学校编码：10384　　分类号＿＿＿密级＿＿＿

学号：24320141152409　　UDC＿＿＿

厦门大学

硕 士 学 位 论 文

基于文档主题结构与语义的中文文本关键词提取算法研究

Research on Keyword Extraction Algorithm for Chinese Text Based on Document Topic Structure and Semantics

许振团

指导教师姓名：　林坤辉　教　授

专 业 名 称：　软　件　工　程

论文提交日期：　2017 年　4　月

论文答辩日期：　2017 年　5　月

学位授予日期：　2017 年　6　月

指 导 教 师：＿＿＿＿＿

答辩委员会主席：＿＿＿＿＿

2017 年 4 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。)

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

迈入二十一世纪，伴随着科技的不断进步和互联网的高速发展，各类的信息资源成倍快速增加。人们迫切地希望能快速的、准确的从庞大的信息源中寻找到对自己真正有用的资料。关键词能够高度归纳文档的内容，并且反映文档的主题，为人们寻找资源提供有力的帮助。

目前大部分的文本资源没有提供关键词。虽然人工标注关键词往往拥有较高的准确性，但是因为标注者的学识储备、理解程度差异以及总结概括能力不尽相同，往往带有较强的主观性。况且，其需要花费较多时间阅读、理解文本，这显然无法满足如今信息资源快速增长的速度。关键词提取技术由此出现，其能很好地处理这个难题。建立统一的标准，借助于计算机的快速处理能力，自动提取关键词，可以大量减少人力、时间消耗，降低主观性的影响。

本文以对中文文本进行关键词提取作为研究对象。阐述了关键词提取的基本概念，并对当前国内外的研究情况进行调研。接着，对基于文档主题结构的方法以及基于语义的方法进行详细研究。文中剖析了中文分词和英文分词两者间的差别，前者更加复杂，对关键词提取影响更大。针对中文分词的新词识别这一难点问题，动态更新分词词典来提高中文分词的准确性。同时，借助于向量空间模型，使用改进算法在连续的文本分段中寻找最优聚类，构建文档的主题结构。对基于文档主题结构的算法进行改进，提取全局关键词。并在此基础上，加入中文词语之间的语义相似度的因素，进一步改进算法，将统计方法与语义相结合，提升关键词提取的效果。

本文以准确率、召回率以及F度量作为评价指标，改进算法与其他算法的对比实验的实验结果表明改进算法能够较好地提高对中文文本进行关键词提取的结果，验证了改进算法的有效性。

关键词：关键词提取；主题结构；语义相似度

Abstract

Into the twenty-first century, with the continuous progress of technology and the rapid development of the Internet, various types of information resources doubled rapidly. People are eager to be able to quickly and accurately from a huge source of information to find own really useful information. Keywords can highly induce the content of the document and reflect the theme of the document, it will provide a powerful help for people to find resources.

Most of the current text resources do not provide keywords. Although manual tagging keywords often have a higher accuracy, it tends to has strong subjectivity because of the differences in the knowledge reserves, the differences of understanding degree and summary ability. Moreover, it takes more time to read and understand the text, which obviously can't meet the rapid growth of information resources today. Keyword extraction technology emerges, which can handle this problem well. Establishing a unified standard, with the help of the computer's fast processing power, automatically extract keywords, which can greatly reduce the human and time consumption and reduce the impact of subjectivity.

In this dissertation, the keyword extraction for Chinese text as the research objects. The basic concept of keyword extraction is expounded, and the research on the research situation at home and abroad is carried out. Then, the method based on the document topic structure and the method based on semantic are studied in detail. This dissertation analyzes the differences between the Chinese word segmentation and the English word segmentation, and the former is more complicated and has a greater impact on keyword extraction. Aiming at the difficult problem of new word recognition in Chinese word segmentation, this dissertation dynamically updates the word segmentation dictionary to improve the accuracy of Chinese word segmentation. At the same time, with the help of vector space model, the improved algorithm is used to find the optimal clustering in the continuous text segment, and the topic structure of the article is constructed. The algorithm based on the topic structure of the document

is improved to extract the global keywords. On the basis of this, adding the semantic similarity between the Chinese words to further improve the algorithm. Combine the statistical methods with semantics to improve the effect of keyword extraction.

In this dissertation, the accuracy rate, recall rate and F metric are taken as the evaluation indexes, and the experimental results of the improved algorithm and other algorithms indicate that the improved algorithm can improve the result of keyword extraction for Chinese text, and the effectiveness of the improved algorithm is verified.

Key words: Keyword Extraction; Topic Structure; Semantic Similarity

目 录

Contents

第一章 绪论

1.1 研究背景与意义

近十几年，互联网的触角伸展到了各个行业，深入基层，各类信息正以惊人的速度快速增加。于 2017 年 1 月公布的《中国互联网络发展状况统计报告》[1]使用比较可靠的数据对中国的互联网络的情况进行解析。到 2016 年 12 月截止，我国的网民规模达到了 7.31 亿，每年以几千万人的速度增长，近五年的网民规模变化情况如图 1-1 所示；互联网近年来的普及程度达成新高，近 53.2%，近五年的互联网普及程度变化情况如图 1-2 所示。我国的域名总数达到 4228 万个，比前一年增长了 36.3%；中国网站也再创新高，达到 482 万个，比前一年增长了 14.1%。快速增长的数字背后显示了日益丰富的信息资源。

图 1-1 中国网民规模（2012-2016 年）

互联网正在改变人们的生活。人们足不出户，通过互联网就可以了解国内国外发生的事情，可以在线与人交流，可以网上购物，可以在线查找信息，可以在线阅读文档等。同时，其也给人们带来了许多不便，其一，面对如此庞大的资源，大量信息杂乱无章的分布，人们在寻找需要的信息的时候，越来越感到力不从心，不知从何处找起；其二，人们不知如何合理地选择信息。因而人们迫切希望能先对资源进行过滤、分类，进而快速且高效地寻找到对自己有价值的资源。

图 1-2 中国互联网络的普及率（2012-2016 年）

在一篇文档当中，关键词是对文档的内容的深度提炼，其一般通过几个词语或者短语来表示。透过文档的关键词，能够洞悉该篇文档描写的主要内容，迅速判断出是否是需要的资源。关键词提取技术可以帮助人们从庞大的数据资源中高效查找与识别出其所需的资源，提高对资源检索的效率。

在多个领域，关键词提取技术为其提供了重要的技术支持。关键词可作为文本的索引，使得人们可以很方便得查找到特定主题的资料。利用搜索引擎（如百度搜索、Google 搜索等）查询资源时，提供合适的关键词将使得搜索结果更加准确，提高查询效率。文本分类指的是按照一定的标准，构建出分类的模型，把待分析文本中提取出来的特征信息放入模型中计算，计算出待分析文本所属的类别，完成分类。文本分类的典型方法是依据文本当中是否包含和类别名称相关的关键词，进而将其归类于所属的类别。而文本聚类会选用文本中的某些特征信息来计算不同文本间的相似程度，将彼此之间具有较高相似性的文本聚拢到一起，并且将彼此之间的相似性较低的文本分散到不相同的簇里。准确的关键词能够优化文本聚类的过程，获取更加符合实际的结果。

当前大部分的文本资源没有标注关键词，如技术文章、微博等。通过调研得知，一共存在两种解决方案，其一，人工标注关键词；其二，关键词提取。人类阅读、理解文本内容后，对其标注关键词，一般拥有比较高的准确性。但是由于标注者的学识储备、对关键词的理解以及总结概括能力不同，带有较强的主观性，提取的关键词不尽相同。况且，使用人力来对文本标注关键词会花费较多的精力

来阅览、理解文本内容，这显然满足不了当前信息资源数量不断翻倍的现状。关键词提取技术由此产生，其可以很好地处理这个问题。建立一致的提取规则，借助于机器的强大功能，自动提取文本的关键词。通过该方法，能够大大减轻人力、时间的花费。

1.2 国内外研究现状

从文本内容当中提取关键词之前需要先对其进行分词。西文的词语与词语之间存在明显的分隔符，如英文使用空格作为词语间分隔符，而中文的字与字之间不存在显式的分隔符，所以和西文文本相比，对中文文本进行关键词提取将更加复杂。通常，中文需要两个或者更多的字才能表达某个意思，构成一个词。因此，为了让机器能够正确处理中文句子，需要先使用中文分词器对句子分词，将句子转换为一个个词语。中文分词结果的好坏直接影响了计算机对于句子的理解以及关键词的提取结果，即中文关键词提取更加依赖于分词。国外所做的西文文本的关键词提取研究无法将其直接应用到中文领域。

我国的研究人员在对中文文本进行关键词提取的领域做出了不少成果。文献[2]中将最大熵模型应用于关键词提取。但是其在选择特征以及估计参数值的时候，存在较大缺陷，关键词提取的结果不佳。文献[3]提出一种使用文献主题的方法，但是它存在较大的局限性，它需要先对语料库做好标引，并只能从元数据标题当中提取出对应的关键词。文献[4]在 TF-IDF（Term Frequency-Inverse Document Frequency）方法的基础上，为了避免单纯使用 TF-IDF 方法造成的误差，文中引入了词语关联度的因素。文献[5]针对中文的新闻网页开展关键词提取研究，文中构建了词汇链，将其应用到关键词提取算法中。文献[6]对基于词语网络的方法进行了改进，为关键节点选择策略定义了两种新的指标，其一，网络平均逆路径长度；其二，有效聚类系数。这样，其能应用到非连通图。并且在网络节点选择的过程中，使用了基于词语相似度的中文分词算法，使得含义相同的词能得到更高的权值。文献[7]通过组合比较有可能的词语成短语，其能表述更具体的意义，更有可能成为关键词，并且使用了同义词词典来识别同义词，再结合词频等统计信息，构造新的词语得分公式来提取关键词。文献[8]中的改进算法应用了高维聚类技术，其首先通过双重分词进行中文分词，使用高维聚类处理，最后检验每个词语，从中提取出最适合的词语作为关键词。文献[9]使用一

定大小的窗口来计算词语共现的频率，当其大于预设的阈值时，将其作为组合词，结合词性、词频-逆文档频率等统计信息，使用贝叶斯模型对数据集开展大量训练，最后提取关键词。

相对而言，我国在该领域的研究开始得比国外晚一些，国外做了更多的研究，在该领域已经获得更多的成就。文献[10]在其改进算法中应用了遗传算法，并且将该算法应用于一个商业的项目中，取得较大的成功。文献[11]选取文本特征，使用朴素贝叶斯技术训练模型，最后使用模型来提取关键词。文献[12]将一种复杂网络理论应用到关键词提取中，得到较好的结果。文献[13]把语言学应用于关键词提取领域中，将统计和语法相结合，文中定义了字符串块，并且加入词频、词性等统计信息，取得了比较好的关键词提取结果。从信息检索当中使用的 PageRank 算法得到启发，文献[14]把词和文档之间的语义关系考虑进来，引入图的概念，其提出的 TextRank 方法把文本中的词作为图的结点，把词和词之间的关联作为图的边，利用经典的 PageRank 算法来计算词语的得分，进而得到关键词评分，根据评分排序，取得了一定的效果。文献[15]把支持向量机这一机器学习方法应用到关键词提取中。但是该方法需要一个足够大的语料库，并且语料库中的文本需要提前标注好关键词，使用其训练模型，然而当前缺乏标准的、统一的中文关键词语料库。根据时间序列，文献[16]把文档分割成多个小文档的集合，对每个小文档集使用 TF-IDF 方法进行关键词抽取，然后，综合考虑所有的小文档提取出来的关键词，进而，提升关键词提取的准确性。文献[17]在对单文本进行关键词提取的时候，使用到了遗传算法，并且使用频率特征以及词语的位置关系来提取关键词。文献[18]把文本当作词和词间的语义关系，组成语义图，使用维基百科来控制词汇量，过滤掉噪声词，并且利用维基百科来对词加权，建立语义关系。但是把维基百科当作外部的资源加入到算法中，在过滤干扰词的时候，同时也有可能过滤到真正的关键词。

1.3 论文研究内容

本文首先介绍了文本关键词提取的基本概念、中文分词与英文分词存在的差异、常用的中文文本关键词提取方法。研究了基于文档主题结构的方法和基于语义的方法，分析其原理以及现有的实现方案。针对中文分词中存在的新词识别的难点，使用动态更新分词词典来不断改善中文分词的效果。对基于文档主题结构