ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 3,May 2018,pp. 530－537
DOI:10. 16337/j. 1004-9037. 2018. 03. 017

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于图聚类的汉越双语新闻话题发现

王禹森　余正涛　高盛祥　周　超　洪旭东

(昆明理工大学信息工程与自动化学院,昆明,650500)

摘　要:跨语言新闻话题发现是将互联网上报道相同事件的不同语言新闻进行自动归类,由于不同语言文本很难表示在同一特征空间下,对其共同话题的挖掘就比较困难。然而类似的新闻事件在不同语言文本表达上具有相同的新闻要素,这些要素之间关联能够体现出新闻事件的关联性,因此,针对汉越新闻话题发现问题,提出基于文档图聚类的汉越双语新闻话题发现方法。首先提取汉越新闻文本新闻要素,借助文本中要素相似度计算汉越文本相关度,构建汉越双语文本图模型,获得新闻文本相似度矩阵;然后,借助图模型中文本间的传播特点,采用随机游走算法对相似度矩阵进行调整,最后利用信息传递算法进行聚类。实验结果表明提出的方法取得了很好的效果。

关键词:汉越双语;事件要素;话题发现;图聚类

中图分类号:TP391　　**文献标志码**:A

Chinese-Vietnamese Bilingual News Topic Detection Methods Based on Graph Clustering

Wang Yusen, Yu Zhengtao, Gao Shengxiang, Zhou Chao, Hong Xudong

(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China)

Abstract: The purpose of cross-language topic discovery is to classify news texts written in different languages by their topics automatically. However, due to the difference in different languages, it's hard to describe these texts on the same feature space, so mining the same topic is not an easy work. When a particular news event is reported, the news elements are the same no matter which language describe it. So news elements can reflect the relevance among different news texts. Therefore, the paper proposed Chinese-Vietnamese bilingual news topic detection methods based on graph clustering. Firstly, Chinese-Vietnamese bilingual news elements are extracted and the similarity of different news texts is calculated by using the news elements' similarity to set up a Chinese-Vietnamese bilingual news graph model. Secondly, through the propagation characteristics of the Chinese-Vietnamese bilingual news graph model, the similarity matrix is adjusted by using the random walk algorithm. Finally, affinity propagation algorithm is used to cluster topic. The experimental result shows that the proposed method is effective.

Key words: Chinese-Vietnamese; events element; topic detection; graph clustering

---

基金项目:国家自然科学基金(61472168,61175068,61672271)资助项目;云南省自然科学基金重点(2013FA130)资助项目;云南省科技创新人才基金(2014HE001)资助项目。
收稿日期:2016-09-24;**修订日期**:2016-10-19

引 言

随着经济全球化,不同国家之间的联系日益紧密,共同关注的事件、话题也越来越多。跨语言新闻话题发现就是针对互联网上不同国家发布的不同语言新闻进行分析处理,获得的不同类别话题的新闻,帮助人们及时掌握当前国际和地区发生的热点事件,以及对同一事件不同国家的不同看法。目前话题发现研究基本都是在单语环境下做的,并取得了很好的成果。单语话题发现方法一般分为以下 3 类:(1)向量空间模型。它通过抽取文本词频、词性和语法结构等特征,将文本表征成多维特征向量,利用向量之间的关系实现文本相似度的计算,从而进行共同话题的挖掘[1-2]。(2)概率模型。它利用新闻文本中词语与话题分布的统计规律,构建话题统计概率模型,分析挖掘新闻文本话题[3-4]。(3)图模型。它提取新闻文档特征及特征之间的关系,如特征词之间关系,建立特征概率图模型,通过图的求解思路分析文本的话题[5]。相比单语环境下的话题发现,双语环境下的话题发现研究较少,其关键问题在于如何跨越语言障碍,目前主要基于以下 3 类方法,(1)基于机器翻译。它将不同语言的新闻文本转化到同一目标语言,在单语环境下进行共同话题的挖掘与分析,机器翻译的准确性对这种方法有着很大的影响。(2)借助双语词典。该方法对文本中的实体,关键词进行翻译,来构造跨语言特征词空间,进行话题发现[6],这种方法忽略了没有互译关系却存在联系的词语,比如“阮富仲”和“越南国家领导人”在词典中是没有互译关系的,却表达相同的意义。(3)基于大规模双语语料[7-9]。如利用概率主题模型,对平行语料或者可比语料进行跨语言主题挖掘,将获得一系列的跨语言主题作为特征空间,这种方法难点在于大规模对齐语料收集整理。

在对汉越跨语言新闻话题发现方面,由于汉越双语新闻采用不同语言进行表征,而不同语言在不同的词空间下,导致不同语言文本很难表示在同一个特征空间上,这给汉越双语新闻话题发现工作带来了挑战。同时,新闻报道中的时间、地点、人物、事情经过和事情发生的原因具有真实性,这些关键内容必须具体、明确,对于同一事件的报道,汉语与越南语新闻在这些新闻要素上一致,这为进行汉越双语的话题发现研究提供了有效的途径。利用新闻要素表征文档,计算要素间相关性,可以计算出文本间的相似度,构成汉越双语新闻图模型,图中节点的紧密程度表示文本相似度高低,这样便将汉越双语新闻话题发现看成图模型的聚类问题来分析。

1 面向汉越新闻报道的图模型

汉越双语新闻图 $G=\{V,E,W\}$,表示汉越双语新闻集合 N 与图的一个映射。V 是汉越双语新闻集合中的新闻文本在图中对应的文本集合,v_i 为汉语文本,v_j 为越南语文本,表示为 $V=\{v_i,v_j\mid 1\leqslant i\leqslant n,1\leqslant j\leqslant m\}$。$E$ 是汉越双语新闻集合中的新闻文本在图中的边,(v_{i_1},v_{i_2}) 为汉语文档间的边,(v_{j_1},v_{j_2}) 为越南语文档间的边,(v_i,v_j) 为汉越双语文档间的边,表示为 $E=\{(v_i,v_j),(v_{i_1},v_{i_2}),(v_{j_1},v_{j_2})\mid i_1\neq i_2,j_1\neq j_2\}$。$W$ 表示图中边的权重,表示为 $W=\{w(i,j),w(i_1,i_2),w(j_1,j_2)\}$,权重由新闻要素相似度决定。新闻的事件要素一般包括时间、地点、人物、经过和原因,可以表示为 When,Where,Who,What 和 Why,其中,时间可以用时间实体来表示,地点可以用地点实体来表示,人物可以用人物实体来表示,经过一般用要素中的动词来表示。规定两个新闻文本间具有连接线必须满足以下条件之一:(1)两篇新闻在时间、地点和人物等要素上有相同的要素对出现;(2)两篇新闻在 What 这个要素上相似度达到 0.5 以上。

计算单语文档间边权重时,考虑新闻文本中的词对于所在新闻文本的重要程度,采用 TF-IDF 方法计算。抽取新闻文本要素,以向量的形式表征一篇新闻文本,每个向量由其特征项及权重表示,构成文

本向量空间。相同语言文档节点间的相似度采用两篇文档空间向量的夹角余弦来计算。

设任意两个节点 $\forall x_i, x_k \in V$，TF-IDF 公式为

$$W_{t,x} = \mathrm{TF}_{t,x} \times \mathrm{IDF}_{t,x} \tag{1}$$

$$\mathrm{TF}_{t,x} = \frac{N}{M} \tag{2}$$

$$\mathrm{IDF}_{t,x} = \log(X/X_N) \tag{3}$$

式中：$W_{t,x}$ 为新闻要素 t 在新闻文本 x 中的权重；$\mathrm{TF}_{t,x}$ 指词语 t 在文档 x 中出现的频率，如式(2)表示一篇有 M 个词的文档含有 N 个新闻要素 t。$\mathrm{IDF}_{t,x}$ 反映新闻要素 t 在所有新闻文档中的常见程度，在一定程度上体现了该新闻要素的区分能力，其中 X 表示所有新闻文档的数目，X_N 表示所有新闻文档中包含新闻要素 t 的文档数。

利用文档向量间的夹角余弦分别计算相同语言文档节点间的权重为

$$\mathrm{Sim}(x_i, x_k) = \cos\theta = \frac{\sum_{t=1}^{n} W_{t,x_1} \times W_{t,x_2}}{\sqrt{\left(\sum_{t=1}^{n} W_{t,x_1}^2\right)\left(\sum_{t=1}^{n} W_{t,x_2}^2\right)}} \tag{4}$$

式中：W_{t,x_1}，W_{t,x_2} 分别为文档 x_1，x_2 中的第 t 个特征项的权重，从而得到相同语言文档间的权重，即 $w(i_1, i_2)$，$w(j_1, j_2)$。

计算汉越双语文档间边权重时，抽取新闻文档要素，将汉越双语文档表征成向量，计算汉语文档向量中新闻要素与越南语文档向量中每个新闻要素的相似度和，从而得到汉越双语文档间的相似度为

$$w(i,j) = \frac{1}{m \times n} \sum_{a=1}^{m} \sum_{b=1}^{n} w(a,b) \tag{5}$$

式中：$w(i,j)$ 为汉越双语文档间边的相似度，即图中边 v_i 与 v_j 之间的权重；$w(a,b)$ 为汉越文档中两个要素的相似度。$w(a,b)$ 具体相似度的计算方法是借助维基百科中具有中越互译关系的概念，不同语言词语会出现在不同的概念页面上，且词语与其他概念之间存在一定的共现关系，首先提取维基百科中汉语越南语具有对应关系的概念集合，构建双语概念特征空间，然后根据词语在相应概念描述文本中出现的词频特征，以及词语与概念在其他概念文本中的共现特征构建词语的概念向量值，最后通过夹角余弦对两个向量进行词语相似度计算[10]。最后可以得到汉越新闻图模型，基本框架如图 1 所示。

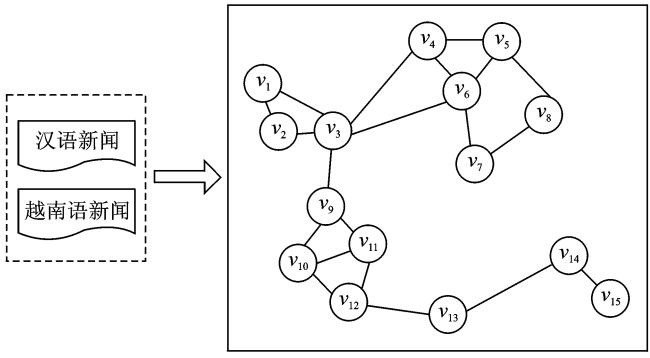

图 1 汉越双语新闻图

Fig. 1 Chinese-Vietnamese bilingual news graph

2 随机游走相似度矩阵计算

汉越双语新闻图的转移概率矩阵可以表示为 $\boldsymbol{p}_z = (p_{ij})$，它是一个 $n \times n$ 矩阵，其中的每一个元素 p_{ij} 表示任意一个顶点 v_i 到其邻居节点 v_j 的转移概率为

$$p_{ij}=\frac{w_{ij}}{\sum_{k} w_{ik}} \tag{6}$$

式中:w_{ij}为新闻文本节点 v_i 与 v_j 的相似度,即图中边的权重;k 为图中以文本节点 v_i 为端点的边的个数;$\sum w_{ij}$ 为所有以文本节点 v_i 为端点的边的权重之和;图中不具有连线关系的文本节点的转移概率为 0。

定义 1 给定图 $G=\{V,E,W\}$,顶点 v_i 到 v_j 的路径是集合 E 中从顶点 $v_0=v_i$ 出发到顶点 $v_{k+1}=v_j$ 结束的一系列边的集$(v_0,v_1),(v_1,v_2),\cdots,(v_{k-1},v_k),(v_k,v_j)$,可表示为 $\mathrm{Path}(v_i,v_j)$,如果有这样一条可以相通的路径就说明顶点 v_i 和 v_j 是相连的。路径上边的权重之和可以表示路径的长度,而顶点 v_i 和 v_j 之间的距离指长度中最大的一个。

采用随机游走模型来度量汉越新闻图中顶点之间的相似度,若两个顶点之间相通的路径越多,则说明两顶点之间的转移概率就越大,顶点之间的相似度就越大。

定义 2 图 G 的 $n\times n$ 的转移概率矩阵为 $\boldsymbol{p}_z$,给定 l 为随机游走的路径长度,则顶点 v_i 到 v_j 的邻近随机游走相似度为

$$s(v_i,v_j)=\sum_{\mathrm{length}(\mathrm{Path}(v_i,v_j))\leqslant l} p(\mathrm{Path}(v_i,v_j)) \tag{7}$$

式中:$\mathrm{Path}(v_i,v_j)$是顶点 v_i 到 v_j 的路径,其长度为 $\mathrm{length}(\mathrm{Path}(v_i,v_j))$,$p(\mathrm{Path}(v_i,v_j))$为转移概率。随机游走相似度矩阵可表示为

$$\boldsymbol{S}_{\boldsymbol{p}_z}=\sum_{\min(\mathrm{length}(\mathrm{Path}(v_i,v_j)))}^{l}\boldsymbol{p}_z \tag{8}$$

式中:$\boldsymbol{p}_z$ 为转移概率矩阵,l 为随机游走的路径长度。过程如算法 1 所描述。

算法 1 汉越新闻图随机游走相似度矩阵算法。

输入:汉越新闻图。

输出:随机游走相似度矩阵。

(1)计算汉越新闻图的转移概率矩阵。

(2)计算汉越新闻图的邻近随机游走相似度。

(3)利用转移概率矩阵计算汉越新闻图的随机游走相似度矩阵。

(4)输出汉越新闻图的随机游走相似度矩阵。

3 基于信息传递的汉越新闻图聚类

利用汉越新闻文本相似度矩阵进行图聚类与一般聚类问题相比存在以下特点:(1)通过随机游走得到的汉越新闻文本相似度矩阵,描述的是节点之间的相关程度,而不是节点之间的欧式距离,故无法直接使用 K-Means 算法进行求解。(2)本文得到的汉越新闻文本相似度矩阵不对称,故无法使用谱聚类的方法进行求解。因此,本文采用信息传递算法[11]对汉越新闻文本图模型进行聚类,整个聚类的过程,利用汉越双语新闻图的随机游走相似度矩阵,通过迭代更新吸引度和归属度两种信息完成聚类,相应的更新公式为

$$r(v_i,v_j)=s(v_i,v_j)-\max_{v_j\neq v_j}(a(v_i,v_j')+s(v_i,v_j')) \tag{9}$$

$$a(v_i,v_j)=\min\{0,r(v_j,v_j)+\sum_{v_i'\in\{v_i,v_j\}}\max\{0,r(v_i',v_j)\}\} \tag{10}$$

$$a(v_j,v_j)=\sum_{v_i'\neq v_j}\max(0,r(v_i',v_j)) \tag{11}$$

式中：$r(v_i,v_j)$为从顶点 v_i 发送到聚类中心 v_j 的数值消息，反映顶点 v_j 是否适合作为顶点 v_i 的聚类中心；$s(v_i,v_j)$为顶点 v_i 和 v_j 的相似度；$a(v_i,v_j)$为从候选聚类中心 v_j 发送到顶点 v_i 的数值信息，反映顶点 v_i 是否选择 v_j 作为其聚类中心。在信息传递聚类算法的每次迭代更新顶点 v_i 的过程中，吸引度 R_i 和归属度 A_i 要与上次迭代所得 R_{i-1} 与 A_{i-1} 的值进行加权更新，更新公式为

$$R_i=(1-l_{\mathrm{am}})\times R_i+l_{\mathrm{am}}\times R_{i-1} \tag{12}$$

$$A_i=(1-l_{\mathrm{am}})\times A_i+l_{\mathrm{am}}\times A_{i-1} \tag{13}$$

其中，$l_{\mathrm{am}}\in[0,1]$通过改变 l_{am} 的值可以改进算法的收敛性。

算法满足以下条件之一，即停止迭代：

(1)达到预先设定的迭代次数；(2)顶点信息改变量低于设定的阈值；(3)所选的聚类中心在连续若干次的迭代中保持稳定的值。

根据 $r(v_i,v_j)+a(v_i,v_j)$的值判断顶点 v_j 能否作为聚类中心，最后，将其他顶点分配到与其最邻近的聚类中心。具体的聚类过程如算法 2 所示。

算法 2 信息传递聚类算法。

输入：邻接随机游走相似度矩阵。

输出：k 个簇 $\boldsymbol{C}_1,\boldsymbol{C}_2,\boldsymbol{C}_3,\cdots,\boldsymbol{C}_k$。

(1)初始化 $r(v_i,v_j)=0, a(v_i,v_j)=0$

(2)迭代执行以下更新过程：

(3)$r(v_i,v_j)=s(v_i,v_j)-\max\limits_{v_j'\neq v_j}(a(v_i,v_j')+s(v_i,v_j'))$

(4) $a(v_i,v_j)=\min\{0,r(v_j,v_j)+\sum\limits_{v_i'\in\{v_i,v_j\}}\max\{0,r(v_i',v_j)\}\}$

(5) $a(v_j,v_j)=\sum\limits_{v_i'\neq v_j}\max(0,r(v_i',v_j))$

(6)$R_i=(1-l_{\mathrm{am}})\times R_i+l_{\mathrm{am}}\times R_{i-1}$

(7)$A_i=(1-l_{\mathrm{am}})\times A_i+l_{\mathrm{am}}\times A_{i-1}$

(8)对于任一个顶点 v_i，如果 $r(v_i,v_i)+a(v_i,v_i)$达到迭代次数或者不再变化，则 v_i 是一个聚类中心。

(9)基于 $s(v_i,v_i)$，将其他顶点 v_j 分配到与它最邻近的聚类中心。

(10)输出 k 个簇 $\boldsymbol{C}_1,\boldsymbol{C}_2,\boldsymbol{C}_3,\cdots,\boldsymbol{C}_k$。

基于以上随机游走算法和信息传递算法，最后得到 k 个簇，认为每一个簇都是一个话题，完成了汉越双语新闻的话题发现任务。

4 实验结果与分析

4.1 实验数据

选取了 180 个中文门户网站和 20 个论坛以及 125 个不同专题的越南语网站。中文新闻包括新华社、人民日报、知名论坛、主流门户网站和越南网站(以每日快讯、越讯社和越共机关等核心平台为主)。在从爬取到的数据中选择训练集时，选取了 5 个话题：两会、朝核、中国反腐、南海争端和叙利亚反恐。因为在这 5 个话题上，越南的各大媒体和中国的各大媒体关注最多。另外，一个话题出现以后，会在一段时间内出现很多关于该话题的新闻报道，所以在进行新闻文档选取的时候只选取近 10 天的新闻数据进行实验。

新闻最核心的是 What，Who，Where，When 和 Why 5 个要素，而这 5 个要素的词性主要对应了动

词、名词、时态词、形容词和数词,因此在对汉语和越南语新闻文本进行分词和词性标注后,将这些词性的词语抽取出来作为新闻要素。对于中文词性标注和命名实体识别,采用 ICTCLAS3.0 工具。利用越南语分词工具[12]对越南语新闻文本进行分词、词性标注等处理,根据处理结果,人工辅助抽取要素。各类新闻数如表 1 所示。

表 1 实验数据集

Tab. 1 Experimental data set

话题类别	汉语新闻/篇	越南语新闻/篇
两会	50	50
朝核	50	50
中国反腐	50	50
南海争端	50	50
叙利亚反恐	50	50

4.2 评价方法

在话题发现研究中,经常会用错检率 F 和漏检率 M 作为评价指标。评价指标的具体含义见表 2。

表 2 评价指标的具体含义

Tab. 2 Meaning of evaluation

类别	相关文档数	不相关文档数
被监测到的文档数目	A	B
没有被监测到的文档数目	C	D

在表 2 中用大写字母 A,B,C,D 来表示某一个话题的检测结果,用 $F=B/(B+D)$ 来表示话题检测的错检率,用 $M=C/(A+C)$来表示话题检测的漏检率。此外,为了综合漏检率和错检率,定义耗费函数(Cost function)为

$$C_{\mathrm{Det}}=C_{\mathrm{miss}}P_{\mathrm{miss}}P(\mathrm{rel})+C_{\mathrm{fa}}P_{\mathrm{fa}}(1-P(\mathrm{rel})) \tag{14}$$

式中:C_{miss}和 C_{fa}为话题检测中漏检和误检的代价,$P(\mathrm{rel})$表示某个新闻报道属于某一类的先验概率,P_{miss}和 P_{fa}为话题检测的漏检概率和误检概率。在 TDT 的标准中,令 $C_{\mathrm{miss}}=1.0$,$C_{\mathrm{fa}}=0.1$,$P(\mathrm{rel})=0.02$。由此可以看出耗费函数越小,话题发现效果越好。

4.3 实验结果与分析

本文通过 3 个不同方法进行汉越新闻话题发现,方法 1 通过基于多策略优化的分治多层聚类算法的话题发现方法,首先得出单语文档下的聚类结果,然后通过机器翻译的方法将其合并;方法 2 采用双语文档主题生成模型(Latent Dirichlet allocation, LDA),利用 Wikipedia 中的 10 000 对汉越双语文档构建可比语料,训练双语主题模型,对不同语言文本进行表示[13]。认为一对文档主题上具有相同的概率分布。本文共设置了 100 个主题,利用获得的双语主题模型来对要聚类的 500 篇新闻进行推断,最后,采用 K-Means 进行聚类。方法 3 采用本文提出的话题发现方法。实验结果如表 3 所示。

表 3 新闻话题发现对比实验结果

Tab. 3 Result of comparative experiments

话题	基于多策略优化的分治多层聚类算法				双语 LDA				本文方法			
	A	*B*	*C*	*D*	*A*	*B*	*C*	*D*	*A*	*B*	*C*	*D*
两会	75	1	25	399	80	0	20	400	90	2	10	398
朝核	80	3	20	397	87	2	13	398	86	0	14	400
中国反腐	65	5	35	395	74	4	26	396	87	2	13	398
南海争端	80	1	20	399	88	4	12	396	80	4	20	396
叙利亚反恐	54	12	46	388	65	6	35	394	81	3	29	397

根据实验结果数据，分别计算每种方法的误检率、漏检率和消耗函数，对比结构如表 4 所示。通过表 4 的实验结果对比可以发现，在给定训练集的 5 个话题下，本文方法通过计算新闻要素的相似度，求得图模型，并通过随机游走算法求得相似度矩阵，在话题发现方面，不论是漏检率、误检率还是最后的耗费函数都要优于基于多策略优化的分治多层聚类算法和双语 LDA 方法。由此可见，本文提出的基于图模型的汉越双语新闻话题发现图聚类模型是可行的。

表 4 误检率、漏检率和消耗函数

Tab. 4 Mistake rate, miss rate and consumption function

方法	误检率	漏检率	耗费函数
基于多策略优化的分治多层聚类算法	0.012 0	0.292	0.007 016
双语 LDA	0.008 0	0.212	0.005 024
本文方法	0.005 5	0.172	0.003 979

5 结束语

在双语环境下进行话题发现是一项比较困难的任务，本文提出基于图聚类的汉越双语话题发现方法，利用双语新闻要素作为跨语言的桥梁，根据不同语言新闻要素之间的关联计算不同语言新闻文本之间的相似度。通过本文的研究可以发现利用新闻要素可以更好地表征一篇新闻文档；此外，利用新闻要素作为跨语言桥梁，建立汉越双语新闻图模型，通过图中节点的紧密程度表示新闻相似程度，采用基于信息传递的汉越双语新闻图聚类算法能够有效地提高话题发现的效果。下一步工作将融合新闻主题句的关联，提高汉越双语话题发现的效果。

参考文献：

[1] Zhang Dan, Li Shengdong. Topic detection based on K-means[C]// International Conference on Electronics, Communications and Control. Ningbo, China:[s. n.], 2011:2983-2985.

[2] 赵华，赵铁军，于浩，等. 基于查询向量的英语话题跟踪研究[J]. 计算机研究与发展，2007，44(8)：1412-1417.
Zhao Hua, Zhao Tiejun, Yu Hao, et al. English topic tracking research based on query vector[J]. Journal of Computer Research and Development, 2007, 44(8):1412-1417.

[3] Guo Xin, Xiang Yang, Chen Qian, et al. LDA-based online topic detection using tensor factorization[J]. Journal of Information Science, 2013, 39(4):459-469.

[4] Phan X H, Nguyen L M, Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large-scale

data collections[C]// International Conference on World Wide Web. Beijing, China:[s. n.], 2008:91-100.
[5] Zhao Wenqing, Hou Xiaoke. News topic recognition of Chinese microblog based on word co-occurrence graph[J]. CAAI Transactions on Intelligent Systems, 2012, 5: 444-449.
[6] Mathieu B, Fluhr C. Multilingual document clusters discovery[C]// Computer-assisted Information Retrieval. Avignon, France: [s. n.], 2004:116-125.
[7] Boyd-Graber J, Blei D M. Multilingual topic models for unaligned text[C]// Conference on Uncertainty in Artificial Intelligence. [S. l.]: AUAI Press, 2012:75-82.
[8] Mimno D, Wallach H M, Naradowsky J, et al. Polylingual topic models[C]// Conference on Empirical Methods in Natural Language Processing. Singapore: ACL, 2009:880-889.
[9] Vúlic I, Smet W D, Tang J, et al. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications[J]. Information Processing & Management, 2015, 51(1):111-147.
[10] 杨启悦, 余正涛, 洪旭东,等. 基于维基百科的汉越词语相似度计算[J]. 南京理工大学学报(自然科学版),2016,40(4):461-466.
Yang Qiyue, Yu Zhengtao, Hong Xudong, et al. Chinese-Vietnamese word similarity computation based on Wikipedia[J]. Journal of Nanjing University of Science and Technology, 2016,40(4):461-466.
[11] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814):972-976.
[12] Cam Tu N, Xuan Hieu P, Thu Trang N. JVnTextPro: A Java-based Vietnamese text processing tool[EB/OL]. http://jvntextpro. sourceforge. net/, 2010-1-1.
[13] Ni Xiaochuan, Sun Jiantao, Hu Jian, et al. Mining multilingual topics from Wikipedia[C]// International Conference on World Wide Web. Madrid, Spain:[s. n.], 2009:1155-1156.

作者简介:

王禹森(1992-),男,硕士研究生,研究方向:信息检索、自然语言处理,E-mail:wys11307@163.com。

余正涛(1970-),男,教授,博士生导师,研究方向:自然语言处理、信息检索和机器翻译。

高盛祥(1977-),女,博士研究生,研究方向:信息检索、机器翻译。

周超(1989-),男,硕士研究生,研究方向:自然语言处理。

洪旭东(1989-),男,硕士研究生,研究方向:信息检索、自然语言处理。

(编辑:陈珺)