山西大学学报(自然科学版)
Journal of Shanxi University(Natural Science Edition)
ISSN 0253-2395,CN 14-1105/N

《山西大学学报(自然科学版)》网络首发论文

题目：基于主题模型的短文本关键词抽取及扩展
作者：曾曦，阳红，常明芳，冯骁骋，赵妍妍，秦兵
DOI：10.13451/j.cnki.shanxi.univ(nat.sci.).2018.05.28.004
收稿日期：2018-05-28
网络首发日期：2018-11-01
引用格式：曾曦，阳红，常明芳，冯骁骋，赵妍妍，秦兵．基于主题模型的短文本关键词抽取及扩展[J/OL]．山西大学学报(自然科学版).
https://doi.org/10.13451/j.cnki.shanxi.univ(nat.sci.).2018.05.28.004

DOI:10.13451/j.cnki.shanxi.univ(nat.sci.).2018.05.28.004

基于主题模型的短文本关键词抽取及扩展

曾曦，阳红，常明芳，冯骁骋，赵妍妍，秦兵

（中国电子科技集团公司第三十研究所，四川 成都，610000）

摘　要：随着互联网不断发展，短文本在互联网数据中的比重不断加大，如何能够快速理解或者检索短文本成为一个重要研究任务。与长文本相比短文本所蕴含的知识更加精炼，因此关键词能更好地表达出短文本内容，但由于篇幅限制，某些短文本会省略一些细节信息，给用户检索信息造成不便。本文主要对短文本关键词抽取及具有丰富文本含义的关键词扩展问题研究，在关键词抽取工作中将文本主题分类信息和词搭配关系引入到传统的 TF-IDF 算法中；在关键词扩展的工作中，通过构建词的特征表示向量，计算文本关键词和类别特征词相似度，从而发现所需扩展的关键词，两方面工作均取得了令人满意的结果。

关键词：关键词抽取；关键词扩展；短文本

中图分类号：TP391.1　　**文献标志码**：A

Topic Model Based Keyword Extraction and Expansion for Short Text

ZENG Xi, YANG Hong, CHANG Mingfang, FENG Xiaocheng, ZHAO Yanyan, QIN Bing

（*China Electronic Technology Group Corporation 30th Research Institute, Chengdu* 610000,*China*）

Abstract： With the development of the Internet, the proportion of short text in Internet data is increasing. An important research task is how to quickly understand or retrieve the short text. Compared with the long text what this contains more refined in short text, therefore, keywords can better express the information of short text. However, due to the limitation of article length, some short text omits some details, causing inconvenience to the user to retrieve information. This paper is mainly about keyword extraction for short text and keyword expansion for rich text meaning. In the keyword extraction, the information of text topic classification and word collocation relationship are introduced to traditional TF-IDF algorithm and the extraction result presents well. In the keyword expansion, through training monolingual word alignment model and LDA topic classification model to build the one dimensional vector about word; and through calculating similarity of keyword and its special topic words to conduct keyword expansion and the result presents well.

Keywords： Keyword Extraction; Keyword Expansion; Short Text

0 引言

关键词（keyword）抽取一直以来都是信息抽取领域内一个重要的研究方向，如同摘要在长文本中所起的重要作用一样，关键词能准确的反映出短文本所要表达的内容，是人们快速了解文档内容、把握主题的重要方式。并且关键词对自然语言处理领域的文本分类和文本

收稿日期：2018-05-28；**接受日期**：2018-10-16
作者简介： 曾曦(1969－)，女，四川成都人，研究员，主要研究领域为认知域对抗技术与应用。
E-mail:zxmm2@163.com

聚类任务有积极作用；同样关键词在信息检索领域也有重要的应用价值。然而在海量的互联网文档中又仅有少部分带有关键词标注，如何给短文本打上一个表意准确的关键词标签成为信息抽取领域的重要问题。

本文提出了一种基于文档主题特征的关键词抽取及关键词扩展方法，系统框架如图 1 所示。首先对短文本进行分词及词性标注等预处理，然后采用 TF-IDF 算法计算出词的初始权重，并且训练短文本的主题模型，得到短文本的分类信息和类别特征词，再采用单语词对齐技术抽取出短文本中的词搭配，之后根据上述信息对关键词权重进行调整，通过阈值筛选出关键词，最后构建词的表示向量，通过计算词与短文本之间的相似度找到与内容信息最贴合的类别特征词作为扩展关键词，建立短文本的关键词集合。

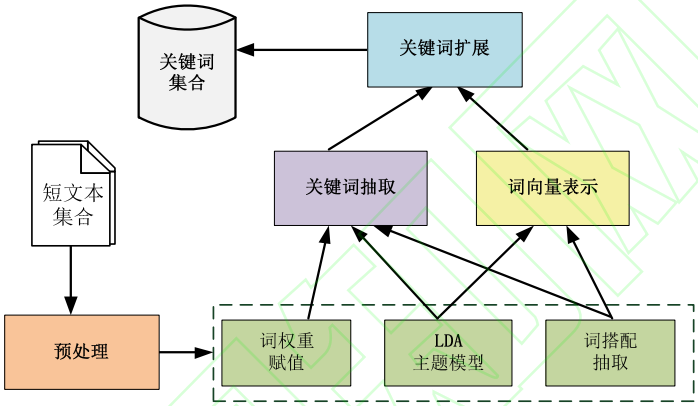

Fig.1 The Architecture of Our System

图 1 系统框图

本文第 1 节介绍相关研究工作，第 2 节介绍关键词抽取的相关内容，第 3 节详细叙述关键词扩展的方法，第 4 节给出实验结果和分析，第 5 节对本文研究工作进行总结。

1 相关研究概述

在关键词抽取研究初期，最常用的方法是通过词的出现频次来获得关键词，然而这种方法所取得的效果并不理想。之后人们采用有监督的机器学习方法来抽取关键词，1999 年 Turney 将关键词抽取问题看成是一个分类问题[1]，通过关键词的出现位置和长度等特征来训练学习，所抽取到的结果要明显优于统计方法得到的结果。Frank 等人将朴素贝叶斯的方法应用在关键词抽取任务中[2]，使得结果有了进一步提升。Hulth 加入了更多的语言学知识[3,4]，如句法特征，在实验结果上获得了一定的成功；但是随着网络数据规模的增加，人工标注数据的工作量变得异常巨大，目前人们主要采用基于图的方法来抽取关键词。2004 年 Mihalcea 和 Tarau 将 PageRank 算法思想带入到了关键词抽取领域[5]，提出了一种基于图的排序算法 TextRank。Litvak 和 Last 将同样用于网页排序的 HITS 算法用于候选关键词排序[6]，在 F 值上取得了一定的提升。Wan 等人通过聚类的方法将相似文档中的知识应用在图模型中[7,8]。Liu 提出基于文档内部信息构建主题的关键词方法[11]，通过计算语义相似度来对候选词进行

聚类，再通过聚类中心词找到合适的关键词，之后 Grineva 将多主题文档的方法应用在构建语义图模型上[9]。Elbeltagy 和 Rafea 创建的 KP-Miner 系统在关键词抽取结果上有着不错的效果[10]。该系统对关键词词频和反文档频率统计提出了更高的要求，并对关键词出现在文章中的位置与其重要性关系进行了分析。2013 年 You 对现有关键词抽取系统进行了总结[12]，并针对前人缺点进行了改进，对候选词的预处理提出了更高的要求。对于图模型的方法而言，训练时间相对较长，无法在短时间内构建索引满足用户需求。

关键词扩展任务可以借鉴查询扩展任务，查询扩展主要为了改善资讯检索召回率，将原来查询语句增加新的关键字来提高查全率和查准率。查询扩展任务分为全局分析[13,14]、局部分析[15-19]、基于用户查询日志[23]和语义相似度计算[24]等几个方面；关键词扩展并不是针对单一的查询语句，而是对大量文本补充关键词，丰富其含义，在构建索引的时候就扩展了数据的内容，而不是在检索的时候扩展查询语句的含义。关键词扩展的方法类似于查询扩展中的全局方法，并采用局部分析中的一些优化策略，使用全部文档蕴涵的相关信息扩展关键词[27-30]；2009 年 Wang 将关键词抽取和扩展应用在聚类任务中[20]，实验结果有一定提升。2014 年 Abilhoa[25]提出一种推文集合的关键字提取方法，它将文本表示为图并应用中心度量来查找相关顶点作为关键词。2017 年 Zhao[26]将神经网络的词向量特征应用于短文本关键词抽取系统，在 textrank 的基础上其实验结果获得一定的提高。与长文本相比短文本的统计特性相对较弱，在抽取关键词任务中所遇到的困难更多。本文所提出的基于主题模型的关键词抽取及扩展方法上与前人有着本质的不同，考虑到了主题分类信息和词搭配信息，关键词抽取效果也更加精确。 并且通过构建词的表示向量来计算词和文本的相似度，从而扩展出关键词，丰富短文本含义。

2 关键词抽取

2.1 概述

本文所采用的基于主题模型的关键词抽取方法主要分为 5 个步骤：（1）预处理，获取初步的候选关键词；（2）关键词赋权，基于改进的 TF-IDF 方法给关键词一个初始权重；（3）LDA 主题模型，根据类别特征词对关键词权重进行调整；（4）词搭配抽取，根据词搭配信息对权重进行调整，（5）根据阈值抽取关键词。图 2 为关键词抽取的一个实例图：

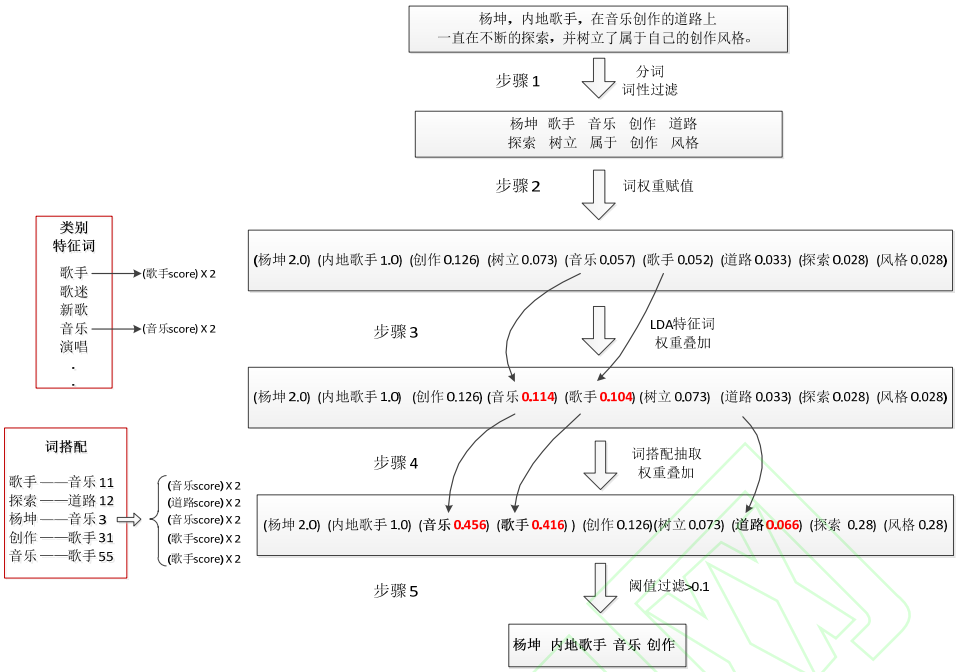

Fig.2 The Process of the Proposed Keyword Extraction

图 2 关键词的抽取过程

2.2 关键词初始权重赋值

本文首先通过文本分词，词性标注和停用词等方法获得候选关键词，如图 2 中步骤 1，去掉“一直”、“属于”等词。

2.2.1. 基于 TF-IDF 的关键词赋权

TF-IDF 是一种统计方法，用以评估字词对于一个文件集或一个语料库中的其中一份文件的重要程度。本文基于汉语中词语长度与词语重要程度存在一定关系，对原有 TF-IDF 算法做出了改进，通过公式(1)对候选关键词打分，获得候选关键词的基本权重值。

$$
\begin{aligned}
Score_{t_i} &= tf_{i,j} \times idf_i \times len(t_i) \\
tf_{i,j} &= \frac{n_{i,j}}{\sum_k n_{k,j}} \\
idf_i &= \log \frac{|D|}{|\{j : t_i \in d_j\}|}
\end{aligned}
\tag{1}
$$

上式中 $Score_{t_i}$ 为 t_i 的最终权重值， $tf_{i,j}$ 表示词频，指的是某一个给定的词语在该文件中出现的频率。 idf_i 表示逆向文档频率，表示是一个词语普遍重要性的度量， $len(t_i)$ 为词语 t_i 的字节长度。 $n_{i,j}$ 是词 t_i 在文件 d_j 中出现的次数，而分母则是文件 d_j 中出现所有字词的出现次数之和； $|D|$ 是语料库中文件总和， $|\{j : t_i \in d_j\}|$ 表示包含词语 t_i 的文件总数，计算结果如图 2 中步骤 2 所示。

2.2.2. 基于规则的关键词赋权

通过观察数据我们发现，在每一条短文本中有一些特殊的字词可以直接作为关键词，这

些字词往往可以直接表达该文本的某些特定信息，因此本文在 TF-IDF 的基础上采用下列规则抽取一些字词作为候选关键词，并直接打上一定分数，用以表达这类关键词的特殊性，规则如下：

(1) 根据书名号或括号抽取书名、歌曲名等作为候选关键词，如“赵薇主演过《还珠格格》《情深深雨蒙蒙》”，其中“还珠格格”、“情深深雨蒙蒙”的权重值如公式(2)所示：

$$Score_{t_i} = 2.0 \tag{2}$$

(2) 根据此类文本的特殊性，抽取一些短标题直接作为候选关键词，如图 2 中的“杨坤”，其权重值如公式(3)所示：

$$Score_{t_i} = 2.0 \tag{3}$$

(3) 根据共现信息将一些词合并成常见短语，常见短语就是人们在日常生活中经常能够看到或者使用到的短语，如图 2 中的“内地歌手”，其权值如公式(4)所示：

$$Score_{t_i} = 1.0 \tag{4}$$

2.3 基于LDA主题模型的关键词赋权

LDA (Latent Dirichlet Allocation，隐含狄利克雷分配[21]）主题模型是近年来在中文信息处理领域发展起来的一种生成主题概率模型，它基于一定的常识性假设：文档集中所有文档均按照一定比例共享隐含主题集合，而隐含主题集则是由一系列相关特征词组成。LDA 模型定义每篇文档均为隐含主题集的随机混合，从而可以将整个文档集特征化成隐含主题的集合。

本文将大规模短文本用 LDA 主题模型进行聚类，通过类别信息来进行关键词表示，为关键词扩展中的相似度计算提供数据；并通过主题模型得到每个类别下的主题特征词，将这些特征词作为关键词抽取中的一个权重打分标准，其具体公式如下：

$$\begin{aligned} &Score1_{t_i} = 2^{\tau_1(t_i)} \times Score_{t_i} \\ &\tau_1(t_i) = \begin{cases} 1 & if \ \ t_i \in S \\ 0 & other \end{cases} \\ &S = < s_0, \cdots s_i, \cdots s_n > \end{aligned} \tag{5}$$

其中 $Score1_{t_i}$ 为词语 t_i 当前权重，$Score_{t_i}$ 为上一节中给词语 t_i 所赋的权重，S 为类别特征词集合。如果候选关键词 t_i 是类别特征词，则权重加倍。权重修改结果如图 2 中步骤 3 所示。因为“音乐”和“歌手”都出现在特征词列表中，所以其权重加倍。

2.4 基于词搭配的关键词赋权

搭配（Collocation）一般被定义为词和词在一起的概率要远大于一般随机出现的概率，在汉语中常用的搭配“影视明星”、“室内装修”等等。本文认为搭配对中的两个词往往具有

一定的语义联系，例如“影视”和“明星”间是存在一定的潜在联系，这些词可以互相表达、相互支持，希望通过这些搭配来形成一种新的关键词抽取方法。

本文采用的搭配抽取模型为单语词对齐模型（MWA, monolingual word alignment），单语词对齐是仿照双语词对齐的一类计算任务，通过统计计算出同一语言中关系相近的不同搭配。Liu[22]分别修改了 IBM model 1, model 2 以及 model 3，使得相同的词之间不能互译，最终抽取出的搭配，来自于三种翻译模型词互译结果的融合。

本文将通过词搭配对关键词权重再次进行调整，因为词搭配中蕴含着一定的语义关系，我们认为一条文本中如果两个候选关键词构成词搭配关系，并且该词搭配的频次超过一定阈值，则认为该词搭配中的候选关键词相比于其它词语更加重要，因为词搭配中的词是存在先后关系的，当一条文本中出现两个候选关键词组成词搭配时，我们指对第二个候选关键词的权重进行加倍，通过找到文本中的不同词搭配，使得部分候选关键词权重发生变化，经过再次排序可以将排名靠前的候选关键词作为关键词输出，其权重变化如公式 6 所示：

$$
\begin{aligned}
&Score2_{\mathrm{t}_j} = 2^{\eta} \times Score1_{\mathrm{t}_j} \\
&\eta = \sum_{i=0}^{n} \tau_i(\mathrm{t}_i, \mathrm{t}_j) \\
&\tau_i(\mathrm{t}_i, \mathrm{t}_j) = \begin{cases} 1 \;\; if \;\; (\mathrm{t}_i, \mathrm{t}_j) \in Pair(T, \mathrm{t}_j) \\ 0 \;\; other \end{cases} \\
&T = < \mathrm{t}_0, \cdots \mathrm{t}_i, \cdots \mathrm{t}_n > \quad \mathrm{t}_j \notin T
\end{aligned}
\tag{6}
$$

其中 t_i 和 t_j 是文本中的候选词，$\tau_1(\mathrm{t}_i, \mathrm{t}_j)$ 为一个二值函数，如果 t_i 和 t_j 构成以 t_j 为第二个词的词搭配，则 t_j 的权重就增加一倍，如果不构成词搭配，则权重无变化。T 为与 t_j 构成搭配对关系的候选关键词集合。

权重修改结果如图 2 中步骤 4 所示，文本中“歌手”和“音乐”组成词搭配，因为词搭配具有先后关系，本文只对词搭配中的第二个关键词进行权重调整，所以“音乐”的权重加倍一次，并且“杨坤”和“音乐”也组成了词搭配关系，所以“音乐”的权重再次翻倍，通过不断叠加，“音乐”的权重变为最初的八倍。

最后重新排序，根据阈值将排序结果靠前的词作为关键词输出。

3 关键词扩展

3.1 词向量表示

词向量表示一直是机器学习问题在自然语言处理领域中的一个重要研究方向，最常用的词表示方法是 Bag-Of-Words，该方法把词表示成一维向量。这个向量的维度是词表大小，其中绝大多数元素为 0，只有一个维度的值为 1，这个维度就代表了当前的词，该表示方法相对简单，但是该方法存在着两个主要问题，一是所需存储的向量维度相对较大；二是存在

很严重的数据稀疏问题。使用该方法计算相似度时还需要统计共现信息，较为繁琐。本文给出一种不同于上述方法的词向量表示机制，并且包含一定的语义信息。

本文所提出的词向量表示方法主要是根据文本类别信息得到的，对文本使用 2.3 节的 LDA 主题模型进行分类，之后将每个 Topic 下的类别特征词用一维特征向量进行表示，该一维向量的维度即文本的分类个数，其元素的含义表示该词是否为该文本类别下的特征词，对于赋值而言，若该类别不含该特征词，则向量中的该元素为 0，若类别特征词中含有该词，则对应的向量维度为该类别下的特征词的概率，基于上述表示机制我们可以得到所有特征词的向量表示，具体形式见公式(7)：

$$\vec{w} = < p_0(\mathrm{w}), \mathrm{p}_1(\mathrm{w}), \cdots p_i(\mathrm{w}), \cdots p_n(\mathrm{w}) > \tag{7}$$

其中 i 是指 LDA 模型的类别体系，w 为主题分类下每个类别中的特征词，$P_i(\mathrm{w})$ 表示词 w 出现在 LDA 模型类别 i 中的概率。

如果只对类别特征词进行词向量表示，所能够被表示的词数量太少，因此本文提出一种词向量传递机制，通过词搭配将类别特征词的向量传递到候选关键词上，使更多的词可以被表示，对此我们构建了一个公式：

$$\begin{aligned} &\vec{w}_{new} = \vec{w}_{old} + \sum_{(w,v_i)\in l(w,v)} \left(\frac{freq(w,v_i)}{freq(w)} \times \vec{v}_i\right) \\ &l(w,v) = <(w,v_0),(w,v_1),\cdots(w,v_n)> \end{aligned} \tag{8}$$

其中 $l(w,v)$ 表示词 w 所有的词搭配组合，(w,v_i) 表示 w 与 v_i 的词搭配，$\overrightarrow{w_{old}}$ 表示 w 初始的词向量，$\overrightarrow{w_{new}}$ 表示新求得的 w 词向量，$\vec{v}_i$ 表示词 v_i 的词向量，$freq()$ 代表统计出现的频次。

3.2 关键词扩展

通过 2.3 节训练的 LDA 模型，我们可以知道每一条文本所属的具体类别，并且每一个类别含有一些特征词。本文所提出的关键词扩展策略是计算文本关键词与类别特征词之间的相似度，再根据排序结果和一些统计规律将相似度排名靠前的类别特征词作为该文本的扩展关键词输出，具体公式如下：

$$\begin{aligned} &P_{w_i,w_j} = \tau_1(w_i,w_j) \times \tau_2(w_i,w_j) \times \sum_{a\in l} sim_{\cos}(\vec{w}_i,\vec{w}_j) \times p(\mathrm{id}) \\ &\tau_1(w_i,w_j) = \begin{cases} 1 \ \ if \ \ freq(w_i) > freq(w_j) \\ 0 \ \ other \end{cases} \\ &\tau_2(w_i,w_j) = \begin{cases} 0 \ \ if \ \ < w_j, w_i > \in Bigram(w_j,w_i) \\ 1 \ other \end{cases} \\ &w_i \in T_i, \ T_i = < w_0, w_1, \cdots >, \qquad w_j \in C_j, \ C_j = < s_0, \mathrm{s}_1, \cdots > \end{aligned} \tag{9}$$

其中 w_i 为文本中的关键词，w_j 为分类体系中的类别特征词，$\vec{w}_i, \vec{w}_j$ 为词的特征向量，

$sim_{\cos}(\vec{w_i},\vec{w_j})$ 是指两个向量 $\vec{w_i},\vec{w_j}$ 的 *cos* 相似度， $p(\mathrm{id})$ 为文本分类概率，其中 $\tau_1(w_i,w_j)$ 和 $\tau_2(w_i,w_j)$ 为两个二值函数，均表示扩展的方向性，其中 $\tau_1(w_i,w_j)$ 代表频次关系，表达含义是将具体的词扩展出泛化的词，而不是将泛化的词扩展出来具体的专有词，例如：万达可以扩展出来商场，反之不可以； $\tau_2(w_i,w_j)$ 代表 Bigram 的顺序，在所有统计的 Bigram 列表中，后者均不能扩展出前者，例如英孚英语，英孚扩展出来英语，而英语扩展出英孚。

4 实验

本文使用 100 万微信公用账号简介作为短文本数据，该数据包含微信公用账号名称及相关简介，如图 2 所示。

4.1 关键词抽取实验

对于从内容中抽取关键词的实验结果，我们采用人工构建测试集方法进行评价，依然按照准确率、召回率和 F 值进行评测。在这里我们将传统的 TF-IDF 算法作为 Baseline，将实验结果与 Wang[20]和 TextRank[5]进行对比，随机抽取 500 条短文本作为测试数据，并人工标注了 4135 个关键词作为关键词抽取的测试集，其实验结果如下：

表 1 关键词抽取对比实验

Table 1 Experiments of Keywords Extraction

Method	Extract		Right		Precision	Recall	F-Measure
	Total	Average	Total	Average			
TF-IDF	4978	10.0	2688	5.4	0.54	0.65	0.5899
Wang	4319	8.6	2505	5.0	0.58	0.61	0.5898
TextRank	3697	7.4	2440	4.88	0.66	0.59	0.6230
Full Model	3808	7.6	2894	5.79	**0.76**	**0.70**	**0.7287**
- Rule	3626	7.2	2438	4.88	0.67	0.59	0.6271
- Topic	3945	7.9	2821	5.64	0.71	0.68	0.6947
-Collocation	3539	7.1	2407	4.77	0.68	0.58	0.6260

通过上表可以看到，在准确率、召回率和 F 值三个测试指标中，本文方法均取得了最优的实验效果，其中 average 是指一条短文本平均能抽取几个关键词；从表(1)可以看到，本文方法所取得准确率和 F 值基本上都比第二名高出 10%左右，并且召回率也有小幅提高；从上述实验结果可以看出，本文所提出的基于词搭配信息的关键词抽取方法是真实有效的，在运用统计知识的基础上考虑到了具有语义联系的词搭配信息，因此取得了相对好的实验结果。最终在 1009713 条实验数据中，共对 978716 条文本抽取到关键词，对于没有获得关键词的文本主要是因为其描述采用英文或者繁体字。

4.2 关键词扩展实验

本文方法 KEK（KEYWORD-EXPEND-KEYWORD）扩展出来的关键词，我们~~依然~~采用准确率、召回率和 F 值进行评测，但是有所不同的是并不构建测试集，因为一篇文本人们通过想象扩展出来的关键词会存在很大的差异性，所以我们采用人工的方法来看文本扩展出的关键词是否正确；由于不存在测试集，在召回率上则更加偏重对扩展能力的评价，在召回率上我们则随机抽取一定量的文本数据，通过统计这些短文本中有多少扩展出新的关键词

来计算召回率，公式如下：

$$Recall(\mathrm{id}) = \frac{expend(\mathrm{id})}{all(\mathrm{id})} \quad (10)$$

expend(id)为扩展出关键词的短文本数量，*all*(id)为参与实验的短文本数量，*Recall*(id)本节召回率计算结果。在本文实验中将 *all*(id)设为 500.

并且针对不同规模的短文本进行对比实验，其具体的实验结果如下：

表 2 不同数据集下的关键词扩展

Table 2 Keywords Expansion of Different Data Sets

Data Sets	Keywords(Precision)			ID(Recall)		Precision	Recall	F-Measure
	Total	Average	Right	Total	Expand			
50000	342	2.14	195	500	160	0.57	0.32	04098
200000	439	2.31	281	500	191	0.64	0.38	0.4768
500000	596	2.56	429	500	233	0.72	0.46	0.5613
1000000	795	2.89	652	500	275	**0.82**	**0.55**	**0.6583**

通过表(2)可以看到，我们在随机抽取的 500 篇文档中给 275 篇短文扩展出了关键字，并且共扩展出 795 个关键字，正确的 652 个，并通过人工测评的方法计算了准确率，从表(2)可以看出，准确率曲线和召回率曲线均成上升趋势，因为训练数据越多，主题模型训练的越充分，分类更加准确，所以关键词扩展的效果越好。

我们还与 Wang 的方法进行了对比，他的方法主要是文本中找到同义词进行替换，在英文领域采用的是 Word-Net 上的同义词替换资源，我们将同样的方法移植到中文上，由于 Word-Net 上没有中文资源，这里我们采用哈尔滨工业大学构建的《同义词词林》进行替换；为了说明关键词抽取的重要性，我们将本文的关键词扩展策略进行修改，提出了一种基于全文本的关键词扩展方法 AWEK（ALL-WORD-EXPEND-KEYWORD），该方法与第三章所阐述的扩展方法略有不同，不在只与文本中的关键词计算相似度，而是将所有候选词作为扩展依据计算相似度，将本文方法与上述两种方法相对比，将 100 万条短文本作为训练语料进行对比实验，实验结果如下表(3)：

表 3 关键词扩展对比实验

Table 3 Experiment of Keywords Expansion

Data Sets	Keywords(Precision)			ID(Recall)		Precision	Recall	F-Measure
	Total	Average	Right	Total	Expand			
Wang	1275	2.95	574	500	432	0.45	**0.86**	0.5908
AWEK	902	2.91	523	500	310	0.58	0.62	0.5993
KEK	795	2.89	652	500	275	**0.82**	0.55	**0.6583**

上表可以看出，在三组实验中，本文方法取得了最优的准确率，并且 F 值也要高出其它方法 5 个百分点，通过该实验说明短文本中如果只采用简单的同义词来扩展关键词，虽然会对很多短文本都打上扩展标签，但是由于同义词扩展出的关键词并一定能具有文本所要表达的含义，所以准确率并不高；而第二种基于全文本的相似度计算扩展方法，由于文本存在着大量噪声词，这些词在做关键词扩展任务中具有很强的干扰作用，使得扩展结果与原文语义发生很大偏差，所以所取得扩展结果也并不理想；而本文方法之所以取得了相对较好的结

果，是因为只基于文本关键词计算相似度，文本中的关键词基本上都与文本语义保持一致，所以扩展出来的关键词不会有太大偏差，效果相对理想。表 4 给出了本文方法的相关实例：

表 4 关键词抽取与扩展实例

Table 4 Examples of Keywords Extraction and Keywords Expansion

Text	Extracted Keyword	Expanded Keyword
吴金洪 冷笑话，热心肠！每日奉送冷笑话，快乐精华分享平台…	冷笑话 吴金洪 热心肠	糗事 爆笑 段子搞笑
韩美专业祛痘 杭州韩美专业祛痘，我们承诺：签约服务，无效退款。	签约 祛痘 杭州 韩美	美妆 护肤 肌肤
湘财证券杭州教工路营业部湘财证券杭州教工路营业部资讯公众平台 投资好助手。	证券 湘财 营业部 投资 咨询	行情 财富 股票
哈尔滨红黄蓝亲子园亲子教育，早期教育、早教、儿童教育、婴幼儿早教、父母交流、亲子交流、一体化教育	早教 婴幼儿 亲子园 教育 父母 红黄蓝 儿童 哈尔滨	母婴 玩具 亲子 素质

5 结论

本文介绍了短文本关键词抽取和扩展的具体方法。在关键词抽取任务中，采用主题分类和词搭配信息抽取关键词，取得了较好的实验结果；在关键词扩展任务中，定义了一种基于 LDA 主题分类结果的词向量表示机制，这种表示机制具有一定的语义信息，并且更加节约空间开销，最终的关键词扩展结果也非常理想；而且本文对搜索引擎系统提出了一条新的改善思路，不同于传统的查询扩展工作，不再只对文本内容构建索引，而是通过关键词标签对其内容进行语义上的丰富，扩大索引集合，以提升搜索引擎系统的查全率和查准率。

参考文献

[1] Turney P. Learning to Extract Keyphrases from Text[R]. National Research Council Canada, Institute for Information Technology, Technical Report 1999.

[2] Frank E, Paynter G W, Witten I H, *et al.* Nevill-Manning. Domain-specific Keyphrase Extraction[C]// In Proceedings of the 16th International Joint Conference on Artificial Intelligence, 1999. 668–673. DOI: 10.1145/1099554.1099628

[3] Hulth A. Improved Automatic Keyword Extraction Given More Linguistic Knowledge[C]//In Proceedings of the 2003 conference on Empirical methods in natural language processing, 2003. 216–223. DOI: 10.3115/1119355.1119383

[4] Hulth A. Reducing False Positives by Expert Combination in Automatic Keyword Indexing[C]//Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003, 367. DOI: 10.1075/cilt.260.41hul

[5] Mihalcea R, Tarau P. Textrank: Bringing Order into Texts[C]//In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004.

[6] Litvak M, Last M. Graph-based Keyword Extraction for Single-document Summarization[C]//Proceedings of the workshop on multi-source multilingual information extraction and summarization. Association for Computational Linguistics, 2008: 17-24. DOI: 10.3115/1613172.1613178

[7] Wan X, Xiao J. CollabRank: towards a Collaborative Approach to Single-document Keyphrase Extraction[C]//Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, 2008: 969-976. DOI: 10.3115/1599081.1599203

[8] Wan X, Xiao J. Exploiting Neighborhood Knowledge for Single Document Summarization and Keyphrase Extraction [J]. *ACM Transactions on Information Systems*, 2010, **28**(2): 1-34. DOI: 10.1145/1740592.1740596

[9] Grineva M, Grinev M, Lizorkin D. Extracting Key Terms from Noisy and Multi-theme Documents[C]//Proceedings of the 18th international conference on World wide web. ACM, 2009: 661-670. DOI: 10.1145/1526709.1526798

[10] El-Beltagy S R, Rafea A. KP-Miner: A Keyphrase Extraction System for English and Arabic Documents[J]. *Information Systems*, 2009, **34**(1): 132-144. DOI: 10.1016/j.is.2008.05.002

[11] Liu Z, Li P, Zheng Y, *et al*. Clustering to Find Exemplar Terms for Keyphrase Extraction[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics, 2009: 257-266. DOI: 10.3115/1699510.1699544

[12] You W, Fontaine D, Barthès J P. An Automatic Keyphrase Extraction System for Scientific Documents[J]. *Knowledge and information systems*, 2013: 1-34. DOI: 10.1007/s10115-012-0480-2

[13] Deerwester S, Dumai S T, Furnas G W, *et al*. Indexing by Latent Semantic Analysis[J]. *Journal of ACM Transactions on Information Systems*, 1990, *41*(6):391-407. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

[14] Qiu Y, Frei H. Concept Based Query Expansion[M]//Korfhage R, Rasmussen E M, Willett P, eds. Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1993. 160~169. DOI: 10.1145/160688.160713

[15] Buckley C, Salton G, Allan J, Singhal A. Automatic Query Expansion Using SMART[J]. Technical Report, TREC-3, 1995. 69-80.

[16] Ricardo B Y, Berthier R N. Modern Information Retrieval[M]. England: Pearson Education Limited, 1999.

[17] Attar R, Fraenkel A S. Local Feedback in Full-text Retrieval Systems[J]. *Journal of the ACM*, 1977,**24**(3):397-417. DOI: 10.1145/ 322017.322021

[18] Xu J X, Croft W B. Improving the Effectiveness of Information Retrieval with Local Context Analysis[J]. *ACM Transactions on Information Systems*, 2000,**18**(1):79-112. DOI: 10.1145/333135.333138

[19] Yahia S B, Jaoua A. Discovering Knowledge from Fuzzy Concept Lattice[C]//Data mining and computational intelligence. Physica-Verlag GmbH, 2001: 167-190. DOI: 10.1007/978-3-7908-1825-3_7

[20] Wang J, Zhou Y, Li L, *et al.* Improving Short Text Clustering Performance with Keyword Expansion[C]//The Sixth International Symposium on Neural Networks (ISNN 2009). Springer Berlin Heidelberg, 2009: 291-298. DOI: 10.1007/978-3-642-01216-7_31

[21] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. *the Journal of machine Learning research*, 2003, 3: 993-1022. DOI: 10.1162/jmlr.2003.3.4-5.993

[22] Liu Z, Wang H, Wu H, *et al*. Collocation Extraction Using Monolingual Word Alignment Method[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. Association for Computational Linguistics, 2009: 487-495. DOI: 10.3115/1699571.1699575

[23] 崔航, 文继荣, 李敏强. 基于用户日志的查询扩展统计模型[J]. 软件学报, 2003, 14(9).

[24] 田萱, 杜小勇, 李海华. 语义查询扩展中词语-概念相关度的计算[J]. 软件学报, 2008, 19(8): 2043-2053.

[25] Abilhoa W D, De Castro L N. A Keyword Extraction Method from Twitter Messages Represented as graphs[J]. Applied Mathematics and Computation, 2014, 240: 308-325. DOI: 10.1016/j.amc.2014.04.090

[26] Zhao D, Du N, Chang Z, *et al*. Keyword Extraction for Social Media Short Text[C]//Web Information Systems and Applications Conference (WISA), 2017 14th. IEEE, 2017: 251-256. DOI: 10.1109/wisa.2017.12

[27] He G X, Fang J W, Cui H R, *et al*. Keyphrase Extraction Based on Prior Knowledge[C]// Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, 2018: 341-342. DOI: 10.1145/3197026.3203869

[28] Mahata D, Kuriakose J, Shah R R, *et al*. Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 634-639. DOI: 10.18653/v1/n18-2100

[29] Chen W, Liu Z, Shi W,*et al.* Keyphrase Extraction Based on Optimized Random Walks on Multiple Word Relations [C]// Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data, 2018: 359-367. DOI: 10.1007/978-3-319-96893-3_27

[30] Figueroa G, Chen P C, Chen Y S. RankUp: Enhancing Graph-based Keyphrase Extraction Methods with Error-feedback Propagation[J]// Computer Speech & Language, 2018, **47**: 112-131. DOI: /10.1016/j.csl.2017.07.004