DOI:10.13833/j.cnki.is.2012.10.010

基于TFIDF和词语关联度的中文关键词提取方法

张建娥

(榆林学院 图书馆, 陕西 榆林 719000)

摘 要:关键词提取技术是文本分类、文本聚类、信息检索等技术的基础,在自然语言处理领域有着非常广泛的应用。结合TFIDF关键词抽取方法的特点和中文具有的自然语言词语间相互关联的特性,提出一种基于TFIDF和词语关联度的中文关键词提取方法。该方法通过引入词语关联度,有效避免了单纯采用TFIDF算法产生的偏差。实验结果表明,该方法的平均召回率与传统方法相比得到明显提升。

关键词:词语关联度;TFIDF;关键词提取
中图分类号:TP391.1 文献标识码:A 文章编号:1007-7634(2012)10-1542-03

A Chinese Keywords Extraction Approach Based on TFIDF and Word Correlation

ZHANG Jian-e

(Library of Yulin University , Yulin 719000,*China)*

Abstract: Chinese Keywords extraction is one of the basic techniques for text classification, text clustering and Information Retrieval. It has been widely used in natural language processing. This paper proposes a key words extraction approach based on both TFIDF and words correlation according to the keywords extraction method of TFIDF and the correlations features of words on Chinese texts. The method can avoid the deviations of TFIDF algorithm by the introduction of word correlation. The experimental results show that the average recall value can be significantly improved compared with traditional methods.

Keywords:word correlation; TFIDF; keywords extraction

1 引 言

关键词提取技术是自然语言处理的重要基础。随着信息科学技术的快速发展以及互联网的普及,网络文本资源呈几何级数不断增长。面对更新日益频繁和规模庞大的文本数据,如何高效准确地实现关键词提取成为影响信息检索系统性能的关键。

目前,不少学者结合中文自然语言的结构和中文关键词的特点,通过结合关键词的频率、位置关系以及词性等特征,从而提高关键词的提取性能。针对网页的内容特征,对经典的TFIDF方法进行改进,文献[1]提出一种综合多因素的关键词提取方法。该方法综合网页中词语的词长、词性以及位置信息进行综合加权。文献[2]利用词语之间的语义的连贯性,结合词频、位置等特征,提出一种基于词汇链的网页关键词提取方法。文献[3-4]通过自然语言表现出的复杂网络特征,根据语言网络的'小世界'特性,提出基于语言网络的关键词提取方法。该方法借用复杂网络的特性,将词语之间的链接关系和位置信息相融合,从而提高关键词的抽取性能。

本文根据网页关键词的词频特性以及词语之间

收稿日期:2011-11-17

作者简介:张建娥(1963-),女,陕西吴堡人,副研究馆员,主要从事信息检索和信息素养教育研究.

相互关联的特性，提出一种基于TFIDF和词语关联度的中文网页关键词提取方法。该方法将TFIDF作为关键词提取的一个特征，并结合中文语言具有的词语间相互关联的特性，从而有效避免了TFIDF方法的偏差，显著提高了中文关键词提取的性能。

2 关键词提取

2.1 TFIDF

设某文档集合 Ω，N 表示 Ω 中全部文档数目。在某一给定文档 d 中，利用TFIDF方法计算给定词条 t 的权重公式[5]。如下

$$\begin{cases} W_{TF-IDF}=TF\times IDF \\ IDF=\log(N/n) \end{cases} \tag{1}$$

式中TF(Term Frequency)表示词条 t 在文档 d 中出现的频率。IDF(Inverse Document Frequency)表示文档的反转频率，n 表示集合 Ω 中包含词条 t 的文档数目。

由公式(1)可知，当词条 t 在文档 d 中出现的频率较高，并且在整个文档集合 Ω 中出现的频率较低时，则该词条的TFIDF权重较高，反映出该词条在该文档中具有较好的代表性，能够用来表达该文档的实际内容。

2.2 词语关联度

汉语语言的词语之间的关联度在全局上显示出高度的连接性，同时在局部具有高度的聚集性[3]。根据自然语言具有的关联特性，可以作为基本特征进行关键词提取。本文利用复杂网络中节点的度与聚集特征表示词语之间的关联度。

设 $V=\{v_1,v_2,\cdots,v_n\}$ 为节点集合，(v_i,v_j) 表示节点 $v_i\in V$ 与 $v_j\in V$ 之间的边。设 $G(V,E)$ 是以 V 为节点集合，以 $E\subset\{(v_i,v_j):v_i,v_j\in V\}$ 为边集合的图。则节点 v_i 的度 D_i 为

$$D_i=|\{(v_i,v_j):(v_i,v_j)\in E,v_i,v_j\in V\}| \tag{2}$$

节点 v_i 的聚集度 K_i 为

$$K_i=|\{(v_j,v_k):(v_i,v_j)\in E,(v_i,v_k)\in E,v_i,v_j,v_k\in V\}| \tag{3}$$

节点 v_i 的聚集度系数 C_i 为

$$C_i=\frac{K_i}{\binom{D_i}{2}}=\frac{2K_i}{D_i(D_i-1)} \tag{4}$$

根据公式(3)和公式(4)计算得到词语关联度特征值为

$$CF_i=\alpha C_i/\sum_{j=1}^{N}C_j+(1-\alpha)D_i/N \tag{5}$$

由以上分析可知，节点的度表示了该节点与其它节点的关联情况。而聚集度系数则表示该节点局部范围的连接特性。节点的度和聚集度系数可以表示当前节点在局部范围的重要程度。将单词或短语作为复杂网络的节点，则词语关联度特征可通过节点的度和聚集度进行表示。

3 基于TFIDF和词语关联度的关键词提取

3.1 基本原理

根据TFIDF算法的基本原理，单纯地通过单词的频率进行关键词提取的方法虽然简单有效，但是在实践中TFIDF并不能在任何场合都表现优秀[6]。TFIDF算法的固有缺陷表现为数据集偏斜[7]，类间、类内分布偏差[8]等。在词语关联度算法方面，由于复杂网络仅仅依靠词语之间的相互关系作为基本特征，忽略了单词的频率特征，容易造成关键词提取的聚集特征并不明显，从而引起关键词提取的误差。

通过上述分析，我们根据TFIDF和词语关联度的特点对两种方法进行结合，提出一种基于TFIDF和词语关联度的关键词提取方法。该方法同时具备了以上两种方法的优点，能够有效对关键词进行区分，同时避免了两种方法各自的缺陷，从而显著提高关键词提取的准确率。采用该方法计算关键词权重的基本公式为：

$$W=\alpha W_{TFIDF}+(1-\alpha)W_{CF} \tag{6}$$

其中，W_{TFIDF} 表示TFIDF特征的权重值，W_{CF} 表示词语关联度特征的权重值，α 为两种特征的插值系数。系数 α 采用EM算法，对得到的关键词计算其召回率，召回率最大时的 α 值作为该类别的插值系数。其召回率计算公式为：

$$recall=tp/(tp+fn) \tag{7}$$

式中 tp 表示正确检索到的关键词数目，$tp+fn$ 表示所有目标关键词的总数。

3.2 方法流程

采用相同插值系数 α 在实际使用过程中的效果不理想，其原因在于统一的 α 不能完全反映词频与词语关联度之间的关系，而且等同对待所有文档无

法区别不同类型文档的特征。因此需要根据文档的类型，对插值系数 α 进行建模。针对训练集文档，首先对文档进行聚类，训练不同类型文档的 α 值。输入测试文档时，对该文档分类确定其 α 系数，并利用公式 (6) 得到最终的关键词。基于TFIDF和词语关联度的关键词提取方法的基本流程如图1所示。

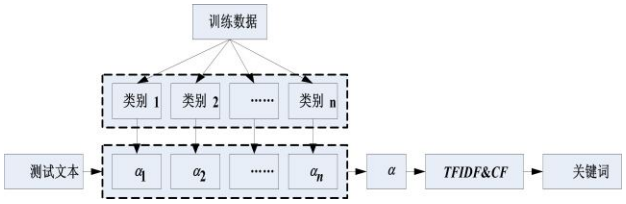

图1 基于TFIDF和词语关联度的关键词提取方法

4 实验结果与分析

4.1 实验设置

本文从互联网中抓取1400个中文网页html文档。通过对网页文本预处理，去除了文档的HTML格式，同时利用ICTCLAS2011分词工具对文档分词。实验采用人工标注的方法对这些文档进行关键词的标注，分别从每个文档中标注得到10个关键词。

根据网页内容的特征，在训练时将1400个网页聚类为7个大类，包含金融、IT、医疗等领域，平均每个类包括200个文档。在不同的领域中，选择前190个文档作为训练集，后10个文档作为测试集。分别对不同领域的文本训练得到 α 的模型，并对测试集数据采用该算法进行关键词提取，并比较单纯采用TFIDF算法和词语关联度算法性能的差异。本文采用召回率标准衡量不同方法之间的性能。

4.2 实验结果

实验利用公式 (6) 对不同领域的训练集分别训练得到相应的插值系数 α。各领域在平均召回率达到最高时对应 α 的值如表1所示。

表1 基于TFIDF和词语关联度不同领域的插值系数 α 表

领域	金融	IT	医疗	体育	旅游	教育	就业
α 值	0.66	0.83	0.75	1.00	0.85	0.85	0.55
平均召回率	59.0%	70.4%	65.8%	70.0%	66.8%	66.5%	66.0%

从表1中可以看出，采用TFIDF和词语关联度方法在训练集上提取关键词，其平均召回率达到65%左右，说明该方法能够有效实现关键词的自动提取。此外，不同领域的插值系数 α 的数值差别较大，表明TFIDF方法和词语关联度方法在不同领域的差异显著，通过调节插值系数能够使关键词召回率达到最优。

为了比较相同领域，系统平均召回率随插值系数 α 的变化过程。我们选取旅游领域绘制关键词平均召回率随着 α 的变化曲线，如图2所示。

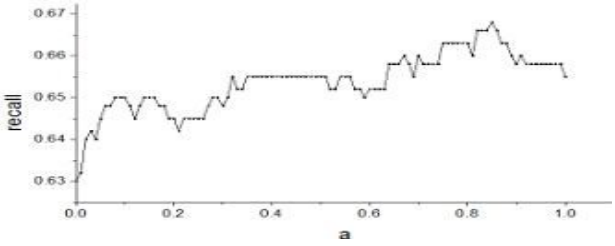

图2 旅游领域平均召回率随插值系数 α 变化曲线

从图2可以看出，随着插值系数 α 从0到1不断变化，关键词的平均召回率也随之变化。根据公式 (6) 可知，当 α 为0时相当于关键词提取的全部权重来自词语关联度特征，当 α 为1时则表示权重全部来自TFIDF特征。根据图2所示曲线，当 α 为0.85时，系统平均召回率达到最大的66.8%，表明基于TFIDF特征和词语关联度特征的关键词提取方法能够有效利用两种方法的优点，从而提高关键词的召回率。

在测试集上，我们对比了单纯采用TFIDF特征和词语关联度特征，以及本文方法的关键词提取性能。以其中一篇测试集文档为例，首先对测试文档进行分类，得到其所在类别的插值系数 α。其次根据插值系数，对测试文档进行关键词提取。实验结果如表2所示。其中TFIDF表示采用TFIDF特征进行关键词提取，CF表示采用词语关联度特征进行关键词提取，TFIDF+CF表示结合两种特征进行关键词提取。

表2 不同关键词提取方法性能比较

提取方法	关键词	召回率
人工选择	跳槽 升值 价值 公司 跳 单位 工作 职位 危机 薪水	—
TFIDF	跳槽 升值 价值 主任 物资 跳 不久 变化 提升 资金	50%
CF	跳槽 升值 价值 主任 都 做 没有 单位 资金 公司	60%
TFIDF+CF	跳槽 升值 价值 主任 物资 跳 公司 单位 资金 确定	70%

由表2可知，采用基于TFIDF和词语关联度关键词提取方法能够有效去除‘都’、‘不久’等无意义单词，从而显著提高查找的召回率。此外，采用该方法的关键词召回率达到最高的70%，表明该方法结合TFIDF和词语关联度的优点，进而提高关键词查找的性能。与人工选择相比，三种 (下转第1555页)

17 IBM. IBM Online Privacy Statement: Protect the Rights and Property of IBM and Others[EB/OL].http://www.ibm.com/privacy/details/us/en/,2012-04-21.
18 R. v. Ward, O.J. No.3116, 2008 ONCJ 355[EB/OL]. http://www.austlii.edu.au/cgi-bin/sinodisp/ai/case s/qld/QCA/2008/222.html?stem=0&synonyms=0&query=title(r.%20and%20.%20ward,2012-8-17.
19 R. V. Verge, O.J.No. 6300. R. v. Friers, O.J. No. 5646, 2008; R. V. McNeice, B.C.J.No.2131; R. V. Spencer, S.J.No. 798, 2009 SKQB 341[EB/OL].http://abbaboutinformation.ca/2010/11/11/case-report-warrantless-search-for-internet-subscribers-data-okayed-by-bcsc/, 2012-08-17.
20 European Commission. Review of the data protection legal framework[EB/OL].ec.europa.eu/justice/policies/privacy/review/index_en.htm,2012-04-17.
21 Randal Picker. Online Advertising, Identity and Privacy 2, Univ. Of Chi. Law & Econ[EB/OL].eWorking Paper No. 475. 2009. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1428065,2012-04-17.
22 Mozy. Decho Corporation Privacy Policy:User Information Decho Collects[EB/OL].http://mozy.com/privacy,2012-04-15.
23 Privacy Statement. Correcting & Updating Your Information[EB/OL].http://web.archive.org/web/20040406201932/salesforce.com/us/statements.jsp?file=privacy&src=web,2012-04-23.
24 Master Subscription Agreement[EB/OL].http://www.salesforce.com/company/msa.jsp#term,2012-05-02.
25 MOZY. Mozy Terms and Conditions: Term and Termination[EB/OL].http://mozy.com/Terms,2012-05-04.
26 MOZY. Mozy Privacy Commitment: Protecting Your Information[EB/OL].http://mozy.com/Privacy/commitment,2012-05-04.
27 IBM. IBM Online Privacy Statement: Information Security and Accuracy[EB/OL]. http://www.ibm.com/privacy/details/us/en/#section_4,2012-04-21.
28 AMAZON. Amazon Web Services Customer Agreement [EB/OL]. http://aws.amazon.com/agreement/,2012-04-21.
29 李 群.个人数据信息隐私权研究在欧美的发展趋势[J]. 档案与建设,2008,(11):45-49.
30 OECD, Briefing Paper for the ICCP Technology Foresight Forum[EB/OL].http://www.oecd.org/dataoecd/39/47/43933771.pdf,2012-04-11.

(责任编辑 徐 波)

(上接第1544页)

方法均得到关键词‘资金”,但该关键词并没用出现在人工选择中。表明人工选择结果具有一定的主观性,容易造成实验结果的偏差。

5 结 语

本文根据自然语言具有的相互关联的特性,提出一种基于TFIDF和词语关联度的中文关键词提取方法。该方法能够结合TFIDF的频率特征以及词语关联度的特征,有效避免单纯采用两种特征时产生的偏差,显著提高关键词查找的召回率。此外,该方法针对不同领域的插值系数进行建模,有效避免了领域之间的差异造成系统性能的误差。实验结果表明,该方法能够融合两种特征的优点,其关键词提取结果的召回率均优于相同条件下单纯采用一种特征的结果。

在目前工作的基础上,我们将进一步研究插值系数α对关键词提取结果的影响,提高不同领域α模型的精度。同时,进一步研究中文词语关联度的特点,将词语的词性以及位置等作为特征引入关键词提取。对于人工选取关键词带有较多主观性因素的问题,我们将对现有网页数据增加标注内容,减少人为因素对实验结果的影响。

参考文献

1 钱爱兵,江 嵐. 基于改进TF-IDF的中文网页关键词抽取——以新闻网页为例[J]. 情报理论与实践, 2008,31(6):945-950.
2 胡学钢. 基于词汇链的中文新闻网页关键词抽取方法[J]. 模式识别与人工智能,2010,123(1):45-51.
3 赵 鹏. 一种基于复杂网络特征的中文文档关键词抽取算法[J]. 模式识别与人工智能,2007, 20 (6):827-831.
4 马 力. 基于小世界模型的复合关键词提取方法研究[J]. 中文信息学报, 2009,23(3):121-128.
5 Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management,1988, 24(5):513-523.
6 施聪莺. TFIDF算法研究综述[J]. 计算机应用, 2009,29(6):167-180.
7 HOW B C, NARAYANAN K. An Empirical Study of Feature Selection for Text Categorization based on Term Weightage [C]. In Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence. Washington, DC:I EEE Computer Society, 2004.
8 张玉芳,彭时名,吕 佳. 基于文本分类TFIDF方法的改进与应用[J]. 计算机工程, 2006,32(19):76-78.

(责任编辑 徐 波)