

文章编号:1007-130X(2014)05-0971-06

一种结合关键词与共现词对的向量空间模型^{*}

唐守忠, 齐建东

(北京林业大学信息学院, 北京 100083)

摘要:提出了一种结合关键词特征和共现词对特征的向量空间模型。首先,通过分词和去除停用词提取文本中的候选关键词,利用文本频率筛选关键词特征。然后,基于获得的关键词特征两两构造候选共现词对,定义支持度和置信度筛选共现词对特征。最后,结合关键词特征和共现词对特征构建向量空间模型。文本分类实验结果表明,提出的模型具有更强的文本分类能力。

关键词:向量空间模型;共现词对;语义相关性;文本分类

中图分类号:TP391.3

文献标志码:A

doi:10.3969/j.issn.1007-130X.2014.05.031

Vector space model based on keywords and co-occurrence word pairs

TANG Shou-zhong, QI Jian-dong

(School of Information, Beijing Forestry University, Beijing 100083, China)

Abstract: A new vector space model is proposed, which uses both keyword and co-occurrence term as the representation features of documents. Firstly, the keyword candidates are extracted from documents by segmenting texts and removing stop words, and the keyword features are filtered by document frequency. Secondly, based on the obtained keyword features, the co-occurrence word pairs are constructed, and support degree and confidence degree are defined to filter the features of co-occurrence word pairs. Finally, the keyword features and the features of co-occurrence word pairs are combined to construct the vector space model. The text-classification experiments show that the proposed model has better ability of text classification.

Key words: vector space model; co-occurrence word; semantical relationship; text classification

1 引言

向量空间模型 VSM(Vector Space Model)是最为经典的文本表示模型,被广泛应用于文本分类、聚类、信息检索等领域。该模型由 Salton G 等人^[1]于 1975 年提出,其基本思想是将文本表示为基于关键词特征的向量,利用 TF-IDF 公式计算关键词特征的权重。VSM 简单高效,但不能表示文本的语义特征:一方面,由于基于关键词之间的相

互独立性假设,VSM 无法表示关键词之间的语义相关性;另一方面,由于完全依赖关键词的字符串匹配,VSM 也无法处理文本中经常出现的同义词和多义词现象。

针对上述问题,本文首先在调研目前 VSM 改进方向的基础上,指出了利用统计语言模型改进 VSM 的优势;然后介绍了统计语言模型中的词共现理论,并总结当前利用词共现信息改进 VSM 的研究工作及其不足;最后利用词共现信息构造“共现词对”特征,定义精确的共现词对特征支持度、置

^{*} 收稿日期:2013-02-25;修回日期:2013-04-24

基金项目:十二五科技支撑课题(2011BAH10B04)

通信地址:100083 北京市清华东路 35 号北京林业大学信息学院 1024 信箱

Address: Mailbox 1024, School of Information, Beijing Forestry University, 35 Qinghua Rd East, Beijing 100083, P. R. China

信度和权重计算方法,并将共现词对特征与 VSM 原有的关键词特征结合,提出了一种结合关键词与共现词对的向量空间模型 KACVSM (Vector Space Model based on Keyword And Co-occurrence word)。文本分类实验对比结果表明了 KACVSM 的有效性。

2 相关工作

针对 VSM 缺乏文本语义表示的不足,有的研究人员提出利用关键短语代替关键词作为 VSM 的表示特征。比如文献[2, 3]利用统计语义平滑机制,提取文本中的关键短语表示文本。文献[4]通过大规模的抽取门户网站上专家手工标引的“关键词”作为关键短语表示文本。文献[5~7]通过改进后缀树模型,提取文本中的关键短语表示网页文本。利用关键短语改进 VSM 的困难在于文本中关键短语难以界定^[8]。

也有研究人员提出利用本体改进 VSM。比如文献[9]首先利用互信息测度来计算关键词之间的相关度,然后利用 WordNet 本体计算两个关键词之间的语义距离,最后结合两者计算关键词的语义权重。文献[10]通过自行构建的领域本体调整 VSM 中关键词的 TF-IDF 权重。文献[11]利用 WordNet 本体改进 VSM 的聚类效果。利用本体改进 VSM 的方法过于依赖诸如 WordNet、领域主题词表等外部语义资源。

统计语言模型^[12]针对特定语料库,通过前期大量的学习和统计,挖掘隐藏的真实信息来增强 VSM 的语义表示能力,是 VSM 语义改进研究的重要方向。关键词的词共现信息是统计语言模型挖掘的重要信息之一,利用词共现信息改进 VSM,比利用难以界定的短语更直观可靠,也无需依赖诸如 WordNet、领域主题词表等外部语义资源。

3 词共现

3.1 词共现理论

自然语言文本中普遍存在词共现现象,即某些关键词经常共同出现在一定的文本范围(如句子、段落或篇章)内,词共现现象中隐含着关键词之间的语义相关性信息。文本集合中任意两个关键词的组合都可被看作一组共现词对,关键词 key_1 和 key_2 构成的共现词对可表示为 (key_1, key_2) 。共现

词对 (key_1, key_2) 的共现频率是指文本集合中同时包含关键词 key_1 和 key_2 的文本数量, (key_1, key_2) 的共现频率越高,表明关键词 key_1 和 key_2 的语义相关性越大。共现词对描述了两个关键词之间的语义相关性,是描述文本语义的最小特征单元。理论上讲,包含 p 个关键词的文本集合中包含 $p(p-1)/2$ 个共现词对,当文本集合中的关键词数量成百或上千时,共现词对的统计量巨大。因而利用共现词对表示文本时,通常定义支持度和置信度两个指标来筛选文本语义表达能力强的共现词对特征。

共现词对 (key_1, key_2) 的支持度定义如下:

$$sup(key_1, key_2) = freq(key_1, key_2) / n$$

其中, n 表示文本总数。 $freq(key_1, key_2)$ 表示共现词对 (key_1, key_2) 的共现频率。

共现词对 (key_1, key_2) 的置信度定义如下:

$$con(key_1, key_2) = \alpha \times con(key_1 | key_1, key_2) + \beta \times con(key_2 | key_1, key_2)$$

其中, $con(key_1 | key_1, key_2)$ 和 $con(key_2 | key_1, key_2)$ 分别为关键词 key_1 和 key_2 的条件置信度,分别对应于在关键词 key_1 和 key_2 出现的条件下,共现词对 (key_1, key_2) 出现的概率。 α 和 β 分别为关键词 key_1 和 key_2 的条件置信度的加权参数。关键词 key_1 和 key_2 的条件置信度计算公式如下:

$$con(key_1 | key_1, key_2) = freq(key_1, key_2) / freq(key_1)$$

$$con(key_2 | key_1, key_2) = freq(key_1, key_2) / freq(key_2)$$

共现词对 (key_1, key_2) 的支持度用于评价其对整个文本集合的区分能力,置信度用于评价关键词 key_1 和 key_2 的语义相关性,置信度计算公式中加权参数 α 和 β 的设置十分关键。

3.2 词共现改进 VSM 相关工作

目前,已有一些利用词共现信息改进 VSM 的工作。例如文献[13]提出了基于共现词组合的 VSM,利用共现词对表示文本,利用布尔值计算二阶共现词的权重。文献[14]统计当前关键词与其前后 n 个关键词组成的长度为 $2n+1$ 的词序列中的词共现信息,生成当前关键词的相关词序列,通过关键词的相关词序列共同包含的关键词数量计算关键词之间的语义相关性。文献[15]通过定义关键词之间的互信息筛选相关性高的共现词,用于扩展 VSM。现有利用词共现信息改进 VSM 的研究,在词共现特征的构造、降维、权重计算方法以及与 VSM 原有关键词特征的结合四个方面不够全面。文献[13]利用共现词对特征表示文本,但没给

出有效的特征降维和特征权重计算方法。文献[14,15]仅采用词共现特征表示文本,舍弃了 VSM 原有的关键词特征。本文提出的 KACVSM 利用共现词对特征表示文本,给出了精确有效的特征降维和权重计算方法,将共现词对特征与 VSM 原有的关键词特征有效结合,综合考虑了上述四个方面。

4 KACVSM

给定文本集合 D , 本文将 KACVSM 的构造流程(图 1 所示)分为文本预处理、关键词特征统计、共现词对特征统计和 KACVSM 向量表示四个步骤。

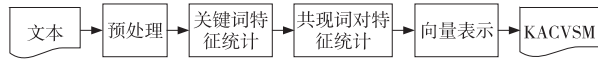


Figure 1 Process of constructing KACVSM

图 1 KACVSM 构造流程图

4.1 文本预处理

针对文本集合 D 中的每个文本,利用 Java 编程语言,调用分词工具进行文本分词,结合停用词表过滤掉停用词,获得候选关键词特征集合。

4.2 关键词特征统计

针对候选关键词特征集合中的每个关键词,首先统计其在所属文本中的词频、在整个文本集合 D 中的文本频率、逆文本频率;然后设定文本频率阈值,以筛选最终有效的关键词特征;最后利用 TF-IDF 公式计算关键词特征的权重。TF-IDF 公式如下:

$$weight(key) = tf(key, d) \times idf(key) \quad (5)$$

$$idf(key) = \log[n/df(key) + 0.01] \quad (6)$$

其中, $tf(key, d)$ 表示词频,即关键词 key 在文本 d 中出现的次数。 $idf(key)$ 表示关键词 key 的逆文本频率。 n 表示文本集合 D 中的文本总数, $df(key)$ 表示文本频率,即文本集合 D 中出现关键词 key 的文本数量, 0.01 为调节参数。

4.3 共现词对特征统计

基于 4.2 节中筛选得到的关键词特征集合,首先两两构造共现词对,获得候选共现词对特征集合;然后针对每个候选共现词对,利用 3.1 节中的方法计算其支持度和置信度以筛选最终有效的共现词对特征;最后计算共现词对特征的权重。关键词的逆文本频率是整个文本集合上的统计量,代表关键词对整个文本集合的区分能力。因而,在计算

候选共现词对 (key_1, key_2) 的置信度时,利用关键词 key_1 和 key_2 的逆文本频率计算加权参数:

$$\alpha = idf(key_1) / [idf(key_1) + idf(key_2)] \quad (7)$$

$$\beta = idf(key_2) / [idf(key_1) + idf(key_2)] \quad (8)$$

TF-IDF 公式是经典的权重计算公式。因而,在计算共现词对特征的权重时,本文延续关键词的 TF-IDF 权重计算方法,提出了如下共现词对特征的 TF-IDF 公式:

$$weight(key_1, key_2) = tf[(key_1, key_2), d] \times idf(key_1, key_2) \quad (9)$$

其中, $tf[(key_1, key_2), d]$ 表示共现词对 (key_1, key_2) 在文本 d 中的词频。关键词的权重是关键词在当前文本中的统计量,因而在计算共现词对 (key_1, key_2) 的词频时,本文采用 $weight(key_1)$ 和 $weight(key_2)$ 进行加权。共现词对 (key_1, key_2) 在文本 d 中的词频计算方法如下:

$$tf[(key_1, key_2), d] = (weight(key_1) \times tf(key_1, d) + weight(key_2) \times tf(key_2, d)) / (weight(key_1) + weight(key_2)) \quad (10)$$

$idf(key_1, key_2)$ 表示共现词对 (key_1, key_2) 的逆文本频率,利用共现词对 (key_1, key_2) 的共现频率进行计算,计算公式如下:

$$idf(key_1, key_2) = \log(n / freq(key_1, key_2) + 0.01) \quad (11)$$

4.4 向量表示

向量表示是指将 4.2 节统计得到的关键词特征和 4.3 节统计得到的共现词对特征线性结合构造文本向量的过程。假设 4.2 节中获得的关键词集合为 $T = \{t_1, t_2, \dots, t_p\}$, 4.3 节中获得的共现词对集合为 $C = \{c_1, c_2, \dots, c_m\}$, 则任意文本 d 的向量表示如下:

$$d = [\omega(t_1), \omega(t_2), \dots, \omega(t_p), \omega(c_1), \omega(c_2), \dots, \omega(c_m)] \quad (12)$$

其中, $\omega(t_i)$ 表示关键词特征 t_i 在文本 d 中的权重,根据 4.2 节中的权重公式计算; $\omega(c_i)$ 表示共现词对特征 c_i 在文本 d 中的权重。

5 实验及结果分析

5.1 实验语料

本文采用复旦大学计算机信息与技术系国际数据库中心自然语言处理小组文本分类语料库进行实验,该语料库共包含 20 类、9 833 篇文本。本实验抽取艺术、计算机、经济、教育、环境、医疗、军

事、政治、体育、交通 10 个类别的数据各 100 篇,共计 1 000 篇文本。每个类别都按照训练集和测试集比例为 7 : 3 切分数据,共得到 700 篇训练文本、300 篇测试文本。

5.2 关键词特征统计

利用 Java 编程语言,调用中科院 ICTCLAS50 分词工具将 1 000 篇文本进行分词,并结合停用词表去除停用词,共获得 33 730 个候选关键词。统计这些候选关键词在其所属文本中的词频、在整个文本集中的文本频率、逆文本频率。表 1 是候选关键词在其文本频率上的分布结果。

Table 1 Distribution of keywords according to document frequency

表 1 关键词特征在文本频率上的分布

文本频率	关键词数量	所占比例/%	累积比例/%
1	17 299	51.3	51.3
2	5 317	15.8	67.1
3	2 516	7.5	74.6
4	1 605	4.8	79.4
5	1 097	3.3	82.7
9	366	1.1	89.1
<16	31 408	—	93.1
<21	31 956	—	94.7
<201	33 679	—	99.8

由表 1 可知,51.3%的候选关键词在文本集合中仅出现了 1 次,0.2%的候选关键词在文本中出现了超过 200 次,这些文本频率过高或过低的关键词特征不具有显著的文本区分能力,不仅会影响文本向量的表示效果,也会增加后续共现词对的统计计算量。本文采用条件 $1 < DF < 201$ 筛选,最终获得 16 380 个关键词特征。

5.3 共现词对特征统计

基于 5.2 节中筛选获得的 16 380 个关键词特征两两构建候选共现词对,获得支持度不小于 0.002 的候选共现词对 2 497 604 个(支持度等于 0.001 的共现词对数量极多且不具有文本表示意义,因而未统计)。按照 4.3 节中的方法计算置信度加权参数,并按照 3.1 节中的方法计算候选共现词对的支持度和置信度。表 2 和表 3 分别是候选共现词对在支持度和置信度上的分布结果。

由表 2 和表 3 可知,55.3%的候选共现词对的支持度等于 0.002,59.8%的候选共现词对的置信度在 0~0.2。支持度或置信度过低的共现词对不具有显著的文本语义表示能力,本文过滤掉支持度小于 0.002 和置信度小于 0.2 的共现词对,共获得

1 002 471 个共现词对特征。

Table 2 Distribution of co-occurrence terms according to support

表 2 共现词对特征在支持度上的分布

文本频率	关键词数量	所占比例/%	累积比例/%
0.002	1 381 827	55.3	55.3
0.003	448 375	18.0	73.3
0.004	228 490	9.2	82.5
0.005	124 646	5.0	87.5
0.006	78 812	3.2	90.7
0.007	52 891	2.1	92.8
0.008	37 575	1.5	94.3
0.009	27 626	1.1	95.4
>0.009	31 408	4.7	—

Table 3 Distribution of co-occurrence terms according to confidence

表 3 共现词对特征在置信度区间上的分布

置信度区间	共现词对数量	所占比例/%	累积比例/%
(0.0,0.1]	594 883	23.8	23.8
(0.1,0.2]	899 547	36.0	59.8
(0.2,0.3]	413 542	16.6	76.4
(0.3,0.4]	203 561	8.2	84.6
(0.4,0.5]	141 307	5.7	90.3
(0.5,0.6]	50 731	2.0	92.3
(0.6,0.7]	95 797	3.8	96.1
(0.7,0.8]	76 678	3.1	99.2
(0.8,0.9]	7 508	0.3	99.5
(0.9,1.0]	14 050	0.6	100.0

5.4 向量表示

基于 5.2 节获得的关键词特征和 5.3 节的共现词对特征,构建 VSM、CTVSM 和 KACVSM 三种文本表示模型。其中,VSM 是传统向量空间模型,仅利用关键词特征表示文本,利用 TF-IDF 计算关键词特征权重;CTVSM 是文献[13]提出的基于共现词对的向量空间模型(CTVSM),仅利用共现词对表示文本,利用布尔值计算关键词特征权重;KACVSM 是本文提出的结合关键词和共现词对的向量空间模型。

5.5 分类实验

基于 5.4 节构建的三种向量空间模型,采用朴素贝叶斯 NB(Naives Bayesian)分类算法,基于 5.1 节中的训练语料构建分类器并分类测试语料,采用常用的正确率(P)、召回率(R)作为评价指标。表 4 为三种模型的朴素贝叶斯分类对比结果。

表 5 表明,KACVSM 的平均分类正确率和召

回率比 VSM 分别高 6.53%和 5.44%,比 CTVSM 分别高 4.67%和 2.82%。艺术、经济、教育、医疗、军事、政治、体育七个类别分类正确率和召回率都有不同程度的提升。这表明,KACVSM 在这几类文本上真正表示了文本的语义特征。另外,KACVSM 在计算机、环境和交通三个类别上的分类正确率或召回率比 VSM 和 CTVSM 低,这是因为这三类文本中的关键词特征的文本区分能力较低,构成的共现词对特征文本语义表示能力较弱,给训练获得的 NB 分类器带来了较强的干扰。

Table 4 Results of NB classification of three kinds of model

表 4 三种模型的朴素贝叶斯分类对比结果 %

	VSM		CTVSM		KACVSM	
	P	R	P	R	P	R
艺术	91.67	73.33	96.07	83.09	100.00	86.67
计算机	84.40	83.33	74.23	90.87	77.78	93.99
经济	62.16	76.67	85.90	77.05	100.00	77.05
教育	96.22	90.30	100.00	85.65	100.00	93.33
环境	60.00	80.00	49.12	80.93	49.12	93.33
医疗	78.79	86.67	79.56	96.67	80.56	96.67
军事	90.78	60.00	86.76	69.99	100.0	65.00
政治	83.87	79.67	84.20	79.81	86.21	83.33
体育	92.86	86.67	96.07	86.12	100.0	88.98
交通	68.57	80.67	65.97	73.33	70.97	73.33
平均值	80.93	79.73	81.79	82.35	86.46	85.17

Table 5 Time consumption of generating NB classifier using VSM

表 5 VSM 的朴素贝叶斯分类器训练速度表

模型	关键词数量/个	训练速度/(文本个数/s)
VSM	16380	2.85

5.6 参数性能分析

本文针对 KACVSM,借鉴文献[13]中的“固定支持度变动置信度”和“固定置信度变动支持度”的方法,考察不同支持度和置信度组合对朴素贝叶斯分类精度和速度的影响。利用 10 个类别文本的分类微平均 F -measure 值作为评价指标。图 1 为固定支持度时,分类微平均 F 值随共现词对置信度阈值的变化。图 2 为固定置信度时,分类微平均 F 值随共现词对支持度阈值的变化。

由图 1 和图 2 可知,当置信度阈值为 0.4 和支持度阈值为 0.005 时,分类效果最好,微平均 F 值分别为 90.74%和 89.97%。当置信度阈值为 0.7 和支持度阈值为 0.009 时,分类效果最差,微平均 F 值分别为 84.97%和 85.12%,但仍然比 VSM 的 80.33%和 CTVSM 的 82.07%要高。另外,无

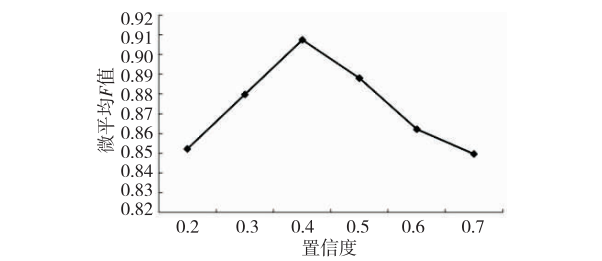


Figure 2 Micro-F1 result with different confidence thresholds based on fixed support

图 2 固定支持度时,微平均 F 值随置信度阈值的变化

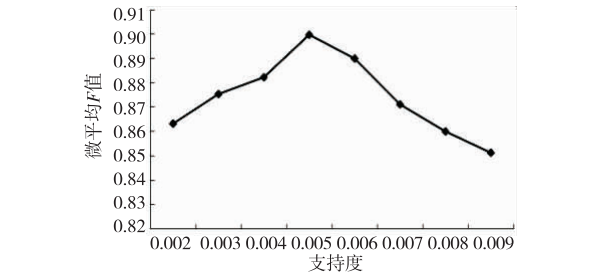


Figure 3 Micro-F1 result with different support thresholds based on fixed confidence

图 3 固定置信度时,微平均 F 值随支持度阈值的变化。论是支持度还是置信度,随着其阈值的不断升高,KACVSM 的分类精度都先升高后降低,因为当阈值较低时,共现词对特征多但语义表示能力普遍较低,当阈值较高时,共现词对的语义表示能力高但数量较少。

表 5 为 VSM 的分类器训练速度。表 6 为固定支持度时,分类器的训练速度随共现词对置信度阈值的变化情况。表 7 为固定置信度时,分类器的训练速度随共现词对置信度阈值的变化情况。

Table 6 Time consumption of generating NB classifier with CTVSM and KACVSM based on fixed support

表 6 支持度固定时,CTVSM 和 KACVSM 的朴素贝叶斯分类器训练速度表

置信度 阈值	共现词 对数量/组	CTVSM 训练速度/ (文本个数/s)	KACVSM 训练速度/ (文本个数/s)
0.2	1 002 471	1.46	1.24
0.3	589 549	1.97	1.57
0.4	384 926	2.07	1.87
0.5	243 362	2.08	1.94
0.6	193 948	2.12	2.05
0.7	98 233	2.14	2.06

由表 5、表 6 和表 7 可知,KACVSM 的分类器训练速度不如 VSM 和 CTVSM,这是由于同时利用关键词特征和共现词对特征表示文本,文本向量的维数增加导致的。但是,分类器的训练速度并没有明显的下滑。相对于 KACVSM 分类精度的提升来说,其速度降低的代价是可以接受的。

Table 7 Time consumption of generating NB classifier with CTVSM and KACVSM based on fixed confidence

表 7 置信度固定时,CTVSM 和 KACVSM 的朴素贝叶斯分类器训练速度表

支持度 阈值	共现词 对数量/组	CTVSM 训练速度/ (文本个数/s)	KACVSM 训练速度/ (文本个数/s)
0.002	1115777	1.40	1.21
0.003	667402	1.87	1.49
0.004	438912	2.01	1.66
0.005	314266	2.04	1.89
0.006	235454	2.08	1.95
0.007	182563	2.10	2.06
0.008	144988	2.11	2.08
0.009	117362	2.11	2.08

6 结束语

本文提出了一种结合关键词特征和共现词对特征的向量空间模型。定义精确有效的共现词对特征的支持度、置信度及权重计算方法,在文本分类实验上证明了所提出的向量空间模型的有效性。但是,本文所提出模型的分类器训练速度有待优化。

参考文献:

[1] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11):613-620.

[2] Zhang Xiao-dan, Zhou Xiao-hua, Hu Xiao-hua. Semantic smoothing for model-based document clustering[C]//Proc of the 6th International Conference on Data Mining, 2006:1193-1198.

[3] Zhou Xiao-hua, Zhang Xiao-dan, Hu Xiao-hua. Semantic smoothing of document models for agglomerative clustering[C]//Proc of the 20th International Joint Conference on Artificial Intelligence, 2007:2922-2927.

[4] Liu Hua. Research of text classification based on key phrases[J]. Journal of Chinese Information Processing, 2007, 21(4):34-41. (in Chinese)

[5] Shi Qing-wei, Zhao Zheng, Chao Ke. Hierarchical clustering of Chinese web pages based on suffix tree[J]. Journal of Liaoning Technical University, 2006, 25(6):890-892. (in Chinese)

[6] Du Hong-bin, Xia Ke-wen, Liu Nan-ping. An improved text clustering algorithm of generalized suffix tree[J]. Information and Control, 2009, 38(3):331-336. (in Chinese)

[7] Wang Jun-ze, Mo Yi-jun, Huang Ben-xiong, et al. Web search results clustering based on a novel suffix tree structure[J]. Autonomic and Trusted Computing, 2008, 5060(23):540-554.

[8] Zhao Jun, Jin Qian-li, Xu Bo. Semantic computation for text retrieval[J]. Chinese Journal of Computers, 2005, 28(12):2068-2078. (in Chinese)

[9] Jing Li-ping, Zhou Li-xin, Ng Michael K, et al. Ontology-based distance measure for text clustering[C]//Proc of the Text Mining Workshop, SIAM International Conference on Data Mining, 2006:1.

[10] Xie Hong-wei, Yan Xiao-lin, Yu Xue-li. Research on web page clustering based on ontology[J]. Computer Science, 2008, 35(9):153-155. (in Chinese)

[11] Zhu Hui-feng, Zuo Wan-li, He Feng-ling. A novel text clustering method based on ontology[J]. Journal of Jilin University(Science Edition), 2010, 48(2):277-283. (in Chinese)

[12] Ponte J M, Bruce C W. A language modeling approach to information retrieval[C]//Proc of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998:275-281.

[13] Chang Peng, Feng Nan. A co-occurrence based vector space model for document indexing[J]. Journal of Chinese Information Processing, 2012, 26(1):51-57. (in Chinese)

[14] Cao Tian, Zhou Li, Zhang Guo-xuan. Text similarity computing based on word co-occurrence[J]. Computer Engineering & Science, 2008, 29(3):52-53. (in Chinese)

[15] Wu Guang-yuan, He Pi-lian, Cao Gui-hong. Vector space model based on word co-occurrence and its application in text classification[J]. Computer Applications, 2003, 23(23):138-140. (in Chinese)

附中文参考文献:

[4] 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.

[5] 史庆伟,赵政,朝柯. 一种基于后缀树的中文网页层次聚类方法[J]. 辽宁工程技术大学学报, 2006, 25(6):890-892.

[6] 杜红斌,夏克文,刘南平. 一种改进的基于广义后缀树的文本聚类算法[J]. 信息与控制, 2009, 38(3):331-336.

[8] 赵军,金千里,徐波. 面向文本检索的语义计算[J]. 计算机学报, 2005, 28(12):2068-2078.

[10] 谢红薇,颜小林,余雪丽. 基于本体的 WEB 页面聚类研究[J]. 计算机科学, 2008, 35(9):153-155.

[11] 朱会峰,左万利,赫枫龄. 一种基于本体的文本聚类方法[J]. 吉林大学学报(自然科学版), 2010, 48(2):277-283.

[13] 常鹏,冯楠. 基于词共现的文档表示模型[J]. 中文信息学报, 2012, 26(1):51-57.

[14] 曹恬,周丽,张国焯. 一种基于词共现的文本相似度计算[J]. 计算机工程与科学, 2008, 29(3):52-53.

[15] 吴光远,何丕廉,曹桂宏. 基于向量空间模型的词共现研究及其在文本分类中的引用[J]. 计算机应用, 2003, 23(23):138-140.

作者简介:



唐守忠(1987-),男,山东东平人,硕士生,研究方向为信息检索。**E-mail:** tang_shouzhong@126.com

TANG Shou-zhong, born in 1987, MS candidate, his research interest includes information retrieval.