题 目：基于关系抽取的答案自动提取系统的设计及实现

学院：******** 专业：******** 学生姓名：**** 学号：**********

项目概述：

1. 选题来源与背景

随着当前互联网的普及，为人们提供了丰富的信息资源，我们能通过搜索引擎快速便捷地获取自己想要的各种信息。传统的搜索引擎是基于文本关键词的搜索，当用户输入问题后，搜索引擎会返回它查询到的与问题关键词相关的所有网页。这种搜索方式的缺点是比较机械化，并没有深入地挖掘用户查询的真正地搜索需求，所使用的关键词的逻辑组合也并不一定能够表述用户真正的检索请求，因此使用传统搜索引擎返回的网页或文档往往比较多，用户很难快速找到自己需要的信息。

与基于文本关键字的传统搜索引擎并不同，问答系统让用户以自然语言的形式提问，并返回准确简短的答案，而不是将大量的相关文本或网页返回给用户。为了满足用户真正的检索需求，方便用户快速地找到所需的答案，国内外很多公司在进行精准自动问答方面的尝试，如微软小冰、百度度秘等，它们综合运用了信息检索、自然语言处理、语义分析、机器学习等技术，系统将会自动理解和分析用户的提问，直接返回用户想要的答案。比如：当用户提问“刘德华的妻子是谁？”，问答系统可以直接给出答案“朱丽倩”，而不是大量的网页文本。而问答系统的答案提取是自动问答系统所需进行的处理中最关键的一步。

问答系统一般主要包括 3 个模块：问题分析、信息检索和答案提取。问题分析模块主要是分析用户提出的问题，理解用户真正的搜索请求，一般主要包括问题领域划分、问题中的关键词提取和关键词的扩展，通过对问题的归一化（多表达归一，最短关键词描述，简化问题）、形式化（转化为结构化查询语言，让机器可以理解和计算）和改写技术（根据上下文语境改写句子，输出替换片段，更好的满足检索要求和扩大相关资源召回），全面理解用户意图；问题分析后将得到的关键词查询集合提交到信息检索模块，通过检索获取与问题相关的网页及文本；答案提取模块是从已检索到的网页及文本中找出相关的简洁答案(一句话或者是一个简单的实体)，并返回给用户。

在问答系统的答案提取方面，目前主要有基于句子相似度的方法、基于潜在语义的方法和基于模式匹配的方法。基于句子相似度的方法是通过计算句子与问题相似度来计算句子权重，从而得到候选答案集合；基于潜在语义的方法是通过统计方法提取并量化同义词和多义词的潜在语义结构，从而提供准确性；基于模式匹配的方法是分析同类问题的答案形式，人工或通过统计来定义不同领域问题的答案句子模式，再通过模式匹配的方式从海量文本中提取新的实体关系对作为答案。问答系统答案提取的准确性对问答系统本身的准确性起着关键作用，本文使用关系抽取的方法解决答案为简单实体类型的领域问题，对于客观答案（或者答案很长，需要整理信息）使用摘要算法提取简要答案提供给用户。

2. 项目与实习的关系

项目来源于本人在百度实习的知识图谱自动问答项目，该项目在百度已有的知识图谱技术的基础上，通过网页库数据挖掘、Indri 检索、nlp 自然语言处理，自动抽取新的知识。在实习期间，本人主要负责 nlp 自然语言处理和知识抽取部分，详细主要包括: 自然文本处理(分词、词性标注、句子依存分析、命名实体识别)、句子特征提取、训练关系分类器、问答系统答案提取以及自动文本摘要。

3. 项目开发意义、必要性与现实价值

答案提取模块是自动问答系统的关键步骤之一，现有解决方案主要有基于句子相似度的方法、基于潜在语义的方法和基于模式匹配的方法。基于句子相似度的方法是通过计算句子与问题相似度来计算句子权重，从而得到候选答案集合；基于潜在语义的方法是通过统计方法提取并量化同义词和多义词的潜在语义结构，从而提供准确性；基于模式匹配的方法是分析同类问题的答案形式，人工或通过统计来定义不同领域问题的答案句子模式，再通过模式匹配的方式从海量文本中提取新的实体关系对作为答案。

- 基于句子相似度的答案提取

基于句子相似度的答案提取是通过计算用户的问题和网页文档中的各个句子的相似度，排序后选取相似度最大的句子作为答案。而两个句子的相似度计算可以使用关键词匹配方法，即通过计算目标问题和文档句子中拥有相同的关键词总数来确定二者的相似程度，使用这样的方法需要考虑共指消解、关系推断等问题。主要包括两个方面，一是实体重名，如“李晨的妻子是谁”，搜索结果为“李晨的老婆是他高中的女友叫赵琼，两人在一起十年了”，但此处李晨指的是主持人李晨而不是演员李晨，再比如“我刚买了两袋苹果”与“我刚买了一台苹果”，两个句子中的“苹果”表述的是不同意义，通过关键词匹配很难处理词的多义性问题；二是人们为了避免重复，习惯用代词、缩略语、简称来指代一些实体全称，如“ICBC 总部在哪里”，使用关键词匹配过程中可能会丢失与“ICBC”同义的“中国工商银行”或者“工商银行”等等。

- 基于潜在语义分析的答案提取

基于潜在语义分析的答案提取的基本理论是文档中的词与词之间存在一些关系，通过构建同义词和多义词的词汇表来解决使用词匹配方式不可避免的共指消解问题，从而提高答案提取时的准确率。然而这种方式并没有分析句子真正的语义，如搜索“王菲的丈夫是谁”，有搜索结果有“王菲和李亚鹏离婚后，谁将成其第三任丈夫？”，此处表述含义为“李亚鹏”已经离婚，但单纯使用关键词或者同义词、多义词的词汇表并不能准确理解这一点，有可能得到错解。

- 基于模式匹配的答案提取

基于模式匹配的答案提取是先通过人工处理得到包含答案的句子，并标注问题分类及答案，形成不同问题分类的问答训练语料。通过人工或统计的方法学习，提取候选答案句子的答案模式，再计算候选答案句子模式置信度和权重，并根据置信度和权重获得相应问题分类的答案句子模式。但是这种方式有一些明显问题，一方面这种方法对不同领域、不同实体属性或关系很难泛化，当需要移植到其他领域时，需要进行

大量的人工标注工作，另一方面由于自然语言的多样性，一个意思可以使用不同的说法进行表达，在不同语境中，对同一实体属性或关系表述差异很大，人工标注过程很难覆盖全面，在实际使用过程中召回率不高。

由于现行搜索方式比较机械化，不一定能够表述用户真正的检索请求，使用传统搜索引擎返回的网页或文档往往比较多，用户很难快速找到自己需要的信息。自动问答能够满足用户真正的检索需求，方便用户快速地找到所需的答案，其必要性和价值也不言而喻。

4. 本设计的主要内容

本文主要针对答案是命名实体的领域，使用构建知识图谱三元组的思想和关系抽取的方法提取答案，另对答案属于长答案的领域问题，本文通过提取相关文档的摘要来作为最终答案。本文主要工作包括：自然文本处理、句子特征提取、训练关系分类器、问答系统答案提取以及自动文本摘要。自然文本处理主要包括分词、词性标注、句子依存分析、命名实体识别，主要给出各个方面的现有解决方案，以及针对问答系统的改进方法。在提取句子特征时面临的主要问题是如何合理地表示文本，使之包括足够的信息能够反映文本的主要特征信息，又不至于过于复杂难以实现，方便在分类器训练及答案提取阶段能有效使用。本文句子特征提取包括词法特征、句法特征和整句特征，主要给出本文系统使用的句子特征及其提取方法。训练关系分类器并提取答案主要介绍了不同机器学习算法和其效果以及答案提取的方法和改进思路。自动文本摘要主要解决部分领域问题答案属于长句子，因此需要从可能包含正确答案的文本中提取有效摘要。

5. 重难点和特色

- 语料获取。由于从 Web 中获取的网页文本大多是半结构化或非结构化的，如果正文提取效果不好，不能准确获取真正有语义的句子，对句子特征提取的准确度影响很大，自然也会影响到后面的答案提取。本文处理的数据基于检索后得到的包含命名实体及其属性或关系的句子或段落，此部分语料的获取是答案提取的基础。由于论文主要目的是解决问答系统答案提取部分，因此对于 Web 数据挖掘获取语料部分只做简要的介绍。
- 中文自然文本处理。从 Web 获取的网页文本中提取到可能包含需要的实体对及其属性或关系的语料集合后，需要对得到的语料进行文本处理，主要包括分词、词性标注、句子依存分析、命名实体识别，其处理结果是本文提取句子词法特征、句法特征和整句特征的前提。
- 机器学习模型选择。问答系统覆盖的领域广，单模型很难覆盖全面，而且往往训练集数据量比较大，特征维度比较大，事实上在处理大数据集时常存在内存占用的问题，本文对不同领域的每种属性或关系建立二分类模型。使用第二章介绍的方法提取文本特征后，将文本特征转化为特征向量，使用不同机器学习算法进行训练和测试，本文使用大量包含指定 SPO 和大量包含指定 SP 但不包含 O 的数据作为训练集，和大量包含 SP 的数据作为测试集。为了提高最终的准确率，本文测试了各种不同的算法，并对结果进行交叉验证。
- 答案提取。答案提取主要包括两个方面，一是对于答案只是命名实体的答案提取，二是对于答案需要总结信息的答案提取。对于命名实体类型，通过第

三章的方法，可得到大量可能表述某实体关系的句子，再通过命名实体进行答案提取；对于长句子类型的答案，需要从可能的答案文档抽取其中的文摘句作为最终答案。

设计方案:

1. 理论基础

1) 关系抽取

知识图谱将杂乱的网页数据构建成一个个结构化的实体，能够为用户提供更加有条理的实体及其属性或关系信息，顺着知识图谱甚至可以探索到更深入、完整和广泛的知识。要实现自动问答，搜索引擎不仅需要理解查询的问题中涉及到的命名实体及其属性或关系，还需要理解查询语句的语义信息。搜索引擎可以通过使用高效的图搜索，在知识图谱中查找与这些实体及属性或关系连接的子图，图搜索结果被进一步提交给图数据库并返回相应的答案给用户。本文通过关系抽取的方法来构建知识图谱。

关系抽取研究在 MUC 评测会议和 ACE 评测会议的引导和推动下，许多先进的信息抽取技术被提出来，并在会议提供的平台上测试。总的来说，这些方法主要分为两类：基于模式匹配的方法和基于机器学习的方法。基于模式匹配的方法需要融合各个领域知识和语言学的知识，通过人工或使用统计编写不同领域的规则集合，构造出特定句子模式，利用模式匹配的方式找到新的关系实例。基于机器学习的方法将关系抽取的属性或关系识别问题转化为分类问题，通过选取有句子代表性的特征，利用不同的机器学习算法训练出不同领域分类器，最终通过训练出的分类器识别实体对之间的属性或关系。

总体来说，基于模式匹配的关系抽取方法同使用编写规则集合，再进行匹配来提取答案的方法一样，存在泛化难度大，在实际使用过程中容易因为覆盖范围不够，导致召回率降低的问题。而基于机器学习的方法主要的解决方案是使用半监督或无监督技术。目前的代表性技术是 Bootstrapping 技术和 Distant Supervision 技术。Bootstrapping 技术从少量的种子实例出发自动抽取新的实例，而 Distant Supervision 技术则充分利用现有的大规模知识库（如 Freebase，谷歌或百度的知识图谱等等），使用非直接监督实例来构建大规模信息抽取系统。

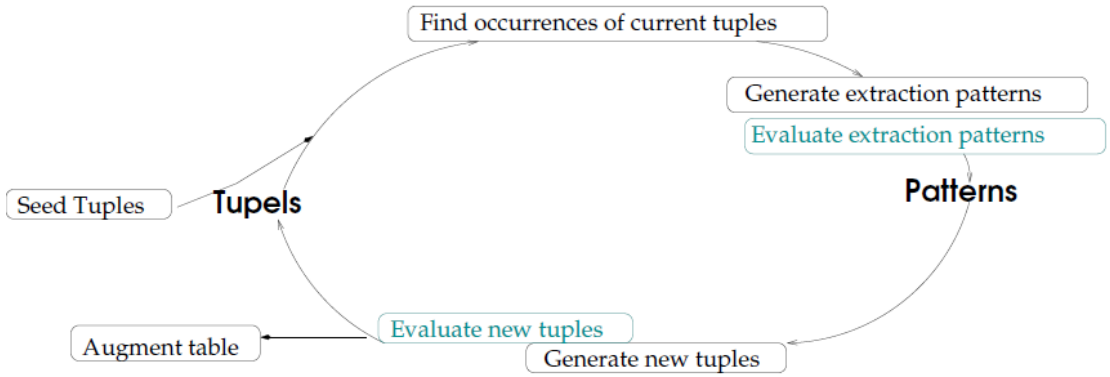
- Bootstrapping 技术

Bootstrapping 技术首先人工选取少量的关系实例作为种子集合，然后利用模式匹配或者训练模型分类器的方法，通过多次迭代，不断获取新的关系实例并添加到此关系实例集合，最终得到足够的关系实例 [4]。

其主要思路（如图 1-1 所示）是：

- 事先定义关系训练集
- 训练关系分类器
- 通过关系分类器去挖掘新的实体关系
- 人工标注或直接加入到原先训练集中
- 重复 2~4

图 1-1



以“人物”和“地点”两个实体抽取“出生”关系为例，首先构建少量<PEOPLE，LOCATION>的关系实例，如<张国荣，香港>等，在已有的语料库中得到与关键词“张国荣”和“香港”有关并表述关系“出生”的句子，通过提取句子模式使用聚类算法得到不同关系的模式，再从互联网或其他语料库中得到关于“出生”关系的句子，使用“出生”这一关系的模式提取新的<PEOPLE，LOCATION>的实体二元组，再将其加入到原来的关系实例集合中，重复迭代提取新的关系模式和二元组。

- Distant Supervision 技术

距离监督方法基本假设为如果两个实体有一个关系，那么包含这两个实体的一句话就可能表达这个关系。利用已有知识库蕴含的事实信息作为基础，训练得到实体关系的训练集，训练分类器，在未标注的其他大规模的语料上，利用分类器从自由文本中挖掘新的实体关系，从而补充现有的知识库[5]。

其主要思路是：

- 将句子关系作为类别，组成一个分类训练样本。如果两个句子表达的 <关系，实体 1，实体 2> 完全一致，则抽取两个句子特征，并将它们合并在一起，组成一个更大的特征向量。
- 训练不同关系的逻辑回归分类器
- 在测试阶段，先对句子中的命名实体进行标注，抽取其中的命名实体对和特征。如果多个句子的命名实体对一样，则将它们的特征合并在同一个特征向量中。然后利用逻辑回归分类器，对关系名称进行识别。

仍以“人物”和“地点”两个实体抽取“出生”关系为例，首先需要基于百度知识图谱或者谷歌知识图谱得到大量<PEOPLE，LOCATION>的关系实例，如<张国荣，香港>等，在已有的语料库中得到与关键词“张国荣”和“香港”有关的句子，通过提取句子特征使用逻辑回归分类其训练不同关系的分类器，再从互联网或其他语料库中得到关于“出生”关系的句子，使用“出生”这一关系的模式提取新的<PEOPLE，LOCATION>的实体二元组，再将其加入到原来的关系实例集合中，重复迭代提取新的关系模式和二元组。

2) 自动文本摘要

自动文本摘要主要分为单文档文摘和多文档文摘，现今单文档文摘主要方法包

括：基于文本特征方法、基于词汇链的方法和基于图排序方法[9]。其根本原理是将文本视为句子的线性序列，将句子视为词的线性序列，通过计算词语权重来计算句子权重，再对按照权重对句子进行排序，选择权重高的句子作为文摘句。对多个类似主题的文档提取文摘，可以针对多文挡集合利用单文档文摘技术来生成多文档文摘，将多文档集当作一个文本，根据位置、词频、标题、段首等信息进行文本单元的抽取。自动文摘技术在新一代搜索引擎问答系统(Q&A)中，将与用户相关的长文档形成简短的文摘交给用户，帮助用户在较少的时间内获得较多的信息，从而大大提高用户获取信息的效率。

- 基于文本特征方法

主要的特征包括词频、句子位置、文档标题等。频繁出现的单词与文章主题有比较大的关联,因此可以根据各单词出现的频率给文中的句子打分,以得分最高的几个句子组成文章的摘要。通过从句子位置特征入手，通过计算文章中段落首末句出现主题句的概率，选取得分最高的若干句子生成摘要。

- 基于词汇链的方法

在文章中描述某个主题的文本块内，使用的词语应该是相关的，这些相关词语构成一条词汇链。所以，词汇链可以视作一个语言片段的标志性主题词语链，不同的词汇链对应了不同的语言片段。其计算句子权重的主要思路是先选取集合中的一个词语，形成的一个词汇链的第一个词语，计算该词语与集合中其他词语的相似度，当相似度大于一定阈值的时候，认为两个词语是在同一个词汇链中；否则，另一个词语作为新一条词汇链的第一个词语；重复上述过程，直到所有的词语都在一个词汇链中为止。再以每个词汇的权重、位置等信息为依据计算每个词汇链的权重，最后对词汇链进行排序，选择大于某个特定阈值的词汇链，包含词汇链上词语的语句作为备选语句集合。

- 基于图排序方法

把文章分解为若干单元(句子或段落等)，每个单元对应一个图的顶点，单元间的关系作为边，最后通过图排序的算法(如 PageRank、manifold ranking 等)得出各顶点的得分，并在此基础上生成文本摘要。

2. 解决问题的方法与步骤

设计与实现了问答系统答案提取部分，系统能够从网页库中提取相关 S 和 P 的数据，并从中抽取答案 O；对于需进行信息整合的数据，系统还实现了单文档自动文本摘要提取。

对于命名实体类型的答案，系统主要分为如下几个模块：

1) 对于某个领域 P，根据已有的 SPO 样本，从网页库中采集相关语料；
2) 文本预处理，进行分词、词性标注、句子依存分析、命名实体识别，提取句子特征；
3) 训练领域 P 的分类器；
4) 使用待抽取的 SP 数据，从网页库中采集相关语料；
5) 使用领域 P 的分类器对语料分类，取得待抽取 SPO 文本；

6） 从待抽取 SPO 文本抽取可能的答案 O；

7） 对可能的答案进行置信度校验，得到最终的答案 O。

对于需进行摘要提取的答案，系统主要分为如下几个模块：

1） 根据待抽取的 SP 数据，从互联网中采集相关的数据；

2） 文本预处理，分段、分句、分词并去除停用词；

3） 计算句子权重，并提取文摘句；

4） 对文摘句进行冗余和润色处理。

本系统的整体架构如图 5-1 所示：

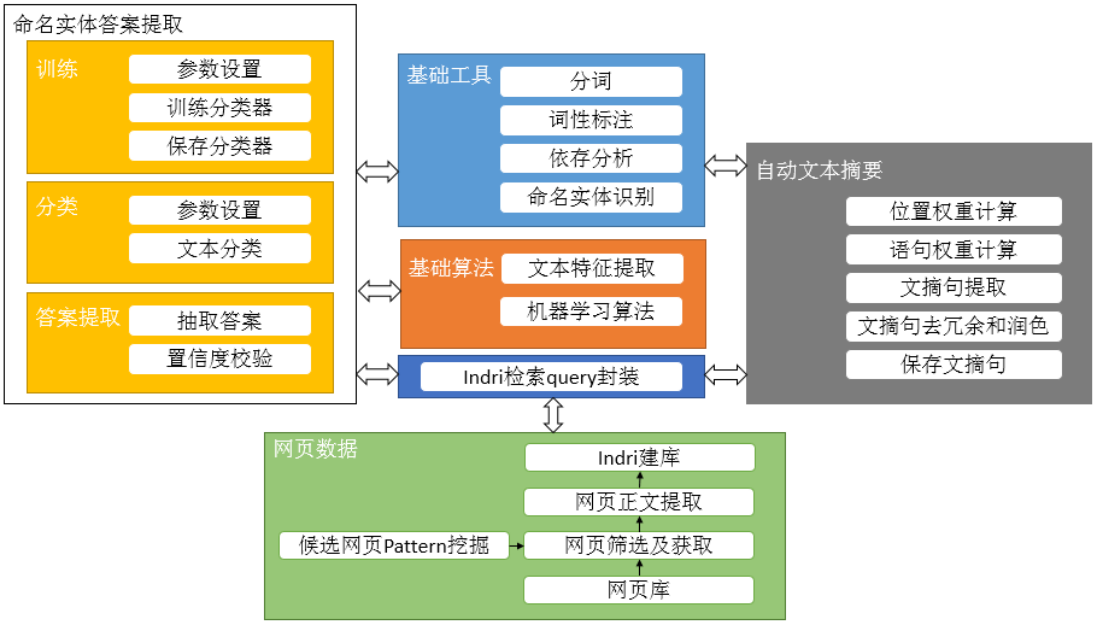


图 5-1

3. 预期成果（效果）

- 文本分类召回率达到 95%，准确率达到 85%
- 命名实体答案提取准确率达到 90%，找回率达到 80%
- 自动文本摘要提取摘要结果经过众包得分 80%

主要参考文献：

1. 郑实福.刘挺.秦兵.李生自动问答综述[期刊论文]-中文信息学报 2002(6)
2. 余正涛.樊孝忠.宋丽哲.高盛祥.YU Zhengtao.FAN Xiaozhong.SONG Lizhe.GAO Shengxiang 汉语问答系统答案提取方法研究[期刊论文]-计算机工程 2006(3)
3. 余正涛;樊孝忠;郭剑毅;基于潜在语义分析的汉语问答系统答案提取[期刊论文]-计算机学报 2006(10)
4. 黄勋，游宏梁，于洋. 关系抽取技术研究综述. 现代图书情报技术，2013，29(11): 30-39
5. Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In: Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Int'l Joint Conf. on Natural Language Processing of the AFNLP. Morristown: Association for Computational Linguistics, 2009. 1003−1011.
6. 张奇 信息抽取中实体关系识别研究[学位论文]博士 2010
7. Etzioni O, Cafarella M, Downey D, Kok S, Popescu A, Shaked T, Soderland S, Weld D and Yates A. Unsupervised Named-entity Extraction from the Web: An Experimental Study [J]. Artificial Intelligence, 2005, 165(1): pp91-134.
8. 马渊 短文本情感分析技术研究[学位论文]硕士 2011
9. 胡侠.林晔.王灿.林立 自动文本摘要技术综述[期刊论文]-情报杂志 2010(8)
10. Gunes E;Radev D R LexRank:Graph-based Centrality as Saliencein Text Summarization 2004
11. H. Zhang (2004). The optimality of Naive Bayes. Proc. FLAIRS.

毕业设计（论文）进度安排：

序号	毕业设计（论文）各阶段内容	时间安排	备注
1	确定选题，问答系统调用、关系抽取调研	2015.7-2016.3	
2	课题申报	2016.3.1-3.6	
3	上传任务书	2016.3.11-2016.3.14	
4	提交开题报告	2016.3.18	
5	开题答辩	2016.3.18	
6	自动文本摘要	2016.4.15	

7	系统设计	2016.5.10	
8	句子特征提取部分	2016.5.11-5.15	
9	文本分类	2016.5.16-5.18	
10	答案提取	2016.5.19-5.25	
11	准确率验证	2016.5.26-6.2	

指导教师意见：

指导教师（审核签名）： _____ **审核日期：** _____年____月____日