文章编号：2095-4980(2015)02-0279-06

基于频繁模式挖掘的中文关键词提取算法

崔诚煜，冉晓旻

(信息工程大学 信息系统工程学院，河南 郑州 450002)

摘　要：针对现有关键词提取算法存在计算复杂、语义信息挖掘较浅等问题，提出一种基于频繁模式挖掘的中文关键词提取算法。该算法采用改进的 FP-增长算法挖掘词共现信息，排除噪音词汇；利用语义相似度算法消除同义词；精简候选词特征，在保证较高准确率和召回率的条件下减少了存储空间和计算量。实验结果表明，该算法所获得的平均 F 值为 59.7%，高于若干经典算法；支持度计数是最重要的影响因素。

关键词：关键词提取；频繁模式挖掘；词共现；FP-增长

中图分类号：TN911.6；TP391　　**文献标识码：**A　　**doi：**10.11805/TKYDA201502.0279

Chinese keywords extraction algorithm based on frequent pattern mining

CUI Chengyu，RAN Xiaomin

(School of Information System Engineering，Information Engineering University，Zhengzhou Henan 450002，China)

Abstract：A keyword extraction algorithm for Chinese documents based on frequent pattern mining is proposed aiming at the problems of existing Keywords Extraction Algorithm(KEA) including high computational complexity and mining shallow semantic information. This algorithm adopts improved FP-Growth technology to extract word co-occurrence information and remove noisy words. It utilizes semantic similarity algorithm to eliminate synonyms and simplify the characteristics of candidates, thus reducing the storage space and the amount of calculation when ensuring the high precision and recall. Experimental results show that the average F value of corpus reaches 59.7%, which is higher than classical algorithms; and that the support threshold is the vital influencing factor.

Key words：Keywords Extraction(KE)；frequent pattern mining；word co-occurrence；FP-Growth

关键词提取(KE)[1]又称文本自动标引，指借助计算机处理技术，从文档中提取出能恰当地代表文档主题的若干词汇或短语。目前文档管理面临海量数据、精确管理等挑战，如何准确而快速地提取出文档的关键词是关键词提取算法研究的主要内容与核心目标。

目前，关键词提取的常用方法包括基于统计信息学的算法[2]、基于语言网络的算法[3]和基于语义的算法[4]等几大类。其中，基于统计信息学的算法利用 TF×IDF、词性和首词位置等关键特征标注关键词，主要工具有 GenEx 系统和 KEA 系统[5]；该类方法具有模型泛化性，但忽略了词汇语义的内在联系。基于统计特征的算法还可以通过机器学习[6]的手段提取文档关键词，首先对已标注关键词的文档提取特征，然后利用 C4.5 决策树、最大熵模型或朴素贝叶斯模型[7]训练样本，对未标注关键词的文档提取关键词；但是不同领域的关键词具有不同的特征，训练成本高，领域移植性差，应用条件较为苛刻，应用范围受到限制。基于语言网络的算法将所有词汇之间的关系形成复杂的语言网络，利用语言网络中的小世界特性，深度考虑了文档中的词共现语义信息，提取效果较好，但是计算复杂度非常高，难以应用于大规模文档；基于语言网络的 TextRank 算法将词汇形成网络，通过随机游走实现关键词提取，具有无需训练，适应性强，计算速度快等优势。基于语义的算法是自然语言处理未来发展的方向，此类算法[8]不局限于词共现信息，可以详尽地描述词汇含义，但是目前发展并不成熟，算法复杂度较高。此外，还有基于极大团理论、词汇链[9]等的关键词提取算法。

收稿日期：2014-04-14；修回日期：2014-05-16

1 频繁模式挖掘基本概念

频繁模式挖掘(Association Rule Mining，ARM)[10]属于数据挖掘范畴，旨在找出反映事物之间的相互依存和关联关系，其目的是在事务数据库的频繁项集和对象中发现关联规则。

定义设 $I=\{i_1,i_2,\cdots,i_m\}$ 为所有项目的集合，D 为事务数据库，事务 T 是一个项目子集 $(T\subseteq I)$。每个事务均具有唯一的事务标识 TID。设 A 是一个由项目构成的集合，称为项集。事务 T 包含项集 A，当且仅当 $A\subseteq T$。如果项集 A 中包含 k 个项目，则称其为 k 项集。如果项集的支持度超过用户给定的最小支持度阈值，就称该项集是频繁项集。从事务数据库中发现频繁项集，进一步地，在满足最小置信度的条件下，可以得到频繁模式，以上称为静态频繁模式挖掘。

频繁模式挖掘算法的效率和成熟度较高，静态模式挖掘经典算法包括 Apriori 算法[11]和 FP-Growth 算法[12]。Apriori 算法思想简单，复杂度较高，其改进算法常采用散列项集技术和事务压缩等手段提高时空效率；FP-Growth 算法只需扫描数据库 2 次，且不需要产生和测试候选集，具有更强的完备性和紧密性，计算复杂度比 Apriori 算法低，然而 FP-Growth 算法需要递归产生大量的条件 FP-Tree，这耗费了大量的存储空间和时间。

研究表明[13]，词汇、短语概念之间的关联关系和语言同义词集是语义信息的表现形式。频繁出现的词共现信息在一定程度上代表了文档词汇的语义关联关系，可应用于关键词提取。本文利用改进的 FP-Growth 频繁模式挖掘技术，挖掘出文档中词汇或短语之间的词共现特征，消除同义词，找出频繁词共现项集。实验结果表明，频繁模式挖掘有利于提高关键词提取的准确率与召回率。

2 基于频繁模式挖掘的中文文档关键词提取算法

2.1 算法基本思想

把事务频繁模式挖掘应用在文本词汇中，需将文本定义为数据库，词汇集合定义为项目集，每个词汇串定义为一个事务，共现的词汇组合建立关联关系。词共现关联关系说明了词汇在文档中具有语义相关性，频繁模式中的词汇具有较为关键的作用，可考虑被采纳为关键词；非频繁的词共现可视为偶然发生的，无意义的，不足以代表文档主要内容。

2.2 算法描述

基于频繁模式挖掘的关键词提取算法主要包括文档预处理、词汇相似度计算、频繁模式挖掘候选关键词集和提取关键词 4 个方面，基本流程图如图 1 所示。

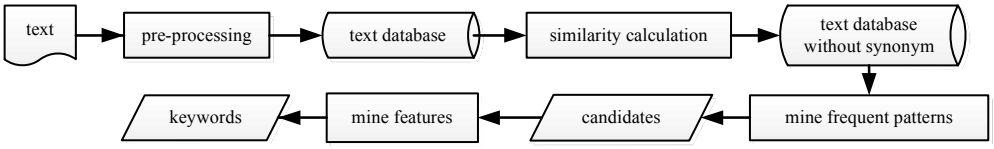

Fig.1 Flow chart of the proposed algorithm

图 1 基于频繁模式挖掘关键词提取算法流程图

2.2.1 文档预处理

对待提取关键词的中文文档 d 进行预处理，包括文档格式处理，分句，分词和未登录词识别以及实词筛选 4 个步骤。

2.2.2 词汇相似度计算

文档中存在大量词义相近的词汇，在抽取关键词过程中，应将词义相同或相近的词汇合并为一个词汇进行处理，这样可以避免同义词汇的重复挖掘和欠挖掘，有效提高关键词挖掘准确度。本文借鉴文献[14]的算法，构建词汇链，逐词计算词汇之间的相似度，将相似度大于某一阈值的词汇归入同一词汇链，将相似度小于该阈值的词汇分别建立词汇链。用词汇链中频度最高的词汇代替该词汇链中的所有词汇。

2.2.3 词共现频繁模式挖掘

经过预处理的文档表现为一系列词汇串，一个词汇串对应一个句子，词汇串中包含若干词汇。将词汇串视为事务数据库 D 中的事务 T，词汇串中的词汇视为项目集合 I 中的项目 i_m，处于同一词汇串中的词汇具有共现关联关系，可进行频繁模式挖掘。

本文借鉴文献[12]和[15]，采用内存压缩的方法对经典 FP-Growth 算法进行改进，建立 FP-Tree。之后,可根据分治策略,对项头表按照自底向上的顺序,把挖掘问题划分为若干子问题,挖掘以某个特定后缀结尾的频繁项集；在挖掘过程中首先从 FP-Tree 中找出该后缀的前缀路径和条件模式库,更新支持度计数与前缀路径，进而得出条件 FP-Tree，基于条件 FP-Tree，可以得到频繁项集。频繁项集中的词汇可视为候选关键词。采用内存压缩的 FP-Growth 算法可有效提高挖掘效率。以下举例说明频繁模式挖掘算法在文本中的应用。

例，给定短文本：“数据挖掘是在大型数据存储库中，自动地发现有用信息的过程。数据挖掘技术用来探查大型数据库,发现大型数据库中先前未知的有用模式。数据挖掘还可以预测未来观测结果。”

将该段文字经过分词和未登录词识别处理,得到如下标注语料：“数据挖掘/n 是/v 在/p 大型数据/n 存储/vn 库/n 中/f，/w 自动/d 地/uv 发现/v 有用/a 信息/n 的/uj 过程/n。/w 数据挖掘/n 技术/n 用来/v 探查/v 大型数据库/n，/w 发现/v 大型数据库/n 中/f 先前/t 未知/a 的/uj 有用/a 模式/n。/w 数据挖掘/n 还/d 可以/v 预测/v 未来/n 观测/vn 结果/n。/w”。

表 1 例文预处理结果

Table1 Results of pre-processing

TID	transaction
1	数据挖掘 大型数据 存储 库
2	发现 信息 过程
3	数据挖掘 技术 探查 大型数据库
4	发现 大型数据库 模式
5	数据挖掘 预测 未来 观测 结果

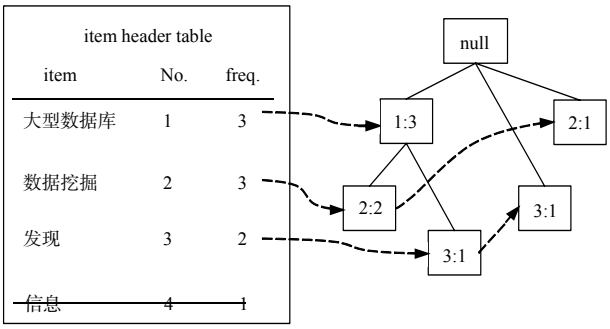

Fig.2 Item header table and correspondent FP-Tree

图 2 例文项头表与对应 FP-Tree

从上述已标注语料中选取可能成为关键词的名词和实义动词，以单句为单位构建事务数据库，如表 1 所示。根据词汇语义相似度计算，建立同义词汇链，可将词汇“大型数据”合并至“大型数据库”中。

找出频繁的词汇单项，最小支持度阈值为 40%。建立项头表，并据此建立 FP-Tree，如图 2 所示。

根据上述 FP-Tree，设定最小支持数为 2，删除所有频繁 1-项集，可得频繁项集结果为{大型数据库，数据挖掘}，其中的词汇可采纳为例文的候选关键词。

2.2.4 从候选关键词中提取关键词

频繁项集包含了所有的候选词汇，对词汇提取特征，包括以下 2 个方面：

1) TF×IDF 值，由于所有文本均已转化为词汇串数据库，因此 TF 值仅指在候选关键词集中的词汇出现频率，IDF 值指在所有文本中出现该词的频率，公式定义如下：

$$W(t,d)=\frac{tf(t,d)\times\log\left(\frac{N}{df(t)}+0.01\right)}{\sqrt{\sum_{t\in D}\left[tf(t,d)\times\log\left(\frac{N}{df(t)}+0.01\right)\right]^2}} \tag{1}$$

式中：$W(t,d)$表示候选词汇 t 在文档 d 的权值；$tf(t,d)$表示候选词汇 t 在文档 d 的频繁项集中出现的频率；N 为文本集总数；$df(t)$为 t 在整个文本集频繁项集出现的文本数。

2) 频繁项目个数，频繁项的项目个数越多，说明其中的词汇具有更强的关联性，需要优先考虑设定为关键词，公式定义如下：

$$n_t=\log\left[L(t,d)-1+0.01\right] \tag{2}$$

式中：n_t 为权值；$L(t,d)$为词汇 t 在文档 d 的频繁项集中所处的最大项集的项数。

综合以上 2 个特征，对其做归一化处理：

$$\tilde{W}(t,d)=\frac{W(t,d)-E\left[W(t,d)\right]}{\sigma\left[W(t,d)\right]} \tag{3}$$

$$\tilde{n}_t=\frac{n_t-E\left[n_t\right]}{\sigma\left[n_t\right]} \tag{4}$$

式中：$E[\cdot]$是求均值运算；$\sigma[\cdot]$是求均方差运算。对于基于改进的 FP-Growth 算法挖掘出的词汇 t，其权重用下式

定义：

$$Weight(t)=\alpha\tilde{W}(t,d)+\beta\tilde{n}_t \tag{5}$$

式中：α 和 β 为权值参数；$\alpha+\beta=1$。对候选关键词汇的权重 $Weight(t)$ 进行排序，形成由大到小的队列 queue，选取队列 queue 中取值最高的 m 个词汇作为文档的关键词。

2.3 算法复杂度分析

本文借助改进 FP-Growth 频繁模式挖掘算法提取关键词，主要优势在于：

1) 频繁模式挖掘只保留频繁项集中词汇作为候选关键词，该词汇集词数远小于文档总词数，$M_{\text{Freq}}\ll M$；从候选关键词中提取文本关键词，删除了影响因子较小的特征，只考虑主要特征 TF×IDF 和频繁项目个数，且改进的 TF×IDF 计算量大幅减小。相比于基于复杂网络方法需要把所有的词汇数据存储起来，以及基于统计特征方法需设大量特征，本文算法节约了存储空间，减小了计算量。

2) 频繁模式挖掘算法近年来发展成熟，改进的 FP-Growth 算法具有较高的完备性与紧密性，避免产生大量的条件 FP-Tree，抽取频繁模式的计算复杂度主要受压缩因子(Compaction Factor，CF)影响，对于文本数据库，压缩空间可观，在支持度计数、项数给定条件下，仅考虑词汇数，复杂度为 $O(n/m)$，其中 m 为事务中平均词数，小于基于复杂网络的抽取算法复杂度 $O(n^3)$。

3 实验仿真

选择计算机、通信、国际新闻和军事 4 个领域文档，每个领域抽取 50 篇文档进行测试。分词采用中科院计算所的 ICTCLAS 系统。手动标注文档中的关键词作为标准关键词，将本文算法与 KEA 开源系统、TextRank 算法、基于复杂网络算法、基于 TF×IDF 的算法和基于统计特征方法作对比。采用精确率(precision)和召回率(recall)评价算法的性能，它们统一于调和平均值 $F_{\beta=1}$。指标定义如下式：

$$precision=\frac{|N_{\text{man}}\cap N_{\text{auto}}|}{|N_{\text{auto}}|} \tag{6}$$

$$recall=\frac{|N_{\text{man}}\cap N_{\text{auto}}|}{|N_{\text{man}}|} \tag{7}$$

$$F_{\beta=1}=\frac{2\times precision\times recall}{precision+recall}=\frac{2\times|N_{\text{man}}\cap N_{\text{auto}}|}{|N_{\text{man}}|+|N_{\text{auto}}|} \tag{8}$$

式中：N_{man} 表示人工标注的关键词的集合；N_{auto} 表示机器系统所标注的关键词的集合。

本文仿真中，词汇相似度阈值设为 0.95，频繁 1-项集构成的项头表最小支持度为 2，挖掘出不包含频繁 1-项集的频繁模式最小支持度为 2。本文经过多次试验，得到最佳候选特征权重分配：$\alpha=0.8$，$\beta=0.2$。关键词个数设为 5。将支持度数提高可有效减少频繁项集数，文本频繁模式挖掘事务数相对于词汇数并不多，因此支持度阈值较低，所挖掘出的树结构不深，频繁 2-项集和 3-项集占绝大多数。

KEA 系统采用机器学习的算法，学习特征包括 TF×IDF、首词位置和词长等；TextRank 窗口值设为 5，采用无向加权图模型；基于复杂网络方法将所有关联关系权值设为 1；基于统计特征方法抽取词频、词位置、词跨度、内聚度和 TF×IDF 等 5 个主要特征；所有机器学习算法每个领域训练文档 38 篇，测试文档 12 篇。实验结果数据如表 2 和表 3 所示。

表 2 若干关键词提取算法开放测试准确率和召回率

Table 2 Precision and recall rate comparison of several algorithms

algorithms	computer	commu.	news	military	average	rate
based on FPM	0.608	0.615	0.634	0.610	0.617	precision
	0.581	0.565	0.618	0.567	0.581	recall
KEA	0.503	0.500	0.518	0.496	0.504	precision
	0.517	0.514	0.526	0.508	0.516	recall
TextRank	0.561	0.552	0.569	0.568	0.561	precision
	0.545	0.538	0.553	0.539	0.544	recall
based on complex network	0.616	0.614	0.641	0.603	0.619	precision
	0.596	0.591	0.616	0.585	0.597	recall
TF×IDF	0.459	0.460	0.475	0.456	0.461	precision
	0.421	0.414	0.448	0.410	0.423	recall
based on statistics	0.603	0.597	0.612	0.582	0.599	precision
	0.561	0.560	0.579	0.555	0.564	recall

从以上图表可知，本文算法性能接近基于复杂网络的算法，均高于其余算法。主要原因是，2 种算法从本质上均是提取文本中的词共现信息，差别仅限于提取方法不同，从词共现信息中提取的特征不同。关键词提取的既有要求是删除代表性弱的词汇，保留影响力大的词汇；在提取过程中，代表性弱的词汇无足轻重，不需考虑。本文算法宗旨是快速找出关联关系密集词汇，删除关系稀疏的词汇，顺应了关键词提取的要求，避免了复杂网络方法中将所有词汇先组网络再统计权重的不便。因此，尽管本文算法的 $F_{\beta=1}$ 比复杂网络算法略低，但计算量明显比复杂网络算法低。从 TF×IDF 算法和基于统计特征算法结果可知，TF×IDF 特征是统计特征中最经典的特征，远强于其余特征，故将其与频繁模式挖掘的提取算法相结合。KEA 系统采用机器学习方法，统计特征丰富，故其性能优于 TF×IDF；由于缺乏领域词典，故其性能不如本文算法和基于复杂网络的算法。TextRank 模型忽视了位置信息的重叠和主题相关性，且目前没有针对特定领域的语义网络，所以准确率不高。

表 3 若干关键词提取算法开放测试 $F_{\beta=1}$ 值

Table3 $F_{\beta=1}$ comparison of several algorithms

$F_{\beta=1}$	computer	commu.	news	military	average
based on FPM	0.594	0.589	0.622	0.583	0.597
KEA	0.510	0.507	0.522	0.502	0.510
TextRank	0.553	0.545	0.561	0.553	0.554
based on complex network	0.606	0.602	0.628	0.594	0.607
TF×IDF	0.439	0.436	0.461	0.432	0.443
based on statistics	0.581	0.578	0.595	0.568	0.580

频繁项集的最小支持度是影响计算复杂度的重要因素,本文采用不同支持度计数做统计,仅针对计算机领域,不考虑文档间的词汇重复，仅做相加处理，统计如表 4 所示。

从上表可知，当频繁项集支持度计数取 2 时，算法效率最高，随着支持度计数取值升高，效果逐渐恶化。主要原因是非频繁项集包含的关键词被过滤掉。尽管频繁 2-项集中仍包含了大量的词汇，但对于抽取特征、存储数据来说，已经大大减小了计算量，相比于复杂网络方法，处理速度有较大提升。

表4 支持度计数S对算法性能影响

Table 4 Influence of support number S on the performance of proposed algorithm

S	2	3	4	6	10
total No.			223 768		
words in FP	91 436	42 872	29 502	7 382	1 784
$F_{\beta=1}$	0.583	0.546	0.451	0.297	0.132

4　结论

本文基于改进的 FP-Growth 频繁模式挖掘技术，提出一种关键词提取算法，将数据挖掘技术应用于自然语言处理领域。该算法在考虑词共现的文本信息基础上，消除同义词，缩减计算量，精简候选关键词及其特征，保持了较好的准确率与召回率，优于 KEA、统计特征和 TextRank 等经典算法。在大规模数据条件下，运算速度优于基于复杂网络的算法。

基于频繁模式挖掘的关键词提取算法可从多方面进行改进。一方面，利用词汇本体描述语言中存在的固有语义信息，可进一步挖掘出文档的深度语义信息，有利于关键词提取；另一方面，文本词汇中存在大量的多层次嵌套概念，将针对多层次概念的频繁模式挖掘应用于关键词提取，能更有效地提取出文本关键词。

参考文献:

[1] 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. (LUO Zhunchen, WANG Ting. Research on the Chinese keyword extraction algorithm based on separate models[J]. Journal of Chinese Information Processing, 2009,23(1):63-70.)

[2] 罗繁明,杨海深. 大数据时代基于统计特征的情报关键词提取方法研究[J]. 情报资料工作, 2013,34(3):64-68. (LUO Fanming,YANG Haishen. On the statistical features based information keyword extraction method in the era of big data[J]. Information and Documentation Services, 2013,34(3):64-68.)

[3] 杨洁,季铎,蔡东风,等.基于联合权重的多文档关键词抽取技术[J]. 中文信息学报, 2008,22(6):75-79. (YANG Jie, JI Duo,CAI Dongfeng,et al. Keyword extraction in multi-document based on joint weight[J]. Journal of Chinese Information Processing, 2008,22(6):75-79.)

[4] 战学刚,吴强. 基于 TF 统计和语法分析的关键词提取算法[J]. 计算机应用与软件, 2014,31(1):47-49. (ZHAN Xuegang, WU Qiang. Keyword extraction algorithm based on TF statistics and syntactic parsing[J]. Computer Application and Software, 2014,31(1):47-49.)

[5] 陈平,周昌乐,练睿婷. 一种改进的 KEA 关键词抽取算法研究[J]. 心智与计算, 2011,5(2):48-54. (CHEN Ping,ZHOU Changle,LIAN Ruiting. An improved approach to keyword extraction using KEA[J]. Mind and Computation, 2011,5(2):48-54.)

[6] 刘佳宾,陈超,邵正荣,等. 基于机器学习的科技文摘关键词自动提取方法[J]. 计算机工程与应用, 2007,43(14):170-172. (LIU Jiabin,CHEN Chao,SHAO Zhengrong,et al. Automatic extraction of key phrases from scientific articles based on machine learning method[J]. Computer Engineering and Application, 2007,43(14):170-172.)

[7] 王锦波,王莲芝,高万林,等. 一种改进的朴素贝叶斯关键词提取算法研究[J]. 计算机应用与软件, 2014,31(2):174-181. (WANG Jinbo,WANG Lianzhi,GAO Wanlin,et al. On an improved naïve Bayesian keyword extraction algorithm[J]. Computer Application and Software, 2014,31(2):174-181.)

[8] 许珂,蒙祖强,林啓峰. 基于语义关联和信息增益的 TFIDF 改进算法研究[J]. 计算机应用研究, 2012,29(2):557-560. (XU Ke,MENG Zuqiang,LIN Qifeng. Improved TFIDF feature extraction algorithm based on semantic association and information gain[J]. Application Research of Computers, 2012,29(2):557-560.)

[9] 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. (SUO Hongguang, LIU Yushu,CAO Shuying. A keyword selection method based on lexical chains[J]. Journal of Chinese Information Processing, 2006,20(6):25-30.)

[10] ZHOU M,WANG T. Improved pattern tree for incremental frequent-pattern mining[J]. Transactions of Tianjin University, 2010,16(2):129-134.

[11] 章志刚,吉根林. 一种基于 FP-Growth 的频繁项目集并行挖掘算法[J]. 计算机工程与应用, 2014,50(2):103-106. (ZHANG Zhigang,JI Genlin. Parallel algorithm for mining frequent item sets based on FP-Growth[J]. Computer Engineering and Application, 2014,50(2):103-106.)

[12] 秦亮曦,苏永秀,刘永彬,等. 基于压缩 FP-树和数组技术的频繁模式挖掘算法[J]. 计算机研究与发展, 2008,45(Z1): 244-249. (QIN Liangxi,SU Yongxiu,LIU Yongbin,et al. A compact FP-Tree and array-technique based algorithm for frequent patterns mining[J]. Journal of Computer Research and Development, 2008, 45(Z1): 244-249.)

[13] 唐俊. 复杂网络在新闻网页关键词提取中的应用[J]. 云南民族大学学报:自然科学版, 2012,21(4):305-312. (TANG Jun. Application of complex networks to keyword extraction of news web pages[J]. Journal of Yunnan Nationalities University:Natural Sciences Edition, 2012,21(4):305-312.)

[14] 席耀一,林琛,李弼程,等. 基于语义相似度的论坛话题追踪方法[J]. 计算机应用, 2011,31(1):93-96. (XI Yaoyi,LIN Chen,LI Bicheng,et al. Method for BBS topic tracking based on semantic similarity[J]. Journal of Computer Application, 2011,31(1): 93-96.)

[15] Rajaraman A, Ullman J D. Mining of Massive Datasets[M]. Cambridge:Cambridge University Press, 2012.

作者简介:

崔诚煜(1989-),男,山西省长治市人,在读硕士研究生,主要研究方向为数据挖掘、文本信息处理、自然语言处理、网络体系结构.email:dongrixinyu.89@163.com.

冉晓旻(1971-),男,郑州市人,副教授,主要研究方向为数据挖掘、网络体系结构.