

基于频繁模式挖掘的维吾尔文智能组词方法

吐尔地·托合提, 维尼拉·木沙江, 艾斯卡尔·艾木都拉*

(新疆大学 信息科学与工程学院, 乌鲁木齐 830046)

(* 通信作者电子邮箱 askar@xju.edu.cn)

摘要:以词间空格作为自然分隔符, 非常容易获取维吾尔文中的词, 但又很难获取结构完整的语义词, 因此多种文本处理效果总是不理想。提出维吾尔文组词的新概念, 将数据挖掘中的频繁模式挖掘方法引入到维吾尔文组词中, 再结合维吾尔文的语言文字特点, 将无先验知识的模式挖掘问题转化为特定模式的匹配问题, 提出了一种快速高效的频繁模式挖掘算法, 来获取语义完整的维吾尔文词。实验结果表明, 通过该算法获取的维吾尔文词, 在结构上是稳定的, 语义上是完整而独立的。

关键词:维吾尔文本; 分词; 组词; 语义词; 频繁模式

中图分类号: TP18 **文献标志码:** A

Intelligent method for word grouping based on frequent pattern mining in Uyghur language

TUERDI Tuoheti, WEINILA Mushajiang, AISIKAER Aimudula*

(School of Information Science and Engineering, Xinjiang University, Urumqi Xinjiang 830046, China)

Abstract: It is very easy to get the words in Uyghur text lines by the natural delimiters such as spaces, but it is difficult to obtain the completely structured semantic words. Therefore, many kinds of text processing methods always seem not to be very effective. This paper put forward a new concept of Uyghur word grouping and introduced the frequent pattern mining method in data mining scheme, and combined the Uyghur language features, turned the pattern mining problem without prior knowledge into a pattern matching with special pattern, and proposed a fast and efficient frequent pattern mining algorithm to obtain the Uyghur words with complete semantics. The experimental results show that, words obtained by this algorithm are stable in structure, and semantically complete and independent.

Key words: Uyghur text; word segmentation; word grouping; semantic word; frequent pattern

0 引言

维吾尔文与中文不同, 是一种拼音文字, 词与词之间以空格隔开, 这些特点上与英文类似。因此, 维吾尔文中从未探索或研究过分词问题, 用完全与英文类似的方法, 以空格作为自然分隔符隔开文本中的词, 直接获取词的集合^[1]。但维吾尔文又与英文不同, 在很多情况下, 由多个维吾尔文词的上下关联组合来表达一个完整的语义, 如果这种上下关联性被简单分词破坏, 其原有的完整语义也就完全丧失。最近的相关研究表明, 采用空格分割的简单方法获取的维吾尔文单词, 不能作为基本语言单位来处理文本。因为, 作为一个完整语义载体的若干个相邻单词被分开, 其原有的语义被淡化甚至完全被丧失, 这样获取的单词就难以在文本标引中发挥词的作用, 因而基于词特征的文本处理效果也总是不理想^[2]。因此, 为了能够从文本中获取结构稳定、语义完整而独立的维吾尔文词的组合, 研究一种有效的组词方法, 是目前维吾尔文文本处理中必须解决的关键问题。一个维吾尔文文本, 从表面上可以被看成一个已经过分词的单词序列, 但从一个完整的语义上观察, 其中部分单词可以充当完整语义载体, 这种词就用空格分割的方法可以直接获取, 也就是分词获取, 而作为

完整语义载体的上下关联单词的稳定组合却需要采用特殊方法来获取, 也就是组词问题。从数据挖掘的角度上看, 这种相邻单词的稳定组合是一种关联模式^[3], 在一个或内容相近的多个文本中会多次出现, 是一种频繁模式^[4]。因此, 可以将维吾尔文组词问题看成数据挖掘中的频繁模式挖掘问题来解决, 其关键是设计一种适合于维吾尔文的频繁模式挖掘算法。

目前已有多种频繁模式挖掘算法, 如古典 Apriori 及其改进算法^[5], FP-tree (Frequent Pattern tree)^[6] 的频繁模式挖掘算法^[7-8], 基于 Top-K 项的频繁模式挖掘算法^[9] 等。但这些算法都是无先验信息的模式搜索算法, 因此都具有较复杂的数据结构或较高的计算量等特性^[10], 而且也不能直接用于维吾尔文中。本文完全从维吾尔文的文字特点出发, 再将无先验信息的频繁模式挖掘问题转化为已知的特定模式搜索问题, 设计出了一种增量式搜索的维吾尔文完整频繁模式挖掘算法, 从而解决了维吾尔文组词问题。

1 维吾尔文本处理中的新概念——组词

1.1 传统分词方法

从文字表面上看, 维吾尔文本是以空格隔开的词序列, 因此维吾尔文中没有把分词看成技术问题, 就以空格作为自然

收稿日期: 2012-04-10; **修回日期:** 2012-05-22。 **基金项目:** 国家自然科学基金资助项目 (61063022; 61262062; 61163033; 61142004); 新疆维吾尔自治区高技术研究发展计划项目 (201212124); 教育部新世纪优秀人才支持计划项目 (NCET-10-0969)。

作者简介: 吐尔地·托合提 (1975-), 男, 新疆乌鲁木齐人, 副教授, 博士研究生, CCF 会员, 主要研究方向: 互联网搜索、挖掘与内容安全; 维尼拉·木沙江 (1960-), 女, 新疆乌鲁木齐人, 教授, 主要研究方向: 信息检索; 艾斯卡尔·艾木都拉 (1972-), 男, 新疆乌鲁木齐人, 教授, 博士生导师, CCF 会员, 主要研究方向: 多语种信息处理。

分隔符简单获取文本中单词,是到目前为止唯一的分词方法。但在很多情况下,由上下关联的若干个维吾尔文词的稳定组合来完整地表达一个名词、专业术语或新词的语义,因为传统的方法丧失了这种单词组合共现而表达的完整语义,因此难以满足维吾尔文文本处理中的分词需求。如表1所示的一个维吾尔文语义词,如按空格隔开提取为两个词,则整词表达的

原有语义就完全丧失,而无空格字符串的英文单词,就不会受到空格分词的影响。因为,带空格字符串形式的一个维吾尔文语义词,以空格分割提取为若干个维吾尔文单词,这就失去了它们的上下关联组合表达的完整语义,因此从这些单词中选取的特征也很难表征文本主题,或难以在文本标引中发挥作用,因而无法提高文本处理效率。

表1 维吾尔文传统(空格)分词与英文对比实例

语种	整词(组)提取	空格分词
英文	Software College	Software College
维文	يۇمشاق دېتال ئالىي مەكتەپ	ئالىي مەكتەپ يۇمشاق دېتال

1.2 维吾尔文组词思路

在统计意义上看,如果某个关联模式在文本中频繁出现,那么该模式就越容易被发现。同样,如果某个关联模式被重复观察到的次数越多,那么该模式构成语义词的可能性就越大。因此,本文从数据挖掘中的关联规则与序列模式挖掘方法中找到了启发,提出了一种基于频繁模式挖掘的维吾尔文智能组词思路。

2 基于频繁模式挖掘的维吾尔文组词算法

根据维吾尔文组词思路,为了能够从文本集中准确获取成为语义词语的频繁模式,严格定义了频繁模式的结构完整性,再针对维吾尔文文字特性及可能出现各种特殊情况,采取了特定措施并设计出了一种适合于维吾尔文的频繁模式挖掘算法。

2.1 完整频繁模式的定义

将文本看成一个长词串 S ,则本文提出的维吾尔文组词方法中,完整频繁模式 P 为满足以下条件的频繁模式:

1) 在词串 S 中出现次数超过两次的维吾尔文单词关联模式,是多词连续的序列模式。

2) 假定 P 出现在 S 中的 n 个不同位置 r_1, r_2, \dots, r_n ,至少存在一对 $\langle i, j \rangle (1 \leq i < j \leq n)$,使得第 $(r_i - 1)$ 个词和第 $(r_j - 1)$ 个词不相同,此时 P 为左最大化的频繁模式;或者至少存在一对 $\langle i, j \rangle (1 \leq i < j \leq n)$,使得第 $(r_i + |P|)$ 个词和第 $(r_j + |P|)$ 个词不相同,此时 P 为右最大化的频繁模式^[1]。简单地说,给定一个频繁模式 P 的两次出现 P_1 和 P_2 ,任何向左或向右扩展一个维吾尔文单词的操作都会使 P_1 和 P_2 变得不相等。也就是说,完整频繁模式就是不可向左和向右扩展的频繁模式。

2.2 频繁模式挖掘中的语言问题

完整频繁模式被定义确定之后,还得考虑以下维吾尔文文字特性,并采取针对性的处理措施,否则很难搜索到完整频繁模式,从而降低组词准确率。

1) 文字拼写不规范性:规范拼写的维吾尔文本中,词与词以一个空格隔开,但也常会出现两个单词被拼接或以多个空格隔开的不规范拼写现象。如:组成语义词“يۇمشاق دېتال”(软件)的两个单词被拼接写成一个单词“يۇمشاقدېتال”,或被多个空格隔开写成“يۇمشاق دېتال”,就变成完全不同的模式串,从而不能匹配成为频繁模式(□表示空格符)。

2) 词缀构形:维吾尔文中的一个词,常会以不同词形(词干+构形词缀)在文本中多次出现,词干是词去掉构形附加成分后剩下的部分,它包含着词的词汇意义。假如说,一个文

本中多次出现同词干“يۇمشاق دېتال”(软件),但不同构形后缀(نى, لار, نىڭ, دىن)的词“يۇمشاق دېتالنى”, “يۇمشاق دېتاللار”, “يۇمشاق دېتالنىڭ”, “يۇمشاق دېتالدىن”,既然这些词在语义上是同一个词,但因为词形不同,不能匹配成为一个频繁模式。

2.3 多词频繁模式挖掘算法

频繁模式挖掘问题可简单描述为:给定一个文本集,要求输出在一个或多个文本中出现次数大于指定次数(本算法中设为2)的子串,即由多个词组成的连续单词序列。

输入文档集 D ,希望得到每一个文本中能够成为语义词的,长度大于2(2个维吾尔文词)的完整频繁模式,算法流程可以分为词干切分、伪文本构造、模式搜索3个步骤,如图1所示。

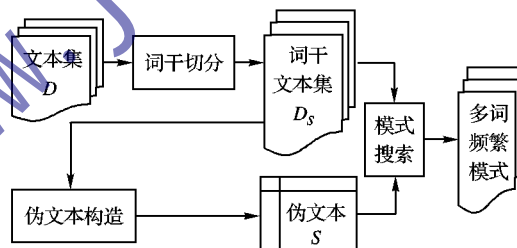


图1 维吾尔文频繁模式挖掘算法流程

1) 词干切分:为了排除词缀构形对模式匹配的影响,首先对文本集 D 中的所有文本进行词干切分,得到词干文本集 D_s 。

2) 伪文本构造:算法依次扫描文本集 D_s 中的每个文本,将维吾尔文标点符号,以及非维吾尔文字母都替换成定界符“\$”。然后把每一个文本视为一个字符串,文本中的词按顺序拼接(去除词间空格),再将全部文本以定界符“\$”连接成一个伪文本 S 。

3) 模式搜索:从当前文本中第一个单词开始往左组词,如 $(W_1 W_2), (W_1 W_2 W_3), (W_1 W_2 W_3 \dots)$,并把当前组合作为搜索窗口,在伪文本 S 中的有效范围 $(\$ \dots \$)$ 内进行模式匹配,因为跨越文本、文本内句子和句子内语段边界的频繁模式是没有意义的(不能成为语义词语),这样降低了我们的频繁模式搜索算法的代价。

为了排除不规范书写(词间无空格或多空格隔开现象)对模式匹配的影响,我们要过滤词间空格并把当前组内多个词拼接为一个词,如将单词组合(يۇمشاق دېتال)拼接为(يۇمشاقدېتال),然后将这个无空格词串作为特定模式,在无空格伪文本 S 内进行快速单模式匹配。但是,我们最后要得到的是以空格隔开的单词组合,因此拼接单词之前要保留原来的单词组合。

因此,本文算法中设有两个增量式窗口:组词窗口 $GWIN$ 和搜索窗口 $SWIN$ 。单词组合先进入组词窗口,然后按顺序被拼接为一个无空格词串并进入搜索窗口,两个窗口的长度,移动或往左最大化是一致的。组词过程中,以 $SWIN$ 的匹配结果来判定当前词串是否为频繁模式,然后将 $GWIN$ 作为组词结果输出。

假设,文本中单词个数为 m , k 为组词窗口 $GWIN$ 的大小, Pos 为单词指针。搜索算法依次从文本集 D_s 中读取一个本文 d_i ($1 \leq i \leq n$),并初始化组词窗口大小($k=2$, 2个单词)及单词指针($Pos=1$, 指向文本中的第一个单词),然后在当前文本 d_i 内从右到左(维吾尔文文字方向),从 Pos 位置开始提取连续且不是常用词及定界符“\$”的 k 个单词 W_{Pos-j} ($j=0, 1, \dots, k-1; k < m-Pos$);将 k 个词拼接成一个子串,放入到搜索窗口 $SWIN$ 中,然后在伪文本 S 中进行单模式匹配。

这样,将文本 d_i 中提取的 k 个单词的无空格词串作为特定模式,在无空格伪文本 S 中进行快速单模式匹配,将无先验信息的频繁模式搜索问题转化为已知的特定模式搜索问题来实现频繁模式的快速搜索,同时也完全排除了维吾尔文不规范书写对模式匹配的影响,从而提高了维吾尔文频繁模式搜索算法整体性能。 $GWIN_k$ 为长度为 k 的组词窗口,则无空格词串的快速频繁模式搜索过程如图2所示。

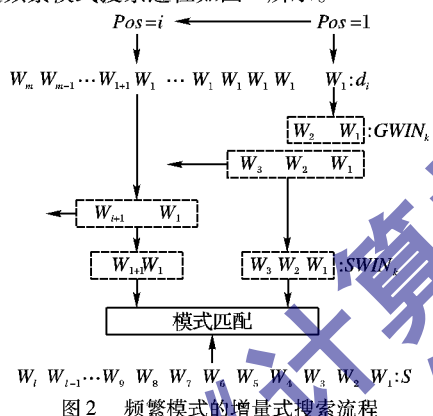


图2 频繁模式的增量式搜索流程

如下情况下,单词指针 Pos 往左移动或跳跃:

1) 当前被指向的是常用词(根据常用词词典判定)或定界符号“\$”,或 $k=1$ 时 $GWIN_k$ 不是频繁模式,则往左移动到下一个单词位置($Pos=Pos+1$);

2) $k>1$ 时, $GWIN_k$ 不是频繁模式,则往左跳跃到新的 Pos 位置($Pos=Pos+k-1$);

2.4 置信度判定及组词输出

在文本 d_i 内,组词窗口的移动与单词指针是同步的。组词窗口每次移动到 Pos 指向的单词位置,窗口大小缩小到 $2(k=2)$,然后在以下条件下继续往左最大化(增大):当前搜索窗口 $SWIN_{k+1}$ 是频繁模式,并且满足一个最小置信度(Minimum Confidence, MC)。也就是说,通过一个最小置信度来预测组词窗口 $GWIN$ 是否可以继续增大^[12]。最小置信度计算公式如下:

$$MC = \frac{SWIN_{k+1}.count}{SWIN_k.count}$$

其中: $SWIN_{k+1}.count$ 表示 $GWIN_{k+1}$ 在 S 中的匹配次数, $SWIN_k.count$ 表示 $GWIN_k$ 在 S 中的匹配次数。在 $GWIN_k(W_1 W_2)$ 的基础上扩展的 $GWIN_{k+1}(W_1 W_2 W_3)$,如 $GWIN_{k+1}$ 的匹配次数突然变小,也就是说置信度太低,说明 $W_1 W_2 W_3$ 不是一个完整频繁模式,是 $W_1 W_2$ 和 W_3 的一种偶然性关联。这种情况下,组词窗口停止最大化,将 $SWIN_k$ 对应的 $GWIN_k$ 作为组词结果输出,并移动组词窗口。以这种方法,边扫描边判定构词置信度,将文

本扫描一遍就得到全部组词结果。

3 组词实验与分析

3.1 实验数据集

为了在更广的范围进行组词实验及评价,从已人工分类的多个领域文本集中,选取3000篇维吾尔文文本作为实验数据。实验数据属于交通、宗教、体育、健康、军事、房产、教育、旅游、经济和电脑等10类文档,每类均为300篇文本。

3.2 评价指标

如果把组成语义词的频繁模式(词组)搜索问题看成是一种特定的信息检索问题,则主要的评价指标是组词准确率(Precision),组全率(Recall)及它们平衡指标 $F1$ 值。

P (组准确率) = 组词正确的词语数 / 实际组词的词语数

R (组全率) = 组词正确的词语数 / 应有的词语数

$F1 = 2PR / (P + R)$

3.3 实验结果及分析

本算法组词中,影响评价指标的一个重要参数是最小置信度 MC 的阈值,因此在 MC 的不同阈值情况下观察了组词评价指标。实验中发现,当 $MC \leq 0.5$ 时的组准确率 P 普遍很低,因此本文只给出了 MC 在 $[0.6, 1.0]$ 取值时的组词结果,如表2。

表2 不同 MC 阈值下的组词结果

类别	$MC=0.6$		$MC=0.7$		$MC=0.8$		$MC=0.9$		$MC=1.0$	
	P	R	P	R	P	R	P	R	P	R
交通	80.1	89.9	84.0	88.1	87.4	83.3	90.8	81.6	92.8	79.1
宗教	81.3	88.6	85.2	87.3	86.9	83.1	89.1	80.8	91.2	78.1
体育	82.0	88.3	87.5	87.7	88.5	82.7	89.3	80.1	92.1	79.2
健康	83.7	89.1	89.6	86.9	91.3	83.2	93.1	80.7	95.1	78.8
军事	81.5	89.1	86.2	88.2	88.6	84.2	91.1	81.5	93.4	78.8
房产	70.5	84.6	75.8	82.6	77.2	77.7	82.2	74.5	84.9	73.3
教育	80.2	91.2	84.6	89.1	86.6	82.1	88.9	78.3	92.8	77.2
旅游	82.5	77.3	85.4	75.2	88.6	71.9	91.2	70.4	93.1	68.8
经济	81.8	88.2	86.9	86.1	88.7	83.3	91.8	81.2	93.4	78.5
电脑	80.4	89.4	86.2	85.9	88.9	81.6	90.2	78.9	93.2	77.4
平均	80.4	87.6	85.1	85.7	87.3	81.3	89.8	78.8	92.2	76.9
$F1$	83.8		85.4		84.2		83.9		83.9	

当 MC 在 $[0.6, 1.0]$ 取值时,组准确率 P 开始逐步提高,但组全率 R 却开始下降。当 $MC=0.7$ 时, $F1$ 值为最高,在 $[0.7, 1.0]$ 取值时, $F1$ 值也相对稳定。 MC 不同取值对评价指标影响如图3所示。

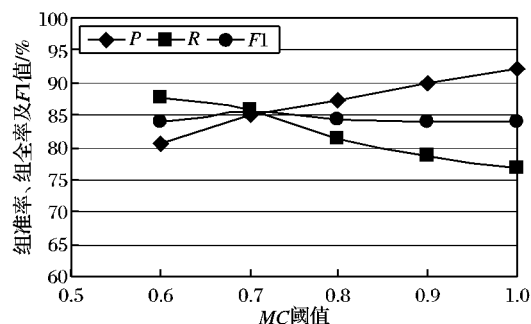


图3 MC 取值变化对评价指标的影响

通过组词对比分析,发现以下两种情况对组词效率影响较大:1) 拼写错误;2) 名词术语的不规范缩写。因此,如果在预处理阶段采取一种自动的方法,对文本进行拼写校对和名词术语规范化,得到较高的组全率是完全可能的,这也是本

(下转第2926页)

显的改进。其中“旅游”这一类提高幅度最大,达到了13.88%，“教育”类提高幅度最低,有6.07%。分析原因是因为很多旅游相关名称如“Boghda köli(天池)”、“Dêhقانlar xatliqi(农家乐)”、“Qanas köli(喀纳斯湖)”、“Menzire noqtisi(景点)”等都是两个词组成的短语,如果将两个词分开处理会降低其类别区分能力。只有两个词组合在一起,它就携带较高的类别信息量,进而对分类的贡献也就比较大。而在教育类中出现频率较高的“Ali mektap(大学)”,“Oqutquchi oqughuchi(师生)”,“Oqutux matëriyali(教材)”,“Oqutqushilar qoxuni(师资)”等各短语中至少有一个单词本来就携带较高的类别信息量,因此对该类而言,短语作为特征项分类的 $F1$ 值提高幅度比起其他类别要低一点。

表3 分类实验结果对比

类别	单词			单词+短语		
	P	R	$F1$ 值	P	R	$F1$ 值
政治	0.7170	0.7600	0.7379	0.9286	0.7800	0.8478
经济	0.6630	0.8133	0.7305	0.7528	0.8933	0.8171
体育	0.7547	0.8000	0.7767	0.9245	0.7656	0.8376
旅游	0.7037	0.7600	0.7308	0.8000	0.9524	0.8696
教育	0.7969	0.9444	0.8644	0.8831	0.9714	0.9251
文化	0.7692	0.8000	0.7843	0.8065	0.9524	0.8734

4 结语

文本分类的进一步改进除了算法方面,应该还立足于影响文本分类最底层、最根本的因素:文本表示中的特征项的选择,提高特征项的完整独立程度。在维吾尔语中短语具有较强文本表示功能,在表示文本时,能将文本的内容特征(如主题类别)鲜明地表示出来。本文提出了基于统计方法的维吾尔语短语抽取算法,将抽取到的短语作为文本特征进行了分类实验。实验结果表明,短语作为文本特征明显提高维吾尔文文本分类的准确率和召回率。

参考文献:

- [1] 苏金树,张博峰,徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报,2006,17(9):1848-1859.
- [2] 刘华. 基于关键短语的文本分类研究[J]. 中文信息学报,2007,

21(4):34-41.

- [3] CAROPRESO M F, MATWIN S, SEBASTIANI F. Statistical phrases in automated text categorization, Statistical Phrases in Automated Text Categorization [R]. Paris: Centre National de la Recherche Scientifique, 2000: 78-102.
- [4] KOSTER C, BENEY J. Phrase-based document categorization revisited[C]// PaIR'09: Proceedings of the 2nd International Workshop on Patent Information Retrieval. New York: ACM, 2009: 49-55.
- [5] 张爱华,荆继武,向继. 中文文本分类中的文本表示因素比较[J]. 中国科学院研究生院学报,2009,26(3):400-407.
- [6] 李钝,曹付元,曹元大. 基于短语的文本情感分类研究[J]. 计算机科学,2008,35(4):132-134.
- [7] 阿力木江·艾沙,吐尔根·依布拉音,艾山·吾买尔. 基于机器学习的维吾尔文文本分类研究[J]. 计算机工程与应用,2012,48(5):110-112.
- [8] 张震,胡学钢. 基于互信息量的分类模型[J]. 计算机应用,2011,36(6):1678-1680.
- [9] VAPNIK V. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995.
- [10] JOACHIMS T. Text categorization with support vector machines: Learning with many relevant features[C]// European Conference on Machine Learning. Berlin: Springer-Verlag, 1998: 137-142.
- [11] 孙建涛,郭崇慧,陆玉昌,等. 多项式核支持向量机文本分类器泛化性能分析[J]. 计算机研究与发展,2004,41(8):1321-1326.
- [12] HSU C-W, LIN C-J. A comparison of methods for multiclass support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2):415-440.
- [13] CHANG C-C, LIN C-J. LIBSVM: A library for support vector machines [EB/OL]. [2011-09-10] <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.
- [14] SEBASTIANI F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34(1):1-47.
- [15] 施聪莺,徐朝军,杨晓江. TFIDF 算法研究综述[J]. 计算机应用,2009,29(S1):167-170.

(上接第2922页)

文下一阶段的研究重点。

4 结语

本文探讨传统分词方法导致的词语义不完整性问题,并提出了一种基于频繁模式挖掘的维吾尔文组词的新思路,并将无先验知识的频繁模式挖掘问题转化为特定模式的匹配问题,设计出了一种简单而高效的智能组词算法。该算法仅需对文本进行一次扫描,不需产生候选模式集,就挖掘出能作为语义词的完整频繁模式集。实验结果证明本算法的正确性和实用性,本算法也可以直接引用到同语系的哈萨克文和柯尔克孜文的组词中,具有一定的推广意义。

参考文献:

- [1] TOHTI T, HAMDULLA A, MUSAJAN W. Research on Web text representation and the similarity based on improved VSM in Uyghur Web information retrieval[C]// CCPR 2010: Chinese Conference on Pattern Recognition. Chongqing: [s. n.], 2010: 984-988.
- [2] 阿力木江·艾沙,吐尔根·依布拉音,艾山·吾买尔,等. 基于机器学习的维吾尔文文本分类研究[J]. 计算机工程与应用,2011,48(5):110-112.

- [3] 刘晓涛,郭福亮. 一种有趣关联模式挖掘方法[J]. 计算机工程,2010,36(11):36-38.
- [4] 朱琼,施荣华. 一种数据流中的频繁模式挖掘算法[J]. 计算机应用,2008,28(6):1463-1466.
- [5] 刘兵. Web 数据挖掘[M]. 1版. 北京:清华大学出版社,2009.
- [6] 宋余庆,朱玉全,孙志挥,等. 一种基于频繁模式树的约束最大频繁项目集挖掘及其更新算法[J]. 计算机研究与发展,2005,17(5):777-783.
- [7] 张锦,马海兵,胡运发. 一种基于 FP-Tree 的频繁模式挖掘自适应算法[J]. 模式识别与人工智能,2005,18(6):763-768.
- [8] 李也白,唐辉,张淳,等. 基于改进的 FP-tree 的频繁模式挖掘算法[J]. 计算机应用,2011,31(1):101-103.
- [9] 敖富江,杜静,陈彬,等. 一种基于混合搜索的高效 ToP-K 最频繁模式挖掘算法[J]. 国防科技大学学报,2009,31(2):90-93.
- [10] 马青霞,李广水,孙梅. 频繁模式挖掘进展及典型应用[J]. 计算机工程与应用,2011,47(15):138-144.
- [11] 花红娟,张健,陈少华. 基于频繁模式树的约束最大频繁项集挖掘算法[J]. 计算机工程,2011,37(9):78-80.
- [12] 肖波,徐前方,蔺志青,等. 可信关联规则及其基于极大团的挖掘算法[J]. 软件学报,2008,19(10):2597-2610.