中文文本的主题关键短语提取技术

杨 玥 张德生
(西安理工大学理学院 西安 710054)

摘 要 在大数据时代,信息量暴增,人们接触最多的信息就是文本信息,每天在互联网上都有无数文本信息被上传或下载。快速掌握这些文本信息内容的重要方法之一就是关键词提取。然而,在传统关键词提取算法中,通常忽略了两个重要的方面:词语长度和文本主题。针对以上两方面问题,提出了提取中文文本的主题关键短语技术。将LDA主题模型与频繁短语发现算法相结合,生成不同长度的频繁候选短语;然后,利用所提的完整性筛选和排序函数对候选短语进行筛选和排序;最后,根据排序结果选择最终的主题关键短语。

关键词 关键词提取,LDA主题模型,频繁短语,完整性筛选,排序函数
中图法分类号 TP311 **文献标识码** A

Technology of Extracting Topical Keyphrases from Chinese Corpora

YANG Yue ZHANG De-sheng
(School of Science, Xi'an University of Technology, Xi'an 710054, China)

Abstract In the big data era, the information is exploding. The most popular information among people connection is text message. On the Internet, there are countless text information upload or download every day. The important way to quickly grasp content of countless text message is extracting keywords. However, the traditional work of extracting keywords from text corpora ignores two problems: the length of keywords and the topic of text corpora. In this paper, a new algorithm which is in consideration of two aspects mentioned above was proposed. This paper combined the LDA topic model and frequent phrases discovery algorithm to generate frequent candidate phrases with different length, at the same time, this paper proposed an algorithm of completeness filter and rank function to filt and rank candidate. Finally, according to the rank list, the real keyphrases were chosen.

Keywords Extracting keywords, LDA topic model, Frequent phrases, Completeness filter, Rank function

1 引言

文本挖掘是数据挖掘中一个非常重要的方面,并且文本是与人们接触最多,也最直观的一种数据类型。文本挖掘的一个核心任务就是信息抽取,而关键词提取是信息抽取中的一个重要技术。关键词提取不仅能够快速掌握文本的主要内容,随着文本挖掘技术的不断发展,关键词提取也有了更广泛的应用,例如利用关键词来进行文本聚类、情感判别等。

与传统关键词提取技术相比,本文提出的中文文本主题关键短语提取技术有以下几点创新和优势:

1)将LDA主题模型与频繁短语发现算法结合,能够在综合文本集中发现不同长度的频繁短语;

2)优化排序函数,加入排序因子,对候选频繁短语进行剔除和排序,使排序结果更加符合人们的理解;

3)主要针对中文文本的关键短语提取,能够在一定程度上消除由于中文分词造成的错误。

本文第2节中主要介绍本文算法的优势与流程步骤;第3节按照步骤分别介绍算法流程;第4节对本文算法进行实证研究,并分析结果。

2 算法综述

本文的主要内容如下:

1)将LDA主题模型与频繁短语发现算法结合,推测综合文本集中隐含的主题以及各个主题中的频繁短语;

2)提出了完整性约束算法,对候选关键短语进行筛选,保留完整短语,剔除不完整短语,避免提取结果中存在关键短语及其子短语同时出现的情况,提高提取精准率;

3)优化了排序函数,使关键短语的排序更符合人们的理解,对主题的主要内容更具代表性;

4)对本文提出的算法进行了实证研究。

本文算法的流程如下:

1)文本数据预处理:中文分词、消除停用词;

2)利用LDA主题模型进行文本聚类,对综合文本集合中的每一篇文档指派一个主题标号;

3)设置最小支持度阈值,在各个主题类中发现频繁短语,生成候选主题关键短语集合;

4)利用完整性约束筛选候选主题关键短语,保留完整候选短语,剔除不完整候选短语;

杨 玥(1991-),女,硕士生,主要研究方向为数据挖掘算法及关键词提取技术,E-mail:yangvic@126.com;张德生(1964-),男,博士,硕士生导师,主要研究方向为数理统计及应用,E-mail:zhangds@xaut.edu.cn。

5)计算所有保留候选短语的 Rank 排序函数得分并排序;

6)根据得分排名,选择前 5 或前 10 个关键短语作为主题关键短语。

3　中文文本主题关键短语提取技术

3.1　数据预处理

对中文文本进行数据预处理的过程主要有两部分:1)中文分词;2)去除停用词。

在形式上,中文文本与英文文本有很大区别,中文的词与词之间没有间隔。为了使计算机能够识别词语,须对中文文本进行分词预处理。分词方法以分词词典为依据,通过文本中的汉字串与词典中的词逐一匹配,完成词语切分。

文本集中几乎都会包含一些没有意义但使用频率极高的词,如"的""就"等。这些词在所有文本中的频率非常相近,从而增加了文本之间的相似程度,给文本聚类或提取关键词均带来了一定困难。解决该问题的方法是利用停词表或禁用词表将这些词语从文本中剔除。

常用的停词包括实词和虚词两种,例如:

虚词:"的""把""被""就"…

实词:"有""会"

3.2　文本聚类——LDA 主题模型

文本数据预处理之后形成一种中间形式,这种形式是对文本进行各项算法的基础。

本文提出的算法的第二步是对文本集中的每篇文档及每个词语进行聚类。在自然语言处理中,常用于文本聚类、文本分类等领域的模型主要有:向量空间模型(Vector Space Model)、统计语言模型(Statistical Language Model)、主题模型(Topic Model)等。其中,主题模型是一个数据挖掘与分析的有力工具,它挖掘出文本中潜在的主题信息。

判断两个文档是否相似,传统意义上是观察两个文本是否会出现共同的词汇。如果两个文本共现的词语交集多,则认为两个文本比较相似,否则认为不相似。比如:

"阿里巴巴上市了。"

"马云成为了中国内地首富。"

上述两个文本具有相似的特点,但却没有共同的词语。如果用传统的文本相似的计算方法将文本映射到向量空间中,计算两个文本词汇的共现,则上述例子文本的相似度为 0。因此,需要从主题词的方面来考虑文本的相似度。

2003 年,D. Blei 等人提出了潜在狄利克雷分配模型 LDA(Latent Dirichlet Allocation)。LDA 主题模型是在向量空间模型和统计语言模型的基础上改进了 PLSI 模型,利用更富有表现力的主体层来表示文本表达式,形成了一种语义上一致的话题模型。LDA 模型是目前非常流行的主题模型之一,被广泛应用于机器学习、自然语言处理、文本挖掘、知识发现等多个领域。

LDA 主题模型的目的是生成主体分布,而这个分布在文本集中是不能直接观测到的隐变量,因此这种主题模型被称为潜在狄利克雷分配。LDA 主题模型包含 3 个基本元素:文档(d)、文档中的词语(w)以及主题(z),这 3 个基本元素通过两层分布联系起来,即"文档-主题"分布和"主题-词语"分布。"文档-主题"这一层分布的意思是一篇文档可以表示成不同主题所构成的概率分布,"主题-词语"层分布的意思则是每个主题又是由各个词语所构成的概率分布。LDA 主题模型的图模型结构(类似贝叶斯网络结构)如图 1 所示。

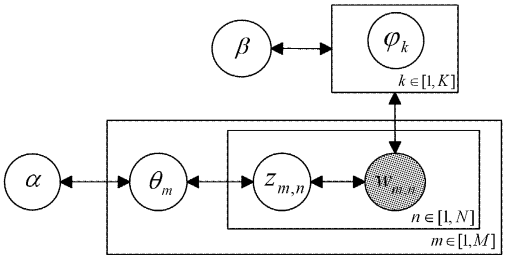

图 1　LDA 主题模型的贝叶斯网络图

图 1 中的箭头表示变量依赖关系,方框表示重复抽样,方框右下角的字母 K,M,N 表示方框中过程重复的次数,在这里的实际意义分别是主题个数、文档个数和文档中的词语个数。阴影圆圈 w 表示可观测到的变量词语,其他白色圆圈都是不可观测到的隐变量。其中,z 表示主题,θ 表示文档-主题分布,α 是主题分布 θ 的先验分布的参数,φ 表示主题-词分布,β 是主题-词分布 φ 的先验分布的参数。通过学习 θ 和 φ 这两个参数,可以计算训练文本集中文档的主题概率分布和主题中的词语概率分布。一般通过 Gibbs 抽样学习这两个参数。Gibbs 抽样是一种迭代方法,其优点是易于实现,而且可以高效地在大规模文本集中抽取主题。

得到 θ 和 φ 这两个分布之后,进而就可以得到综合文本集合中每篇文档对应每个主题的权重,按照权重大小给文档指派主题标号。由此,可以将一个文本集合按照主题聚类成几个子集。

3.3　生成候选——频繁短语发现算法

频繁短语发现算法的前提是:

定义 1(短语)　本文认为短语在计算机中可以定义为,一组连续且在文本中频繁共现的词语的集合。

性质 1(向下封闭性)　如果一个短语不是频繁的,那么包含它的任何超集一定都不是频繁的。因此,包含这个非频繁短语的更长的短语就会被剔除,并且不会再做任何拓展。

性质 2(前缀性)　如果一个短语是频繁的,那么以它为前缀组成的更长的短语也有可能是频繁的。

算法 1　频繁短语发现算法

```
输入:同一主题标号的文档语料库 C,最小支持度阈值 ζ
输出:字典 f(主题中的频繁短语:频率)
  dictionary f←ϕ //创建一个字典 f 并初始化
  dictionary index←ϕ //创建一个字典 index 并初始化
  for i←1 to |C| do
    index[C[i]]←index[C[i]]∪i
  end for
  while Any |index.value| of key>=ζ do
    index′←ϕ //创建一个字典 index′并初始化
    for u∈index.key do
      if |index[u]|≥ζ then
        f[u]←|index[u]|
        for j∈index[u] do
            if j+1>=|C|
              continue
            end if
            u′←u⊕C|j+1|
            index′[u′]←index′[u′]∪{j+1}
```

```
            end for
          end if
      end for
        index←index′
end while
return f
```

3.4 筛选排序

3.4.1 完整性约束筛选

完整性约束的实质是条件概率问题，如果短语 p 的超集为 $p'=p\cup\{w_i\}$，这里的$\{w_i\}$表示超集 p'中除去子集 p 的词语的集合，那么条件概率 $P=(e_t(p')|e_t(p))$表明当短语 p 出现时超集 p'也出现的概率。这个条件概率越大，说明短语 p 不完整的可能性越大。在同一个主题的文本集合中，这实际上是出现频次的比较问题。通常情况下，随着短语长度的增加，它在文本中出现的频次会减少。那么，这个条件概率问题可以转变为：短语 p 在文本集合中出现的频次与超集 p'出现的频次相比较，如果 $f_t(p)-f_t(p')\leqslant\gamma$，也就是短语 p 与它的超集 p'在文本中出现的频次之差小于等于 γ，那么说明短语 p 可能不够完整，它的超集 p'可以作为更完整的短语。这里的 γ 是一个可接受的完整程度调节阈值。通过大量实验，$\gamma\approx(f_t(p))\times50\%$。

按照这种思想，依据完整性指标筛选完整短语的步骤如下：

步骤 1 寻找并收集候选短语中所有长度为 1 的词语的所有前缀为该词的超集，存入 Map $S^{(1)}$；

步骤 2 选择 $S^{(1)}$中超集个数在 2 个或 2 个以上的项，并将这些项的键与值存入 $S^{(2)}$；

步骤 3 计算 $S^{(2)}$中所有键对应的值中每个集合的大小，并找到值中最大的超集，将最大超集本身作为键，它的所有子集作为值，存入 $S^{(3)}$；

步骤 4 在 $S^{(3)}$中键相同的项中，合并不重复的子集，然后剔除其余键及其对应的值；

步骤 5 统计全部键所对应的值中包含的每个集合的频次，并进行比较，依据 $f_t(p)-f_t(p')\leqslant\gamma$，$S^{(3)}$ $\gamma\approx(f_t(p))\times50\%$的规则找到完整短语；

步骤 6 在字典 f 中，将筛选出的完整短语的所有子集和超集删除，形成候选完整短语集合 S。

3.4.2 构造排序函数

排序函数的目的是使前面从候选关键短语中保留下来的完整短语按照人们理解的好坏以及对主题的代表程度进行排序。为此，本文构造了 4 个指标来对关键短语进行排序。

(1)覆盖性

覆盖性，实际上也就是频率。候选短语在文本集中出现的频率能够等价于该候选短语在文本集中的覆盖程度，这是所有排序算法中最基本的规则。本文利用在同一主题的文本中短语出现的频率作为量化覆盖性的度量手段：

$$\pi_t^{cov}(p)=P(e_t(p))=\frac{f_t(p)}{|D_t|} \tag{1}$$

其中，p 表示某个短语，$e_t(p)$表示事件：在主题 t 中出现短语 p，$f_t(p)$表示短语 p 在主题 t 中出现的次数，D_t 表示主题标号为 t 的所有文档中语料的集合，$|D_t|$表示主题标号为 t 的所有文档中语料集合的大小。因此，覆盖性实际上是用短语在某一个主题中出现的频率来量化的。

(2)纯洁性

如果一个短语只在某一个主题中是频繁的，而在其他主题中不是频繁的，那么就说这个短语在这个主题中是纯洁的；相反，则说这个短语不是纯洁的。例如，在数据挖掘主题中“支持向量机”这个短语的纯洁性肯定比“支持”这个短语的纯洁性高。虽然在其他主题中“支持”这个短语的出现频率不一定会非常低，但是“支持向量机”这个短语如果仅在数据挖掘主题的文本中频繁出现，而在其他主题的文本中没有频繁出现，那么可以认为“支持向量机”在数据挖掘主题中具有较高的纯洁性。度量纯洁性的手段依旧是频率，即通过比较一个短语出现在某一个主题 t 的语料集合中的频率和出现在其他主题 $t'=0,1,\cdots,K(t'\neq t)$的语料中的频率来度量这个短语的纯洁性。$e_t(p)$表示在主题 t 中出现短语 p 的随机事件，用 $e_{t'}(p)$表示在其他主题 t'中出现短语 p 的随机事件，$|D_{t'}|$表示主题为 t'的文本大小，则纯洁性的量化度量形式为：

$$\begin{aligned}\pi_t^{pur}(p)&=\log\frac{P(e_t(p))}{\max\limits_{t'\neq t}P(e_{t'}(p))}\\&=\log\frac{f_t(p)}{|D_t|}-\log\max_{t'\neq t}\frac{f_{t'}(p)}{|D_{t'}|}\end{aligned} \tag{2}$$

这个比较的结果越大，说明短语 p 在主题 t 中出现的频率越高于在其他主题 t'中出现的频率，那么短语 p 对这个主题的纯洁性就越高，越有可能代表这个主题的意义，从而成为这个主题的关键短语。相反，如果这个比较结果越小，甚至为负，说明短语 p 在主题 t 和其他主题 t'中出现的频率越接近，甚至在其他主题中出现的频率更高，那么这个短语 p 对主题 t 的纯洁性就越低，越不能作为主题 t 的关键词来提取。

(3)短语性

由于本文提出的算法提取的短语实际上是一组词语的集合，这样一组词语可能只是同时出现在文本中，而并不是固定搭配或短语，因此这个指标主要是为了评价一组词语是否能够合适地构成一个短语。例如，一个长度为 2 的候选短语 $p=\{w_1,w_2\}$，如果组成这个候选短语的两个词语单独在文本中出现的频率远远高于这两个词语共同出现的频率，那么可以认为这个候选短语作为短语的可能性非常小。相反，如果在文本中当 w_1 出现时，另一个词语 w_2 以很大的概率同时出现，并且这个频率远大于 w_1 单独出现的频率，那么可以认为这个候选短语作为真正短语的可能性是非常大的。量化表达式为：

$$\begin{aligned}\pi_t^{phr}(p)&=\log\frac{P(e_t(p))}{\prod\limits_{w\in p}P(e_t(w))}\\&=\log\frac{f_t(p)}{|D_t|}-\sum_{w\in p}\log\frac{f_t(w)}{|D_t|}\end{aligned} \tag{3}$$

(4)短语长度

本文认为短语或固定搭配在文本中包含的信息要比单个词语包含的信息多，但这也不是绝对的。有文献提出，提取短语的最佳长度是 3 或 4，如果短语长度更长，或许也会起到相反的作用，因此本文用以下表达式来作为量化短语长度的指标：

$$\pi^{len}=\begin{cases}\dfrac{len(p)-Elen}{M}, & len(p)\leqslant Elen\\[2ex]\dfrac{Elen-len(p)}{M}, & len(p)>Elen\end{cases} \tag{4}$$

其中，$len(p)$表示短语 p 的长度，$Elen$ 表示一个期望的最佳

长度.M 表示最长短语的长度,目的是对指标作归一化处理。

将覆盖性、纯洁性、短语性、短语长度这 4 个指标综合成为一个 Rank 函数时,KERT 算法中利用了 KL 散度的概念,利用概率的乘积来综合。而 KL 散度表示的是两个分布之间的差异程度,本文认为在这里利用 KL 散度来综合指标并不是最好的办法。本文选择加权综合的方法:

$$Rank=\frac{1}{|\omega_1\pi_t^{cov}+\omega_2\pi_t^{pur}+\omega_3\pi_t^{phr}+\omega_4\pi^{len}|} \tag{5}$$

其中,$\omega_1,\omega_2,\omega_3,\omega_4$ 分别表示 $\pi_t^{cov},\pi_t^{pur},\pi_t^{phr},\pi^{len}$ 的权重,$\omega_1+\omega_2+\omega_3+\omega_4=1$。权重的大小表示用户对这 4 个指标的重视程度,将对更重要的指标设置相对大的权重,对相对不太重要的指标设置相对小的权重。

由于在 Rank 函数中 π_t^{phr} 和 π_t^{len} 是负值,为了保证 Rank 结果非负,并且不影响排序,Rank 函数采用了绝对值倒数的形式。

4　实证研究

4.1　数据来源

本文算法适用于任何长度的中文文本集,内容可以是新闻报道、学术论文、文学、医学等领域的综合文本。在本文的实验中选择了学术论文摘要、新闻报道与政治评论文章 3 方面的综合文本,详细信息如表 1 所列。

表 1　实验数据说明

数据来源	文本数/篇	总字数	总词数
学术论文	38	7841	3165
新闻报道	12	2647	850
政治评论	25	11564	4541
总计	75	22052	8556

4.2　实验结果

为了证明本文算法能够在一定程度上消除中文分词误差对提取结果的影响,政治评论类文本中的“居民健康卡”词组,分为“居民”“健康”“卡”。在其他两类文本中分词也存在误差:“数字化医院”分为“数字化”“医院”等。

由于排序函数中有 4 个权重,根据不同需求可对 4 个权重分别进行调整。表 2 列出其中一组权重组合的实验结果。

表 2　以“短语性”为主的权重组合

$\omega(pur)=0.1,\omega(phr)=0.5,\omega(\mathrm{cov})=0.1,\omega(len)=0.3$					
Topic 0		Topic 1		Topic 2	
不忘初心 继续前进	3.2	绵阳市	3.4	中文文本 关键词提取	1.9
发展	3.1	平台建设	3.3	实验结果表明	1.7
中国特色 社会主义	2.8	居民健康卡	3.1	主题	1.6
改革	2.4	管理	2.9	挖掘	1.6
中华民族	2.3	数字化医院	2.8	统计特征	1.5
时代	2.2	服务	2.7	TF-IDF 算法	1.5

从实验结果中可以看出:

1)在分词过程中产生的分词错误,能够在提取结果中被合并,如“居民健康卡”、“数字化医院”等;

2)本文提出的算法可以正确地将综合文本集按照不同主题进行聚类,并提取主题下不同长度的关键词;

3)提取结果中不存在关键短语及其子短语同时出现的现象。

同类算法中,将本文算法与 2014 年明尼苏达大学Marina 提出的 KERT 算法进行比较。

表 3 列出 KERT 算法在同一数据中的提取结果。

表 3　KERT 算法提取结果

min_sup=10,8 γ=0.5					
Topic 0		Topic 2		Topic 1	
发展	0.50	绵阳市	0.48	关键词提取方法	0.44
继续	0.29	居民健康卡	0.47	关键词	0.39
不忘初心 继续前进	0.08	费用	0.37	提出基于	0.29
前进	0.08	患者	0.32	实验结果表明	0.17
中国	0.08	医院	0.29	结果表明	0.17
时代	0.05	数字化	0.25	TP-IDF	0.09
发展	0.05	医疗机构	0.18	分类算法	0.04

从提取效果上可以明显看出,本文算法优于 KERT 算法。KERT 算法虽然也能够提取出有效的关键短语,但是在提取结果中存在关键短语与其子短语同时出现的情况,例如:“继续”“前进”“不忘初心继续前进”等。这种情况虽然并没有直接影响人们对文本内容的理解,但是会对整体的提取结果产生影响。

除了提取结果看上去的优化,本文还进行了精准率、召回率和 F_1 值的比较,结果如表 4 所列。结果表明,本文所提算法明显优于 KERT 算法。原因在于,KERT 算法提取的关键短语存在其子短语同时出现的情况,并且存在一些并不是固定短语但 KERT 并不能完全识别出来的情况,因此提取的关键词和短语的个数往往比较多;而本文算法避免了这两种情况。在相同的支持度阈值条件下,本文算法提取的关键词和短语总数大大减少,因此精准率肯定高于 KERT 算法。

表 4　本文算法与 KERT 算法的评价对比

	精准率			召回率			F_1 值		
	T0	T1	T2	T0	T1	T2	T0	T1	T2
本文算法	73	75	76	0.4	1.0	0.4	0.88	2.09	1.9
KERT 算法	62	50	56	0.4	1.0	0.4	0.87	2.07	1.9

在召回率方面,本文算法和 KERT 算法的实验数据完全相同,提取出关键短语的个数也基本相同,因此召回率也基本相等。但是从召回率的数值上可以看出,召回率这个指标并不适合于关键词提取算法的评估,因为提取出的关键词的个数与文本总词数之间并没有实际的关系。从 F_1 值结果来看,本文算法和 KERT 算法的 F_1 值相差不大,但总体上本文算法偏高。

结束语　未来将在以下几个方面进行改进:

(1)改进 LDA 主题模型

目前,主题模型在关键词和关键短语上的应用越来越流行,自 LDA 主题模型被提出之后,相继有许多学者对其进行了改进,其中一些改进是专门针对关键短语提取设计的,例如 phraseLDA 主题模型,它可以在文本建模过程中直接对短语进行聚类。这将作为今后的重点改进方向。

(2)结合信息论知识对短语进行切分

本文的短语提取和合并方法使用的是单纯的统计知识。事实上,多领域知识交叉应用会得到更精准的提取结果,因此可以将信息论的知识和本文统计知识相结合,以改善提取效果。

(3)将本文算法应用在其他应用中

目前文本聚类和关键短语提取技术的应用非常广泛,可以进一步拓展应用。

参 考 文 献

[1] FELDMAN R, DAGAN I. Knowledge discovery in textual databases[C]// International Conference on Knowledge Discovery & Data Mining. 1995: 112-117.

[2] 刘静. 面向中文微博关键词提取技术研究[D]. 长沙: 中南大学, 2014.

[3] TAN P N, STEINBACH M, KUMEA R V. Introduction to Data Mining[M]. Beijing: China Machine Press, 2010.

[4] LUO S M, WANG Z K, WANG Z P. Big-Data Analytics: Challenges, Key Technologies and Prospects[J]. ZTE Communication, 2013(2): 11-17.

[5] 陈晓云. 文本挖掘若干关键技术研究[D]. 上海: 复旦大学, 2005.

[6] rickjin. 通俗理解 LDA 主题模型[EB/OL]. http://blog.csdn.net/v_july_v/article/details/41209515? utm_source=tuicool & utm_medium=referral. 2014.

[7] DAMILEVSKY M, WANG C, DESAI N, et al. Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents[C]// SDM. 2014.

[8] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003(3): 993-1022.

[9] HOFMANN T. Probabilistic Latent Semantic Indexing[C]// ACM Proceeding of the 1999 ACM SIGMOD International Conference on Management of Data. New York: ACM, 1999: 50-57.

[10] 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 Processing: Volume 1. Assciation for Computational Linguistics, 2009: 248-256.

[11] LIU J L, SHANG J B, WANG C, et al. Mining Quality Phrases from Massive Text Corpora[C]// ACM Proceeding of The 2015 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2015: 1729-1744.

[12] ZHAO W X, JIANG J, YANG J H, et al. Topical Keyphrase Extraction from Twitter[C] // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 2011: 379-388.

[13] HAN J, PEI J, YIN Y. Mining Frequent Patterns without Candidate Generation[C]// Proceedings of the 2000 ACM-SIGMOD International Conference on Management of Data. New York: ACM, 2000: 1-12.

[14] LIU Z, HUANG W, ZHENG Y, et al. Automatic keyphrases extraction via topic decomposition[C]// EMNLP. 2010.

[15] 郝峰. 文本关联分析中频繁项集挖掘的研究与改进[D]. 太原: 太原理工大学, 2008.

[16] 蔡鹏飞. 基于概率图模型的关联规则更新方法与实现[D]. 昆明: 云南大学, 2013.

[17] 李艳美. 基于贝叶斯网络的数据挖掘应用研究[D]. 西安: 西安电子科技大学, 2008.

[18] 徐文海, 温有奎. 一种基于 TF-IDF 方法的中文关键词抽取算法[J]. 情报理论与实践, 2008, 31(2): 298-302.

[19] YAN X, GUO J, LIU S, et al. Learning topics in short texts by non-negative matrix factorization on term correlation matrix [C]// SDM. 2013.

[20] RAJARAMAN A, ULLMAN J D. Mining of Massive Datasets [M]. Cambridge: Cambridge University Press, 2012.

[21] 章志刚, 吉根林. 一种基于 FP-树和数组技术的频繁模式挖掘算法[J]. 计算机工程与应用, 2014, 50(2): 103-106.

(上接第 402 页)

索系统中交互设计的基本原则，并对其进行示例验证。交互有效性判定是一个难题，本文给出了一种交互度量，我们将进一步充实有效交互的理论成果，给出交互有效性的定量的判定度量方式；在本文的基础上，可以结合已有对信息语义关联网的研究成果，提出新的探索式搜索交互方法，将本文思路应用到探索式搜索中，检验其实践效果。

参 考 文 献

[1] 任磊, 魏永长, 杜一, 等. 面向信息可视化的语义 Focus+Context 人机交互技术[J]. 计算机学报, 2015, 38(12): 2488-2498.

[2] FURNAS G W. A fisheye follow-up: Further reflections on Focus+Context[C] // Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems. Montreal, Canada, 2006: 999-1008.

[3] ZHANG Y, LI Y, WANG H. A Study on Exploratory Medical Search Behavior Based on Cognitive and Log Analysis[J]. Library & Information Service, 2014, 58(11): 36-42.

[4] TAN B, LV Y, ZHAI C X. Mining long-lasting exploratory user interests from search history[C] // ACM International Conference on Information and Knowledge Management. 2012: 1477-1481.

[5] SUN H, JIANG C, DING Z, et al. Topic-Oriented Exploratory Search Based on an Indexing Network[J]. IEEE Transactions on Systems Man & Cybernetics Systems, 2015, 46(2): 1.

[6] ZHAO Q, WANG C JIANG C. Hsim: A Novel Method on Similarity Computation by Hybrid Measure[C]// International Conference on Information and Communication Systems. IEEE, 2015: 160-165.

[7] MAO Y, SHEN H, SUN C. A Social-Knowledge -Directed Query Suggestion Approach for Exploratory Search[C]// International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery(Cyberc 2011). Beijing, China, 2011: 1-8.

[8] KOTOV, ALEXANDER, BENNETT, et al. Modeling and analysis of cross-session search tasks[C]// Paper Presented at the Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011.

[9] QVARFORDT P, GOLOVCHINSKY G, DUNNIGAN T, et al. Looking ahead: query preview in exploratory search[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2013: 243-252.

[10] QU Y. Supporting Ideation by Integrating Exploratory Search, Browsing, and Curation[C]// ACM on Conference on Human Information Interaction and Retrieval. ACM, 2016: 361-363.