LDA 模型和 LM 模型。由表 2 可见，当 λ 取值在 0 到 1 之间时，得到的实验结果均比分别采用两种方法时要好，而当 $\lambda=0.6$ 时，将两者结合达到了最好的效果。

表 2 LM 与 LDA 组合时不同 λ 取值对实验结果的影响

λ	0(LDA)	0.1	0.2	0.3	0.4
F1 值	0.1779	0.2047	0.2062	0.2102	0.2104
0.5	0.6	0.7	0.8	0.9	1(LM)
0.2135	0.2144	0.2132	0.2136	0.2103	0.1993

结束语 本文针对部分科技文献关键词信息缺失的问题，引入社会化标签推荐的方法，采用语言模型、LDA 模型、PAT 模型，在训练集上建立关键词和组成摘要的特征词之间的关系。在给定未标注的科技文献摘要时，预测其可能的关键词。同时，本文还将语言模型和 LDA 模型进行了线性组合用于关键词标注。实验结果表明，本文所采用的方法得到的关键词预测结果能够较好地描述文献内容。而定量的评价结果显示，将语言模型和 LDA 模型相结合，取得了最好的效果。同时，本文选取的关键词预测的方法，训练语料简单易得，同时独立于具体的语言，可以很好地移植到其他数据当中。本文推荐结果的准确率可能还有较大提升空间，如所推荐的关键词仅仅是关键词罗列，不能反映文献的研究领域与子领域等。因此，本文今后的研究方向为如何改善推荐结果，及对训练集合的关键词建立层次关系，进而实现对未标注文本的层次化的关键词推荐等。

参 考 文 献

[1] 索红光，刘玉树，曹淑英. 一种基于词汇链的关键词抽取方法[J]. 中文信息学报，2006，20(6)：25-30

[2] 罗准辰，王挺. 基于分离模型的中文关键词提取算法研究[J]. 中文信息学报，2009，23(1)：63-70

[3] 张雪英，Krause J. 中文文本关键词自动抽取方法研究[J]. 情报学报，2008，27(1)：512-520

[4] 李素建，王厚峰，俞士汶，等. 关键词自动标引的最大熵模型应用研究[J]. 计算机学报，2004，27(9)：1192-1197

[5] Heymann P, Ramage D, Garcia-Molina H. Social Tag Prediction [C]// Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 2008: 531-538

[6] Krestel R, Fankhauser P, Nejdl W. Latent Dirichlet Allocation for Tag Recommendation[C]// Proceedings of the third ACM Conference on Recommender Systems. New York: ACM, 2009: 61-68

[7] Sun Ke, Wang Xiao-long, Chengjie Sun, et al. A language model approach for tag recommendation[J]. Expert Systems with Applications, 2011, 38: 1575-1582

[8] Yin Da-wei, Xue Zhen-zhen, Hong Liang-jie, et al. A probabilistic model for personalized tag prediction[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2010: 959-968

[9] Shepitsen A, Gemmell J, Mobasher B, et al. Personalized recommendation in social tagging systems using hierarchical clustering [C]// Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2008: 259-266

[10] Ponte J M, Croft W B. A language modeling approach to information retrieva[C]// Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1998: 275-281

[11] Zhai Cheng-xiang, Lafferty J. A study of smoothing methods for language models applied to information retrieval [J]. ACM Transactions on Information Systems, 2004, 22(2): 179-214

[12] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022

[13] Steyvers M, Griffiths T. Probabilistic topic models. In Latent Semantic Analysis[C]// Landauer T, Mcnamara D, Dennis S, et al., eds. A Road to Meaning. Lawrence Erlbaum, 2005

[14] Wei Xing, Croft W B. LDA-Based Document Models for Ad-hoc Retrieval[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information and Retrieval. New York: ACM, 2006: 178-185

[15] Griffiths T L, Steyvers M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(1): 5228-5235

[16] Steyvers M, Smyth P, Rosen-Zvi M, et al. Probabilistic author-topic models for information discovery[C]// Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2004: 306-315

[17] Song Yang, Zhuang Zi-ming, Li Hua-jing, et al. Real-time automatic tag recommendation[C]// Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2008: 515-522