基于 Lucene 的中文是非问答系统的设计与实现*

罗东霞，卿粼波，吴晓红

（四川大学 电子信息学院，四川 成都 610065）

摘 要：针对中文是非问句，设计并实现了基于 Lucene 的问答系统，主要包括问句预处理、索引创建和答案整理三部分。问句预处理部分，引入句法成分权重和命名实体权重改进 TextRank 算法，得到一种提取问句核心词的方法。在索引创建部分，针对本地的多源数据进行文档融合创建索引，降低数据多样性带来的复杂度。在答案整理部分，对查询索引结果进行答案判决，输出肯定或否定含义的答案。实验结果表明，数据融合能有效减少索引创建耗时，改进 TextRank 的核心词提取方法准确率明显高于 TextRank，系统具有较为不错的性能。

关键词：是非问答；Lucene；TextRank；核心词提取

中图分类号：TP391.1　　**文献标识码**：A　　**DOI**：10.19358/j.issn.2096-5133.2020.11.012

引用格式：罗东霞，卿粼波，吴晓红. 基于 Lucene 的中文是非问答系统的设计与实现[J].信息技术与网络安全，2020，39(11)：74-78.

Design and implementation of Chinese yes-no question answering system based on Lucene

Luo Dongxia，Qing Linbo，Wu Xiaohong

（College of Electronic Information，Sichuan University，Chengdu 610065，China）

Abstract：A Chinese yes-no question answering system based on Lucene around Chinese yes-no questions is designed and implemented，and this system includes three parts：question preprocessing，index creation，and answer sorting. In the first part，introducing the syntactic component weights and named entity weights to improve the TextRank algorithm，a method for extracting the core words of the question sentence is obtained. In the second part，the document fusion is created for the multi-source data to reduce the complexity which is caused by data diversity. In the last section，the query index results are judged by the answer，and then the answers with positive or negative meanings are output. The experimental results show that data fusion can effectively reduce the index creation time，and the accuracy rate of the improved TextRank core word extraction method is significantly higher than TextRank，which means the system has good performance.

Key words：yes-no question answering；Lucene；TextRank；core word extraction

0 引言

随着人工智能技术的飞速发展，传统搜索引擎已不能满足用户需求，自动问答系统逐渐成为信息检索领域的研究热点，并具有广泛应用前景[1]。自动问答系统指允许用户以自然语言的形式描述问句，并将简洁答案返回给用户的一种信息检索系统[2]。

近年来，自动问答系统相关的研究和应用十分广泛。2011 年，IBM 公司的深度问答系统首次将自然语言处理与深度学习结合起来，使得众多机构和企业纷纷效仿。2013 年 3 月，京东上线京东 JIMI 客服机器人，提供客户常规咨询服务；2016 年 10 月，百度推出百度医疗大脑，实现健康在线咨询[3]。但目前关于中文自动问答系统的研究多是围绕特指问句，其开放性的回答方式不适用于是非问句的二值答案。例如，对 JIMI 提问：“京东自营满 88 包邮对吗？”，JIMI 的答案是京东自营商品包邮的详细说明，而非是非问句要求的“对”或“不对”的二值答

* 基金项目：四川省科技计划项目(2018HH0143)

案。中文是非问答系统的设计与实现，能够弥补目前中文自动问答仅能作答特指问句的不足，帮助用户快速获取简洁的答案，对自动问答系统的研究和应用有着极其重要的意义。

本文利用 Lucene 设计并实现一种中文是非问答系统，主要工作包括：(1)引入句法成分权重和命名实体权重，改进 TextRank 算法[4-5]，提出一种问句核心词提取方法；(2)针对 MySQL、Neo4j 和本地新闻文件中的多源数据，提出一种多源数据融合索引创建方法，减少索引创建耗时；(3)查询索引并对索引结果判决，获得是非问句的二值答案。

1 是非问答系统设计

本系统主要包括三个模块：多源数据融合创建索引、问句预处理和答案整理。其中，问句预处理包括问句分词、问句关键词组 $K_w[T_1, T_2, \cdots, T_n]$($T_i(i=1 \cdots n)$表示关键词)与核心词 C_w 的提取；多源数据融合创建索引对本地数据进行文档融合，统一创建索引；答案整理将索引结果 LR 与 C_w 和 K_w 进行匹配，匹配成功则判定为肯定答案，反之为否定答案。系统框图如图 1 所示。

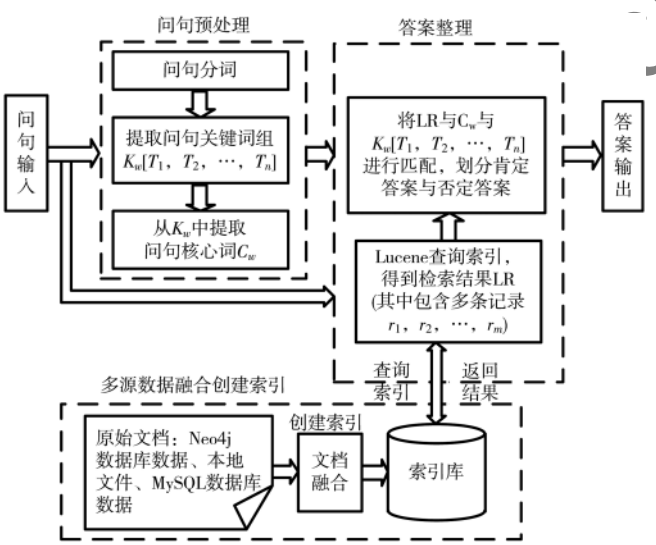


图 1 是非问答系统框图

2 问句预处理

问句预处理指对问句的理解和分析，其任务是根据问句确定检索和答案提取策略[6]。具体到本系统，问句预处理包括问句分词、问句关键词组提取和问句核心词提取。

2.1 问句预处理流程

中文问句分词指将构成问句的汉字序列切分成一个一个单独的词的过程[7]。关键词指体现个人需求的词汇，多个关键词组成关键词组 $K_w[T_1, T_2, \cdots, T_n]$。问句核心词 C_w 指 K_w 中关键度最高的词，其高低由提取关键词时的词语得分 W_s、词语句法成分权重 W_c 和命名实体权重 W_e 共同决定。HanLP 是由一系列模型与算法组成的自然语言处理工具包，适用于 Java 生产环境[8]，本文采用 HanLP 进行问句分词，分词结果包含词语和词性。TextRank 算法是一种基于图的排序算法，不需要提前训练，本文采用该算法提取问句关键词。问句预处理流程如图 2 所示。

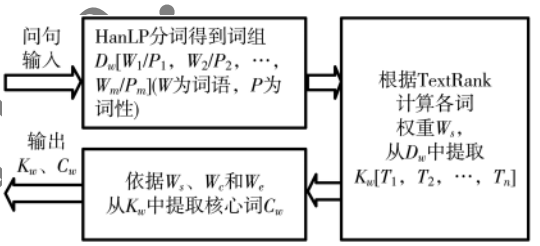


图 2 问句预处理流程图

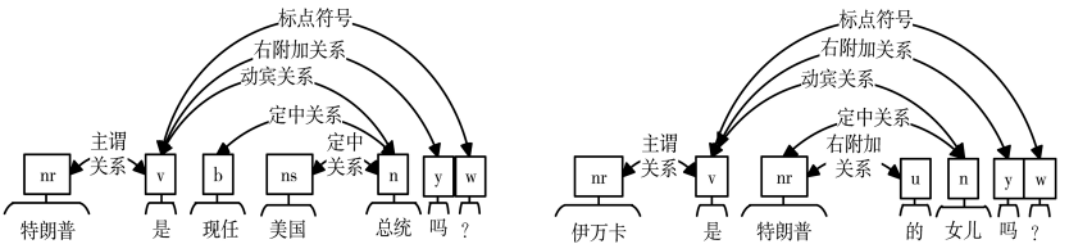
2.2 改进 TextRank 的核心词提取方法

核心词指关键词组中关键度最高的词，不同句法结构关键词组 K_w 大小不同。表 1 中是两种不同的句型结构及其对应的关键词组的大小。其中，“n”指名词，包括人名、地名和专用名词等；“v”指动词，“np”是名词性短语，“u”是助词，“y”是语气词。

词语在句中充当的成分不同，对句子有效信息的贡献度也不同[9-10]。利用依存句法分析，将词语的句法成分权重 W_c 加入关键度计算中。表 1 两例句的依存句法分析如图 3 所示，其中“nr”指人名，“b”指区别词，“ns”指地名，“w”指标点符号，“u”指助词。在非省略句中，句子的主谓宾对句子贡献较大，但谓语动词一般不作为句子关键词，主语和

表 1 句型结构与关键词组大小

	句型结构	关键词组大小	例句	关键词组
结构 1	n+v+np+y	2	特朗普+是+现任+美国+总统+吗？	[特朗普，总统]
结构 2	n+v+n+u+n+y	3	伊万卡+是+特朗普+的+女儿+吗？	[伊万卡，特朗普，女儿]



(a)结构 1 例句依存句法分析图　　(b)结构 2 例句依存句法分析图

图 3　例句依存句法分析图

宾语多数情况下具有相同重要性。句末语气词和标点对句子内容贡献较小，暂不考虑两者的句法成分权重。结构 1 例句主干“特朗普(主语)+是(谓语)+总统(宾)”的 W_c 大小为：

$$W_c(\text{特朗普})=W_c(\text{总统})>>W_c(\text{是}) \tag{1}$$

“现任”和“美国”作定语分别从性质和范围修饰定中短语中心语“总统”，W_c 依次为：

$$W_c(\text{总统})>W_c(\text{美国})\geqslant W_c(\text{现任}) \tag{2}$$

最终结构 1 例句中各词的 W_c 大小依次为：

$$W_c(\text{特朗普})=$$
$$W_c(\text{总统})>W_c(\text{美国})\geqslant W_c(\text{现任})>>\ W_c(\text{是}) \tag{3}$$

与结构 1 例句不同，结构 2 例句中“特朗普”从领属修饰“女儿”，表示“女儿”从属于“特朗普”，则：

$$W_c(\text{特朗普})>W_c(\text{女儿}) \tag{4}$$

助词作辅助之用，赋予其较小的 W_c，结构 2 例句中各词 W_c 为：

$$W_c(\text{特朗普})>W_c(\text{伊万卡})=W_c(\text{女儿})>>W_c(\text{是})$$
$$\approx W_c(\text{的}) \tag{5}$$

考虑数据源中存在较多命名实体信息，引入命名实体权重 $W_e(0\leqslant W_e<1)$，对词语关键度做到更好的区分度。对于任意一个词 T_i，若该词是数据库中的命名实体，则为词分配一个命名实体权重 $W_e(T_i)$，反之将该词的实体权重置为零。对于是非问句关键词组 K_w 中任意一个词 T_i，其关键度 $W(T_i)$ 由 TextRank 关键词得分 $W_s(T_i)$、句法成分权重 $W_c(T_i)$ 和命名实体权重 $W_e(T_i)$ 三部分组成，具体表现为：

$$W(T_i)=\alpha\times W_s(T_i)+\beta(0.4\times W_c(T_i)+0.4\times W_e(T_i)) \tag{6}$$

经实验统计，当 $\alpha=0.4$，$\beta=0.6$ 时，核心词提取正确率最高，因此令 $\alpha=0.4$，$\beta=0.6$。$W_s(T_i)$ 的计算形式为[10]：

$$W_s(T_i)=(1-d)+d\times\sum_{T_j\in \mathrm{In}(T_i)}\frac{W_{ji}}{\sum T_k\in \mathrm{Out}(T_j)}W_s(T_j) \tag{7}$$

其中 d 为阻尼系数(一般取值 0.85)，W_{ji} 指图中任两点(任意两相连词语)T_i、T_j 相连边的权重，$\mathrm{In}(T_j)$ 为指向点 T_j 的点集合，$\mathrm{Out}(T_i)$ 为点 T_i 指向的点集合。是非问句的核心词 C_w 为关键度 $W(T_i)$ 最高的词语，即：

$$W(C_w)=\max(W(T_i)) \tag{8}$$

3 多源数据融合创建索引与答案整理

多源数据融合创建索引将 Neo4j 数据库中节点与关系的信息、本地新闻文件信息和 MySQL 数据库中存储的标准问答对信息进行文档融合，降低数据多样性带来的复杂度，减少索引创建的时间。答案整理将查询索引的结果 LR 与问句预处理输出的关键词组 K_w 和核心词 C_w 进行匹配判决，最终得到符合是非问句答语标准的二值答案。

3.1 Lucene 全文检索系统模型

全文检索指计算机索引程序通过扫描文章的每一个词，对每个词建立一个索引，指明该词在文章中出现的次数和位置。当用户进行查询时，检索程序就根据事先建立的索引进行查找，并将查找的结果反馈给用户的检索方式。Lucene 是一款高性能的 Java 全文检索工具包，它提供了完整的查询引擎、索引引擎和部分文本分析引擎[11]。

3.2 多源数据文档融合创建索引

文档(Document)是 Lucene 全文检索中索引创建和查询的基本单位，一篇文档包含不同类型的信息，分别存储在不同的域(Field)里，不同的文档可以有不同的域，每个文档有独立的编号。对多源数据进行文档融合能够降低数据多样性带来的复杂度，为索引创建提供更便利省时的方法。文档融合流程如图 4 所示，主要处理如下：

(1)使用 Cypher 图数据库查询语言对 Neo4j 进行节点查询，为每个节点创建一个 Document 对象，将查询得到的属性添加到 Document 对象中；

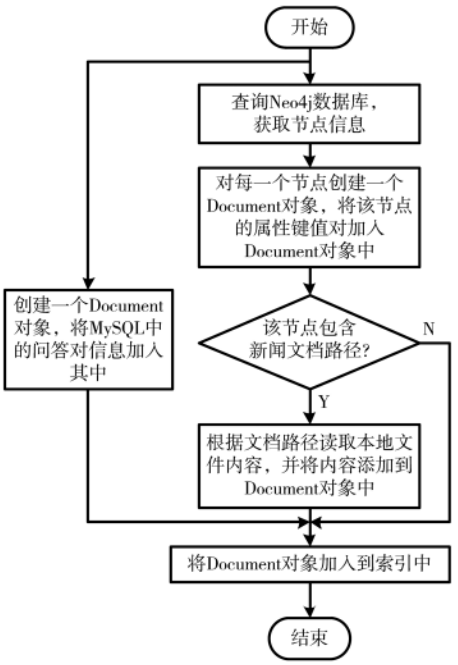


图 4 文档融合流程图

(2)若该节点包含新闻文件信息,获取节点中存储的新闻文件的路径,通过 I/O 操作从本地文件系统中读取文件内容,并将该内容添加至 Document 对象;

(3)通过 SQL 语句查询 MySQL 中的标准问答对信息,并将该信息添加到一个新的 Document 对象中。

融合后的文档对象如图 5 所示,其中 Document 1 至 Document k 为 Neo4j 数据库中实体和新闻文件的文档对象,域"fullText"为实体节点的所有属性键值数据,域"newsText"为节点相关新闻文件数据;Document $k+1$ 是 MySQL 数据库中的标准问答对信息,域中存储每一对问题-答案信息。

图 5 文档对象

Lucene 自带的中文分词器中,SmartChineseAnalyzer 对中文支持较好,使用简便[12],分词效果能满足系统需求,故采用 SmartChineseAnalyzer 分析文档。分析文档得到的语汇单元经 IndexWriter 对象加入索引库,至此多源数据融合创建索引完成。

3.3 答案整理

答案整理包括索引查询和答案判决。根据问句内容查询索引,返回检索结果 $LR(r_1, r_2, \cdots, r_m)$。LR 是命中文档的详细信息,而非是非问句的具体答案,需对检索结果进行答案判决,输出肯定或否定的答案。答案判决首先遍历 LR,判断检索记录 r_i 是否为 C_w 指向的实体,若是则进行下一步判断,否则取下一条 r_{i+1} 继续进行此步判断,直到遍历结束;判断该条记录下是否包含 $K_w[T_1, T_2, \cdots, T_n]$ 中所有词汇,且词汇顺序与 $K_w[T_1, T_2, \cdots, T_n]$ 中关键词顺序一致,若是则判定为肯定答案,否则为否定答案。

4 实验结果与分析

为验证本系统的可行性,根据上述设计方案实现了一个中文是非问答系统。实验数据来源于本地 Neo4j 数据库、MySQL 数据库和本地新闻文件。Neo4j 数据库中包括人物、组织和事件等共计 12 451 个实体信息,人物之间的亲属关系和组织与领导人关系等共 25 319 条。MySQL 数据库中包含 1 000 对是非问句及对应正确答案。本地文件中存储 149 篇 PDF 格式的实体相关新闻文件。构建前文结构 1 和结构 2 类型的问句共 1 000 条作为核心词提取和系统功能的测试数据,人工标注测试问句的核心词和问题的正确答案。为了定量分析系统效果,实验用准确率进行核心词提取和系统整体功能的分析。准确率定义如下:

$$P=(P_{\text{num}}\times 100\%)/R_{\text{num}} \tag{9}$$

式中,P 为准确率,P_{num} 为正确提取核心词的问句数量或正确得到答案的问句数量,R_{num} 表示所有测试问句数量。

(1)实验 1:核心词提取对比分析。使用1 000 条测试数据,对改进 TextRank 的核心词提取方法和传统 TextRank 提取的方法进行对比分析,对比结果表明改进 TextRank 的核心词提取方法能够较为准确地实现问句核心词的提取,提取准确率有显著提升。对比结果如表 2 所示。

表 2 核心词提取对比

	问句数	正确数	P/%
TextRank	1 000	324	32.4
改进 TextRank	1 000	987	98.7

(2)实验 2:多源数据创建索引对比分析。对本地多源数据进行文档融合,将融合创建索引、直接创建索引以及三种数据源分别创建索引耗时作统计分析,结果表明文档融合能有效减少创建索引的时间,对降低多源数据索引创建复杂度有一定作用。时间对比结果如表 3 所示。

表 3 创建索引时间对比

(ms)

	融合	直接	Neo4j 数据	新闻文件	MySQL 数据
时间	14	16	10	5	1

(3)实验 3:系统整体功能测试与分析。针对1 000 条测试数据,结合核心词提取准确率和系统输出准确率对系统整体功能进行评估,测试结果如表4 所示。表 4 显示,在 1 000 条测试数据上,系统正确回答数为 954,正确率为 95.4%。其中,当核心词提取正确时,系统回答正确率为 95.6%,而核心词提取出错时,问答的准确率仅有 76.9%,核心词提取的正确与否对系统回答的准确性有较大影响。

表 4 系统测试结果

	句子数	正确数	P/%
系统测试	1 000	954	95.4
核心词测试	1 000	987	98.7
核心词正确	987	944	95.6
核心词错误	13	10	76.9

5 结论

本文设计并实现了基于 Lucene 的中文是非问答系统,基于 TextRank 算法引入句法成分权重和命名实体权重实现是非问句核心词提取,捕获是非问句的关键要素。对分类存储的多源数据进行文档融合,降低数据多样性带来的复杂度。测试证明,改进 TextRank 算法能有效提取问句核心词,多源数据融合创建索引比直接创建耗时更少,且系统在回答是非问题时不会出现答非所问的情况,回答结果较为满意。

参考文献

[1] ALLAM A M N, HAGGAG M H. The question answering systems: a survey[J]. International Journal of Research and Reviews in Information Sciences, 2012, 2(3): 211-221.

[2] LIU L, ZENG Q T. An overview of automatic question and answering system[J]. Journal of Shandong University of Science & Technology, 2007(4): 79-82.

[3] 饶林尚,吴怡,冯前进.面向医疗设备的深度问答系统的设计[J].计算机应用与软件,2019,36(6):171-176.

[4] LU G M, XIA Y L, WANG J M, et al. Research on text classification based on TextRank[C]. International Conference on Communications, Information Management and Network Security, Shanghai, China, 2016.

[5] ZHANG H, LIU Y. Improved text rank algorithm research based on topic weighting[C]. 2015 International Conference on Industrial Informatics, Machinery and Materials, Chiengmai, Thailand, 2015.

[6] 刘朝涛.中文问答系统中的句型理论及其应用研究[D].重庆:重庆大学,2010.

[7] 李庆虎,陈玉健,孙家广.一种中文分词词典新机制——双字哈希机制[J].中文信息学报,2003,17(4):14-19.

[8] HE H. HanLP: Han language processing[EB/OL]. [2020-05-01]. https://hanlp.hankcs.com/.

[9] 薛征,廖闻剑.基于位置权重和实体识别的关键词提取[C].中国电子学会第十六届信息论学术年会,中国,北京,2009.

[10] 徐立.基于加权 TextRank 的文本关键词提取方法[J].计算机科学,2019,46(S1):142-145.

[11] 赵肄江,李锋.基于 Lucene 的在线答疑系统的构建[J].当代教育理论与实践,2018,10(1):66-69.

[12] 董杨.基于 Lucene 的全文检索技术研究与应用[D].西安:西安理工大学,2017.

(收稿日期:2020-05-05)

作者简介:

罗东霞(1995-),女,硕士研究生,主要研究方向:智能系统与设计。

卿粼波(1980-),男,博士,副教授,主要研究方向:图像处理、图像/视频编码通信、嵌入式系统。

吴晓红(1970-),女,博士,副教授,主要研究方向:智能系统与设计、图像处理与模式识别。

版权声明