

文章编号:1001-9081(2015)04-0996-05

doi:10.11772/j.issn.1001-9081.2015.04.0996

互信息改进方法在术语抽取中的应用

杜丽萍¹, 李晓戈^{1*}, 周元哲¹, 邵春昌²

(1. 西安邮电大学 计算机学院, 西安 710121; 2. 中央民族大学 理学院, 北京 100081)

(*通信作者电子邮箱 lixg@xupt.edu.cn)

摘要:为了确定改进互信息(PMI^k)方法的参数 k 取何值时能够克服互信息(PMI)方法过高估计两个低频且总是一起出现的字串间结合强度的缺点,解决术语抽取系统采用经过分词的语料库时由于分词错误导致的某些术语无法抽取的问题,以及改善术语抽取系统的可移植性,提出了一种结合 PMI^k 和两个基本过滤规则从未经过分词的语料库中进行术语抽取的算法。首先,利用 PMI^k 方法计算两个字之间的结合强度,确定2元待扩展种子;其次,利用 PMI^k 方法计算2元待扩展种子分别和其左边、右边的字的结合强度,确定2元是否能扩展为3元,如此迭代扩展出多元的候选术语;最后,利用两个基本过滤规则过滤候选术语中的垃圾串,得到最终结果。理论分析表明,当 $k \geq 3 (k \in \mathbb{N}_+)$ 时, PMI^k 方法能克服PMI方法的缺点。在1GB的新浪财经博客语料库和300MB百度贴吧语料库上的实验验证了理论分析的正确性,且 PMI^k 方法获得了比PMI方法更高的精度,算法有良好的可移植性。

关键词:术语抽取;专业术语;知识获取;互信息

中图分类号: TP391.1 **文献标志码:**A

Application of improved point-wise mutual information in term extraction

DU Liping¹, LI Xiaoge^{1*}, ZHOU Yuanzhe¹, SHAO Chunchang²

(1. College of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an Shaanxi 710121, China;

2. College of Science, Minzu University of China, Beijing 100081, China)

Abstract: The traditional Point-wise Mutual Information (PMI) method has shortcoming of overvaluing the co-occurrence of two low-frequency words. To get the proper value of k of improved PMI named PMI^k to overcome the shortcoming of PMI, and solve the problem that the term extraction cannot be obtained from a segmented corpus with segmentation errors, as well as maintaining the portability of term extraction system, combining with the PMI^k method and two fundamental rules, a new method was put forward to identify terms from an unsegmented corpus. Firstly, 2-gram extended seed was determined by computing the bonding strength of two adjoining words by PMI^k method. Secondly, whether the 2-gram extended seed could be extended to 3-gram was determined by respectively computing the bonding strength between the seed and the word in front of it and the word located behind it, and then getting multi-gram term candidates iteratively. Finally, the garbage of term candidates were filtered using the two fundamental rules to obtain terms. The theoretical analysis shows that PMI^k can overcome the shortcoming of PMI when $k \geq 3 (k \in \mathbb{N}_+)$. The experiments on 1 GB SINA finance Blog corpus and 300 MB Baidu Tieba corpus verify the theoretical analysis, and PMI^k outperforms PMI with good portability.

Key words: term extraction; technical term; knowledge acquisition; Point-wise Mutual Information (PMI)

0 引言

术语抽取在中文信息处理领域中是一项重要的基础性研究课题。随着科技、经济、文化的快速发展,各个学科领域中的术语也发生了很大变化,为了及时了解学科的发展动态,术语抽取的需求应运而生。

术语抽取方法总体上有两种:基于规则的方法和基于统计的方法^[1-2]。目前,主流方法是将两者结合起来使用,即基于统计与规则相结合的方法。

统计部分通常是通过计算字串间的结合强度来判定两个字串是否可以组成一个术语^[3]。一般情况下,互信息(Point-

wise Mutual Information, PMI)方法能够很好地反映字串之间的结合强度,但PMI方法的缺点是过高地估计了低频且总是相邻出现的字串间的结合强度^[3-4],例如,“倜”和“傥”、“蝙”和“蝠”等在语料库中低频且总是相邻出现,这些字串的PMI值非常高,包含这些低频字串的垃圾串的PMI值也非常高,例如“是倜”和“傥”、“的蝙”和“蝠”等。针对此问题,有些研究者将PMI方法与其他方法相结合进行术语抽取。文献[4]采用PMI方法和log-likelihood方法结合进行术语抽取,第一步采用PMI方法和log-likelihood方法结合确定2元待扩展种子,过程中预先设定了PMI的阈值、词频阈值、log-likelihood阈值等过滤2元待扩展种子;第二步采用log-likelihood方法

收稿日期:2014-10-30;修回日期:2015-01-13。

基金项目:国家自然科学基金资助项目(61373116);西安邮电大学研究生创新基金资助项目(ZL2013-31)。

作者简介:杜丽萍(1987-),女,陕西宝鸡人,硕士研究生,主要研究方向:自然语言处理、文本数据挖掘;李晓戈(1962-),男,浙江杭州人,教授,主要研究方向:自然语言处理、数据挖掘、机器学习;周元哲(1974-),男,陕西西安人,讲师,硕士,主要研究方向:自然语言处理、机器学习;邵春昌(1987-),男,山东淄博人,硕士研究生,主要研究方向:自然语言处理、数据挖掘、机器学习。

把2元待扩展种子扩展成多元术语。文献[5]提出了使用 PMI^2 和 PMI^3 的改进 PMI 方法来抽取术语。文献[6]利用 PMI 方法衡量字串间的结合强度,再结合NC-value方法融入词语上下文信息来提高三字以上长术语的抽取精度。文献[7]采用互信息方法F-MI抽取结构简单的质词,在此基础上,进一步抽取结构复杂的合词。文献[8]提出了一种语言文法信息与互信息相结合的抽取方法,可以选择合适的 PMI 值来满足具体的应用要求。文献[9]也提出了向 PMI 方法中引进 k 个联合概率因子来克服 PMI 方法的缺点,称为 PMI^k 方法(但未确定 k 值)。本文通过抽象语料库中低频且总是相邻出现字串的数学特征,从理论上证明当且仅当向 PMI 方法中引进3个及以上的联合概率因子时 PMI^k 方法能够克服 PMI 方法的缺点。

规则部分通常是根据汉语构词法知识建立规则库,过滤掉不符合构词法的候选术语^[2]。文献[3]首先采用 PMI 方法基于大规模领域语料计算子串间的相关度,取出相关度大于预先设定阈值的子串作为候选术语;其次,通过对大量的候选术语分析,人工建立了普通词语搭配前缀、后缀库,去掉候选术语中的普通词汇,例如“十分高兴”“非常重要”等词;最后,通过总结词性构成规则库,过滤掉不符合词性规则的候选术语,得到最终的术语抽取结果。规则方法最大的缺点是需要人工去挖掘特定领域的构词法规则,这会严重影响系统对不同领域语料的适用性。

许多研究者^[3,6-8,10-13]都是从经过某种分词器处理的语料库中进行术语抽取,这样做的不足之处是:有些术语由于分词错误导致无法识别。例如下面句子:

- 1) 开始集中仓位位于兴业银行。
 - 2) 上升持续时间最长的行情为主升浪行情。
 - 3) 除了被牢牢套死的散户外,……
- 经过ICTCLAS分词结果是
- 1) 开始/v 集中/v 仓/ng 位于/v 兴业/nz 银行/n 。/w
 - 2) 上升/v 持续/vd 时间/n 最/d 长/a 的/u 行情/n 为主/v 升/v 浪/n 行情/n 。/w
 - 3) 除/v 了/u 被/p 牢牢/d 套/v 死/a 的/u 散/a 户外/s ,/w …

无法从这些句子中识别出“仓位”“主升浪”“散户”等术语。

本文提出了一种基于统计与规则相结合的术语抽取方法:利用 PMI^k 方法,以逐字扩展的方式从未经过分词器处理的语料库中抽取候选术语,再结合两个基本规则过滤候选术语,最后剔除候选术语中的核心词汇(包含在核心词典中的词,如ICTCLAS中的核心词典)得到最终的术语抽取结果。

1 PMI方法和改进方法 PMI^k

1.1 PMI方法

定义1 PMI方法定义如下:

$$PMI(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

其中: $p(x)$ 、 $p(y)$ 分别表示字串 x 、字串 y 的概率, $p(x, y)$ 表

示字串 x 和字串 y 的联合概率, $PMI(x, y)$ 表示字串 x 和字串 y 的相关度(也称为PMI值)。

定义2 给定字串 str_1 、 str_2 , $p(str_1)$ 、 $p(str_2)$ 分别代表 str_1 、 str_2 在语料库中出现的概率。对于充分小的正数 δ ,如果字串 str_1 和 str_2 在语料库中的概率满足 $p(str_1) \leq \delta$ 、 $p(str_2) \leq \delta$ 且它们总是相邻出现,则称字串 str_1 和 str_2 为低频共现字串;否则,称为非低频共现字串。

特别地,低频共现字串 str_1 和 str_2 及它们的组合字串 str_1str_2 的概率满足:

$$p(str_1) = p(str_2) = p(str_1str_2) \leq \delta$$

非低频共现字串 str_1 和 str_2 及它们的组合字串 str_1str_2 的概率满足:

$$\delta < p(str_1str_2) \leq p(str_1)$$

$$\delta < p(str_1str_2) \leq p(str_2)$$

根据定义2,给出方法 f 对低频共现字串敏感的定义如下:

定义3 设低频共现字串集合:

$$Low = \{(str_1, str_2) | str_1 \text{ 和 } str_2 \text{ 为低频共现字串}\}$$

非低频共现字串集合:

$$Comm = \{(str_1', str_2') | str_1' \text{ 和 } str_2' \text{ 为非低频共现字串}\}$$

衡量两个字串相关度的方法为 f :

如果 $\forall (str_1', str_2') \in Comm$,总存在 $(str_1, str_2) \in Low$,使得通过方法 f 计算得到的字串 str_1 和 str_2 的相关度总是大于字串 str_1' 和 str_2' 的相关度,则称方法 f 对低频共现字串敏感。

定理1 PMI方法对低频共现字串敏感,即对于非低频共现字串集合 $Comm$ 中的任意两个字串 c 和 d ,总存在低频共现字串集合 Low 中两个字串 a 和 b ,使得字串 a 和 b 的PMI值大于字串 c 和 d 的PMI值。

证明 对于 $\forall (a, b) \in Low$,有

$$p(a) = p(b) = p(a, b); p(a) \leq \delta$$

则

$$PMI(a, b) = \log \frac{p(a, b)}{p(a)p(b)} = \log \frac{1}{p(a)} \geq -\log \delta$$

对于 $\forall (c, d) \in Comm$,有

$$\delta < p(c, d) \leq p(c); \delta < p(c, d) \leq p(d)$$

则

$$PMI(c, d) = \log \frac{p(c, d)}{p(c)p(d)} \leq \log \frac{1}{p(c) \text{ or } p(d)} < -\log \delta$$

所以,对于 $\forall (c, d) \in Comm$, $\exists (a, b) \in Low$,

s.t. $PMI(a, b) > PMI(c, d)$

故PMI方法对低频共现字串敏感。证毕。

1.2 改进方法 PMI^k

PMI^k 方法是通过在PMI方法中引进一个或者多个字串 x 与 y 的联合概率因子 $p(x, y)$,来克服PMI方法对低频共现字串敏感的缺点。 PMI^k 方法的定义如下:

定义4 PMI^k 算法^[9]定义如下:

$$PMI^k(x, y) = \log \frac{p^k(x, y)}{p(x)p(y)}; k \in \mathbb{N}_+$$

其中: $p(x)$ 、 $p(y)$ 分别表示字串 x 、 y 的概率, $p(x, y)$ 表示字

串 x 和 y 的联合概率, $PMI^k(x, y)$ 表示字串 x 和 y 的相关度(也称 PMI^k 值)。

特殊地,当 $k = 1$ 时, PMI^k 方法即 PMI 方法。

定理 2 当且仅当正整数 $k \geq 3$ 时, PMI^k 方法能解决对低频共现字串敏感的缺点,即对于低频共现字串集合 Low 中任意两个字串 a 和 b ,存在非低频共现字串集合 $Comm$ 中的两个字串 c 和 d ,使得字串 c 和 d 的 PMI^k 值大于字串 a 和 b 的 PMI^k 值,其中, PMI^k 方法中的参数 k 必须大于等于 3。

证明 分为 $k = 1, k = 2$ 和 $k \geq 3$ 三种情况证明:

1) 当 $k = 1$ 时, PMI^k 方法即 PMI 方法, 证明如定理 1。

2) 当 $k = 2$ 时, 对于 $\forall (a, b) \in Low$, 有

$$p(a) = p(b) = p(a, b)$$

则

$$PMI^2(a, b) = \log \frac{p^2(a, b)}{p(a)p(b)} = \log 1$$

对于 $\forall (c, d) \in Comm$, 有

$$p(c, d) \leq p(c)$$

$$p(c, d) \leq p(d)$$

则

$$PMI^2(c, d) = \log \frac{p^2(c, d)}{p(c)p(d)} \leq \log 1$$

所以 $k = 2$ 时, 对于 $\forall (c, d) \in Comm$, 存在

$$\exists (a, b) \in Low$$

s. t. $PMI^2(a, b) > PMI^2(c, d)$

3) 当 $k \geq 3$ 时, 对于 $\forall (a, b) \in Low$, 有

$$p(a) = p(b) = p(a, b); p(a) \leq \delta$$

确定 2 元待扩展种子

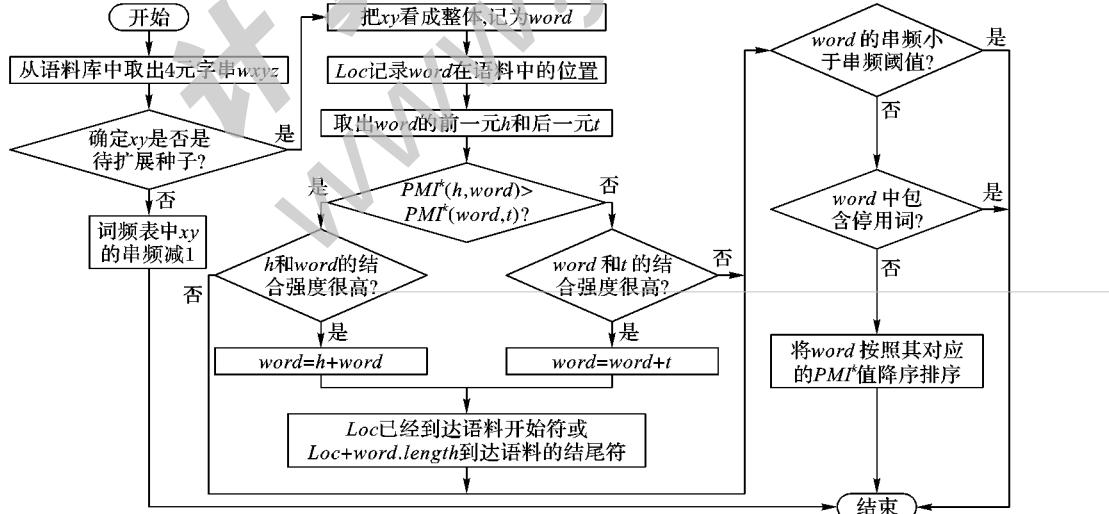


图 1 术语抽取算法流程

2.1 确定 2 元待扩展种子

对于语料库中每一个 4 元字串 w, x, y, z (w, x, y, z 均为单字),首先计算两个值 $mean_1$ 和 $mean_2$:

$$mean_1 = \frac{1}{2}(PMI^k(w, x) + PMI^k(x, y))$$

$$mean_2 = \frac{1}{2}(PMI^k(x, y) + PMI^k(y, z))$$

如果 $PMI^k(x, y) > PMI^k(w, x) + mean_1$ 并且 $PMI^k(x, y) > PMI^k(y, z) + mean_2$, 则认为字串 xy 是一个词或者词的

一部分的可能性大于 wx 和 yz , 将字串 xy 确定为待扩展种子;否则,认为字串 x, y 各自成词或是词的边界,将 xy 在词频表(存放语料库中所有 1 元及以上字串的串频)中的串频减 1。

2.2 2 元待扩展种子扩展成 $2 - n$ 元

取出 xy 的前一个字 h 和后一个字 t , 分为两种情况:

1) 如果 $PMI^k(h, xy) > PMI^k(xy, t)$, 则认为 h 和 xy 组成术语的可能性要大于 xy 和 t 的可能性,令

$$mean = \frac{1}{2}(PMI^k(h, xy) + PMI^k(x, y))$$

如果 $PMI^k(h, xy) + mean > PMI^k(x, y)$, 则认为 h 和 xy 可组成术语 hxy , 继续迭代地逐字扩展; 否则, 输出字串作为候选术语。

2) 如果 $PMI^k(h, xy) \leq PMI^k(xy, t)$, 则认为 xy 和 t 组成术语的可能性要大于 h 和 xy 的可能性, 令

$$mean = \frac{1}{2}(PMI^k(x, y) + PMI^k(xy, t))$$

如果 $PMI^k(xy, t) + mean > PMI^k(x, y)$, 则认为 xy 和 t 可组成术语 xyt , 继续迭代地逐字扩展; 否则, 输出字串作为候选术语。

2.3 过滤规则

利用可存在性过滤规则实施过滤的方法: 如果候选术语集合中的术语 $word$ 在词频表中的串频小于串频阈值 $Threshold$, 则将 $word$ 从候选术语集合中过滤掉; 否则, 保留 $word$, 进行下一步过滤。

利用停用词表过滤规则实施过滤的方法: 如果候选术语集合中的术语 $word$ 中包含字串停用词表中的词, 则将 $word$ 从候选术语集合中过滤掉; 否则, 保留 $word$ 。

2.4 术语判定

术语判定分为两个步骤:

1) 从抽取的术语结果中剔除掉包含在核心词典中的词语;

2) 人工判定结果。

3 实验与分析

3.1 实验数据

1) 1 GB 新浪财经博客语料, 用于抽取财经领域的术语。

2) 300 MB(约 1000 万字)的 2013 年百度贴吧语料, 用于

表 1 1 GB 新浪财经博客语料实验结果

k	候选术语量	核心词汇量	术语量	准确率/%			
				前 500 条	前 1000 条	前 1500 条	前 2000 条
1	269 460	25 812	243 648	31.2	25.50	22.53	21.20
2	133 908	26 261	107 647	31.8	28.20	27.13	27.35
3	61 248	25 537	35 711	63.4	56.10	51.53	48.95
4	41 061	23 241	17 820	75.0	67.60	62.33	59.50
5	30 138	20 052	10 086	81.2	74.70	69.53	67.05
6	23 609	17 129	6 480	85.6	78.90	76.00	74.05
7	19 355	14 822	4 533	91.2	84.20	83.00	81.85
8	16 616	13 167	3 449	94.0	90.30	88.67	88.95
9	14 693	11 885	2 808	94.6	91.99	90.47	91.30
10	13 442	11 025	2 417	95.2	92.40	91.60	92.60

表 2 300 MB 百度贴吧语料实验结果

k	候选术语量	核心词汇量	术语量	前 1000 条中正确术语量	准确率/%
1	26 631	6 912	19 719	686	68.60
2	14 099	6 672	7 427	738	73.80
3	8 930	6 265	2 665	822	82.20
4	6 897	5 342	1 555	833	83.30
5	5 672	4 647	1 025	848	84.80
6	4 847	4 100	747	692	92.64
7	4 344	3 739	605	580	95.87
8	3 935	3 416	519	499	96.15
9	3 680	3 216	464	449	96.77
10	3 467	3 046	421	410	97.39

抽取网络专用语。

3) 停用词典: 包含 702 个停用词(选自哈尔滨工业大学停用词表), 用于过滤候选术语集合中的垃圾串。

4) ICTCLAS 核心词典: 共收集了 79 836 个词语, 是目前比较规范的词典之一, 用于判定专业术语。

3.2 实验结果

对于大规模语料库, 难以获得语料中所有出现的术语, 因此, 无法计算系统抽取结果的召回率。本文只选取准确率作为评测系统抽取结果的指标。准确率的计算公式如下:

$$\text{准确率} = \frac{\text{正确术语条数}}{\text{术语条数}} \times 100\%。$$

针对 1 GB 新浪财经博客语料库实验, 在给定串频阈值 $Threshold = 6$ 情况下, PMI^k 方法的参数 k 分别取 1~10 的 10 个正整数值进行术语抽取实验。由于 PMI^k 方法的参数 k 值不同, 抽取的候选术语条数也不同, 为了有效对比不同 k 值下的抽取结果, 根据术语对应的 PMI^k 值对术语降序排序, 分别取前 500, 1000, 1500, 2000 条结果作评测。表 1 描述了实验结果, 其中, 候选术语指经过术语抽取系统抽取的结果条数; 核心词汇指候选术语中包含在 ICTCLAS 核心词典中的词汇条数; 术语指从候选术语中剔除掉包含在 ICTCLAS 核心词典中的词汇后剩余的词汇条数。

针对 300 MB 的百度贴吧语料实验, 实验条件设置同财经语料实验相同。表 2 描述了实验结果, 其中候选术语、核心词汇、术语的意义跟表 1 相同, 正确术语分为两种情况统计: 当术语条数大于等于 1 000 条时只选取术语结果的前 1 000 条(按 PMI^k 值降序排序)统计其中的正确术语; 当术语条数小于 1 000 条时, 统计全部术语结果中的正确术语。

表 3~4 分别列举了新浪财经博客语料库和百度贴吧语料库的抽取结果的前 10 条。

表 3 新浪财经博客语料前 10 条实验结果

序号	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
1	萎愈肝	片仔癀	升浪							
2	惆惋	道扬镳	蓝筹	主升						
3	锱铢必	简意赅	散户	博客	点击	点击	点击	点击	点击	点击
4	嗟诺酮	火如荼	博客	点击	短线	短线	短线	短线	短线	短线
5	名遐迩	饕餮	主升	短线	博客	博客	博客	博客	选股	
6	耄耋	偃旗息	鼎砾	散户	权重	权重	权重	选股	博客	
7	马齐喑	何鸿燊	私募	蓝筹	散户	选股	选股	选股	权重	权重
8	虎作伥	草甘膦	狼啸	权重	蓝筹	散户	操盘	操盘	操盘	操盘
9	如弊屣	沆瀣	短线	涨停	涨停	居前	涨停	涨停	涨停	涨停
10	醐灌顶	贵研铂	点击	私募	居前	涨停	散户	均线	均线	均线

表 4 百度贴吧语料前 10 条实验结果

序号	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
1	晦涩难	南海保镖	真朱	真朱	大神	大神	大神	大神	大神	大神
2	非贪婪	赫卡特	寒云	大神	楼主	楼主	楼主	楼主	楼主	楼主
3	周子琦	青年范儿	大神	楼主	真朱	真朱	真朱	真朱	真朱	真朱
4	嚙嚙	刘易雯	蛋疼	窗体	窗体	控件	控件	控件	控件	控件
5	金针菇	徽太尉	窗体	百度	控件	窗体	窗体	窗体	窗体	窗体
6	啰嗦	满智勇	良化	良化	百度	百度	百度	百度	百度	百度
7	耦合度	寒云似雾	百度	控件	良化	良化	良化	良化	良化	良化
8	肝肠	童鞋	楼主	寒云						
9	蜀黍	叮叮	控件	蛋疼	蛋疼	蛋疼	蛋疼	贴吧	贴吧	
10	吧头衔	云似雾	菜鸟	菜鸟	菜鸟	贴吧	贴吧	蛋疼	蛋疼	

3.3 结果分析

由表 1 ~ 2 可看出:当 $k \geq 3$ 时,选术语及术语条数相比 $k = 1$ (即 PMI 方法)时大幅减少,准确率显著提高。

表 1 中,前 500 条、前 1000 条、前 1500 条、前 2000 条的准确率,均随着 k 值增大而提高。当 $k \geq 3$ 时,前 500 条结果的准确率比 $k = 1$ 时至少提高了 32.2%;前 1000 条结果的准确率比 $k = 1$ 时至少提高了 30.6%;前 1500 条结果的准确率比 $k = 1$ 时至少提高了 29%;前 2000 条结果的准确率比 $k = 1$ 时至少提高了 27.75%。故当 PMI^k 方法的参数 $k \geq 3$ 时 PMI^k 方法显著改善了 PMI 方法在术语抽取中的效果。

从表 3 ~ 4 可看出,当 PMI^k 方法的参数 $k \geq 3$ 时抽取的术语和 $k = 1, k = 2$ 时差异较大。在 $k = 1$ 和 $k = 2$ 的结果中,排名在前的字串中均包含低频的字或词,例如垃圾串“锱铢必”“道扬镳”“晦涩难”“徽太尉”中分别包含“锱铢”“镳”“晦涩”“徽”等低频字串,且这些字串的搭配词语固定,该现象反映出 PMI 方法、 PMI^2 方法对低频共现字串敏感的缺点;在 $k \geq 3$ 的结果中,均没有出现低频共现字串,说明 $k \geq 3$ 时 PMI^k 方法克服了 PMI 方法的缺点,与理论证明相一致。

新浪财经博客语料和百度贴吧语料的抽取术语结果中大多为 2 元术语,并且随着 PMI^k 方法参数 k 值不同,结果中 2 元、3 元、4 元及 5 元术语所占的比重也不同,共同特点是 2 元术语均占到 70% 以上。

4 结语

本文解决了 3 个问题:1)从理论上证明了 PMI^k 方法的参数 $k \geq 3$ 时能够克服 PMI 方法的缺点,并在新浪财经博客语料

库和百度贴吧语料库上验证了该结论。2)通过两个基本规则过滤候选术语,使系统具有良好可移植性。系统不仅在比较规范的新浪财经博客语料上取得了较好的效果,在百度贴吧这种用词不规范、噪声大、领域广的网络语料上也取得了较好的效果。3)提出了采用未经过分词器处理的原始语料库以逐字扩展的方式进行术语抽取,解决了由于分词错误而导致有些术语无法识别的问题。

本文只验证了 PMI^k 方法的参数 k 取值大于等于 3 时能够克服 PMI 方法的缺点,但没有确定究竟当参数 k 取何值时系统的结果能达到最优(根据具体应用综合考虑结果的精度、数目等因素),下一步工作是研究 PMI^k 方法的参数 k 取值与语料库规模、语料特征等因素的关系,提出一种自适应地确定参数 k 值的方法;另外,改善系统抽取长术语的能力也是下一步工作的一方面。

参考文献:

- [1] PAULO J L, CORREIA M, MAMEDE N J. et al. Using morphological, syntactical, and statistical information for automatic term acquisition [C]// Proceedings of the Third International Conference on Advances in Natural Language Processing, LNCS 2389. Berlin: Springer-Verlag, 2002: 219 - 227.
- [2] ZHU Q, LENG F. Existing problems and developing trends of automatic term recognition[J]. Library and Information Service, 2012, 56(18): 104 - 109. (祝清松, 冷伏海. 自动术语识别存在的问题及发展趋势综述[J]. 图书情报工作, 2012, 56(18): 104 - 109.)
- [3] ZHANG F, XU Y, HOU Y, et al. Chinese term extraction system based on mutual information[J]. Application Research of Computers, 2005, 22(5): 72 - 74. (张峰, 许云, 侯艳, 等. 基于互信息的中文术语抽取系统[J]. 计算机应用研究, 2005, 22(5): 72 - 74.)

(下转第 1005 页)

- 2013, 33(11): 3080 – 3083. (史庆伟, 李艳妮, 郭朋亮. 科技文献中作者研究兴趣动态发现[J]. 计算机应用, 2013, 33(11): 3080 – 3083.)
- [2] LIU T, LIU B, XU Z, et al. Automatic domain-specific term extraction and its application in text classification[J]. Acta Electronica Sinica, 2007, 35(2): 328 – 332. (刘桃, 刘秉权, 徐志明, 等. 领域术语自动抽取及其在文本分类中的应用[J]. 电子学报, 2007, 35(2): 3283 – 332.)
- [3] HAN H, ZHU D, WANG X. Technical term extraction method for patent document[J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(12): 1280 – 1285. (韩红旗, 朱东华, 汪雪锋. 专利技术术语的抽取方法[J]. 情报学报, 2011, 30(12): 1280 – 1285.)
- [4] NICAM K, LAFFERTY J, McCALLUM A. Using maximum entropy for text classification[EB/OL]. [2010-10-10]. <http://www.ka-malnigam.com/papers/maxent-ijcaiws99.pdf>.
- [5] NEWMAN S, ASUNCION A, SMYTH P. Distributed inference for latent Dirichlet allocation[J]. Neural Information Processing Systems, 2007, 32(7): 55 – 65.
- [6] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 43(12): 993 – 1022.
- [7] BLEI D M. Probabilistic topic models[J]. Communications of the ACM, 2012, 55(4): 77 – 84.
- [8] LI W, SUN L, ZHANG D. Text classification based on labeled-LDA model[J]. Chinese Journal of Computers, 2008, 31(4): 620 – 627. (李文波, 孙乐, 张大鲲. 基于 Labeled-LDA 模型的文本分类新算法[J]. 计算机学报, 2008, 31(4): 620 – 627.)
- [9] BLEI D M, McAULIFFE J M. Supervised topic models[EB/OL]. [2010-10-10]. <https://www.cs.princeton.edu/~blei/papers/BleiMcAuliffe2007.pdf>.
- [10] RAMAGE D, HALL D, NALLAPATI R, et al. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora [C]// EMNLP 2009: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. New York: ACM Press, 2009, 1: 248 – 256.
- [11] ROSEN Z, CHBMUDUGUNTA C, GRIFFITHS T. The author-topic model for authors and documents[C]// UAI 2004: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. New York: ACM Press, 2004: 487 – 194.
- [12] ROSEN Z, CHBMUDUGUNTA C, GRIFFITHS T. Learning author-topic models from text corpora[J]. ACM Transactions on Information Systems, 2010, 28(1): Article No. 4.
- [13] WANG J. Academic social network based on statistical topic model and applications[D]. Wuhan: Central China Normal University, 2013. (王建文. 基于统计主题模型的学术网络建模与应用[D]. 武汉: 华中师范大学, 2013.)
- [14] SHAN B, LI F. A survey of topic evolution based on LDA[J]. Journal of Chinese Information Processing, 2010, 24(6): 43 – 49. (单斌, 李芳. 基于 LDA 话题演化研究方法综述[J]. 中文信息学报, 2010, 24(6): 43 – 49.)
- [15] HEINRICH G. Parameter estimation for text analysis[EB/OL]. [2010-10-10]. <http://www.arbylon.net/publications/text-est.pdf>.
- [16] GUO X, XIANG Y, CHEN Q, et al. LDA-based online topic detection using tensor factorization[J]. Information Science, 2013, 39(4): 459 – 469.
- [17] HU P, LIU W, JIANG W, et al. Latent topic model for audio retrieval[J]. Pattern Recognition, 2014, 47(3): 1138 – 1143.

(上接第 1000 页)

- [4] PANTEL P, LIN D. A statistical corpora-based term extractor[C]// Proceedings of the 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, LNCS 2056. Berlin: Springer-Verlag, 2001: 34 – 46.
- [5] PAZIENZA M T, PENNNACCHIOTTI M, ZANZOTTO F M. Terminology extraction: an analysis of linguistic and statistical approaches [C]// Proceedings of the NEMIS 2004 Final Conference on Knowledge Mining, SFSC 185. Berlin: Springer-Verlag, 2005: 255 – 279.
- [6] LIANG Y, ZHANG W, ZHOU D. A hybrid strategy for high precision long term extraction[J]. Journal of Chinese Information Processing, 2009, 23(6): 26 – 30. (梁颖红, 张文静, 周德福. 基于混合策略的高精度长术语自动抽取[J]. 中文信息学报, 2009, 23(6): 26 – 30)
- [7] HE T, ZHANG Y. Automatic Chinese term extraction based on decomposition of prime string[J]. Computer Engineering, 2006, 32(23): 188 – 190. (何婷婷, 张勇. 基于质子串分解的中文术语自动抽取[J]. 计算机工程, 2006, 32(23): 188 – 190.)
- [8] SUN J, JIA M, LIU Z. On a text-oriented concept extraction technique[J]. Computer Application and Software, 2009, 26(9): 28 – 30. (孙继鹏, 贾民, 刘增宝. 一种面向文本的概念抽取方法研究[J]. 计算机应用与软件, 2009, 26(9): 28 – 30.)
- [9] BOUMA G. Normalized (pointwise) mutual information in collocation extraction[EB/OL]. [2013-10-10]. <https://svn.spraakdata.gu.se/repos/gerlof/pub/www/Docs/npmi-pfd.pdf>.
- [10] HU A, ZHANG J, LIU J. Chinese term extraction based on improved C-value method[J]. New Technology of Library and Information Service, 2013(2): 24 – 29. (胡阿沛, 张静, 刘俊丽. 基于改进 C-value 方法的中文术语抽取[J]. 现代图书情报技术, 2013(2): 24 – 29.)
- [11] ZHOU L, SHI S, FENG C, et al. A Chinese term extraction system based on multi-strategies integration[J]. Journal of China Society for Scientific and Technical Information, 2010, 29(3): 460 – 467. (周浪, 史树敏, 冯冲, 等. 基于多策略融合的中文术语抽取方法[J]. 情报学报, 2010, 29(3): 460 – 467.)
- [12] ZHOU L, ZHANG L, FENG C, et al. Terminology extraction based on statistical word frequency distribution variety[J]. Computer Science, 2009, 36(5): 177 – 180. (周浪, 张亮, 冯冲, 等. 基于词频分布变化统计的术语抽取方法[J]. 计算机科学, 2009, 36(5): 177 – 180.)
- [13] YAN X, LIU Y, FANG Q, et al. Domain-specific terms extraction based on Web resource and user behavior[J]. Journal of Software, 2013, 24(9): 2089 – 2100. (闫兴龙, 刘奕群, 方奇, 等. 基于网络资源与用户行为信息的领域术语提取[J]. 软件学报, 2013, 24(9): 2089 – 2100.)