信息科学技术专著丛书

大规模英文语义树构建技术

冶忠林　崔宝阳　马子恒　杨燕琳　著

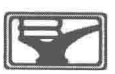
CONSTRUCTION TECHNOLOGY OF
LARGE-SCALE ENGLISH SEMANTIC TREE

北京邮电大学出版社
www.buptpress.com

信息科学技术专著丛书

大规模英文语义树构建技术

冶忠林　崔宝阳　马子恒　杨燕琳　著



北京邮电大学出版社
www.buptpress.com

内 容 简 介

本书主要介绍一种基于多特征联合建模的大规模英文语义树构建技术，将复杂网络的思想引入语言建模中，并通过特设算法构建了语言的空间和结构关系，确定了大规模英文语料中的关键词，从而构建了语义树。本书将语义树按语义、词性及现实关联等不同标签对进行划分，并一一列举。本书的研究成果可为语言研究学者提供丰富的数据支撑。此外，构建的语义树与常识判断存在较大差距，而对这种现象的研究正是语言学家需要解决的问题。

图书在版编目(CIP)数据

大规模英文语义树构建技术 / 冶忠林等著. -- 北京 : 北京邮电大学出版社，2022.7
ISBN 978-7-5635-6667-9

Ⅰ. ①大… Ⅱ. ①冶… Ⅲ. ①自然语言处理 Ⅳ. ①TP391

中国版本图书馆 CIP 数据核字(2022)第 110704 号

策划编辑：姚 顺 刘纳新 **责任编辑**：廖 娟 **责任校对**：张会良 **封面设计**：七星博纳

出版发行：北京邮电大学出版社
社　　址：北京市海淀区西土城路 10 号
邮政编码：100876
发 行 部：电话：010-62282185 传真：010-62283578
E-mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：唐山玺诚印务有限公司
开　　本：787 mm×1 092 mm 1/16
印　　张：15.25
字　　数：362 千字
版　　次：2022 年 7 月第 1 版
印　　次：2022 年 7 月第 1 次印刷

ISBN 978-7-5635-6667-9 **定价**：58.00 元

前　言

自然语言作为现实世界中人与人之间交流的重要工具，数千年来一直在不断发展，在历史的长河中，各个国家的语言都形成了其独特的体系。随着第三次科技革命的兴起，计算机成了科技革命的主要工具，如何使计算机理解人类语言，并处理、使用这些语言，都是自然语言处理的主要研究方向。

关键词提取技术作为自然语言处理研究领域的重要研究分支之一，在各个领域都有所应用，其在文本分类、内容检索方面都属于核心技术。关键词提取技术通过大量的文本知识，且根据提取关键词的侧重方向不同，对文档进行较为精准的分类，或是将提取的关键词作为文档的检索标签，大大提高了知识检索的效率。关键词提取技术在搜索引擎、智能问答、文本生成、情感计算等多方面也有所应用。

从过去的基于统计的方法到如今基于深度学习的方法，我们对数据的分析角度和理念发生了变化。基于统计的方法视数据为最宝贵的资源，系统通过优化后的统计学习模型能在有限的数据中挖掘出更多的特征，供后续的机器学习任务使用和分析。现如今数据虽然很重要，但是重要的不是海量的、含有噪声的数据，而是经过预处理和清洗之后的数据，也称之为“知识”。

在深度学习时代，传统的经验和学习方法已经不能满足任务需求。众多学者认为，数据的海量复现能够弥补模型的部分缺陷，例如精度等；而对大数据的处理更关注的是效率。如何在海量的数据中挖掘出重要的知识一直是各个领域面对的重要挑战。本书主要研究的内容是如何在海量的文本中挖掘出最重要的词语，进而为重要的关键词构建出语义树，并通过分析语义树来分析语言的使用规则和模型的侧重点。

从海量的语料中挖掘出最重要的关键词，首先想到的是基于词频的方法，该方法的缺点也很明显——出现频率最高的词语不一定是英语语料中最重要的词，而且该方法并未考虑词语的搭配和语法等。目前，采用深度学习方法进行命名实体识别、关系抽取、智能问答等的方法较多，但尚未发现识别和发现语料中关键词的方法。该任务的目标是在海量的英语语料中发现多个最重要的关键词，通过对这些关键词语义树的分析，可以归纳出关键词的隐含特征。本书正是基于以上现状，设计适用于大规模数据下的关键词提取技术，并构建其对应的语义树结构，试图从中挖掘现代英文语义环境及特点。通过对语义树的构建及展示，为读者展现如今英文语言的使用习惯和用法规则。本书也从构建的语义

树中抽取部分树进行分析与研究，为研究语言学的读者提供一些思路。

在本书的撰写过程中，作者参考了众多学者的学术专著和研究论文，在本书中均以参考文献的形式进行了标识，如有不慎遗漏，亦表示诚挚的歉意。

本书的完成受国家重点研发计划(2020YFC1523300)、青海省自然科学基金青年项目(2021-ZJ-946Q)、青海省重点研发与转化计划(2020-GX-112)、青海师范大学中青年科研基金项目(2020QZR007)资助。

由于作者水平有限，在撰写过程中难免存在错误和疏漏，敬请读者批评指正(作者邮箱：zhonglin_ye@foxmail.com)。

作　者

2021 年 11 月 3 日

目　录

第 1 章　关键词提取技术概况 …… 1

1.1　关键词提取技术简介 …… 1
1.2　关键词提取技术 …… 3
1.2.1　基于统计的关键词提取技术 …… 3
1.2.2　基于网络的关键词提取技术 …… 4
1.2.3　基于深度学习的关键词提取技术 …… 5
1.3　关键词提取技术总结 …… 6

第 2 章　相关技术介绍 …… 7

2.1　BERT 模型 …… 7
2.2　*K*-Means 聚类算法 …… 8
2.3　DeepWalk 及其矩阵分解形式 …… 9
2.4　结构网络 …… 11
2.5　语义网络 …… 11

第 3 章　大数据英文语义树构建技术 …… 13

3.1　语料处理与聚类 …… 13
3.1.1　语料处理 …… 13
3.1.2　词汇聚类 …… 14
3.2　基于结构与语义的词汇影响力网络 …… 14
3.3　构建节点影响力树 …… 15
3.4　词频与节点影响力加权计算 …… 19
3.5　构建语义树 …… 19

第 4 章　语义树分析 …… 21

4.1　基于语义划分的语义树分析 …… 21
4.2　基于词性划分的语义树分析 …… 23
4.2.1　根据词性划分的语义树 …… 24
4.2.2　根据情感色彩划分的语义树 …… 25
4.3　基于现实关联划分的语义树分析 …… 27

第 5 章　语义树展示 …… 30

5.1　including …… 31
5.2　school …… 32
5.3　laguna …… 33
5.4　valley …… 34
5.5　company …… 35
5.6　returned …… 36
5.7　verb …… 37
5.8　polygon …… 38
5.9　quality …… 39
5.10　bi …… 40
5.11　phalanx …… 41
5.12　salvador …… 42
5.13　ye …… 43
5.14　treble …… 44
5.15　catholic …… 45
5.16　merle …… 46
5.17　streptomyces …… 47
5.18　forage …… 48
5.19　low …… 49
5.20　artillery …… 50
5.21　hydrogen …… 51
5.22　physically …… 52
5.23　andean …… 53
5.24　major …… 54
5.25　inventive …… 55

5.26 gastrointestinal …… 56
5.27 reversible …… 57
5.28 born …… 58
5.29 rabbi …… 59
5.30 pandemic …… 60
5.31 bluish …… 61
5.32 album …… 62
5.33 siena …… 63
5.34 governorship …… 64
5.35 acetate …… 65
5.36 oblong …… 66
5.37 undesirable …… 67
5.38 experience …… 68
5.39 ten …… 69
5.40 confidence …… 70
5.41 epidemiology …… 71
5.42 winston …… 72
5.43 held …… 73
5.44 hypotheses …… 74
5.45 boulogne …… 75
5.46 white …… 76
5.47 philosophy …… 77
5.48 john …… 78
5.49 coronavirus …… 79
5.50 ps …… 80
5.51 anti …… 81
5.52 furlong …… 82
5.53 quran …… 83
5.54 da …… 84
5.55 punk …… 85
5.56 error …… 86
5.57 slave …… 87
5.58 de …… 88
5.59 betting …… 89

5.60 taro …… 90
5.61 dopamine …… 91
5.62 chesterfield …… 92
5.63 sports …… 93
5.64 rama …… 94
5.65 football …… 95
5.66 college …… 96
5.67 bach …… 97
5.68 female …… 98
5.69 fruit …… 99
5.70 ade …… 100
5.71 ho …… 101
5.72 stereotype …… 102
5.73 catchy …… 103
5.74 roman …… 104
5.75 alphonse …… 105
5.76 passerine …… 106
5.77 ab …… 107
5.78 land …… 108
5.79 annual …… 109
5.80 narrow …… 110
5.81 force …… 111
5.82 parish …… 112
5.83 lez …… 113
5.84 wrongdoing …… 114
5.85 hong …… 115
5.86 united …… 116
5.87 java …… 117
5.88 north …… 118
5.89 historian …… 119
5.90 silica …… 120
5.91 costal …… 121
5.92 hair …… 122
5.93 divergence …… 123

5.94 embroidery …… 124
5.95 pc …… 125
5.96 manifesto …… 126
5.97 pregnancy …… 127
5.98 receptive …… 128
5.99 juan …… 129
5.100 thuringia …… 130
5.101 rani …… 131
5.102 president …… 132
5.103 cathode …… 133
5.104 desk …… 134
5.105 station …… 135
5.106 frog …… 136
5.107 expel …… 137
5.108 moto …… 138
5.109 michael …… 139
5.110 gradient …… 140
5.111 limburg …… 141
5.112 enzyme …… 142
5.113 independence …… 143
5.114 central …… 144
5.115 team …… 145
5.116 choctaw …… 146
5.117 alaskan …… 147
5.118 appropriation …… 148
5.119 jura …… 149
5.120 eastwards …… 150
5.121 attribution …… 151
5.122 ford …… 152
5.123 planar …… 153
5.124 band …… 154
5.125 vologda …… 155
5.126 country …… 156
5.127 jones …… 157

5.128 ph …… 158
5.129 bia …… 159
5.130 billie …… 160
5.131 pearson …… 161
5.132 moray …… 162
5.133 guerilla …… 163
5.134 fish …… 164
5.135 erect …… 165
5.136 reopen …… 166
5.137 bacon …… 167
5.138 cholesterol …… 168
5.139 altarpiece …… 169
5.140 actor …… 170
5.141 set …… 171
5.142 intestinal …… 172
5.143 undisputed …… 173
5.144 amplifier …… 174
5.145 ha …… 175
5.146 franc …… 176
5.147 batman …… 177
5.148 god …… 178
5.149 netted …… 179
5.150 tang …… 180
5.151 isotopes …… 181
5.152 romans …… 182
5.153 lucknow …… 183
5.154 corinthian …… 184
5.155 bale …… 185
5.156 angry …… 186
5.157 congolese …… 187
5.158 fascism …… 188
5.159 dialect …… 189
5.160 antibiotic …… 190
5.161 black …… 191

5.162 war …… 192
5.163 zoologist …… 193
5.164 tijuana …… 194
5.165 antelope …… 195
5.166 mould …… 196
5.167 capable …… 197
5.168 reliant …… 198
5.169 quarantine …… 199
5.170 lander …… 200
5.171 dragoon …… 201
5.172 fibrous …… 202
5.173 avian …… 203
5.174 matrix …… 204
5.175 habsburg …… 205
5.176 tahiti …… 206
5.177 bandy …… 207
5.178 direct …… 208
5.179 piety …… 209
5.180 theatrically …… 210
5.181 match …… 211
5.182 neurology …… 212
5.183 impedance …… 213
5.184 pineapple …… 214
5.185 prefecture …… 215
5.186 socialist …… 216
5.187 american …… 217
5.188 ora …… 218
5.189 parallel …… 219
5.190 holotype …… 220
5.191 southern …… 221
5.192 prelude …… 222
5.193 hydroxyl …… 223
5.194 west …… 224
5.195 commune …… 225

5.196 species …… 226
5.197 form …… 227
5.198 days …… 228
5.199 lyceum …… 229
5.200 alienation …… 230

参考文献…… 231

第 1 章　关键词提取技术概况

1.1　关键词提取技术简介

随着大数据时代的到来，数据的种类及规模也在快速增长，从大规模数据中提取关键信息数据并使用，一直都是研究的难点。其中，语言数据作为人际交流的重要工具，如何将其智能化处理一直都是研究的热点[1]。对语言数据的应用及处理的研究覆盖了许多学科与领域，例如文学中，对各种语言结构及语义的研究，对文章的情感、深意的解读；计算机科学中，如何使计算机理解并使用自然语言、识别语音并处理等。随着对自然语言研究得越来越深入，关键词提取作为自然语言处理领域中重要的研究方向，主要应用于文本处理、信息检索等方面。在现实世界中，一般都会对大量的文档或数据采取分类别存储或查找的方式，这样不仅节省了大量的时间，还能够做到合理有序的存储。对文档或数据分类就是基于其核心内容进行分类的，对文档或数据核心内容的抽取所用的就是关键词提取技术。

同时，随着自然语言处理的快速发展，关键词作为表达文档主题意义的最小单位[2-3]，在自然语言处理领域中的各个方面都发挥着极为关键的作用。例如，在图书管理领域，一直以来，书本是通过人力来进行分类布置的，随着书籍种类的增多和数量的增长，完全通过人力来完成这项工作变得极为烦琐和复杂。而关键词提取技术可以利用计算机快速提取书籍的核心内容及特征，并依此将内容相近的书籍自动归类，大大节省了人力、物力。在生活节奏加快的现代社会，人们面对各种各样的文章，很少有耐心通过一篇一篇地阅读来寻找所需的内容，通过关键词提取技术可以提取每篇文章的主要内容，对每篇文章进行关键词标注，使读者通过快速浏览文章关键词快速地寻找所需要的内容，例如知网就是通过对每篇文章的关键词标注，为浏览者提供相关关键词标签的筛选，从而做到快速定位。如图 1-1 所示。

面对庞大且繁杂的数据，如何从文档或者词库中提取其核心词语一直是众多学者研究的目标。基于人工的方法费时费力，且每个人的评定方法和标准也很难做到统一。因此，实现计算机对关键词的自动提取显得尤为重要。1957 年，Luhn 提出的基于词频的关键词抽取技术[4]，开启了研究关键词提取技术的热潮，众多学者不断参与其中，经过近 65

年的发展，关键词抽取技术衍生出了各种各样的方法。从最初基于统计的方法，到现如今基于深度学习的方法，关键词抽取技术不断汲取各个领域的方法、策略以提升提取的准确性和有效性。

图 1-1　知网标签检索

基于统计的关键词提取技术。该方法的核心思想在于统计语言的各种固有属性，对统计的属性进行数字化处理，从而得到词的重要程度，进行排序后取重要程度最高的几个词作为关键词。最为经典的就是基于词频的方法，通过统计每个词汇在语料、文章中出现的次数来抽取关键词，该方法认为词频越高的词汇其重要程度越高。基于统计的关键词提取方法简单有效，但有学者通过研究发现有些关键词并不会在语料中多次出现[5]，该方法忽略了词频低但重要的词。

基于网络的关键词提取技术。该方法将词看作节点，通过语料或文章中词汇的某些关系构建节点间的边得到对应的网络。用网络节点的评价指标来评价节点的重要性，将关键词提取问题转化为网络关键节点识别问题。经典算法为 TextRank 算法[6]，该算法由 PageRank 算法[7]演化而来，通过固定大小的滑动窗口构建词汇的共现网络，然后使用 PageRank 算法不断迭代计算直到收敛，最后将值进行排序，得到节点重要性排序。但该算法主要依据语料自身的结构关系进行网络的构建，缺少外词汇本身的语义知识，仅考虑语言的结构而忽略了词汇的语义信息。

基于深度学习的关键词提取技术。这是当今研究的热点，其核心思想是首先通过高速发展的深度学习建立对应的语言模型，然后通过不断的学习、训练得到语料中词的语义特征，最后通过这些特征来判别词汇中较为关键的词汇。比较经典的就是通过构建神经

网络得到词的向量表示，即词嵌入。Word2Vec[8]就是典型的词嵌入模型之一，通过大规模语料的训练，得到词的向量表示，从一定意义上而言，词向量代表的就是词的语义信息，通过计算词向量之间的余弦相似度，可以得到词之间的语义关联程度。利用这种方式结合构建网络的方法可以充分考虑词之间的语义关系。但基于深度学习的方法大多数需要大规模的数据，当数据规模过小时，很难得到可靠的结果。

1.2　关键词提取技术

1.2.1　基于统计的关键词提取技术

基于统计的关键词提取技术的优点很明显：对文档、语料的要求不高，不需要对语料进行复杂的处理，也不需要外部知识的监督标注，实现简单，易于统计，且往往有着不错的效果。虽然其忽略了某些关键而统计信息得分过低的词，但是基于统计的思想作为提取的方法的一种属性，与其他方法相结合，往往可以得到更好的效果。

（1）基于词频的关键词提取技术

基于统计的关键词提取思想中最为经典的方法就是基于词频的关键词提取技术。统计词频的方法简单实用，该方法认为对于任意文档语料，出现次数越多的词，其为文章核心词汇的概率越高。例如，对于一篇内容为战争的文档，文档中“战争”“炮火”等相关词汇出现的概率及频率大概率会大于“学生”“老师”等词出现的概率及频率。通过分析人们的写作习惯可以了解到，一篇文章往往是围绕一个主题的，与主题相关的词汇就会在文章中反复出现，与主题无关的词则极少出现。所以，对于大多数文档语料都可以通过词频筛选核心词汇。

基于词频的关键词提取技术就是通过上述思想得到，对词频的统计方式有多种，不同的方式侧重点不同，其中最简单的方式就是直接统计，然后对词频降序排序得到词汇重要度序列。算法具体步骤如下。

步骤 1：对文档进行预处理，然后做分词处理。

步骤 2：统计文档中所有词的词频。

步骤 3：将词汇按词频大小降序排序，取排名高的词汇。

（2）TF-IDF

TF-IDF（Term Frequency-Inverse Document Frequency）[9]是基于词频的关键词提取方法的衍生，其中 TF 为词频，IDF 为逆文本频率指数。该算法认为，如果一个词在一篇文章中出现的频率很高，而在其他文章中出现的频率很低，那么说明这个词具有辨识度，在这篇文章中极为重要，可以作为文章的关键词。

TF-IDF 算法在基于词频统计的基础上充分考虑了词汇的特殊性，不仅考虑到词汇在文章中的重要程度，而且进一步通过外部文档来寻找高频词汇中具有特殊性的词。通过该算法提取的关键词可以很好地区分不同文章的核心内容，算法易于实现，但是需要外部文档进行比照。其核心为通过计算归一化处理后的 TF、IDF 值，进而计算得到 TF-IDF

值并降序排序得到词汇重要度序列。TF 计算公式为

$$\mathrm{TF}_i = \frac{F_i}{F_{\max}} \tag{1-1}$$

其中，F_i 为 i 词在输入语料中出现的次数，$F_{\max}$ 为输入语料中词频最高的词的出现次数。IDF 计算公式为

$$\mathrm{IDF}_i = \log \frac{D}{D_i} \tag{1-2}$$

其中，D 为外部文档数量，D_i 为出现 i 词的文档数量。TF-IDF 计算公式为

$$\text{TF-IDF} = \mathrm{TF} \times \mathrm{IDF} \tag{1-3}$$

算法具体步骤如下。

步骤 1：对文档进行预处理，然后做分词处理。

步骤 2：统计文档中所有词的词频并计算每个词的 TF 值。

步骤 3：统计文档中所有词在外部语料的词频并计算其 IDF 值。

步骤 4：计算所有词的 TF-IDF 值，对其进行降序排序，取排名高的词。

1.2.2 基于网络的关键词提取技术

复杂网络起源于哥尼斯堡七桥问题[10]，在复杂网络的研究中，对关键节点的识别一直是研究的重点方向之一[11-12]，该问题与自然语言处理的关键词提取问题相似。Cancho 和 Sole[13]第一次通过研究表明：人类语言也是一种复杂网络，其具有复杂网络的小世界特性与无标度特性。于是，有学者提出通过某种方式将文档构建为网络[14]，将词汇看作节点，将词汇之间的某种关系看作边，使用关键节点识别技术来解决关键词提取问题，根据节点相关属性的值或节点在网络中起到的作用来判断词节点的重要程度，为解决关键词提取问题提供新的研究思路。

而对网络关键节点的识别有多种方式，要从中找到适用于语料网络的关键词识别方法，其中 PageRank 算法作为经典算法之一，从中衍生了多种用于关键词提取的算法，如 TextRank、SingleRank 等。

(1) TextRank

在网络关键节点识别技术中，较为经典的为 PageRank 算法。该算法主要应用于重要网站的识别，即对网站进行重要性排名。TextRank 算法就是 PageRank 算法在关键词提取问题上的应用，首先通过构建语料的共现网络，然后对构建的网络使用 PageRank 算法，使其对节点进行重要程度排名，排名靠前的节点对应的词即为关键词。该算法应用于图模型的合理性已经得到了证明[15]。

该算法提出者认为网络中的节点之间会相互影响，网络中各个节点之间传递的影响力是不同的，通过给节点设定一个 PR 初始值后不断迭代，动态模拟网络的动态行为，直到节点的 PR 值收敛停止，PR 值越大的节点越重要。简单来说，如果一个网站有越多的网络中存在该网站的链接，那么这个网络一定越重要(受欢迎)。且根据该思想，每个网络的重要程度及影响力是不同的，PageRank 就通过不断的迭代计算直至收敛，从而获得每个网络的重要程度评分。TextRank 算法虽然充分考虑了词汇之间的结构关系，但是忽