广西科学院学报
Journal of Guangxi Academy of Sciences

2018,34(1):39～45
Vol. 34, No. 1 February 2018

网络优先数字出版时间:2018-02-27 DOI:10.13657/j.cnki.gxkxyxb.20180227.002
网络优先数字出版地址:http://kns.cnki.net/kcms/detail/45.1075.N.20180227.1044.004.html

藏文文本聚类及其相关技术综述*

A Summary of Tibetan Text Clustering and Its Related Technologies

李玖一,于洪志,徐 涛**
LI Jiuyi, YU Hongzhi, XU Tao

(西北民族大学,中国民族语言信息技术重点实验室,甘肃兰州 730030)
(China National Languages Key Laboratory of Information Technology, Northwest University of Nationalities, Lanzhou, Gansu, 730030, China)

摘要:藏文作为一门古老的语言有其独有的规则和特点。随着网络的普及,互联网用户中的藏族同胞迅速增加,网络上的藏文文本也越来越多。利用藏文文本聚类来提供更高效的管理和更良好的用户体验成为近年的研究热点。本文首先介绍了藏文文本聚类的应用背景和相关概念,然后介绍了藏文文本特点和藏文文本聚类的相关技术,讨论了藏文文本建模和聚类算法,最后对藏文聚类发展和应用进行了总结和展望。

关键词:藏文文本 聚类算法 文本建模
中图分类号:TP391 **文献标识码**:A **文章编号**:1002-7378(2018)01-0039-07

Abstract: As an ancient language, Tibetan language has its own unique rules and characteristics. With the popularization of the Internet, the number of Tibetan compatriots among Internet users has increased rapidly, and there are more and more Tibetan texts on the Internet. Using Tibetan text clustering to provide more efficient management and user experience has become a hot topic in recent years. This article first introduces the application background and related concepts of Tibetan text clustering. Then it introduces the characteristics of Tibetan texts and the related technologies of Tibetan text clustering. After that, the Tibetan text modeling and clustering algorithms of Tibetan text are discussed. Finally, the development and application of Tibetan language clustering are summarized and prospected.

Key words: Tibetan text, clustering algorithms, text modeling

0 引言

在网络发展日新月异的今天,国内广大藏族地区的藏族人民生活和互联网的关系越来越紧密,藏文也随之繁荣起来。随着藏文网络数据与日俱增,热点监测、话题追踪与主题识别等问题亟需高效智能的方法加以解决。同时,世界范围内,各种技术和势力风起云涌,网络信息传播速度之快、范围之广前所未有,这对于青藏地区的安定和发展有至关重要的影响。正是这些因素促使藏文文本处理成为近年来中文信息处理中的热门领域。

藏文是一种古老的拼音文字,属于汉藏语系[1],目前现行的藏文已有 1 300 年的历史。藏语中的词汇和语法十分丰富,还有复杂的敬语体系,加之藏语

收稿日期:2017-12-20
修回日期:2018-01-09

作者简介:李玖一(1994－),女,硕士研究生,主要从事自然语言处理研究,E-mail:lee.91@foxmail.com。

*"民族特色农产品多语言网络交易展示平台关键技术集成与应用示范"项目(2015BAD29B01)资助。

**通信作者:徐 涛(1987－),男,博士,副教授,主要从事自然语言处理研究,E-mail:alfredxly@163.com。

语法的规则体系严密[2]，使得早期的藏文自然语言处理工作大都是基于规则实现。然而随着近年来互联网用户的与日俱增，网络数据大爆炸式地涌现，信息时代数据所具有的多样性和时效性等特点同样在藏文上突显出来。这就对藏文文本信息处理的效率和方式提出了更高的要求。

聚类作为一种高效的无监督学习方法，被广泛应用于各个领域。从影响商务决策到图像识别，无一不体现出聚类在知识发现领域中的优势。在文本挖掘的应用中，聚类是下一阶段分析的预处理环节[3]。文本聚类可以将文档分成若干不同的主题，提高网页搜索效率；在电子档案管理的过程中，利用聚类可以进行民意分析、政策倾向、民生热点等深层的文本挖掘[4]；在微博、新闻热点等话题追踪中，文本聚类可以高效辅助舆情分析[5]。基于聚类在文本处理应用中的种种优势，藏文文本聚类研究应运而生。

1 聚类相关概念及划分

聚类是一种常用文本挖掘方法，是无监督的机器学习过程，即不需要标注语料等工作进行人工干预[6]。聚类将经过预处理的数据通过合适的算法聚合成不同的子集(簇)。在各个簇的内部数据对象是相近的，和其他簇相比则不相近。因此内容相似的文本会出现在同一簇中，而来自不同的簇的文本其相似性则较低。

聚类的具体过程：将已知的文本集进行分词、命名实体识别和去停用词等预处理，经过预处理的文本包含了更多有价值的、主要的文本信息，将这些文本经过特征提取以后进行文本建模，通过数学的方法计算不同文本的相似度，再结合文本的特点选择适合的聚类算法得到聚类划分(图1)。

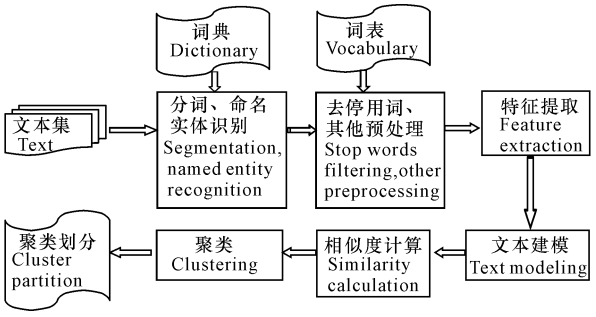

图1 聚类整体流程

Fig. 1 The whole process of text clustering

常用的聚类方法有很多种。按照基本原理可以分为基于划分的方法、基于模型的方法、基于蚁群的方法以及层次聚类、SOM网络。划分法的代表算法有k-means方法，其总体思想是先初始k个划分，然后不断迭代计算文档所属类型，最终收敛得到聚类结果。这种方法简单易行，但结果受初始点和极端值的影响较大。自组织神经网络(Self-Organization Map，SOM)通过训练神经网络，较好地确保了文档之间的特性，目前在文本聚类上有较广泛的应用，但也存在确定参数困难和运行复杂的缺点。

2 藏文文本聚类关键技术

2.1 藏文分词

藏文分词是藏文信息处理的基础。分词即将原始文本语料分割成若干独立的词的操作。藏文在语法体系中属于逻辑格语法体系[7]。尽管和汉语、日语等语言有所差别，但它们在词边界上都具有共同特点，即词与词之间没有像英语一样的天然间隔，在进行文本处理之前必须要做好分词工作。分词的结果在一定层面上决定了后面聚类的效果。

20世纪末，扎西次仁[8]通过最大匹配算法实现了一个藏文分词演示系统，藏文分词工作正式进入人们的视野。2003年，陈玉忠等[9]率先提出了根据格助词和藏文接续特征的分词方法，该方法结合藏文字词形态特点和语法知识库，使得藏文分词具有实用性。才智杰[10]结合藏文文本中紧缩词的添加规则，得到经过复原的藏文文本，该方法使得分词效果得到显著提升。祁坤钰[7]从藏文特点出发，通过对语法逻辑、语义以及音势等藏语语言理论的深入研究，设计了匹配切分分词方案。该方法按藏文格划分层级，可以更好地理解文本词的主次关系。

以上这些藏文分词方法虽然涉及一些如词频之类基于统计的方法，但主要是以藏语语法规则作为切入点。随着近年来网络数据的激增，藏文文本也出现了大规模的增长，大量的语料为机器学习在藏文信息处理中的应用注入了新的动力。

2011年，史晓东和卢亚军[11]率先将基于隐马尔科夫模型(HMM)的汉语分词系统移植到藏文处理中，同时对命名实体识别和训练语料方面做了改进，使得分词结果有了进一步的提高。刘汇丹等[12]实现了SegT藏文分词系统，该方法是先对临界词做识别，同时对格助词进行分块处理，再通过逆向最大匹配算法(RMM)分词，并对紧缩词进行识别，达到了很好的处理结果。

另外，在结合条件随机场方法上，藏文分词研究也有许多成果。何向真等[13]通过有针对性地判断

各个音节在词中所处位置来对藏文文本进行切分,在同等实验条件下,对最大间隔马尔可夫网络、条件随机场(CRF)、最大熵等模型下实现的分词系统进行对比,得到的结果表明条件随机场更适合处理分词问题。李亚超[14]采用基于统计的方法对紧缩词进行识别,并运用最大熵与条件随机场相结合的方式,对藏文进行命名实体识别,取得了很好的分词效果。洛桑嘎登等[15]在条件随机场的基础上,利用语法规则进行再处理,分词效果得以提升。

2.2 命名实体识别

命名实体识别在藏文自然语言处理中扮演重要角色,重点是为了将文本中地名、人名等专有名词自动辨别出来,并对这些词进行分类[16]。从语言分析的角度来说,这一步骤是难度最高、最影响后续文本处理效果的处理工作。国内外对于藏文这一方面的研究较少,目前主要的藏文命名实体识别方法有以下3种:

(1)基于规则的藏文命名实体识别。祁坤钰[17]以大规模藏文语料库为基础,对藏语名词进行充分地分析和分类描述,建立了高质量的词表。金明等[18]设计了一个基于隐马尔可夫模型和语法规则的方法,用于藏文命名实体识别。窦嵘等[19]研究了藏族人名的用字特征后,结合命名规则与词典,通过基于统计的方法给出了藏文人名识别模型。然而运用语言学规则进行识别是最传统的模式,对于未登录词、歧义问题处理能力较弱。

(2)基于条件随机场的藏文命名实体识别。这种方式更加灵活,通用性强,但是需要大量语料作为支撑。藏文发展至今,语料规模较之以前有了很大的提高,但相比于其他流行语言还是略显不足。因此,基于条件随机场的藏文处理往往还有人工干预的处理步骤。何向真等[13]采取大量藏文文本语料进行机器学习,从而提出了最大熵和条件随机场相结合的方法。康才畯等[20]从字的粒度上对藏文人名进行识别和划分,该方法充分考虑了藏文的上下文特征,以判断人名在文本中的位置。

(3)基于感知机的藏文命名实体识别。华却才让等[21]提出基于音节特性的感知机方法,用于识别藏文中的专有名词,该训练方法运用藏文紧缩格辨认音节来达到藏文命名实体自动识别的目的。

2.3 停用词过滤

在文本中往往会出现一些高频但是对文章主旨无太大贡献的词,这样的词即是停用词。停用词过滤又称去停用词,这一过程对于文本数据的清洗十分重要,因为它使得文本向量空间的维度减小,从而节省存储空间。停用词选择方式通常有词频、文档频次、熵计算等。然而停用词的选取往往需要语言学科知识作为支撑,对不同的应用场景所选取的策略也应作必要的更新[22]。才让三智[23]构建了藏语格助词、自由虚词和不自由虚词等虚词知识库,并对虚词的语法分类问题和语言模型的结构框架进行阐述,同时还指出藏语中虚词归类和虚词兼类的问题,这两个问题的处理对藏文文本划分和歧义理解有直接影响。阿雅娜[24]也指出了少数民族的语言特点对停用词处理的影响。珠杰等[25]在分析停用词处理对比实验结果后,总结出语法知识和自动处理相结合的停用词选取方法。

3 藏文文本建模

3.1 特征提取与文本建模

经过预处理的文本,去除了与文本内容和主题相关性不大的元素,大大压缩了文本空间,但对于具体的聚类算法,这样的集合仍然是高维的。因此在不影响结果的前提下,需要尽可能地减少需要处理的词,将最有效的特征选取出来以降低样本空间。特征的选取方法有很多,如文档频率(Document Frequency)[26]、互信息(Mutual Information)[26]、信息增益(Information Gain)[26]、卡方检验(CHI)[26]、期望交叉熵[27]、联合熵[28]等。贾会强等[29]引进G函数来表示特征词的词频,同时通过基于词形特性的提取方式,验证了其提取特征词的效率和降低文本向量空间维度的效果,实验结果有效地缩减了向量维度,同时提高了分类的准确性。

欲对藏文文本聚类,首先要解决的问题是将藏文文本转化为算法可以接受的形式,这一步就是文本建模。现行的文本建模主要是采用向量空间模型(VSM)、词向量、后缀树、主题模型以及本体等方法。江涛和于洪志[30]提出一种面向藏文聚类的文本建模方法,该方法基于词向量构建出藏文文本表示模型,在此基础上得到的结果均优于基于主题模型和VSM的聚类结果(图2)。

徐涛等[31]结合空间向量模型来切分藏文文本中的句子,将卡方统计量应用于文本词项和对比词项相关程度的计算。经过实验对比,在对比词项组H所占文档词项的比例 H_p 为35%时,得出的准确率最高,可以使相似性高低结果明显区分。该方法与传统模型相比,对藏文文本有更好的识别度和区分度。

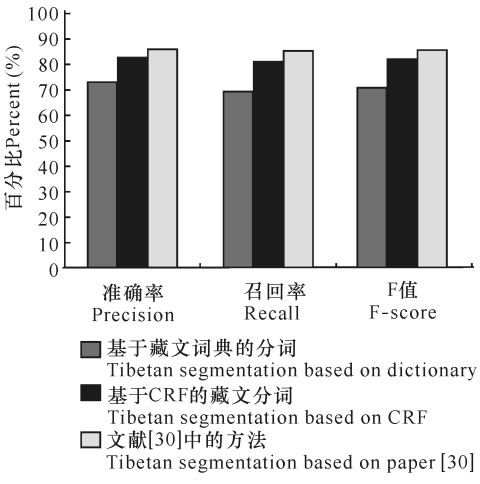

图2 藏文文本处理实验结果对比[30]

Fig. 2 The comparison of experimental results in Tibetan text processing[30]

3.2 相似度计算

文本聚类的依据是各文档之间的相似性,藏文文本聚类也不例外。文本的相似与否需要通过相似度来度量。目前针对藏文提出的文本相似度计算方法少之又少,主要是依据藏文文本特点建模后,将现有的通用相似度评价方法移植过来。根据对语句分析的深度划分,主要有两种策略:基于统计学的策略,对文本做完全句法和语义分析的策略[32]。前者所用到的计算方法主要有基于VSM、广义向量空间(GVSM)、隐语义索引模型以及其他数学手法[33],后者包括基于语言或本体模型等方法[34]。从粒度上来讲,则有短语或词之间的相似性计算、句子之间的相似性计算、段落间的相似性计算以及文档间的相似性计算等。通常对于表示成VSM的文档和文档集合,相似性计算主要包括余弦法、Jaccard系数法以及欧氏距离函数法等。

利用字或词向量的余弦公式计算相似度,是较常用的一种方法。文档向量采用词作为向量特征单元,通过余弦公式计算两个向量之间的夹角,计算公式如下:

$$D(di,dj)=\cos(di,dj)=\frac{\vec{di}\cdot\vec{dj}}{|\vec{di}|\times|\vec{dj}|}=\frac{\sum_{k=1}^{n}w_{i,k}\times w_{j,k}}{\sqrt{\sum_{k=1}^{n}w_{i,k}^2}\times\sqrt{\sum_{k=1}^{n}w_{j,k}^2}},\quad 1)$$

其中 di 和 dj 分别表示两个被比较的单位向量,这里表示词向量,则 $w_{i,k}$ 表示第 i 个文档集合中第 k 个词的权值,$w_{j,k}$ 表示第 j 个文档集合中第 k 个词的权值,n 表示文档中词的总个数。

余弦法实现简单,但忽略了上下文所包含的信息。Jaccard系数法考虑文本中全部不重复的词语,考察两篇文本中所有词语(去重)的交集(记为interSectionSize)与所有词语(去重)的并集(记为unionSize)的比值。见公式2)。

$$SIM=\frac{interSectionSize}{unionSize}。\quad 2)$$

尽管这样的算法一定程度上避免了“一叶障目”的问题,但也在一词多义和多词一义的处理上有局限性。

分段融合的意思是将文本按照段落、句子的层级划分,首先计算句子的相似度,再融合为上一层级——段落的相似度,最后将其融合形成篇章的相似度。邬明强[35]在分析已有的中文分段融合相似度计算方法的基础上,提出了藏文分段融合的相似度计算方法。通过选择段落权重构建相似度矩阵,利用向量夹角预先公式求出段落相似度;而后再根据段落对文章意思的贡献,对文本相似度做精确化处理。目前对藏文文本的相似度计算领域中,还没有进行句子间的深层句法分析并在此基础上计算相似度的报道。

4 藏文文本聚类方法及应用近况

经过前期的处理之后,此时的藏文文本已经成为结构化的语料,因此藏语本身对聚类算法并无太大的影响。常见的聚类算法包括但不限于以下这些方法,如k-means[36]、层次聚合聚类(AHC)[37]、SOM[38]、基于蚁群的方法等。

目前藏文文本聚类的研究主要集中在应用方面,而不是聚类算法本身。在互联网飞速发展的大背景下,藏族网民的增速尤为突出。藏文聚类能用于迅速归类文档的重要信息,同时为藏族地区的网络舆情状况监控提供技术支持。藏文文本聚类的应用主要集中在两个方面:新闻热点和网络舆情监测。

江涛[39]首次将聚类方法用于藏文网络舆情分析,提出藏文热点发现的模型,开发出用于藏文舆情分析的热点发现模块。邓竞伟等[40]结合网络舆情的复杂网络传播特征对藏文热点话题展开挖掘,并根据热点聚类,从而开发出自动追踪事件的功能。但藏文网络新闻受众规模不大,语料也少,因此曹晖等[41]引入子话题的概念,对藏文新闻这类规模较小的语料对象有良好的效果。这种方法得出的聚类结果避免了文本先后次序的影响,但是在训练语料分布不均衡的时候,结果容易受到影响。康健等[42]则考虑了群体智能具备而传统聚类不具备的特点,将

这一技术应用于藏文网页文本聚类,提出了基于群体智能的半结构化藏文 Web 文本聚类(SCAST)算法。这种算法最终获得的聚类正确率相比于 k-means 有大幅提升,具有快速、高效的特点。但由于其基本的模型采用空间向量模型,所以该模型的一些如忽略上下文和语义关系的劣势它也同样存在。袁斌等[43]根据藏文语法特点和句子结构选取特征,构建了语义空间,并对几种常用的分析方法进行整合以及实验对比,使得藏文情感分类效果得到优化。康健[44]引入蚁群算法,同时结合群体智能和多智能体系统,设计了藏文热点舆情管理系统,将藏文舆情管理系统化。

5 展望

本文对藏文文档聚类中比较重要的几个关键技术:藏文文本预处理、文本建模、相似度计算等做了概述,同时也介绍了近年来藏文文本聚类算法的研究情况和应用发展情况。藏文文本聚类的发展还处于初级阶段,目前对于藏文文本预处理环节,诸如藏文字词的处理正逐渐走向成熟,但在语义分析、文本建模等深入的语言处理方面还有许多值得深究的地方。在未来信息化更加普及的情况下,会有越来越多的藏文文本数据急速增长,将有越来越多的地方需要用到藏文文本聚类分析。具体的应用需求和侧重点不同,根据算法优缺点所选取的聚类方法也会有所不同,因此并不存在一个适用于所有藏文文本的聚类算法。加上藏文有其特有的语音语法特点,因此藏文聚类在自然语言理解、信息检索、信息提取等更高层面的研究还有很长的路要走。

参考文献:

[1] 白玛玉珍.藏文文字特征提取方法的研究[J].电脑知识与技术,2013,9(28):6362-6364.
BAIMA Y Z. Tibetan text feature extraction method [J]. Computer Knowledge and Technology, 2013, 9 (28):6362-6364.

[2] 尼玛扎西.藏文信息处理技术的现状、存在的问题及其前景[J].西藏大学学报,1997,12(2):1-5.
NYIMA TRASHI. Present situation, existing problems and prospect of Tibetan information processing technology[J] . Journal of Tibet University, 1997, 12(2): 1-5.

[3] 刘远超,王晓龙,徐志明,等.文档聚类综述[J].中文信息学报,2006,20(3):55-62.
LIU Y C, WANG X L, XU Z M, et al. A survey of document clustering[J]. Journal of Chinese Information Processing, 2006, 20(3): 55-62.

[4] HATZIVASSILOGLOU V, KLAVANS J L, HOLCOMBE M L, et al. Simfinder: A flexible clustering tool for summarization[C]. Proceedings of the Workshop on Summarization in NAACL 01. Pittsburg, 2001.

[5] 邓竞伟,邓凯英,李永生,等.基于藏文网络的舆情传播模型[J].计算机系统应用,2013,22(3):209-211.
DENG J W, DENG K Y, LI Y S, et al. Opinion spreading models on Tibetan networks[J]. Computer Systems & Applications. 2013. 22(3): 209-211.

[6] INGERSOLL G S, MORTON T S, FARRIS A L. Taming text: How to find, organize and manipulate it [M]. USA: Manning Publication, 2013: 156.

[7] 祁坤钰.信息处理用藏文自动分词研究[J].西北民族大学学报:哲学社会科学版,2006(4):92-97.
QI K Y. On Tibetan automatic participate research with the aid of information treatment[J]. Journal of Northwest University for Nationalities: Philosophy and Social Science, 2006(4): 92-97.

[8] 扎西次仁.一个人机互助的藏文分词和词登录系统的设计[C]//中国少数民族语言文字现代化文集.北京:民族出版社,1999:322-327.
TASHI TSERING. Design of a word segmentation system for word segmentation and word registration [C]//Essays on the modernization of Chinese minority languages. Beijing: Nationalities Publishing House, 1999: 322-327.

[9] 陈玉忠,李保利,俞士汶,等.基于格助词和接续特征的藏文自动分词方案[J].语言文字应用,2003(1):75-82.
CHEN Y Z, LI B L, YU S W, et al. An Tibetan segmentation scheme based on case-auxiliary words and continuous features[J]. Applied Linguistics, 2003(1): 75-82.

[10] 才智杰.藏文自动分词系统中紧缩词的识别[J].中文信息学报,2009,23(1):35-37,43.
CAI Z J. Identification of abbreviated word in Tibetan word segmentation[J]. Journal of Chinese Information Processing, 2009, 23(1): 35-37, 43.

[11] 史晓东,卢亚军.央金藏文分词系统[J].中文信息学报,2011,25(4):54-56.
SHI X D, LU Y J. A Tibetan segmentation system-Yangjin[J]. Journal of Chinese Information Processing, 2011, 25(4): 54-56.

[12] 刘汇丹,诺明花,赵维纳,等. SegT :一个实用的藏文分词系统[J].中文信息学报,2012,26(1):98-103.
LIU H D, NUO M H, ZHAO W N, et al. SegT: A

practical Tibetan word segmentation system[J]. Journal of Chinese Information Processing, 2012, 26(1): 98-103.

[13] 何向真,李亚超,马宁,等.基于音节标注的藏文自动分词研究[J].计算机应用研究,2015,32(7):1989-1991.
HE X Z, LI Y C, MA N, et al. Study on Tibetan automatic word segmentation as syllable tagging[J]. Application Research of Computers, 2015, 32(7): 1989-1991.

[14] 李亚超.基于条件随机场的藏文分词与命名实体识别研究[D].兰州:西北民族大学,2013.
LI Y C. Study on the Tibetan word segmentation and named entity recognition with conditional random fields[D]. Lanzhou: Northwest University for Nationalities, 2013.

[15] 洛桑嘎登,杨媛媛,赵小兵.基于知识融合的CRFs藏文分词系统[J].中文信息学报,2015,29(6):213-219.
LUOBSANG KARTEN, YANG Y Y, ZHAO X B. Tibetan automatic word segmentation based on conditional random fields and knowledge fusion[J]. Journal of Chinese Information Processing, 2015, 29(6): 213-219.

[16] CHINCHOR N. MUC-7 named entity task definition [C]//Proceedings of the 7th Message Understanding Conference. Virginia, 1998.

[17] 祁坤钰.基于语料库的藏文名词分类与统计研究[J].西北民族大学学报:自然科学版,2012,33(3):44-49.
QI K Y. Classification and statistical study of Tibetan nouns based on corpus[J]. Journal of Northwest University for Nationalities: Natural Science, 2012, 33(3): 44-49.

[18] 金明,杨欢欢,单广荣.藏语命名实体识别研究[J].西北民族大学学报:自然科学版,2010,31(3):49-52.
JIN M, YANG H H, SHAN G R. The studies of named entity recognition for Tibetan[J]. Journal of Northwest University for Nationalities: Natural Science, 2010, 31(3): 49-52.

[19] 窦嵘,加羊吉,黄伟.统计与规则相结合的藏文人名自动识别研究[J].长春工程学院学报:自然科学版,2010,11(2):113-115.
DOU R, JIA Y J, HUANG W. Automatic recognition of Tibetan name with the combination of statistics and regular[J]. Journal of Changchun Institute of Technology: Natural Science Edition, 2010, 11(2): 113-115.

[20] 康才畯,龙从军,江荻.基于条件随机场的藏文人名识别研究[J].计算机工程与应用,2015,51(3):109-111.
KANG C J, LONG C J, JIANG D. Tibetan names recognition research based on CRF[J]. Computer Engineering and Applications, 2015, 51(3): 109-111.

[21] 华却才让,姜文斌,赵海兴,等.基于感知机模型藏文命名实体识别[J].计算机工程与应用,2014,50(15):172-176.
HUA Q C R, JIANG W B, ZHAO H X, et al. Tibetan name entity recognition with perceptron model[J]. Computer Engineering and Applications, 2014, 50(15): 172-176.

[22] 化柏林.知识抽取中的停用词处理技术[J].现代图书情报技术,2007(8):48-51.
HUA B L. Stop-word processing technique in knowledge extraction[J]. New Technology of Library and Information Service, 2007(8): 48-51.

[23] 才让三智.藏语虚词知识库的构建研究[D].兰州:西北民族大学,2015.
CAIRANG S Z. Research on knowledge base construction for Tibetan function words[D]. Lanzhou: Northwest University for Nationalities, 2015.

[24] 阿雅娜.蒙古文停用词表和词干提取对蒙古文文本分类的影响[D].呼和浩特:内蒙古大学,2009.
AYA NA. The impact of Mongolian stop-list and stemming on Mogolian text categorization[D]. Hohhot: Inner Mongolia University, 2009.

[25] 珠杰,李天瑞.藏文停用词选取与自动处理方法研究[J].中文信息学报,2015,29(2):125-132.
ZHU J, LI T R. Research on Tibetan stop words selection and automatic processing method[J]. Journal of Chinese Information Processing, 2015, 29(2): 125-132.

[26] YANG Y M, PEDERSON J O. A comparative study on feature selection in text categorization[C]// Proceedings of ICML-97, 14th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 1997: 412-420.

[27] MLADENIC D, GROBELNIK M. Feature selection for unbalanced class distribution and naive bayes [C]//Proceedings of the Sixteenth International Conference on Machine Learning. Bled: Morgan Kaufmann Publishers Inc, 1999: 258-267.

[28] 顾益军,樊孝忠,王建华,等.中文停用词表的自动选取[J].北京理工大学学报,2005,25(4):337-340.
GU Y J, FAN X Z, WANG J H, et al. Automatic selection of Chinese stoplist[J]. Transactions of Beijing Institute of Technology, 2005, 25(4): 337-340.

[29] 贾会强,刘晓丽,于洪志.基于词性特征提取的藏文文本分类方法研究[C]//CCF NCSC 2011——第二届中国计算机学会服务计算学术会议论文集.[不详:不详],2011.
JIA H Q,LIU X L,YU H Z. Research of feature extraction methods based on part of speech in Tibetan documents classification[C]//CCF NCSC 2011－The second academic conference on computer association of China computer society.[S. l. :s. n.],2011.

[30] 江涛,于洪志.一种面向藏文聚类的文本建模方法[J].西北民族大学学报:自然科学版,2016,37(3):24-28.
JIANG T,YU H Z. A method of text modeling for Tibetan clustering[J]. Journal of Northwest University for Nationalities:Natural Science,2016,37(3):24-28.

[31] 徐涛,于洪志,加羊吉.基于改进卡方统计量的藏文文本表示方法[J].计算机工程,2014,40(6):185-189.
XU T,YU H Z,JIA Y J. Tibetan document representation method based on improved chi-squared statistic [J]. Computer Engineering,2014,40(6):185-189.

[32] 李彬,刘挺,秦兵,等.基于语义依存的汉语句子相似度计算[J].计算机应用研究,2003(12):15-17.
LI B,LIU T,QIN B,et al. Chinese sentence similarity computing based on semantic dependency relationship analysis[J]. Application Research of Computers,2003(12):15-17.

[33] 金博,史彦军,滕弘飞.基于语义理解的文本相似度算法[J].大连理工大学学报,2005,45(2):291-297.
JIN B,SHI Y J,TENG H F. Similarity algorithm of text based on semantic understanding[J]. Journal of Dalian University of Technology,2005,45(2):291-297.

[34] 吴夙慧,成颖,郑彦宁,等.文本聚类中文本表示和相似度计算研究综述[J].情报科学,2012,30(4):622-627.
WU S H,CHENG Y,ZHENG Y N,et al. A survey on text representation and similarity calculation in text clustering[J]. Information Science,2012,30(4):622-627.

[35] 邬明强.基于分段融合的藏文文本相似度计算方法研究[D].兰州:西北民族大学,2016.
WU M Q. Research on the calculation method of similarity based on the fusion of Tibetan text segment [D]. Lanzhou: Northwest University for Nationalities,2016.

[36] JAIN A K,DUBES R C. Algorithms for clustering data[M]. Upper Saddle River, NJ: Prentice Hall, 1988.

[37] SNEATH P H,SOKA R R. Numerical taxonomy [M]. London:Freeman,1973.

[38] KOHONEN T. Self-organized formation of topologically correct feature maps[J]. Biological Cybernetics, 1982,43(1):59-69.

[39] 江涛.基于藏文 Web 舆情分析的热点发现算法研究[D].兰州:西北民族大学,2010.
JIANG T. Study on hot topic detection based on the analysis of Tibetan public opinion [D]. Lanzhou: Northwest University for Nationalities,2010.

[40] 邓竞伟,邓凯英,李永生,等.基于藏文网络的舆情传播模型[J].计算机系统应用,2013,22(3),209-211.
DENG J W,DENG K Y,LI Y S,et al. Opinion spreading models on Tibetan networks[J]. Computer Systems & Applications,2013,22(3),209-211.

[41] 曹晖,孟祥和.基于藏文新闻文本话题检测的聚类算法研究[J].华中师范大学学报:自然科学版,2014,48(1):37-41.
CAO H,MEND X H. The research of clustering algorithm for topic detection based on Tibetan news texts [J]. Journal of Huazhong Normal University:Natural Science,2014,48(1):37-41.

[42] 康健,乔少杰,格桑多吉,等.基于群体智能的半结构化藏文文本聚类算法[J].模式识别与人工智能,2014,27(7):663-671.
KANG J,QIAO S J,GESANG D J,et al. A semi-structure Tibetan text clustering algorithm based on swarm intelligence[J]. Pattern Recognition and Artificial Intelligence,2014,27(7):663-671.

[43] 袁斌,江涛,于洪志.基于语义空间的藏文微博情感分析方法[J].计算机应用研究,2016,33(3):682-685.
YUAN B,JIANG T,YU H Z. Emotional classification method of Tibetan micro-blog based on semantic space[J]. Application Research of Computers,2016,33(3):682-685.

[44] 康健.基于 Multi-agent 和群体智能的藏文网络舆情管理研究[D].成都:西南交通大学,2015.
KANG J. Research on multi-agent and swarm intelligence Tibetan network public opinion management [D]. Chengdu:Southwest Jiaotong University,2015.

(责任编辑:陆 雁)