基于中文短文本的关键词自动提取对比实验研究

张礼明 张鑫 张朝钿 高泉泽 骆家焕

摘 要：关键词提取是文本挖掘的重要内容之一，本文针对短文本关键词提取算法研究，重点对比了基于 TextRank 词语关系网络方法与基于语义相似网络的方法。实验结果表明，基于语义相似网络的方法在查准率、召回率以及 F1 值上均高于基于 TextRank 的方法。相比于词语结构关系，基于词语语义关系更能体现文档的主题，更能有效地挖掘出主题词。

关键词：关键词提取；TextRank；语义相似度；短文本；文本挖掘

comparing different method of keyword automatic extraction based on Chinese short text

Liming Zhang, Xin Zhang, Chaodian Zhang, Quanze Gao, Jiahuan Lu

Extract : keyword extraction is a crucial part of text mining, this article is concentrated on the study of keyword extracting algorithm for short text, mainly comparing the difference between method based on TextRank word relation network and method based on semantic similarity network. The result shows that, method based on semantic similarity network is better than method based on TextRank word relation network in precision ratio, recall ration and F1 value. As for word structure relation, method based on TextRank word relation network is more capable of representing the document's topic, and more efficiently abstracting the topic word.

Keywords: keyword extraction, TextRank, semantic similarity, short text, text mining

1 引言

关键词提取是文本信息处理的一项重要的技术，是在文本自动摘要、文本自动分类、主题提取以及文本检索的基础工作之一。关键词能够反映文本的主题内容，是文本更为简略的摘要，用户可以根据关键词快速和粗略地获取文本的重要信息，关键词也能作为文本代表性的标签，可以帮助用户迅速地从大量的文档集合中找到用户需要的或者与其相关的文档。

在如今信息爆炸的今天，文本的形式层出不穷，但按文本的语料规模来看，主要分为长文本和短文本，而短文本的表现形式主要有短信、微博、短讯新闻、e-mail 邮件等。但相对于具有语料规模充足，语言规范性强的长文本，短文本通常仅包含 50 多个词，文本长度短、信息量少，特征关键词不足以表示文本。

因此，对于短文本的关键词提取面临着诸多的挑战。

王立霞等人将文关键词自动提取分为 3 类[1]：（1）基于统计特征的方法；（2）基于词语网络分析的方法；（3）基于语义相似度的方法。基于统计特征的方法，典型的有 TF-IDF 模型，该方法简单高效，但效果较差，出现数据集偏斜[2]，主题漂移等情况，尤其难以适应短文本关键词提取。基于词语网络分析的方法，例如基于PageRank改良的TextRank关键词抽取[3],以及利用社会网络分析方法，根据词语共现关系，将词语映射到词语网络，根据网络中的居间度推算关键词的权重[4]。基于语义相似度的方法，主要依靠同义词林，同义词链，计算词语语义、词语概念之间的关联度，如文献[1],引入同义词概念提高关键词提取的效果。

文本主要对比基于 TextTank 词语网络分析的方法与基于语义概念的方法在短文本中的关键词提取效果，以中文文本为处理对象，，并在两种方法均融入词语统计特征，以结巴分词工具的提取关键词作为对照。实验结果表明，

后者在准确率，召回率均高于前者。

2 文本预处理

中文文本预处理的主要工作流程分为：中文分词，去除停用词，词性标注。中文分词，主要使用词典分词的方法，按照一定策略将待分析汉字串与词典中的词条进行匹配，比较成熟的方法有正向最大匹配法、逆向最大匹配法、双向最大匹配法等[5]；去除停用词，主要是指去除文本普遍出现且对关键词提取产生较大噪音的词语，例如“我”，“你”，“也许”等；词性标注，是指为分词结果中的每个单词标注一个正确的词性的程序，也即确定每个词是名词、动词、形容词或其他词性的过程。本文的实验主要采用结巴分词工具（jieba）实现中文分词功能，借助《百度停用词表》和《哈工大停用词表》去除停用词，利用结巴分词工具为分出的词语做词性的标注。词语 W_i 的词性值 pos_i 为 W_i 所属词性的重要度，参照文献[1]，对词性权重的定义为：

$$pos_i=\begin{cases}0.5 & W_i\text{为形容词}\\ 0.3 & W_i\text{为副形词}\\ 0.6 & W_i\text{为名形词}\\ 0.6 & W_i\text{为成语}\\ 0.7 & W_i\text{为简称略语}\\ 0.6 & W_i\text{为习用语}\\ 0.3 & W_i\text{为动词}\\ 0.2 & W_i\text{为动语素}\\ 0.4 & W_i\text{为副动词}\\ 0.6 & W_i\text{为名动词}\\ 0.8 & W_i\text{为名词}\end{cases}$$

3 基于 TextRank 的词语网络分析法

3.1 TextRank 的基本思想 PageRank

TextRank 源于 RageRank 网页排名的算法。PageRank 根据链接关系对网页“投票”排名。一个页面的“得票数”由所有链向它的页面的重要性来决定，到一个页面的超链接相当于对该页投一票。一个页面的 PageRank 是由所有链向它的页面（“链入页面”）的重要性经过递归算法得到的。具体的如公式（1）[6]如下：

$$PR(P)=(1-d)+d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)} \qquad (1)$$

PR(p)表示网页 p 的页面级别;T_i(i =1，2，…, n)表示指向网页 p 的其他网页;d 为用户随机到达一个网页的概率,介于 0 到 1 之间(通常为 0. 85);$C(T_i)$为网页 T_i 向外指出的链接数目;$PR(T_i)/C(T_i)$表示网页 p 的链入网页 Ti 给予 p 的 PR 值。通常，我们设每个网页的初始 PR 值为 1,由公式递归计算各个网页的 PR 值，直到该值趋于稳定[7]。

TextRank 基于 RageRank 思想，以词语作为单个网页，根据词语共现关系构建词语网络，计算每个词语的 TextRank 值并排序，从而提取出关键词。

3.2 基于 TextRank 的关键自动提取算法

本文参照文献[8]的方法，对文本按照句子分割，进行分词和词性标注，并保留切分后的名词、动词、形容词等重要词语，作为词语网络上的节点，根据词语的相邻关系构建词图的边，形成候选关键词图。

在计算 TextRank 的过程中，本文融入词频因子，以词频作为初始的 TextRank 值，以此来增加高频词的重要影响力。

具体算法步骤如下：

输入 文档 D，关键词提取个数 k

输出 文档 D 的关键词

（1）对文档 D 进行分词，去除停用词和词性标注，获得候选词语列表 *CandidateWords*。

（2）去除 *CandidateWords* 中的停用词后，保留形容词、副形词、名形词、动词、动形词，获得词语集合 W。

（3）将文档 D 按照句子作为窗口，分成多个子文档集合$\{d_1, d_2, d_3, \cdots, d_n\}$,统计词语集合 W 中的词语在子文档集合中的文档频率 df。

（4）根据词语在同一子文档中的共现次数作为词语连接的权重，构建无向加权词语网络 G。

（5）根据词语连接关系，以词语的 df 值作为 TextRank 的初始值，按 3.1 节的方法计算词语的 TextRank 值直至收敛，对最后的数值进行排名。

（6）根据排名，选取前 k 个做关键词列表。

4 基于语义概念的关键词提取方法

4.1 同义词链与词语相似度

同义词链是指文档中根据上下文关系确

定词义相同或相近的词的集合[9]，多个同义词链形成的集合即是同义词林。常见的词语间语义相似度的计算方法是根据同义词词典中词语编码距离来求得。文本采用的是哈工大的《同义词词林》扩展版，根据编码的距离判断词语间的语义距离。其中，每个词有若干个编码，每个编码由5层代码和1位标志位描述，即编码 $Code_i$ 描述为 $Code_i=X_{i1}X_{i2}X_{i3}X_{i4}X_{i5}F_i$ 。5层代码分别描述大类、中类、小类、词群和原子词群，1位标志位为“=”、“#”或“@”，其中“=”表示同义；“#”表示同类，属于相关词语；“@”表示词语自我封闭、独立，在词典中既没有同义词，也没有相关词。不同标志位有不同的权重，标志位越前权重越大，本文参照文献[1]的方法，定义一个权重数组如下：

$$weights = [w_1, w_2, w_3, w_4, w_5, w_F] \quad (2)$$

其中 $w_1=1.0$，$w_2=0.5$，$w_3=0.25$，$w_4=0.125$，$w_5=0.06$，$w_F=0.03$。

根据词语之间不同编码的距离，选择最小编码距离作为词语语义距离 $Dis(W_i, W_j)$：

$$Dis(W_1,W_2) = \min_{i,2,\ldots,m,j=1,2,\ldots,n} Dis(Code_{1i}, Code_{2j})$$

而词语之间的相似度 $Sim(W_i, W_j)$ 的定义为：

$$Sim(W1,W2) = \frac{a}{Dis(W_1,W_2)+a} \quad (4)$$

其中，a是一个可调节的参数，表示当相似度为0.5时的词语距离值。a控制语义相似度Sim的取值范围，a越大，语义相似度越不灵敏。本文参看文献[1]对a的取值为5，语义相似度的取值范围为[0.33,1]。

4.2 词语语义相似度网络与中间中心性分析

根据词语之间的语义相似度形成词语相似度网络，语义较为相似的词语联系紧密，不相似的则较为疏远。根据词语网络的特点，可利用社会网络分析理论挖掘出主题词语。本文采用的是中间中心性（Between Centrality）的方法。

中间中心性[10]，用来测量的是一个点在多大程度上位于图中其他点的“中间”，即节点位于其他节点间最短路径的次数。一个度相对比较低的节点可能起到重要的中介作用，因而处于网络的中心。一个点的中间中心性测量的是该点对应的行动者在多大程度上成为“掮客”或者“中间人”，能在多大程度上控制他人。顶点 V_i 的中间中心度 bc_i 定义[1]为：

$$bc_i = \sum_{m,k=1}^{n} \frac{g_{mk}(v_i)}{g_{mk}} \quad (5)$$

其中，n 为图 G 的顶点数目；gmk 表示顶点 Vm 和 Vk 之间的最短路径数；$gmk(Vi)$ 表示顶点 Vm 和 Vk 之间的最短路径是否通过顶点 Vi，通过 Vi 则为1，否则为0。通过计算词语的中间中心度，可以找出文档中主题相关的词语。

4.3 基于语义相关的短文本关键词提取算法

本文参照文献[1]，融入语义特征并结合词语的统计特征。

其算法的具体步骤如下：

输入 文档D，目标提取关键词的个数k

输出 文档D的关键词列表

（1）对文档D进行分词和词性标注，获取候选词语列表 *CandidateWords*。

（2）去除 *CandidateWords* 中的停用词后，保留形容词、副形词、名形词、动词、动形词，并记词语的具体词语词性 *pos*，获得词语集合W。

（3）根据词语集合W和词语间的词语相似度构建词语网络G。

（4）计算图G中所有顶点的中间中心度 *bc*。

（5）计算W中词语的词频特征值 *tf*。

（6）将pos，bc和tf加权获得词语的关键度，词语Wi的关键度计算函数为：

$$score(W_i) = \alpha pos_i + \beta bc_i + \gamma tf_i \quad (6)$$

其中，参照文献[1]，实验中参数α为0.5，β 为0.6，γ 为0.3

5 实验结果与分析

5.1 实验设置和数据源

为了进一步对比出基于TextRank与基于语义相关的关键词自动提取算法的有效性，本文使用python仿真了上述两种算法，编程环境为Ubuntu 14.04，编译工具为pycharm 4.0。本文的数据源主要涉及的领域有数据挖掘研究领域和社会舆论领域。数据挖掘研究领域以中国知网为例，获取了共200条数据，涉及的范围有聚类研究、分类研究、关联规则研究、离群点检测研究，分别各50条数据。数据的内

容包括论文的摘要，论文提供的关键词列表。社会舆论领域以微博为例，获取共50条微博新闻，通过人工标注的方法选取出关键词列表。

5.2 实验结果分析

本文使用两种不同的关键词提取算法，提取关键词个数k的选取为4、5、6,评价算法性能指标为查准率（P）、召回率（R）以及两者调和的平均值F1度量值。

$$P = \frac{\text{已知的关键词集合} \cap \text{提取出的关键词集合}}{\text{已知的关键词集合}} \quad (7)$$

$$R = \frac{\text{已知的关键词集合} \cap \text{提取出的关键词集合}}{\text{提取出的关键词集合}} \quad (8)$$

$$F1 = \frac{2PR}{P+R} \quad (9)$$

其中，已知的关键词集合是指通过人工事先标注或论文摘要提供的关键词数目，提取出的关键词集合是指通过本文叙述的算法提取的关键词数目。本文使用了数据挖掘研究领域、社会舆论领域这两组不同的数据源进行实验，并以开源结巴分词工具提供的关键词提供算法作为参照，该关键词提取算法主要是有监督的提取算法，利用其庞大的语料库以及词性规则。实验结果如图2、图3所示：

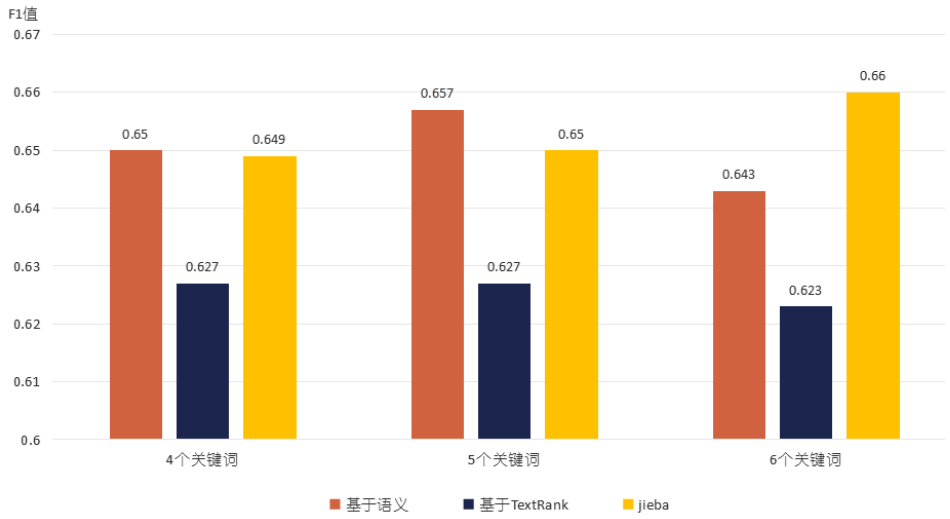

图2 数据挖掘研究领域关键词提取F1值对比情况

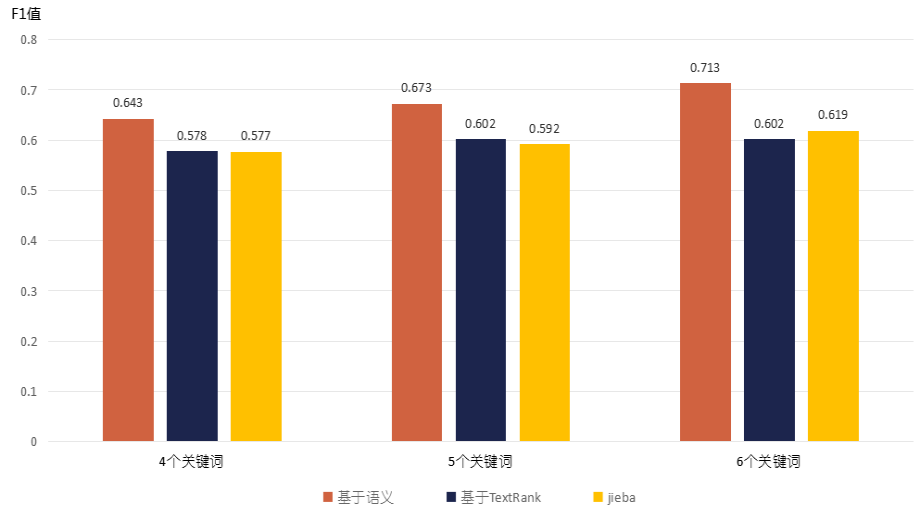

图3 微博新闻领域关键词提取F1值对比情况

从图表可以看出，关键词个数取值为4、5、6的情况下，基于语义的关键词提取算法在查准率、召回率和F1测度值均比基于TextRank的关键词提取要高。在数据挖掘研究领域的关键词提取效果上，总体上基于语义的方法与结巴分词工具的提取方法不相上下，个别上稍逊

与后者，而基于TextRank的方法都比前两者要差，原因是可能因为数据挖掘研究领域专业术语较多，导致出现较多的非登录词，以及关键词列表主要以组合词语居多，影响了分词和提取的效果。而在微博新闻的数据源分析中，基于语义的方法都比较基于TextRank和结巴分词工具要好。由此，验证出基于语义的关键词提取方法的性能要优于基于TextRank的关键词提取方法。具体的代码以及数据集可以查询附录。

6 结语

针对当前主流的中文短文本关键词提取方法，本文主要以基于TextRank与基于语义的方法为研究，阐述两者算法的思想，以及做相关的实验对比。实验结果表明基于语义的关键词提取效果从各方面都优于基于TextRank的提取效果。下一步的工作是研究如何结合这两者的算法思想，并融合到新的关键词提取方法当中，以取长补短，进一步提高关键词提取的精度。

参考文献

[1]王立霞，淮晓永，基于语义的中文文本关键词提取算法，计算机工程[J]，2012，28(1)，01-05.

[2]张建娥，基于 TFIDF 和词语关联度的中文关键词提取方法，情报科学[J]，2012,30(10)，1542-1545.

[3]夏 天，词语位置加权 TextRank 的关键词抽取研究，现代图书情报技术[J]：知识组织与知识管理版，2013,237（09），30-34.

[4]张 敏，耿焕同，王照法。一种利用 BC 方法的关键词自动抽取算法研究，小型微型计算机系统[C]，2007,28（1）：189-192.

[5]奉国和，郑伟，国内中文自动分词技术研究综述，图书情报工作[J]，2011,55（2），41-45

[6]He Qie, Wang Ling. A hybrid particle swarm optimization with a feasible-based rule for constrained optimization [J].Applied Mathematics and Computation，2007 ， 186:1407-1422

[7]李稚楹，杨 武，谢治军，PageRank 算法综述，计算机科学[J]，2011,38(10)，185-188.

[8]顾益军，夏 天，融合 LDA 与 TextRank 的关键词抽取研究，现代图书情报技术[J]，知识组织与知识管理版，2014,289（8），41-47.

[9]刘 群，李素健，基于同义词链的中文关键词提取算法，计算机工程[J]，2010,36（19）:193-95

[10]苏娜，张志强，《社会网络分析在学科研究趋势分析中的实证研究—以数字图书馆领域为例》，情报理论与实践[J]，2009，

附 录：

[1] TextRank的python代码实现链接：

https://github.com/tracyzhangxin/keyword/tree/master/key-master/textrank

[2] 基于语义的python代码实现链接：

https://github.com/tracyzhangxin/keyword/tree/master/key-master/semantic

[3] 实验所用到的数据源：

https://github.com/tracyzhangxin/keyword/tree/master/key-master/DataSet

[4] 参考文献的收集整理链接：

https://github.com/tracyzhangxin/keyword/tree/master/key-master/paper