ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2012/06(03)-0193-15
DOI: 10.3778/j.issn.1673-9418.2012.03.001

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-51616056

问答系统研究综述*

毛先领+, 李晓明
北京大学 信息科学技术学院, 北京 100871

A Survey on Question and Answering Systems*

MAO Xianling+, LI Xiaoming

School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
+ Corresponding author: E-mail: mxl@net.pku.edu.cn

MAO Xianling, LI Xiaoming. A survey on question and answering systems. Journal of Frontiers of Computer Science and Technology, 2012, 6(3): 193–207.

Abstract: Recently, question and answering systems have attracted lots of attention. Given a question, the goal of question and answering is to return a concise, exact answer. According to the format of data, question and answering can be divided into three categories: the structural data based question and answering, the free-text based question and answering, the question-answer pairs based question and answering. This paper describes and summarizes the characteristics and related researches of these three categories respectively. Then, it discusses the future work of question and answering.

Key words: question and answering (QA); traditional question and answering (TQA); community-based question and answering (CQA); information retrieval; answer extraction

摘 要：近年来, 问答系统被大量广泛地研究。问答系统的目标是给定一个问题, 能够得到简短、精确的答案。根据处理数据的不同, 将问答系统分为三类: 基于结构化数据的问答系统、基于自由文本的问答系统、基于问题答案对的问答系统。对这三大类系统的特点、面临的问题和相关的研究分别进行了叙述和总结。最后, 讨论了问答系统未来可能的研究方向。

关键词：问答系统(QA); 传统问答系统(TQA); 基于社区的问答系统(CQA); 信息检索; 答案抽取

---

*The National Natural Science Foundation of China under Grant No. 60933004 (国家自然科学基金).
Received 2011-04, Accepted 2011-12.

文献标识码：A 中图分类号：TP393

1 简介

随着互联网的普及，互联网上的信息越来越丰富，人们能够通过搜索引擎方便地得到自己想要的各种信息。但是搜索引擎存在很多不足，主要有两个方面：一是返回结果太多，导致用户很难快速准确地定位到所需的信息；二是搜索引擎的技术基础，即关键字匹配，只关注语言的语法形式，没有涉及语义，同时用户采用简单的查询词很难准确地表达信息需求，使得检索效果一般。

满足信息需求的方式除了搜索引擎外，还有另外一种服务方式——问答。与搜索引擎系统不同，问答系统(question and answering，QA)不仅能用自然语言句子提问，还能为用户直接返回所需的答案，而不是相关的网页。显然，问答系统能更好地表达用户的信息需求，同时也能更有效地满足用户的信息需求。

1.1 问答系统的定义

关于问答系统的内涵和外延，很多的研究者都给出了各自的定义。例如 Molla 等人[1]在 2007 年把问答系统定义为一个能回答任意自然语言形式问题的自动机。虽然定义很多，并且各种定义之间略有不同，但是一般都认为问答系统的输入应该是自然语言形式的问题，输出应该是一个简洁的答案或者可能答案的列表，而不是一堆相关的文档。例如用户向问答系统提交一个问题，“电话是什么时候发明的?”,系统应该返回一个精简的答案，“1876”。

1.2 问答系统的一般处理流程

给定一个问题，问答系统的处理流程一般如下：首先分析问题，得到问题的句子成分信息、所属类别和潜在答案类型等信息；然后根据问题分析得到的信息在数据集中得到可能含有答案的数据，这缩小了进一步精确分析的范围；在得到的小范围数据中采用各种技术提取答案或者答案集合；最后将答案返回给用户。

1.3 问答系统研究的基本问题

对应问答系统的处理流程，问答系统一般包括三个主要组成部分：问题分析、信息检索和答案抽取。这表明问答系统研究包含三个基本问题：如何去分析问题；如何根据问题的分析结果去缩小答案可能存在的范围；如何从可能存在答案的信息块中抽取答案。

在问答系统的不同发展阶段，对于这三个基本问题的解决方法随着数据类型的变化在不断变化，从而形成了不同类型的问答系统。

1.4 问答系统的复杂性

一般情况下，问答系统的复杂程度可以从三个维度来衡量：问题、数据、答案。对于问题维度，问答系统可以分为限定领域(指系统能接受的问题只能是关于某个特定的主题)的问答系统和开放领域(指系统能接受的问题可以是任意主题的问题，没有任何限制)的问答系统；对于数据维度，问答系统可以分为处理结构数据(或半结构数据)的问答系统(例如关系数据)和处理无结构数据(例如文本)的问答系统；对于答案维度，问答系统可以分为抽取式(所谓抽取，是指答案是从数据或者文本中抽取出来的，例如文本片段)问答系统和产生式(所谓产生，是指答案是通过一定的规则或者内在的编码生成出来的，例如对话)问答系统。所以问答系统根据这三个维度各自采取的形式拥有不同的复杂性。一般地，开放领域的问答系统比限定领域的问答系统复杂，处理无结构数据的问答系统比处理结构数据的问答系统复杂，同时抽取式的问答系统比产生式的问答系统复杂。因此衡量和分析问答系统的复杂性，可以从问题、数据、答案三个维度来评价；同时问答系统根据问题、数据、答案三个维度的不同而属于不同类别。

2 问答系统的发展历程

图灵测试可能是对问答系统最早的构想。本文

依据问答系统处理的数据格式，将问答系统的发展历史划分为三个阶段：基于结构化数据的发展阶段、基于自由文本数据的发展阶段、基于问题答案对(question-answer pairs)数据的发展阶段。其中基于结构化数据的发展阶段又可以划分为人工智能(artificial intelligence, AI)阶段和计算语言学阶段两个子阶段。其大概的时间划分、特点和代表系统分别叙述如下：

在 20 世纪 60 年代，由于人工智能的发展，研究人员试图建立一种能回答人们提问的智能系统。这个时期主要是限定领域、处理结构数据的问答系统，被称为 AI 时期，主要都是 AI 系统和专家系统，代表系统有 BASEBALL[2]和 LUNAR[3]。

在 20 世纪 70 年代和 80 年代，由于计算语言学的兴起，大量研究集中在如何利用计算语言学技术去降低构建问答系统的成本和难度。这个时期被称为计算语言学时期，主要集中在限定领域和处理结构数据，代表系统是 Unix Consultant[4]。

进入 20 世纪 90 年代，问答系统进入开放领域、基于文本的新时期。由于互联网的飞速发展，产生了大量的电子文档，这为问答系统进入开放领域、基于文本的时期提供了客观条件。特别是在1999 年 TREC(text retrieval conference)的 QA track 设立以来，极大地推动了问答系统的发展。

随后，网络上出现了常问问题(frequent asked questions, FAQ)数据，特别是在 2005 年末以来大量的社区问答(community based question answering, CQA)数据(例如 Yahoo！Answer)出现在网络上，即有了大量的问题答案对数据，问答系统进入了开放领域、基于问题答案对时期。

由于各个阶段处理的数据格式和形式不同，导致各个阶段解决问答系统的三个基本问题的方法和技术各不相同。本文将分别对这三个阶段各自的问题进行叙述 [1)]。

3 基于结构化数据的问答系统

基于结构化数据的问答系统的主要思想是通过分析问题，把问题转化为一个查询(query)，然后在结构化数据中进行查询，返回的查询结果即为问题的答案。从其基本思想可知，这种方法一般只能用在限定领域。

主要数据处理流程如下：

(1) 根据问题特点来分析问题，产生一个结构数据的查询语言格式的查询(对应于问答系统的问题分析部分)。

(2) 将产生的查询提交给管理结构数据的系统(如数据库等)，系统根据查询的限制条件筛选数据(对应于问答系统的信息检索部分，即缩小答案可能存在的范围)。

(3) 把匹配的数据作为答案返回给用户(对应于问答系统的答案抽取部分，由于数据库查询的精确匹配特性，“抽取”的动作不明显)。

系统结构如图 1 所示。

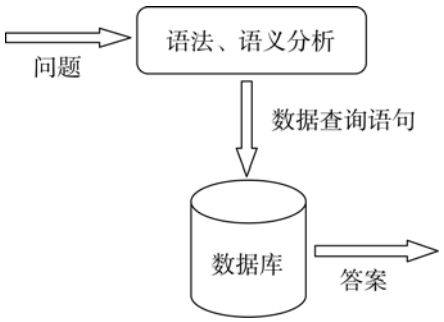

Fig.1 The architecture of QA system based on structural data

图 **1** 基于结构化数据的问答系统的体系结构

构建一个基于结构数据的问答系统有两个关键问题：一是需要构建一个特定领域的比较完备的结构数据库；二是准确、高效地把问题转化为查询语言形式的查询。

基于结构数据的问答系统有两个相关研究大量

1) 这种阶段划分旨在强调方法的时代特征，并不意味着这种方法已经过时或者消亡，很多旧有方法在现在仍然被大量使用，只是这种技术在目前没有大突破，例如基于结构化数据的技术在今天仍然在医疗等限定领域内被大量使用。

涌现的时期：一是人工智能时期；二是计算语言学时期。下面分别对这两个时期进行叙述。

3.1 人工智能时期

在 20 世纪 60 年代左右，由于人工智能领域的蓬勃发展，研究者试图通过问答系统来体现人工智能，导致很多的问答系统诞生，最著名的问答系统是 BASEBALL 和 LUNAR。BASEBALL 系统能回答关于美国篮球联赛一个赛季的相关问题[2]；LUNAR 系统能回答关于阿波罗月球探测任务取回的岩石样本分析结果的相关问题[5]。两个系统在各自特定的领域取得了巨大的成功。特别是，LUNAR 系统曾在1967 年的一个月球科学会议上被展示，在没有提前告知可以允许的短语或者不允许的短语的情况下，它能够回答地理学家所提出的问题中的 90%[6]。LUNAR 和 BASEBALL 只是这个时期问答系统中最突出的两个系统，还有大量的类似系统。Simmons 在一篇综述[7]中描述了早期各种这类形式的问答系统。大多数的这类早期系统(包括 LUNAR 和 BASEBALL)都只是聚焦在非常局限的领域内的“玩具系统”(toy system)。这些早期系统仅仅采用少量的数据集合作为内在的信息源，同时依赖人去消除数据集合中的句子歧义，或者转换这些句子为简单的英语句子。

这个时期的问答系统强调句法解析和领域知识，而且系统构建花费巨大、很脆弱(特别是在领域知识边界的地方)，因此为了降低构建系统的代价，在 20 世纪 80 年代慢慢转为计算语言学时期。

3.2 计算语言学时期

在 20 世纪 70 年代和 80 年代期间，计算语言学理论得到了大力研究和开发。计算语言学研究的蓬勃发展推动了问答系统研究向着比早期系统所能处理领域更复杂的领域发展。这类研究的主要目标是用问答系统作为一个应用框架，在这个框架内，各种自然语言处理的理论能被测试；同时，通过自然语言技术的应用，极大地降低了从电子文本构建数据库的人力代价和成本。这个时期有大量的项目，其中 Berkeley Unix Consultant[8]是最著名的。

Berkeley Unix Consultant 项目(UC)是为 Unix 操作系统开发的一个帮助系统，这个帮助系统集成了计划(planning)、自然语言处理(natural language processing，NLP)和知识表示(knowledge representation)等方面的研究。首先，分析用户的问题，将问题对应的含义通过知识表示形式进行编码；然后，UC 通过用户模型分析和目标分析来猜测用户的实际信息需求；最后，答案被裁减成适合用户的专家特性和目标形式。UC 系统提供的对话实例是令人印象深刻的，然而没有来自现实生活中的实际对话被展示出来，因此很难说在 UC 系统开发中应用到的方法和理论在实际应用中是足够鲁棒的(robust)。下面是来自 UC 的样例：

% UC
Welcome to UC (Unix Consultant) version 3.23
To a UC ' # ' prompt, please type in your questions about the Unix file system in English.
To leave, just type a 'AD' or '(exit)'.
Hi. How can I help you?
How can I delete a file?
Use rm.
For example, to delete the file named foo, type 'rm foo'.
What does rwho do?
Rwho is used to list all users on the network, the users’ tty, the users’ login time, and the users’ idle time.

3.3 小结

20 世纪 90 年代末以前的问答系统大都属于限定领域、强调领域知识、构建花费巨大、很脆弱的系统。这类系统对于能回答的问题的准确度比较高，但是对于不属于这个领域的问题就无能为力。同时构建代价非常大，即使利用计算语言学技术来帮助构建系统，这个代价仍然是非常大的。在 20 世纪 90年代，由于互联网的快速发展，产生了大量的电子数据，问答系统从人工智能角度转为从信息检索角度看待问答系统。特别是 TREC-8(1999)中加入 QA track 之后，进一步促进了这类新的问答系统研究范式的发展，即问答系统研究进入了开放域研究的时代。

4 基于自由文本的问答系统

基于自由文本(free-text based)的问答系统属于开放域问答系统，它只能回答那些答案存在于这个文档集合中的问题。信息检索评测组织 TREC 自1999 年开始每年都设立 QA track 的评测任务，同时其他评测组织如 NTCIR 和 CLEF 也设置有问答系统评测的任务，这些评测任务极大地推动了这类问答系统的相关研究[9−10]。

基于自由文本的问答系统的体系结构如图2 所示。

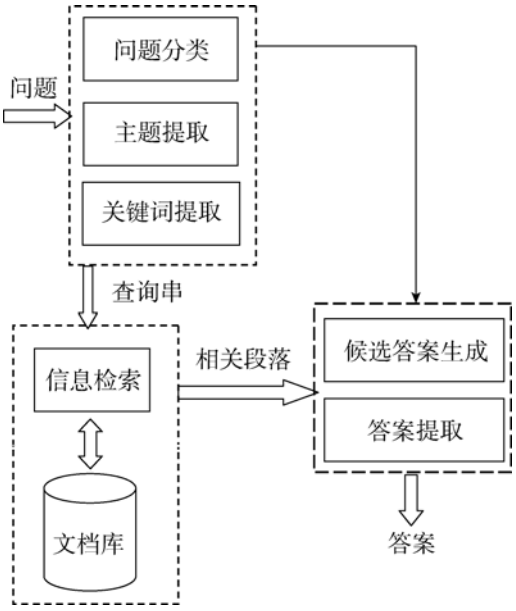

Fig.2 The architecture of QA system based on free text

图 **2** 基于自由文本的问答系统的体系结构

问题分析部分所需要完成的功能包括问句类型分析、问句主题识别、问句指代消解和问句语法分析等。问句分类是问答系统中一个很重要的环节，它需要将问句根据它的答案类型分到某一类别中，之后的检索和提取会根据问句类别采用不同的措施。在现有基于自由文本的问答系统的解决方案中，很多都根据精细问句类型和精细实体答案识别的对应关系来提取答案，因此问句分类的性能非常关键。找出问句的主题，首先需要找出与主题相关的文档和段落，以便于后续的处理。某些系统交互式地回答用户的提问，用户的问题中会出现指代词，因此需要根据上下文明确指代词在问题中的具体所指。有的系统通过对问句进行语法分析，来匹配问句的语法结构和包含答案句子的语法结构。

信息检索部分的功能是根据问句来构造查询，利用一定的检索模型找到可能包含答案的文档或者段落。一般有两个步骤：首先检索出可能包含答案的文档；然后从检索出来的文档中抽取出可能包含答案的段落。这里涉及到的问题包括：采用什么样的信息检索模型；如何构造查询；如何对这些段落进行排序；如何追求查全率和查准率之间的折衷；检索阶段的性能和最终的系统总性能有什么样的关系等。

答案抽取部分是问答系统的最后一个部分，它的输出就是问句的最终答案。它分析检索获得的文档或者段落，从中抽取能够回答问题的答案。在抽取答案时，问句类型直接决定如何生成候选答案集合。

下面将分别对图 2 中的各个部分的相关研究进行叙述。

4.1 问题分析

问题分析部分主要是用于分析和理解问题，从而协助后续的检索和答案提取，一般具有问句分类、问句主题提取两个主要研究内容，下面分别进行叙述。

4.1.1 问句分类

问句分类是根据问句的答案类型对问句进行分类，它是问句分析最重要的功能之一。目前大多数这类问答系统都利用答案类型来指导后续步骤，尤其是答案抽取策略，例如对于问人物的问题，答案抽取会利用人物的各种特征来提取答案候选集合。一般问句分类可以通过疑问词直接决定问句的类型，如含有“谁”的问句可以认为需要的答案是人物。这种方法准确度不高，例如在英语中对 What、How 这样的疑问词，可以对应非常多的答案类型，同时还有一些问句不包含疑问词(如“列出中国所有的省份”)，或者问句包含了多个疑问词(如英语中含有定语从句、宾语从句等情况)。Li 等人[11]提出了更加详细的分类。该问句体系有 6 个大类(缩略语、描述、实体、人物、地点、数量)，在 6个大类下面又分了 50 个小类，如在数量类里面有距离、钱等小类。问句分类的任务就是把一个问句分类到已有的分类

结构中一个或几个类(软分类)。问句分类的方法主要包括模式匹配方法和机器学习方法两类。模式匹配方法为每一种问题类型建立一个模式集合，对于一个问句，只要与某种问题类型对应的模式相匹配，就被认为是这种类型的问题。机器学习方法[11−12]首先定义一个问题的特征集合，然后在训练数据上得到一个分类器，就可以对新的问句进行分类了。Zhang 等人[12]使用表层 *n*-gram 特征，在 *K* 最近邻算法、决策树、朴素贝叶斯等分类方法中，支持向量机(support vector machine, SVM)效果较好。Li 等人[11]采用更深层次的特征，包括语法(词性、词组)和语义(解释、近义词)信息，首先顶层分类器把问句归类在一个大类别中，然后根据该大类内的分类器把它分到小类别中，分类效果较好。

4.1.2 问句主题提取

信息检索部分需要选择问题中的一些关键词作为查询词，很多时候会调整查询，为了保证高相关性，查询词都应该包含问题的主题。一般通过对问题进行句法分析，获得问题的中心词，然后选取中心词及其修饰词作为问题的主题。因此，如何选取合适的中心词是该类方法的核心。Cui 等人[13]提出了一种基于外部资源选取主题的方法。它利用搜索引擎返回的结果去计算各种词组合的点互信息，如果词组合的点互信息高于某个阈值就被认为是词组，那么词组的序列就被认为是问题的主题。

在问题分析中还有查询关键词如何产生的问题，但是由于关键词提取和检索模型密切相关，本文将在信息检索部分介绍这个问题。

4.2 信息检索

信息检索的主要目的是缩小答案的范围，提高下一步答案抽取的效率和精度。

对于信息检索部分，最简单的方法是去掉问题中的停用词和问句相关的词(如疑问词)生成查询，然后利用已有的检索模型进行检索，把返回的结果作为答案提取部分的输入。这种方法很难获得较好的效果，而文档检索的效果会直接影响到问答系统的整体性能[14]。如果检索系统的相关性精度较差，那么会有大量无关文档需要后续处理，而答案提取通常采用复杂的自然语言处理技术，这必然导致系统整体效率低下；如果检索系统的召回率较低，那么很多包含答案的文档或者段落没有被返回，显然含有正确答案的文档或者段落越少，提取出正确答案的可能性也越小，这会导致系统整体性能较差。

信息检索一般分为两个步骤：

(1) 文档检索，即检索出可能包含答案的文档。

(2) 段落检索，即从候选文档中检索(抽取)出可能包含答案的段落。

下面分别叙述这两个步骤。

4.2.1 文档检索

文档检索是给定一个由问题产生的查询，通过某个检索模型去得到相关的文档。有两个问题需要处理：一是检索模型的选取；二是查询的生成。下面分别对这两个问题进行论述。

信息检索领域常用的模型包括布尔模型、向量空间模型、语言模型、概率模型等。实验发现在文档检索中，简单的布尔模型、概率模型与改进的向量空间模型的效果相当[15−16]。

无论采用哪种检索模型，检索系统的输入应该是由问句生成的查询。最简单的方法是删除问句中的停用词，其余的词作为关键词。当问句很长时，关键词就会很多，若采用布尔模型，检索返回的文档太少，召回率就很低；当问句很短时，关键词就会很少，检索返回的文档太多，相关性文档的精度就很低。针对这种查询词选择的不确定性，可以对查询关键词进行调整。若关键词太多、查询限制太严格，就去掉一些；若关键词太少、查询限制宽松，就加上一些。Moldovan 等人[17]采用这种迭代式调整技术，多次查询，根据返回文档的多少，调整查询，决定是否增删关键词以及是否采用词形、句法或者语义级别的扩展形式。

问答系统的目的是找出一个问题的答案，而上述方法是找一个和问题相关的文档，因此如果从一个问句推测它的答案中可能包含的关键词，用这些关键词来进行查询，会得到更好的效果。对于特定类别的问题，可以从训练数据中学习得到这类问题的回答模式，根据得到的模式从问句生成包含答案

关键词的查询[18]。

4.2.2 段落检索

段落检索就是从候选文档集合中检索出最有可能含有答案的段落[16](自然段落或者文档片段),进一步缩小答案存在的范围。

Tellex 等人[16]细致地考察了八种最好的段落检索算法。实验结果表明，基于密度的算法可以获得相对较好的效果。所谓基于密度的算法，就是通过考虑查询关键词在段落中的出现次数和接近程度来决定这个段落的相关性。目前表现比较好的段落检索算法有三个：MultiText 算法[19]、IBM 的算法[20−21]和 SiteQ 算法[22]。这三个算法，虽然在设计和实现细节上有很大的差异，但是都使用了 IDF 值的总和，都考虑了邻近关键词之间距离的因素。

但是基于密度的算法只考虑了独立的关键词及其位置信息，没有考虑关键词在问句中的先后顺序，也没有考虑语法和语义信息。Cui 等人[23]提出了一种基于模糊依赖关系匹配的算法。这种算法把问题和答案都解析成为语法树，并且从中得到词与词之间的依赖关系，然后通过依赖关系匹配的程度来进行排序。实验结果表明，这种方法的检索效果比基于密度算法(SiteQ)好。

4.2.3 富信息索引

传统检索方法一般只需要处理关键词，而问答系统需要处理更多的语法、语义信息。因此，部分问答系统也把语法、语义等信息添加到索引中，丰富了传统的索引，以提高检索效果。Radev 等人[24]把一些关键词或者词组的属性放入索引，这样构造的查询包含关键词和答案的属性要求。例如，对于一个问时间的问题，构造的查询包含关键词和时间属性，返回的段落中要求包含时间。Bilotti 等人[25]把查询变成一个结构化的查询，表达查询词和段落中应该包含的某些词的属性。为了解决问句关键字的顺序问题, Katz 等人[26]把句子解析为〈主,谓,宾〉三元组的形式，然后加入索引。另外, Chu-Carroll 等人[27]还索引了句子中词和词组的语义关系。

4.3 答案抽取

答案抽取的主要目的是得到用户想要的答案，满足用户需求。一个问答系统通过问句分析和文档段落检索可以获得问题答案的段落集合，答案提取是从这些段落中获取正确的答案。为了提取答案，一般有两个步骤：

(1) 生成候选答案集合。

(2) 提取答案。

4.3.1 候选答案集合的生成

通过问题分析，已经获得问句的类别。目前问答系统能处理的问题类型一般都是事实类型的问题，大多数的问题类型对应的答案比较短，可能是实体名，如人名、地点等，可能是抽象名词，如人类、学科、树木、植物等，也可能是数字，如距离、速度等。对于这类问题，可以通过找到相应类型的词、词组或者片断来回答。目前，自然语言处理领域命名实体的识别已经能够达到非常好的效果,如隐马尔可夫模型(hidden Markov model, HMM)或者条件随机域模型(conditional random field, CRF)。对于实体名词词表，除了利用 WordNet 等字典之外，还可以采用一些概念性名词和具体名词作为训练的种子，用 Bootstrap 方法从文档集或者 Web 中找到这种连接概念性名词和具体名词的模式，再根据这种模式提取更多的具体名词，多次迭代可以发现更多的〈概念名词,具体名词〉词对和相应的模式[28]。另一种简单的方法是直接利用 Web 资源(如 Wikipedia)中具体名词列表。对于抽象名词，通常是构建一个名词列表，若片段中含有这个列表中的词，就作为答案返回。对于数字度量，可以通过正则表达式来获取，譬如距离的一个模式是数字跟上距离单位，如 5 米。通过在文本中匹配相应问句类型的短语，就构成了候选答案集。

4.3.2 答案提取

在得到候选答案集合以后，主要有四种方法获得问句的最佳答案，下面分别进行叙述。

(1) 基于表层特征的答案提取。常用的表层特征是答案周围段落的一些特征，如段落和查询的相关程度、查询词之间的距离、查询词和候选答案的距离等。一般来说，段落相关程度越高，查询词之间以及查询词和候选答案之间距离越接近，则该候

选答案越可能是问题的答案。另一个常用特征是该候选答案出现的次数。对于一个比较大的文档集，一个问题的答案可能反复出现，出现的次数越多，则它越可能是正确答案[29]。

(2) 通过关系抽取答案。表层特征没有考虑语法、语义的因素，容易出错，特别是词相同，但词序不同的情况。Light 等人[30]指出这种基于实体识别和表层特征的方法的性能上限是 70%。为了克服基于表层特征抽取答案的缺陷，不同的改进方法被提出。其中一种方法是把问句和文本中的句子转换成三元组，三元组的基本构成是〈主语,谓语,宾语〉，删除句子中的修饰成分，就可以从文本三元组中获得答案而不产生混淆[26]。另一种改进方法是建立问题到答案的逻辑表示[31]，逻辑表示是介于句法解析表示和深层语义表示之间的一种表示形式，它可以通过对解析获得的句法树进行一些规则计算获得，表达了主语、宾语、前置词、复杂的名词性短语、附属的形容词或副词之间的关系。

(3) 通过模式匹配抽取答案。模式可以通过手工定制或机器学习获得。一般手工定制的模式扩展性和覆盖率都比较低，主流方法是在训练数据上自动学习得到模式。如 Cui 等人[13]提出了一种软模式的方法，来处理定义类问题答案的抽取。

(4) 利用统计模型抽取答案。近年来，研究人员开始尝试用统计模型对答案提取进行建模。目前两个有代表性的建模方法为：一种是噪音信道模型[32]，该模型把问句看成目标信息，把答案看成源信息，假设源信息需要通过一个包含噪音的信道，则转换概率可以通过一组训练数据(问题答案对集合)训练获得。另一种是用无向图模型对答案提取过程进行建模[33]。

5 基于问题答案对的问答系统

5.1 FAQ 和 CQA 的对比

基于问题答案对的问答系统研究主要有两个发展阶段：一是基于常问问题(FAQ)列表的问答系统研究阶段；二是基于社区问答(CQA)的问答系统研究阶段。在早期，很多公司或者组织会在自己的网站上专门设有常问问题列表，以便用户直接得到想要的答案。在网络上存在大量的这类 FAQ 列表，这种数据形式很快就引起了研究者的注意，因此针对这类数据的各种研究也就逐渐出现。FAQ 具有量大、问题质量高和组织好等优点，但是在特定领域问题数目相对较少，这个缺点制约了基于 FAQ 的问答系统的应用范围。自 2005 年末以来，一种新的问题答案对形式的数据开始大量出现，即 CQA 数据，不仅问题答案对的数量大，而且在特定领域问题答案对数目也特别多，同时还在不断增加。相对 FAQ 问题答案对，CQA 数据中的问题答案对的质量参差不齐，而且用语不规范，有很多口语和省略语。相较于 FAQ 问题答案对，CQA 数据还多了一个社会网络存在于问题答案对数据中，这使得对 CQA 的研究比对 FAQ 的研究有了丰富的内容，如用户的投票对答案质量的判断是一个重要的特征等。FAQ 和 CQA 两种数据的比较，如表 1 所示。

Table 1 The comparison of FAQ and CQA

表 **1** **FAQ** 和 **CQA** 的比较

问答系统	质量	用语	总量	特定领域数量	社会网络
FAQ	高	规范	大	小	无
CQA	良莠不齐	口语化、不规范	巨大	大	有

因为对 CQA 的研究基本上覆盖了对 FAQ 的研究，所以下面将主要对基于 CQA 的问答系统的相关研究进行叙述。

5.2 三个基本研究问题

每种类型的问答系统都有三个基本研究问题：问题分析、信息检索、答案抽取。下面将针对这三个问题，对基于问题答案对的问答系统的相关研究进行叙述。系统架构如图 3 所示 2)。

2) 体系结构中还有用户互动的一个模块，用于用户提出问题，其他用户回答问题和给答案打分等，由于非重点部分，在图3中略。

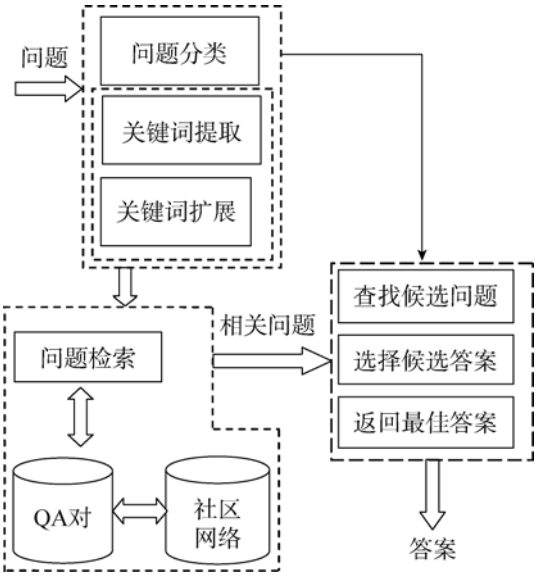

Fig.3 The architecture of QA system based on question-answer pairs

图 **3** 基于问题答案对的问答系统的体系结构

5.2.1 问题分析部分

在问题分析部分，和基于自由文本的问答系统的问题分析部分基本一样，这里就不再赘述。下面只叙述几个不同的研究点。

(1) 问题主客观的判断

基于结构数据的问答系统和基于自由文本的问答系统一般都只能处理客观、事实类型(factoid)的问题。然而在 CQA 数据中有大量的主观类型的问题，而对于主观类型的问题和客观类型的问题有不同的处理方式，例如主观问题没有标准答案，而且答案可以多个，然而客观问题却只能有一个标准答案。因此，在问题分析部分，首先判断问题属于主观问题还是客观问题，这对后面的信息检索、答案提取都是非常有帮助的。显然，主客观判断是一个分类问题，关键在于选取用于分类的特征。Li 等人[34]通过选择各种特征组合作为一个有监督的机器学习算法的特征来解决这个问题，实验表明能在选择词、词性和 *n*-gram 作为特征的情况下达到 74.2%的准确率。Li 等人[35]在上述工作的基础上，进行了更加细致的研究，把数据集合分为健康、艺术等类别，然后分别测试各个类别下算法的精确度。它主要采用 Co-training 的数据挖掘思想来提高分类的效果，其选取的特征和文献 [34] 中选取的特征是一致的。

(2) 问题的紧急性

在 CQA 数据中，由于存在一个社区网络，有用户的互动参与，当问题的紧急程度不一样时，系统应该有不一样的处理方式。因为有些问题很长时间回答都没有影响，但是有些问题需要尽快地得到回答，见表 2。对于紧急性的问题，系统应对它进行快速处理，或者尽量把它放在明显的位置让其他用户进行回答；对于紧急程度不高的问题，就不用这样进行特别处理。对于问题的紧急性判断处理的好坏，将直接影响到用户对一个 CQA 问答系统的评价，因此非常重要。显然，把紧急性判断作为分类问题来处理，关键在于特征选取。Liu 等人[36]采用基于 SVM 和决策树的有监督分类算法，采用的特征是问题的文本(主要采用 1 000 个出现频率最高的词)、问题的目录以及问题的答案(主要选择频率最高的词)，实验效果最好能达到 85%。

Table 2 The examples of the urgency questions

表 **2** 问题紧急性的例子

紧急程度	实例
紧急	这是煤气泄漏吗？我能打开煤气煮饭吗？刚才我回到家，没有开煤气，但是听到煤气罐似乎有丝丝的声音，请求帮助，我很害怕！谢谢！
不紧急	这里有孙燕姿的歌迷吗？两个月后她在北京的演唱会，你们去吗？
似乎紧急	我想买一个智能手机，请问 iphone 和黑莓，应该买哪个？谢谢！

(3) 研究 query 生成问题

由于 CQA 问题答案对数量巨大，微软研究院的 Lin[37]指出用户只提出一个查询(query)，然后通过这个查询生成其对应的问题，就可以在 CQA 数据中找到相应的答案返回给用户。文中没有给出具体的解决方案，但这是一个待研究的有趣问题。

5.2.2 信息检索部分

在信息检索部分，与基于自由文本的问答系统不同，基于问题答案对的问答系统已经有了问题和对应的答案，不必在文本中搜寻答案，因此在检索

部分只需找到和问题类似的问题，然后返回答案或者相似问题列表即可。研究适合问题答案对的检索模型和两个问题的相似性判断是最关键的两个问题。

(1) 研究问题答案对的检索模型

在经过问题分析之后，需要通过信息检索部分把相关的问题检索出来，然后才能在答案抽取部分抽取合适的答案。

Duan 等人[38]采用基于最小描述长度(minimum description length, MDL)的树剪切(tree cut)模型去识别问题的主题和焦点，然后利用语言模型的方法去匹配问题主题和焦点来缩小检索范围，提高检索精度。Xue 等人[39]首先构建一个相似问题的集合，然后训练一个翻译模型，对问题部分采用基于翻译模型的语言模型，对答案部分采用 Query Likelihood 模型，这两者合并在一起构成了一个检索模型。Bian 等人[40]提出了对于事实类型问题的检索框架，属于半监督的方法。实验发现，其中的文本特征和社区特征对于检索效果的影响很大；同时，其他特征，如用户的反馈，也能提高训练排序函数的效果。Wang 等人[41]通过类比推理技术去建模问题与答案之间的关系，提出了对于答案的排序方法。实验表明这种方法的平均排序倒数(mean reciprocal rank，MRR)值能达到 78%，比采用朴素贝叶斯和 Cosine 的方法效果都好。

(2) 研究问题答案对的相似性

除了从传统的信息检索的角度来看待信息检索部分，还可以从问题对相似性的角度来看这个问题。其实传统的信息检索角度，其本质也是把相似的问题检索出来。首先把与给定问题相似的问题找到，然后再在相似的问题中寻找最好的答案。

Wang 等人[42]通过将每一个问题句子都生成其 syntactic tree 的形式，然后比较问题的 syntactic tree 的相似性来判断问题的相似性。实验表明，这种问题的 syntactic tree 特征对于问题的相似性比较具有明显的价值。还有一些研究[43−44]主要依据这样一个假设：“如果两个问题的答案相似，那么这两个问题是相似的”，因此作者把问题的所有答案作为单一文档，然后用语言模型去计算答案之间的相似性，从而去判断两个问题之间的相似性。Akiyoshi 等人[45]采用简单的基于关键字的检索方法去缩小搜索范围，然后再利用 BBS 文章的结构特征(即 thread 结构，一个 thread 包含一个问题网页和一组答案网页)，基于这种结构特征，可以计算一篇问题文章和商业搜索引擎返回的词之间的相关性，最后来判断其相似性。作者在一个 BBS 站点上进行了实验，结果表明比基于 Cosine 的相似度方法提高了 30%的准确率。

5.2.3 答案抽取部分

在答案抽取部分，由于问题答案对已经有了答案，答案抽取最重要的工作就是判断答案的质量。

在相似的问题或者相关的问题得到之后，由于 CQA 中每个问题都有很多答案，虽然大部分问题都会有一个标记为 Best Answer 的答案，但是这个 Best Answer 未必就是最好的答案。因为形成这个 Best Answer 的机制有可能是投票或者系统自动标注得到的，所以非常有必要研究怎么从问题的众多答案中选择一个最好的答案，这就需要研究答案的质量。

Liu 等人[46]提出了问题和答案的分类体系，特别是对 open 和 opinion 类型的问题进行了研究，以提高这两种类型问题的答案质量，并通过实验得出这两种类型的问题占总问题数目的 56%～80%的结论。Jeon 等人[47]利用非文本特征来预测答案的质量，主要用到的特征有答案的接受率、答案长度、提问者的自评、回答问题的人的积极程度、回答问题的人的回答问题偏好、用户打印答案的次数、用户拷贝答案到他们博客的次数、用户的推荐次数、服务提供商的推荐、是否是 Sponsor 的答案、点击次数、答案的数目、用户的不推荐数目等。作者系统地分析了这些非文本特征对于预测答案质量的作用，最后应用最大熵方法和核密度方法去预测答案的质量。

除了上述选取问题的一个答案来回答问题的方法外，Liu 和 Tomasoni 等人[46,48]利用大多数 CQA 问题都有多个答案的特性，把答案抽取的问题转换为

多文档摘要的问题，从而达到较好的效果。

5.3 CQA 特有性质及其对应的研究

除了上述关于问答系统的三个基本研究问题之外，由于 CQA 的特殊性(如有社区特性等)，还有很多研究是关于 CQA 特有性质或者特有问题的，下面将分别进行叙述。

5.3.1 Vote Spam

在 CQA 中用户可以对某个答案进行投票来决定其优劣，例如在 Yahoo!Answer 中，用户可以投赞成(大拇指朝上)或者反对(大拇指朝下)来对一个答案进行投票评价。在这个投票机制中没有任何责任或者机制约束，完全是自由的，这可能为某些用户恶意投票所利用，从而产生垃圾投票。因为问题的质量评价或者排序很大程度上需要依靠投票的质量，所以研究哪些投票是垃圾投票或者良好投票非常重要。Bian 等人[49]提出了一种基于机器学习的排序框架，它集成了用户交互和内容相关性，主要特征有内容特征和用户交互特征，通过排序的方法把垃圾投票的答案排在了较为靠后的位置，从而达到去除垃圾的效果。

5.3.2 用户问题的信息满意度

对于社区网络，用户的感受非常重要。在 CQA 中，最重要的感受就是提出一个问题，CQA 能不能给用户满意的答案，这就是问题的满意度。为了了解用户对于问题的满意度，Liu 等人[50]把用户的先验信息集成到检索模型中，从而实现了个性化的检索，即对于每个用户提出的问题，其对某个答案的满意度是依赖于这个用户的历史记录的。Liu 等人[51]第一次提出了在问答社区中预测问题提出者的满意度这个研究问题，并提出了利用文本内容、结构特征和社区特征的预测模型来解决这一问题。

5.3.3 CQA 用户权威性或者专家网络

在一个由用户与问题组成的社区网络中，这些用户对于每个问题的影响都是不同的。有些用户回答其他用户问题的答案很多都被评为了 Best Answer，这说明这个用户比较有影响力或者有权威性。而另外一些用户就不一样了，可能从注册以来就只回答了几个问题，而且回答的效果都不好。很明显这两类用户具有不同的特性，那么由这两类用户回答的答案显然应该具有不同的权重。因此，在用户与问题网络中发现权威用户非常重要。Jurczyk 等人[52]建立了用户-问题网络和问题-答案网络，然后提出通过链接分析技术 HITS(hyperlink-induced topic search)来发现 CQA 社区中的权威用户。Zhang 等人[53]观察到知识共享网络如 CQA 和其他在线网络(www)在拓扑结构上不同，因此利用这种拓扑结构的不同，提出了一种社会网络分析方法来发现专家网络，并在一个实际的论坛上面测试了 PageRank 和 HITS 算法以及作者自己的算法，实验表明作者的算法优于其他两个著名算法。

5.3.4 在 CQA 背后运作的规律

CQA 作为一个客观的对象，必然有其运作和发展的规律，如果能挖掘出隐藏在它背后的运作规律，对于深刻地理解 CQA 和研究 CQA 都是具有非常重要的意义的。

Jain 等人[54]对用户提出和回答问题的过程用博弈论模型进行了建模研究。Rodrigues 等人[55]研究了目前网络上的知识共享现象(如 Wikipedia 或者 Yahoo!Answer)中存在的一个问题，即用户个体的目标也许会从知识共享变化到社会目标(例如想拥有很高的社区影响力，因此不断地回答问题等)。作者通过把 CQA 的内容分为社会化和非社会化两大类，然后采用社会网络分析技术去分析 CQA 中的内容是属于知识共享还是社会化的内容。这种方法可以被应用到在线社区的动态监测上。

5.3.5 论坛里面的问题答案对提取问题

CQA 系统除了通过 Yahoo!Answer 这类专门的门户网站生成大量的问题答案对之外，在大量的论坛中，其实也散落了大量的问题答案对，它们可以丰富和完善问题答案对。有一些研究是专门针对怎么从论坛中提取问题答案对的，例如 Cong 等人[56]利用基于序列标签的分类方法去检测在论文 thread 中的问题，然后通过基于图的传播方法去检测问题对应的答案，实验表明效果较好。

6 结束语

由于电子数据数量的变化和形态的变化，导致产生了相应的问答技术。本文主要对问答系统进行了总结和分类，并分别叙述了三大类问答系统中存在的重要问题和相应研究。

虽然针对不同的数据类型有不同的技术，但是这些技术的目的都是一致的。如何将已存在的三大类问答系统中的各种技术融合起来去回答问题，特别是融合各种数据类型，如多媒体数据、视频、图像等，来综合地提高问答的性能，是未来提高性能的一个重要方向。

目前网络上已经存在了大量的问题答案对，但大部分研究都集中在某个特定的 CQA 站点来进行研究。如果把网络上所有的问题答案对收集到一起，利用数据源之间的相互印证和相互补充，结合统计等技术手段，相信对于问答系统性能的提升将有很大的帮助。

虽然目前问答系统领域已经有大量的研究，但是大部分的方法都不能处理语义问题，还不能真正地理解问题，因为没有真正地理解，所以性能始终不能有明显的提升。在计算机真正能理解自然语言以前，问答系统性能质量的提升最有希望的方式可能还是依赖于人工的参与。但这种参与不是采用对于一个问题，人工在后台进行回答的方式，而是类似 CQA 这类 Web2.0 类型的新的服务形式，使得人们乐于奉献自己的精力和时间。这种群体智慧类型的新型服务及其基于这些服务产生的数据的挖掘和学习，将是问答系统性能跃升最有希望的途径，也是未来可能的研究方向。

References:

[1] Molla D, Vicedo J L. Question answering in restricted domains: an overview[J]. Computational Linguistics, 2007, 33(1): 41–61.

[2] Green B F, Wolf A K, Chomsky C, et al. BASEBALL: an automatic question answerer[C]//Proceedings of the Western Joint IRE-AIEE-ACM Computer Conference. New York, NY, USA: ACM, 1961: 219–224.

[3] Woods W A. Lunar rocks in natural English: explorations in natural language question answering[J]. Linguistic Structures Processing, 1977, 5: 521–569.

[4] Wilensky R, Chin D N, Luria M, et al. The Berkeley UNIX consultant project[J]. Computational Linguistics, 1988, 14(3): 35–84.

[5] Wood J A, Dickey J S, Marvin U B, et al. Lunar anorthosites and a geophysical model of the moon[C]//Proceedings of the Apollo 11 Lunar Science Conference, Houston, TX, Jan 5–8, 1970. New York: Pergammon Press, 1970: 965–988.

[6] Hirschman L, Gaizauskas R. Natural language question answering: the view from here[J]. Natural Language Engineering, 2001, 7(4): 275–300.

[7] Simmons R F. Answering English questions by computer: a survey[J]. Communications of the ACM, 1965, 8(1): 53–70.

[8] Wilensky R, Chin D N, Luria M, et al. The Berkeley UNIX consultant project[J]. Artificial Intelligence Review, 2000, 14(1/2): 43–88.

[9] Lin J, Demner-Fushman D. Will pyramids built of nuggets topple over?[C]//Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL ’06), Morristown, NJ, USA, 2006. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006: 383–390.

[10] Lin J, Demner-Fushman D. Automatically evaluating answers to definition questions[C]//Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT ’05). Stroudsburg, PA, USA: Association for Computational Linguistics, 2005: 931–938.

[11] Li Xin, Roth D. Learning question classifiers[C]//Proceedings of the 19th International Conference on Computational Linguistics (COLING ’02). Stroudsburg, PA, USA: Association for Computational Linguistics, 2002: 1–7.

[12] Zhang D, Lee W S. Question classification using support vector machines[C]//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’03). New York, NY, USA: ACM, 2003: 26–32.

[13] Cui Hang, Kan M-Y, Chua T-S. Unsupervised learning of

soft patterns for generating definitions from online news [C]//Feldman S I, Uretsky M, Najork M, et al. Proceedings of the 13th International Conference on World Wide Web (WWW 2004), May 17–20, 2004. New York, NY, USA: ACM, 2004: 90–99.

[14] Collins-Thompson K, Callan J, Terra E, et al. The effect of document retrieval quality on factoid question answering performance[C]//Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’04). New York, NY, USA: ACM, 2004: 574–575.

[15] Moldovan D, Pasca M, Harabagiu S, et al. Performance issues and error analysis in an open-domain question answering system[J]. ACM Transactions on Information Systems, 2003, 21(2): 133–154.

[16] Tellex S, Katz B, Lin J, et al. Quantitative evaluation of passage retrieval algorithms for question answering[C]// Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’03). New York, NY, USA: ACM, 2003: 41–47.

[17] Moldovan D I, Harabagiu S M, Pasca M, et al. The structure and performance of an open-domain question answering system[C]//Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL ’00). Stroudsburg, PA, USA: Association for Computational Linguistics, 2000: 563–570.

[18] Agichtein E, Lawrence S, Gravano L. Learning to find answers to questions on the Web[J]. ACM Transactions on Internet Technology, 2004, 4(2): 129–162.

[19] Clarke C, Cormack G, Kisman D, et al. Question answering by passage selection (multitext experiments for TREC-9) [C]//Proceedings of the 9th Text Retrieval Conference (TREC-9), 2000.

[20] Ittycheriah A, Franz M, Zhu W-J, et al. IBM’s statistical question answering system[C]//Proceedings of the 9th Text Retrieval Conference (TREC-9), 2000.

[21] Ittycheriah A, Franz M, Roukos S. IBM’s statistical question answering system—TREC-10[C]//Proceedings of the 10th Text Retrieval Conference (TREC 2001), 2001.

[22] Lee G G, Seo J, Lee S, et al. SiteQ: engineering high performance QA system using lexico-semantic pattern matching and shallow NLP[C]//Proceedings of the 10th Text Retrieval Conference (TREC 2001), 2001.

[23] Cui Hang, Sun Renxu, Li Keya, et al. Question answering passage retrieval using dependency relations[C]//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’05). New York, NY, USA: ACM, 2005: 400–407.

[24] Radev D R, Prager J M, Samn V. Ranking suspected answers to natural language questions using predictive annotation[C]//Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP ’00). Stroudsburg, PA, USA: Association for Computational Linguistics, 2000: 150–157.

[25] Bilotti M W, Ogilvie P, Callan J, et al. Structured retrieval for question answering[C]//Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’07). New York, NY, USA: ACM, 2007: 351–358.

[26] Katz B, Lin J. Selectively using relations to improve precision in question answering[C]//Proceedings of the EACL-2003 Workshop on Natural Language Processing for Question Answering, 2003: 43–50.

[27] Chu-Carroll J, Prager J, Czuba K, et al. Semantic search via XML fragments: a high-precision approach to IR[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’06). New York, NY, USA: ACM, 2006: 445–452.

[28] Mann G S. Fine-grained proper noun ontologies for question answering[C]//Proceedings of the 2002 Workshop on Building and Using Semantic Networks (SEMANET ’02), Morristown, NJ, USA, 2002. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002: 1–7.

[29] Lin J, Katz B. Question answering from the Web using knowledge annotation and knowledge mining techniques[C]//Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM ’03). New York, NY, USA: ACM, 2003: 116–123.

[30] Light M, Mann G S, Riloff E, et al. Analyses for elucidating current question answering technology[J]. Natural Language Engineering, 2001, 7(4): 325–342.

[31] Moldovan D, Harabagiu S, Girju R, et al. LCC tools for question answering[C]//Proceedings of the 11th Text Retrieval Conference (TREC 2002), Department of Commerce, National Institute of Standards and Technology, 2002.

[32] Echihabi A, Marcu D. A noisy-channel approach to question answering[C]//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL ’03), Morristown, NJ, USA, 2003. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003: 16–23.

[33] Ko J, Nyberg E, Si L. A probabilistic graphical model for joint answer ranking in question answering[C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’07). New York, NY, USA: ACM, 2007: 343–350.

[34] Li Baoli, Liu Yandong, Ram A, et al. Exploring question subjectivity prediction in community QA[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’08). New York, NY, USA: ACM, 2008: 735−736.

[35] Li Baoli, Liu Yandong, Agichtein E. CoCQA: co-training over questions and answers with an application to predicting question subjectivity orientation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP ’08). Stroudsburg, PA, USA: Association for Computational Linguistics, 2008: 937– 946.

[36] Liu Yandong, Narasimhan N, Vasudevan V, et al. Is this urgent?: exploring time-sensitive information needs in collaborative question answering[C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’09). New York, NY, USA: ACM, 2009: 712–713.

[37] Lin C-Y. Automatic question generation from queries[C]// Workshop on the Question Generation Shared Task, 2008.

[38] Duan Huizhong, Cao Yunbo, Lin C-Y, et al. Searching questions by identifying question topic and question focus[C]//Proceedings of Association for Computational Linguistics (ACL), 2008: 156–164.

[39] Xue Xiaobing, Jeon J, Croft W B. Retrieval models for question and answer archives[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’08). New York, NY, USA: ACM, 2008: 475–482.

[40] Bian Jiang, Liu Yandong, Agichtein E, et al. Finding the right facts in the crowd: factoid question answering over social media[C]//Proceedings of the 17th International Conference on World Wide Web (WWW ’08). New York, NY, USA: ACM, 2008: 467–476.

[41] Wang Xinjing, Tu Xudong, Feng Dan, et al. Ranking community answers by modeling question-answer relationships via analogical reasoning[C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Iinformation Retrieval (SIGIR ’09). New York, NY, USA: ACM, 2009: 179–186.

[42] Wang Kai, Ming Zhaoyan, Chua T-S. A syntactic tree matching approach to finding similar questions in community-based QA services[C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’09). New York, NY, USA: ACM, 2009: 187−194.

[43] Jeon J, Croft W B, Lee J H. Finding similar questions in large question and answer archives[C]//Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM ’05). New York, NY, USA: ACM, 2005: 84–90.

[44] Jeon J, Croft W B, Lee J H. Finding semantically similar questions based on their answers[C]//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’05). New York, NY, USA: ACM, 2005: 617– 618.

[45] Akiyoshi M, Iwai K, Komoda N. A retrieval method for similar Q&A articles of Web bulletin board with relevance index derived from commercial Web search engine [C]//Proceedings of the 10th International Conference on Information Integration and Web-Based Applications & Services (iiWAS ’08) . New York, NY, USA: ACM, 2008: 583−586.

[46] Liu Yuanjie, Li Shasha, Cao Yunbo, et al. Understanding and summarizing answers in community-based question answering services[C]//Proceedings of the 22nd International Conference on Computational Linguistics-Volume1 (COLING ’08). Stroudsburg, PA, USA: Association for Computational Linguistics, 2008: 497–504.

[47] Jeon J, Croft W B, Lee J H, et al. A framework to predict the quality of answers with non-textual features[C]//Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’06). New York, NY, USA: ACM, 2006: 228–235.

[48] Tomasoni M, Huang Minlie. Metadata-aware measures

for answer summarization in community question answering[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL ’10). Stroudsburg, PA, USA: Association for Computational Linguistics, 2010: 760–769.

[49] Bian Jiang, Liu Yandong, Agichtein E, et al. A few bad votes too many?: towards robust ranking in social media[C]//Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb ’08). New York, NY, USA: ACM, 2008: 53–60.

[50] Liu Yandong, Agichtein E. You ’ve got answers: towards personalized models for predicting success in community question answering[C]//Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (HLT-Short ’08). Stroudsburg, PA, USA: Association for Computational Linguistics, 2008: 97–100.

[51] Liu Yandong, Bian Jiang, Agichtein E. Predicting information seeker satisfaction in community question answering[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’08). New York, NY, USA: ACM, 2008: 483–490.

[52] Jurczyk P, Agichtein E. Discovering authorities in question answer communities by using link analysis[C]//Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM ’07). New York, NY, USA: ACM, 2007: 919–922.

[53] Zhang Jun, Ackerman M S, Adamic L. Expertise networks in online communities: structure and algorithms[C]//Proceedings of the 16th International Conference on World Wide Web (WWW ’07). New York, NY, USA: ACM, 2007: 221–230.

[54] Jain S, Chen Yiling, Parkes D C. Designing incentives for online question and answer forums[C]//Proceedings of the 10th ACM Conference on Electronic Commerce (EC ’09). New York, NY, USA: ACM, 2009: 129–138.

[55] Mendes Rodrigues E, Milic-Frayling N. Socializing or knowledge sharing?: characterizing socail intent in community question answering[C]//Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM ’09) . New York, NY, USA: ACM, 2009: 1127–1136.

[56] Cong Gao, Wang Long, Lin C-Y, et al. Finding question-answer pairs from online forums[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’08). New York, NY, USA: ACM, 2008: 467–474.

MAO Xianling was born in 1983. He is a Ph.D. candidate at Peking University. His research interests include information retrieval and Web data mining, etc.

毛先领(1983—)，男，贵州遵义人，北京大学博士研究生，主要研究领域为信息检索，网络数据挖掘等。

LI Xiaoming was born in 1957. He received his Ph.D. degree in computer science from Stevens Institute of Technology (USA) in 1986. Now he is a professor and Ph.D. supervisor at Peking University, and the senior member of CCF, IEEE and ACM. His research interests include information retrieval, Web data mining, computer parallel and distributed processing, etc.

李晓明(1957—), 男，湖北荆州人, 1986 年于美国史蒂文斯理工学院获得博士学位，现为北京大学计算机科学技术系教授、博士生导师，中国计算机学会副理事长，中国电子学会常务理事，教育部高等学校计算机专业教学指导委员会主任委员, IEEE 和 ACM 高级会员，主要研究领域为信息检索，网络数据挖掘，计算机并行与分布处理等。