硕 士 学 位 论 文

社区问答系统中问句检索技术的研究

The Study on Question Retrieval Technology in Community Question Answer System

作 者 姓 名：杨海天

学科、 专业：计算机应用技术

学　　号：21109229

指 导 教 师：王健 副教授

完 成 日 期：2014.05

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：社区问答系统中问句检索技术的研究

作 者 签 名 ：_____ 日期：_____年____月____日

摘　　要

互联网技术的发展给人们日常生活带来便利的同时，也使人们淹没在信息的海洋中，很难找到自己所关心和需要的信息。随着 web2.0 的飞速发展，面对传统搜索引擎暴露出来的诸如不能对于专业的问题进行有效的检索、无法给用户带来交互式的体验等问题，近年来出现的社区问答（CQA）系统在一定程度上弥补了这些缺陷，正在给用户带来全新的搜索体验。在社区问答系中，人们可以自由地提出自己的问题，并由其他用户回答。由于任何人都可以在上面提问和回答，Yahoo! Answers 等社区问答系统建立几年来已经积累了大量的历史问答对，如何有效的利用这些问答对成为众多学者研究的焦点。问句检索的研究就是为了能够有效地利用这些历史的问答对信息，快速找到与用户所关心的问题相同或相近的原有问题，缩短用户得到想要的答案的等待时间。但是，由于自然语言中存在大量的同义词、语义特性和丰富的句法特征，所以从社区问答系统中找到相似的问句并不是一项轻松的任务。

本文主要是对问句检索进行研究，主要是解决了问句检索过程中的存在的三个问题，首先是解决了问句检索过程中缺少语义信息造成的问句歧义性问题，由于自然语言中存在大量的同义词、语义特性和丰富的句法特征，所以单纯的仅依靠词本身的特征很难解决问句检索的问题。针对这个问题，本文提出一种基于特征融合的社区问答问句相似度计算方法，它主要是利用问句本身的统计特征、词序特征、语义特征和问句对应的答案特征相结合来解决问句检索问题。

其次是解决了问句检索过程中效率问题，在解决检索效率问题中，本文提出一种融合问句类别信息和问句对应答案类别信息的问句检索模型，该模型主要是考虑了问句的类别信息和问句所对应答案的类别信息，利用类别信息来过滤掉不相关的问句，从而提高问句检索的效率和性能。

最后解决的问题是由于误分类对检索结果造成影响的问题，针对这个问题，本文提出一种融合问句主题信息和问句对应问句答案主题信息的问句检索模型，该模型主要考虑了问句本身的主题信息和问句所对应的答案主题信息，利用主题信息对相似的问句类别进行合并，从而减轻误分类对检索结果的影响。最终将这三种模型分别在 Yahoo! Answers 网站上抽取的真实标注数据集上进行实验，并通过多角度的对比实验表明，针对各自要解决的问题，本文提出的模型取得了良好的性能。

关键词：社区问答；搜索引擎；问句检索；问句相似度

The Study on Question Retrieval Technology in Community Question Answer System

Abstract

The development of Internet technology brings convenience to people's daily life, and also makes people drown in the ocean of information; people find it is hard to find their own concerned and needed information. With the rapid development of Web2.0, and more and more problems exposed in traditional search engines, such as not giving professional problems effective retrieval and unable to give users an interactive experience. Community Question Answering (CQA) emerged in recent years makes up for these shortcomings to a certain extent, and nowadays provides users a new search experience. In the Community Question Answering system, people can put forward their own questions freely, and answered by other users. Since anyone can ask and answer questions on it, some Community Question Answer systems such as Yahoo! Answers have accumulated a lot of question-answer pairs. How to use these question-answer pairs effectively becomes the focus of many scholars. The study of question retrieval is to use the former question-answer pairs effectively to find the same or similar questions of the user concerned one, so that to short the waiting time of users. However, as there is a large amount of synonyms, semantic features and syntactic characteristics in natural language, it is not an easy task to find the similar questions in the Community Question Answering systems.

This paper mainly focuses on question retrieval, and concentrate to solve the three problems of the questions in the process of retrieval. The first is to solve the questions ambiguity problem caused by lacking of semantic information during the process of retrieval. In natural language, there is a large amount of synonyms, semantic features and syntactic characteristics so that just only depending on the characteristics of the word itself is difficult to solve the retrieval problem of the question. Aiming at this problem, we propose a similarity calculation method of CQA questions based on feature fusion, which mainly uses statistical characteristics, word order features, semantic features, and the answer features related to the question to solve questions retrieval problem.

The second is that it will promote the efficiency for problem questions in the process of retrieval. In the section of solving the problem of retrieval efficiency, this paper puts forward a kind of retrieval model which fuses category information of the questions together with category information of the questions' corresponding answer. This model is mainly

considered question category information and the corresponding answer category information, using category information to filter out irrelevant questions, so as to improve the efficiency and performance of the questions retrieve.

The last point is the problem caused by misclassification which impacts the retrieval result. Aiming at this problem, this paper puts forward a kind of retrieval model which fuses questions' topic information together with the topic information for corresponding questions' answers. This model is mainly considered question topic information and the topic information for corresponding questions' answer, making use of topic information to merge the similar questions categories, so as to reduce side-effect caused by the misclassification for retrieval results. Finally, with these three models, we experiences separately on the actual annotation data sets from Yahoo! Answers, and through comparative experiments from the multiple perspectives, it shows that, these three models we have mentioned, which respectively focus on their own task, achieved good performance.

Key Words: Community Question Answer; Search Engine; Question Retrieval; Question Similarity

目　　录

1 绪论

1.1 研究背景

随着科学技术的发展，尤其是随着 web2.0 的普及，当今时代已经成为一个网络的时代，互联网逐渐的走进千家万户，人们生活的方方面面也逐渐离不开互联网。人们可以轻松自由的通过互联网，对自己所关心和感兴趣的问题进行关注和追踪。但互联网给人们生活带来方便的同时，也带来了一些信息过载的现象，使人们淹没在爆炸式的信息海洋中，很难快速直接有效的获取到人们所关心和关注的问题。并且随着互联网融入人们的生活，人们对信息的需求和质量都在逐渐的提高。这就造成国外各大互联网公司对信息检索领域激烈的竞争[1-2]，因为这里面蕴藏了一个全球性巨大的商机。国内互联网公司，如百度、360 和搜狗等对信息检索领域的竞争也是一个有力的证明。

搜索引擎的出现和发展，使人们能够更好的利用互联网的资源来查找信息，并且具备了与用户交互的功能，深受用户的喜欢。从表面上看，好像是解决了人们在互联网中面对的信息过载问题。但是，搜索引擎有着自己天然的弊病，即它对用户提出的问题，只是返回一些相关的文档，并没有对其进行智能的整理，给用户提供最精确的信息。因此用户仍然需要花费大量的精力在这些相关的文档中，寻找对自己真正有用的信息。并且在对特定专业领域问题的检索时，更是表现出了传统搜索引擎的缺陷。

为了弥补传统搜索引擎暴露出了缺陷和不足，在相当长的时间内，问答系统成为众多学者研究的焦点。问答系统之所以成为研究的焦点，主要是因为相对于传统的搜索引擎，问答系统提供给用户的是直接的答案，并不是一些相关的文档。这样就省去了用户去相关文档中查找自己所需要信息的时间，从而更受用户的欢迎。但是由于早期的问答系统完全依靠机器理解来自动的产生答案，这种形式产生的答案比较机械和呆板，同时答案的质量也比较不理想，不能满足用户真实的需求。在这种情况下，社区问答系统应运而生。社区问答系统给用户带来的是一种前所未有的搜索体验，由于在社区问答系统中，人们没有任何拘束，可以自由提出自己所感兴趣或关心的问题，并且可以回答别人提出的问题。社区问答系统这种由现实生活中的人给出问题答案的方式，直接克服了传统问答系统由机器自动产生答案质量不理想的缺点，给广大用户带来全新的体验。

由于早期国内的互联网发展速度比较缓慢，因此社区问答系统在国内出现的比较晚，但在国外社区问答很早就出现了。早在 2000 年，基于用户参与交互并相互学习的社区问答在 Empas 就已经形成，并且它的这种架构对后面社区问答系统的发展有着深远的影响。到 2002 年，韩国社区问答 Naver 正式上线，它采用的架构是基于知识共享和

搜索引擎相结合，这种架构成为后来社区问答系统的基础原型。随后，在 2005 年 7 月，基于 Naver 原型的社区问答 Yahoo! Answers 正式上线，并迅速成为目前最大的社区问答系统。在 Yahoo! Answers 中用户首先要注册，注册成为会员以后，可以登录到系统，并且在系统中可以提问自己所关心和感兴趣的问题，同时也可以对别的用户提问的问题进行回答。为了能够鼓励更多的用户在社区问答系统中踊跃的提问和回答，系统对于经常提问和回答的用户给予相应的积分来进行奖励，并且根据积分的多少评选出相应的活跃用户。国内互联网巨头百度和新浪，在 2005 年才意识到社区问答系统的重要性。并且迅速发展自己的社区问答系统，也就是大家所熟悉的百度知道和新浪爱问。近些年随着国内互联网的蓬勃发展，众多互联网公司都意识到了社区问答系统的重要性，因此短短的几年里就迅速涌现出了一批优秀的社区问答系统，如天涯问答、搜搜问问、雅虎知识堂等等。

经过几年的发展社区问题系统中已经积累了大量的问题与答案，即“问答对”。如何能够有效地利用这些历史的问答对信息，快速找到与用户关心问题相同或相近的原有问题[3]，缩短用户得到想要的答案的等待时间成为众多研究者所关心的问题。问句检索在 CQA 中主要是针对用户提出来的新问题，在历史问答对中检索出与用户最相关的问题，从而减少用户等待的时间，给用户带来更好的体验。因此问句检索在社区问题系统中的研究占有举足轻重的地位。

1.2 国内外研究现状

目前用户参与度比较高的社区问答系统有：雅虎问答、百度知道，它们允许用户提出自己的问题，并且从其他用户那里得到详细复杂的答案。然而回答社区问答系统中的问题需要依赖用户自身的能力和自愿性，所以社区问答系统中还有相当一部分比例的问题没有答案。为了降低这些没有答案问题的比例和减少用户等待答案的时间，社区问答系统中的问句检索任务就变得极为重要。Liu 等人[4]提出一种新的混合方法，根据目标问句的类别信息，同时考虑用户的主题相关性，用户的声誉等，来有效的发现社区问答系统中的专家用户。Toba H 等人[5]采用一种层次结构的分类器在问答对语料库中去挖掘高质量的答案，在挖掘高质量答案过程中除了利用分类器外也涉及到了问句的检索。Shtok 等人[6]采用统计的方法，利用过去已经解决的问题，来尽可能多地回答还没有解决的问题，并取得了不错的效果。Riahi 等人[7]使用统计主题模型来研究问句主题，然后采用分段主题模型来研究社区问答系统中专家用户发现的问题，并与 LDA（Latent Dirichlet Allocation）相比，在性能上有了一定的提高。Chen 等人[8]利用机器学习的方法根据文本和元数据特征建立预测模型来预测用户问题的意图，并根据用户意图推荐相关

的答案。上述社区问答系统中的其他研究都涉及到了问句检索任务，所以做好社区问答系统中问句检索任务是十分必要的。

由于早期自动问答系统（frequently asked question ，FAQ）的兴起，早期的学者对于问句检索的研究基本都集中于自动问答系统，经过一段时间的研究，对于问句的检索已经产生了一些高效的方法。Burke 等人[9]采用词语相似度和语义相似度相结合并采用向量空间模型的方法来解决问句之间的检索问题。Berger 等人[10]通过学习多种统计的方法来解决问句之间的词稀疏问题，从而提高问句检索的性能。Jijkoun 等人[11]采用无监督学习的方法从 FQA 网页中抽取出问答对，然后采用向量空间模型进行问句检索。Riezle 等人[12]采用翻译模型来解决问句的检索问题。上述的这些方法大部分集中于解决问句之间的词不匹配问题，仅仅利用了词的表面信息，并没有挖掘句子背后隐藏的主题类别信息。Song 等人[13]提出一种基于语义信息和统计信息相融合的方法来解决 FAQ 中问句相似度检测的问题，该方法主要是较好的解决了数据集稀疏问题，在考虑问句本身的统计信息的同事把问句隐藏的语义信息引入到句子相似度计算中，取得了不错的效果。但是，Song 等人[13]提出的这个方法中的统计信息仅仅利用了词表面的信息，并没有深入的对不同的词，表示相同的意思，和不同的词在一个句子中所属的地位不同等问题进行有效的解决。

随着近期社区问答系统的蓬勃发展，尤其是一系列国际会议（ACL，EMNLP，COLING，SIGIR，WWW，CIKM）对社区问答中问句检索研究的支持，近期学者们的研究大部分集中于社区问答系统中问句检索的研究，并且涌现出了一些新的方法。Jeon 等人[14-15]比较了多种检索模型（向量空间模型，语言模型和翻译模型）等在社区问答系统中问句检索的性能。他们随后[15]提出了基于翻译的语言模型，并且把问题的答案信息也融入到了自己的模型中。Lee 等人[16]提出基于问答对，通过问答对中的重要特征词来加强翻译模型，来改善翻译的概率。Bernhard 等人[17]提出使用平行的语料集作为训练数据集，来通过不同的语义资源，去确定一个词的翻译概率。Duan 等人[3]提出通过识别问句的主题和问句的焦点来进行问句检索。Surdeanu 等人[18]提出一种通过 Yahoo! Answers 中的多种特征融合来对答案进行排序，进而进行问句检索。Wang 等人[19]提出一种只利用一种特征信息对候选的答案经行排序而不是利用多种特征的融合。

以上这些方法虽然在社区问答问句检索中表现出了良好的性能，但是这些方法仅仅以特征向量为处理对象，难以表示结构化的特征，存在数据稀疏问题。接下来 Collins 等人[20]提出了一种树核方法，通过计算两棵句法树之间的相同树片段的数量来比较句法树之间的相似度，但没有区别节点的深度特征和句法成分特征。Wang 等人[21]通过使用

树核对结构化特征进行建模，从而计算两个句子之间的相似度问题，但在相似度计算过程中没有考虑语义信息。

1.3 本文主要工作

在社区问答系统中，问句是核心，人们可以在社区问答系统中提出自己的问题，并由其他用户回答。由于任何人都可以在上面提问和回答，因此社区问答系统成立几年来已经积累了大量的问答对。如何有效地利用这些已有的问答对来回答用户提出的问题成为研究的焦点，本文正是关于这一焦点做的研究。

在问句检索方面，本文首先提出一种基于特征融合的社区问答问句相似度计算方法，该方法主要是利用问句本身的统计特征、词序特征、语义特征和问句对应的答案特征相结合来解决问句检索问题。实验在 Yahoo! Answers 上抽取的真实标注数据集上进行，对比实验结果表明，该方法在性能上得到了较好的结果。在随后的研究中发现问句的类别在问句的检索中起到了举足轻重的作用，首先对于一个查询问句，如果能先识别其类别，然后把问句对应到相应的类别中去检索，不仅在性能上而且在效率上都会有明显的提高。因此，本文提出一种融合问句类别信息和问句对应答案类别信息的问句检索模型，该模型主要是考虑了问句的类别信息，利用问句的类别信息对语言模型进行平滑。同时考虑问句所对应的答案信息，因为我们研究发现，在社区问答系统中，问句通常比较短，但问句所对应的答案信息通常都比较长，所以问句的答案信息能更加清楚的表达问句本身的含义。因此我们把问句的类别信息和问句所对应的答案信息融合起来，对比实验结果表明本文提出的方法取得了较好的性能。

最后，我们发现 Yahoo Answer 中的问句分类是一个很大的研究领域，对于上面提出的融合问句类别信息和问句答案类别信息的模型，能获得较好的前提是基于分类器的准确率。但分类问题在社区问答系统中是一个大规模的层次分类问题[22-23]，目前还没有出现特别有效的分类方法来解决这个问题。因此本文又提出了一种融合问句主题信息和问句对应答案主题信息的问句检索模型，该模型主要是除了考虑问句本身的主题信息外，又考虑了问句对应答案的主题信息。利用主题信息对相似的问句类别进行合并，从而减轻误分类对检索结果的影响。对比实验结果表明，该方法能够有效的减轻误分类对检索结果的影响，从而表现出了更好的性能。

1.4 本文组织结构

本论文共分为五章，首先对社区问答系统国内外目前研究现状进行了详细的介绍，然后对目前社区问答系统中存在的典型检索模型进行了概述，最后对于社区问答系统中

的问句检索，提出三种模型，并详细介绍了三种模型的算法思想、框架结构并进行相应的对比实验分析。具体章节安排如下。

第 1 章，综述了本论文所研究的社区问答中问句检索相关技术的背景以及目前国内外研究现状，阐述了本文提出的方法内容，并安排了本论文的章节结构。

第 2 章，详细介绍了社区问答系统中问句检索的相关技术和实现方法，主要包括向量空间模型、BM25 模型、语言模型、翻译模型和基于翻译的语言模型，并分别阐述了这五种模型从问句的不同用度来实现问句的检索。

第 3 章，详细介绍了本文提出的基于特征融合的问句相似度计算方法，该方法主要利用问句本身的词序特征、统计特征、语义特征和问句对应答案的答案特征相结合来解决问句检索问题。同时对实验所用的数据集进行了详细的介绍，并与目前典型的一些问句检索方法进行了对比实验，最后对实验结果进行了详细的分析。

第 4 章，详细介绍了本文提出的融合问句类别信息和答案类别信息的问句检索模型，该模型主要是利用了问句的类别信息和问句对应的答案类别信息分别对语言模型进行平滑，最后线性结合两个平滑的结果，对问句进行检索。同时对实验所用的数据集进行了详细的介绍，并与目前典型的一些问句检索模型进行了对比实验，最后对实验结果进行了详细的分析。

第 5 章，详细介绍了本文提出的融合问句的主题信息和问句对应答案主题信息的问句检索模型，该模型首先使用 LDA 分别提取问句集合中的主题信息和问句对应答案集合的主题信息，然后利用问句的主题信息和问句所对应答案的主题信息分别对语言模型进行平滑，最后线性结合两个平滑结果，对问句进行检索。同时对实验所用的数据集进行了详细的介绍，并与目前典型的一些问句检索模型进行了对比实验，最后对实验结果进行了详细的分析。

最后对本文所研究的问句检索进行了详细的总结，并提出了下一步的研究方向。

2 问句检索相关技术及实现方法

社区问答系统中积累了大量的“问答对”，并且随着时间的推移，问答对的数量还在急剧的增加。因为每天都有大量的用户参与到社区问答系统，并且提出很多现实生活中的相关问题，同时在社区问答系统中也有相当一部分用户去回答这些问题。但是回答社区问答系统中的问题需要依赖用户自身的能力和自愿性，所以社区问答系统中还有相当一部分比例的问题没有答案。为了降低这些没有答案问题的比例和减少用户等待答案的时间，问句检索的任务就变得极为重要。问句检索在 CQA 中主要是针对用户提出来的新问题，在历史问答对中检索出与用户最相关的问题。下面分别介绍几种最常见的问句检索模型。

2.1 向量空间模型

传统的向量空间模型（vector space model，VSM）主用是用于文本检索中，以其框架清晰简单且效果较好而著称。随着社区问答系统的兴起，很多学者把这一模型应用到了社区问答系统的问句检索中[15,24]，并且取得了一定的效果。但向量空间模型自身也有一定的缺陷性，它假设问句中词与词之间相互独立，彼此之间没有任何关联。之所以这样假设，主要是为了降低问句相似度计算的复杂程度，但却忽略了词与词之间的语义相关性[25]，同时也忽略了词自身所具有的语义信息[26]。

在社区问答系统中，给定一个查询 q 和在历史问答对中的问句 d，采用向量空间模型相似度计算公式如下：

$$P(q \mid d) = \frac{\sum_{t \in q \cap d} w_{q,t} \times w_{d,t}}{\sqrt{\sum_t w^2_{q,t}} \sqrt{\sum_t w^2_{d,t}}} \tag{2.1}$$

$$w_{q,t} = \ln(1 + \frac{N}{f_t}), w_{d,t} = 1 + \ln(f_{d,t}) \tag{2.2}$$

这里 $w_{q,t}$ 表示单词 t 在文档集合中的 IDF(inverse document frequency)，$w_{d,t}$ 表示单词 t 在文档 d 中的词频。N 表是问句集合中的所有问句的数量。f_t 表示有多少个问句中含有单词 t，tf_d 表示词在问答 d 中出现的频率。

2.2 BM25 模型

BM25 模型在问句检索中的应用主要是弥补了 VSM 模型没有考虑到句子长度的缺陷[27]。在文献[24]中，在历史问答对语料库中，一个查询 q 和历史问答语料库中的问句 d 之间的相似得分可有如下公式计算：

$$P(q \mid d) = \sum_{t \in q \cap d} w_{q,t} \times w_{d,t} \tag{2.3}$$

$$w_{q,t} = \ln(\frac{N - f_t + 0.5}{f_t + 0.5}) \frac{(k_2 + 1)tf_{q,t}}{k_2 + tf_{q,t}} \tag{2.4}$$

$$w_{d,t} = \frac{(k_1 + 1)tf_{d,t}}{K_d + tf_{d,t}}, K_d = k_1((1 - b) + b\frac{|d|}{W_A}) \tag{2.5}$$

这里 k_1， b，和 k_2 是参数，分别为 1.2，0.75 和无穷。|d|是问句 d 的长度，w_A 是问句集合中的所有问句的平均长度。

2.3 语言模型

2.3.1 一元语言模型

一元语言模型通常假设文档中每个特征词都是独立的，它主要是关注于一个简单采样对一个特征词的最大似然概率。为了避免零概率问题，在这里使用 Jelinek-Mercer 平滑[28]，主要是因为采用这种平滑，不仅能得到较好的结果而且在计算时间上也有很小的消耗。采用 Jelinek-Mercer 平滑的语言模型公式如下:

$$P(q \mid d) = \prod_{w \in q} P_{LM}(w \mid d) \tag{2.6}$$

$$P_{LM}(w \mid d) = (1 - \lambda)P_{ml}(w \mid d) + \lambda P_{ml}(w \mid C) \tag{2.7}$$

$$P_{ml}(w \mid d) = \frac{\#(w, d)}{|d|}, P_{ml}(w \mid C) = \frac{\#(w, C)}{|C|} \tag{2.8}$$

这里 C 表示问句集合，λ 是一个平滑参数，$\#(w, d)$ 表示特征词 w 在问句 d 中出现的频率。|d|和|C|分别表示候选问句 d 和查询问句 q 的长度。

2.3.2 翻译模型

早期的学者[15,16,17,24,29]一致认为翻译模型在问句检索中可以取得卓越的性能。该模型主要是把特征词转移概率应用到了语言模型的框架中。在文献[15，29]中，采用翻译模型来计算候选问句 d 和查询问句 q 的相似度得分公式为：

$$P(q \mid d) = \prod_{w \in q} P_{TR}(w \mid d) \tag{2.9}$$

$$P_{TR}(w \mid d) = (1-\lambda)[\sum_{t \in d} P(w \mid t) P_{ml}(t \mid d)] + \lambda P_{ml}(w \mid C) \tag{2.10}$$

$$P_{ml}(t \mid d) = \frac{\#(t,d)}{\mid d \mid} \tag{2.11}$$

这里 $P(w \mid t)$ 表示特征词 t 到特征词 w 的翻译概率，Jeon 等[15]将设一个特征词自身到自身的概率是 1，即 $P(t \mid t) = 1$。但单纯基于特征词的翻译模型不能够提供更多的上下文信息，没有上下文信息就很难解决词歧义性问题，为了解决这个问题，近期的学者在翻译模型的基础上又提出了新的模型。Zhou 等人[30]提出了基于短语的翻译模型来用于问句检索。基于短语的翻译模型在模型翻译过程中把短语作为一个整体，因此能获得更多的上下文信息。这将有助于缓解数据的稀疏性和词的歧义性。Singh 等人[31]通过扩充词汇把更多的语义信息融入到基于特征词的翻译模型中来解决词汇的稀疏问题。

2.3.3 基于翻译的语言模型

Xue 等[29]提出把语言模型和翻译模型采用线性结合的方法把两个模型建立到统一的框架中去，并把框架取名为基于翻译的语言模型。文献[29]展示了基于翻译的语言模型比单独的语言模型和单独的翻译模型在问句检索中取得了更好的性能，并提出了这个模型的公式为：

$$P(q \mid d) = \prod_{w \in q} P_{TRLM}(w \mid d) \tag{2.12}$$

$$P_{TRLM}(w \mid d) = (1-\lambda)[\alpha \sum_{t \in d} P(w \mid t) P_{ml}(t \mid d) + (1-\alpha) P_{ml}(w \mid d)] + \lambda P_{ml}(w \mid C) \tag{2.13}$$

这里参数 α 控制翻译部分的权重，α 越大翻译部分的权重越大，反之越小。

2.4 本章小结

本章主要介绍了问句检索的相关技术和实现方法，主要介绍了五种检索模型：向量空间模型、BM25 模型、语言模型、翻译模型、基于翻译的语言模型。五种模型各有优缺点，有其适应的领域情况。向量空间模型和 BM25 模型主要是以其简单、快速著称，但相对于后三种模型效果一般。语言模型、翻译模型和基于翻译的语言模型，这三个模

型中，后两种模型其实从公式上可以看出与语言模型非常相似，只是基于翻译的模型用到了平衡语料，即要有翻译对（即原问句和对应的翻译问句），所以效果上要好于语言模型。但翻译对的获取是极其不容易的。

本文提出的主要检索模型也是基于语言模型的，从效率方面考虑，本文提出了融合问句类别信息和问句对应答案类别信息的语言模型，即对于一个查询，先确定其类别，然后在其相应的类别中检索，这样就过滤掉在不相关类别中的检索，从而提高检索效率。从性能上考虑，本文提出了融合问句主题信息和问句答案主题信息的语言模型，相对于融合问句类别信息和问句对应答案类别信息的语言模型，融合问句主题信息和问句答案主题信息的语言模型，其实是在类别的基础上，考虑了类别背后隐藏的主题信息，也相当于对于对相似类别之间进行合并，合并为一个更大的主题。这样就可以减轻基于类别的误分类问题对检索结果的影响，从而表现出更好的性能。

3 基于特征融合的问句相似度计算方法

本章主要从以下几个方面来阐述基于特征融合的问句相似度计算方法：方法思想的起源、方法用到的五种模型(词序相似度模型、改进的统计模型、问题的主题和焦点确定模型，语义模型和基于答案信息的模型)，并通过实验设计和对比结果分析来论证方法的有效性。

3.1 引言

近年来社区问答系统（Community Question Answering, CQA）已经成为在线寻求帮助信息的有效方法，除了使用通用的网络搜索引擎，如今人们可以有另外的一种选择那就是社区问答系统，如 Yahoo! Answers 等。人们可以在社区问答系统中提出自己的问题，并由其他用户回答。由于任何人都可以在上面提问和回答，Yahoo! Answers 等成立几年来已经积累了大量的问答对。针对传统搜索引擎暴露出来的缺点，不能对特定领域和专业性的问题进行回答，而社区问答系统正好弥补了这方面的问题，所以大量的研究者投身于社区问答系统的研究中。

社区问答系统中存在大量的问答对，如何有效地利用这些已有的问答对来回答用户提出的问题成为研究的焦点。本章正是关于这一焦点做的研究，并在相似度计算过程中充分利用问题的答案信息。由于自然语言中存在大量的同义词、语义特性和丰富的句法特征，所以从 CQA 系统中找到相似的问句并不是一项轻松的任务。比如“How can I lose weight in a few months？”和“Are there any ways of losing pound in a short period？”都是关于寻求减肥方法的问题，但是它们几乎不含有任何相同的词汇，并且具有不同的句法特征，传统的利用词频的方法很难解决上述的问题。又比如“How many London to New York flights are there in a day？”和“How many New York to London flights are there in a day？”都是询问一天飞机的航班情况，并且所有的单词都一样，不过一个是询问从伦敦到纽约，另一个是询问从纽约到伦敦，这里单纯利用传统的基于词频的方法是很难解决的。

针对上述的问题和社区问答系统的特殊性，本章提出一种将问句本身的词序特征、统计特征、语义特征和问句所对应的答案特征进行融合的方法，目的在于更加充分地挖掘句子之间的关系，从而有效地解决句子相似度计算的问题。

3.2 检索模型概述

3.2.1 算法思想

首先从 Yahoo! Answers 网站中抽取“问答对”语料集，然后建立问题集索引和问题对应的答案集索引。在建立索引时，只对已经解决的问题建立问答对索引，即保证索引中每一个问题都有一个最佳的答案。对于任意的一个查询 q，在我们的模型中首先使用改进的统计模型（ISM）和词序相似度线性结合，得到最初结果。对于每个查询 q 返回排名最前的 100 个结果，其次再用 3.2.4 节介绍的（QTFD）方法对每个问句确定其句子中的关键词，对确定的关键词赋予较高的权重，并重新用改进的统计模型进行检索，这次返回排名最前的 1000 个结果，接着再使用 3.2.5 节介绍的语义模型对这次检索的结果再次进行检索，最后利用 3.2.6 节介绍的融合答案信息的方法重新检索，得到的最终结果。根据最终结果，从相应的答案索引中提取出最佳答案返回给用户。算法的流程图如图 3.1 所示。

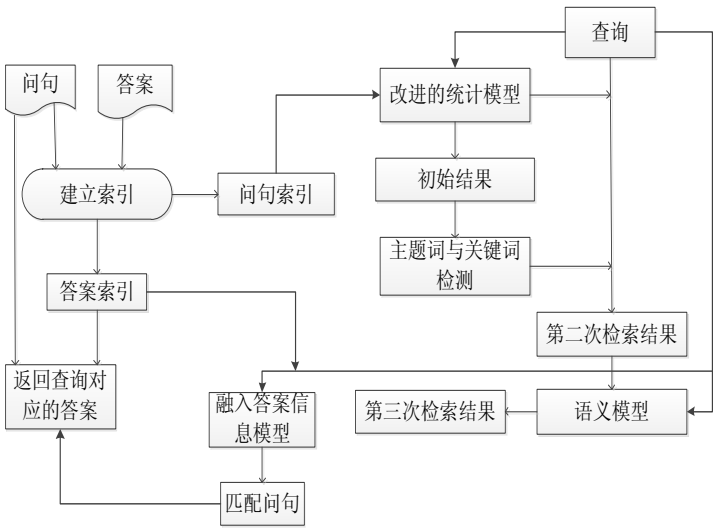

图 3.1 算法流程图

Fig. 3.1 Overview of algorithm

3.2.2 词序相似度

通过引言中的例子可以看到词序对于区分问句是十分有必要的，本章采用的词序相似度(OrdSim)方法是文献[32]提出来的，它主要是反映两个问句中所含同义词或相同词在位置关系上的相似程度，同时以两个句子中所含相同词或同义词的相邻顺序逆向的个数来衡量两个问句的相似程度。

设Q_1和Q_2为两个句子，$OnceWord(Q_1,Q_2)$为Q_1、Q_2中都出现且只出现一次的单词集合，$Pfirst(Q_1,Q_2)$表示$OnceWord(Q_1,Q_2)$中单词在Q_1中的所在位置序列构成的向量，$Psecond(Q_1,Q_2)$表示$Pfirst(Q_1,Q_2)$中的分量按对应单词在Q_2中的词序排序生成的向量，$RevOrd(Q_1,Q_2)$为$Psecond(Q_1,Q_2)$各相邻分量的逆序个数。则Q_1、Q_2的句序相似度计算公式如下：

$$OrdSim(Q_1,Q_2)=\begin{cases}1-\dfrac{RevOrd(Q_1,Q_2)}{|OnceWord(Q_1,Q_2)|-1}, & 当|OnceWord(Q_1,Q_2)|>1\\ 1, & 当|OnceWord(Q_1,Q_2)|=1\\ 0, & 当|OnceWord(Q_1,Q_2)|=0\end{cases} \quad (3.1)$$

3.2.3 改进的统计模型

大家所熟悉的文档相似度计算一般使用向量空间模型（VSM：Vector Space Model），这是因为文档中含有大量丰富的单词，采用向量空间模型把文档中的每个单词表示成一维向量，从而构成的高维向量可以很好地表示文档。但对于社区问答系统中的问句相似度计算，由于社区问答系统中的问句一般由很少的单词构成，如果仍采用高维空间来表示句子，则会存在较大的稀疏问题，从而得到很不理想的结果。针对向量空间模型在问句检索中的不足，Song 等人[13]提出的统计模型（SM：Statistical Model）SM，它的定义为： $QS = Q_1 \cup Q_2$，其中 QS 表示问句 Q_1 和问句 Q_2 的并集（即 QS 表示两个句子中只出现一次的单词集合）。V_1 和 V_2 表示 Q_1 和 Q_2 所对应的向量，维度等于 QS 中单词的个数，分量遵循如下两个规则：

（1）对于 QS 中的单词，如果在句子 Q_i（i=1,2）中不存在，则对应 V_i（i=1,2）中的分量为 0。

（2）对于 QS 中的单词如果在句子 Q_i（i=1,2）中存在，则对应的 V_i（i=1,2）中的分量为该单词在句子 Q_i（i=1,2）中出现的频率。

对于 V_1 和 V_2 采用 cosine 来计算其相似度，计算公式如下：

$$Sim_{statistic} = \frac{v_1 \cdot v_2}{\|v_1\| \cdot \|v_2\|} \quad (3.2)$$

本章中所采用的统计模型（ISM：Improved Statistical Model）主要是在以上基础上做了以下改进：

（1）利用前面提到的语言学知识，对句子中的动词和名词比句子中其他词性的词赋予较高的权重。

（2）利用 wordnet 对句子中的单词进行同义词扩充。

（3）使用了如下与公式（3.2）有所不同的相似度计算方法，计算公式如下：

$$Sim_{statistic}=\frac{v_1 \cdot v_2}{||v_1||+||v_2||} \tag{3.3}$$

3.2.4 问题的主题和焦点确定

众所周知在问句检索过程中，问句的主题信息将起到至关重要的作用，例如我们在查询关于“apple”的问题，可能是关于一种水果问题，也可能是关于苹果公司产品的问题，对于单纯的利用词本身信息很难确定。但如果我们能确定问句背后的相关主题，就可以很容易确定是关于水果还是关于苹果公司的产品。本章中问题主题和焦点信息的确定(QTFD：Question Topic and Focus Determination)采用文献[33]中的方法。对每个查询 Q 首先采用 3.2.3 节介绍的 ISM 模型进行检索，把检索出来的前 N 个问句和 Q 作为一个局部数据集，并在此数据集中对 Q 中的每个单词 w 计算局部 $tf \cdot idf$，计算公式为：

$$loc\_tf \cdot idf = tf_{wi} * \log(loc\_docNum)/loc\_df_{wi} \tag{3.4}$$

然后再把语料库中所有的问句和查询 Q 作为一个全局数据集，并在此数据集中对查询 Q 的每个单词 w 计算全局 $tf \cdot idf$，计算公式为：

$$glob\_tf \cdot idf = tf_{wi} * \log(glob\_docNum)/glob\_df_{wi} \tag{3.5}$$

最后根据 $rank(glob\_tf \cdot idf)$ 和 $rank(loc\_tf \cdot idf)$ 的比值对 q 中的每个单词 w 进行排序，排序公式为：

$$wRank = rank(glob\_tf \cdot idf)/rank(loc\_tf \cdot idf) \tag{3.6}$$

其中 $rank(glob\_tf \cdot idf)$ 和 $rank(loc\_tf \cdot idf)$ 的值为 Q 中的每个单词 w 分别根据 $loc\_tf \cdot idf$ 和 $glob\_tf \cdot idf$ 在 Q 中的排序值。

最终对于每个查询 Q，根据 $wrank$ 值排序，取前 50%的词作为问句的主题信息词，采用 3.2.3 节介绍 ISM 模型重新进行检索，并在这次检索过程中对问句中的主题信息词赋予较高的权重。在这里对于查询 Q 中的主题信息词的权重，我们将确定好的主题信息词的原始 $wrank$ 值乘以 100 作为新的主题信息词的权重。之所以采用这种策略，主要是因为根据 $glob\_tf \cdot idf$ 和 $loc\_tf \cdot idf$ 比值得出来的结果比较小，直接用这个值作为主题信息词的权重不足以显示主题信息词的重要性，所以才对 $wrank$ 值乘以 100 作为主题信息词的权重，从而更好表示主题信息词的重要性。

3.2.5 语义模型

本章所介绍的语义模型（SEM：Semantic Model）使用 WordNet 作为语义资源。WordNet 是由同义词词集构成的分层的结构化网络，同义词词集之间通过指针相互连接。在 WordNet 中两个单词之间的距离越近则认为它们之间的语义相似性越大，反之，则认为语义相似性越小。故可使用如下公式（3.7）来计算两个单词的语义相似度。对于给定两个特征词 w_1 和 w_2，利用 WordNet 提供的最短路径接口函数来得到两个同义词间的最短路径，然后进行归一化得到两个特征词之间的语义相似度，其语义相似度公式如公式（3.7）所示：

$$Sim(w_1,w_2)=\frac{1}{dis(w_1,w_2)+1} \tag{3.7}$$

其中 $Sim(w_1,w_2)$ 表示 w_1 和 w_2 之间的语义相似度，$dis(w_1,w_2)$ 表示 w_1 和 w_2 WordNet 中的语义的最短距离。公式（3.7）很好地表现了两个单词的相似度随语义距离的增大而减小的特点。另外，规定相同单词之间的语义距离为零，此时的相似度为 1。

由上述方法计算出词语的语义相似度后，接下来我们利用特征词之间的语义相似的来计算问句之间的语义相似度，对于给定的两个问句 Q_1 和 Q_2，首先对问句 Q_1 和 Q_2 进行去停用词和词干化操作，然后再采用二分图[13]的方法来进一步计算两个句子之间的相似度，不仅要计算出问题 Q_1 对问题 Q_2 的相似度，而且还要计算出问题 Q_2 对问题 Q_1 的相似度，然后取两者的平均值作为最终问题 Q_1 对问题 Q_2 的语义相似度。其计算公式如公式为：

$$Sim_{semantic}=\frac{1}{2}\left(\frac{\sum_{a_i\in Q_1}\max ssim(a_i,Q_2)}{||Q_1||}+\frac{\sum_{b_j\in Q_2}\max ssim(b_j,Q_1)}{||Q_2||}\right) \tag{3.8}$$

其中 Q_1 和 Q_2 表示给定的两个问句，a_i 和 b_j 分别表示 Q_1 和 Q_2 的特征词，$||Q_1||$ 和 $||Q_2||$ 分别表示 Q_1 和 Q_2 中特征词的个数，max 的定义如下：

$$\begin{cases}\max ssim(a_i,Q_2)=\max\left(sim(a_i,b_1),sim(a_i,b_2),\cdots,sim(a_i,b_{|Q_2|})\right)\\ \max ssim(b_j,q)=\max\left(sim(b_j,a_1),sim(b_j,a_2),\cdots,sim(b_j,a_{|Q_1|})\right)\end{cases} \tag{3.9}$$

3.2.6 基于答案信息模型

在社区问答系统中对于已解决的问题，每个问题都有最佳的答案匹配，这些最佳答案都是由提问者选出或者其他用户投票选出的，因此具有很高的可信度。同时最佳答案一般是由社区里的用户经过精心思考回答得出的，并且这些用户一般是在某个领域有一定特长的，因此所给出的答案信息比较精确，并且一般给出的答案都比问题要长，所以

问句的答案比句子含有更丰富的信息。本章基于答案信息模型(BAIM: Based on Answers Information Model)思想来源于文献[34]的方法（Q+ACV：Question Integrated With Answer Context Vector），利用问题的答案对句子进行扩展，使句子具有更丰富的信息。众所周知传统的查询扩展通常采用基于基础查询日志[35]和相关反馈[36]方式，这两种方式主要是以提高召回率为目标，而本章采用的查询扩展的目的主要是为了更好地表示句子信息，从而更有利于计算句子的相似度。查询扩展的方法如下：

令 x 表示一个新的问题，则 x 的上下文向量 QCV(x)的计算过程如下：

(1) x 作为 Q/A 问答对集合 C 中的问题；

(2) R(x)表示由 x 在 Q/A 问答对集合 C 中检索出来的前 N 个问答对的集合 $(p_1, p_2, ..., p_n)$；

(3)对每个 p_i 计算向量 v_i ,其中向量 v_i 的每一维为 p_i 中单词的权重；

(4)对每一个 v_i 取其前 m 个最高的权重的单词，构成新的向量 v_i 。

则 QCV(x)的计算公式为：

$$QCV(x)=\frac{1}{n}\sum_{i}^{n}\frac{v_i}{||v_i||} \tag{3.10}$$

上述查询扩展方法步骤（3）的权重计算使用信息检索领域最常用的 $tf \cdot idf$ 方法，每个Q/A问答对 p_j 中每个单词 t_i 的权重 w_{ij} 为：

$$w_{ij} = tf_{ij} \times \log(\frac{N}{df_i}) \tag{3.11}$$

其中 tf_{ij} 为 t_i 在 p_j 中出现的频率，N为Q/A问答对集合中问答对的总数，df_i 是包含 t_i 问答对的个数。

最后使用 cosine 值来计算问题 x 和问题 y 的相似度，公式如下：

$$\cos(QCV(x),QCV(y))=\frac{\sum_{i=1}^{n} cx_i * cy_i}{\sqrt{\sum_{i}^{n} cx_i^2}\sqrt{\sum_{i}^{n} cy_i^2}} \tag{3.12}$$

其中 cx_i 和 cy_i 分别是向量 QCV(x)和向量 QCV(y)中每一维的权重。

本节的 BAIM 是在图 3.1 所示框架的流程图中，根据第三次检索得到的结果从答案索引中把对应的答案给抽取出来，然后直接从上述查询扩展方法的第三步开始执行的。

3.3 实验设计

3.3.1 实验数据

本次实验语料是取自 Yahoo! Answers 的一个数据集。由于数据集中存在一定的数据稀疏性，就是有些类别的问句数量比较多，有些类别的问句数量非常少。为了不影响实验结果和保持数据集的平衡性，我们选取相对比较热门的问句类别。首先根据问句类别中问句的数量对每个问句类别进行排序，在这里选取问句类别中问句数量最多的 60 个类别。我们所选取的问句类别中，每个类别的问句数量均在 1000 个以上。然后对选取的 60 个问句类别，再选取问句数最多的前 30 个类别作为相关类别，这些类别中的问句作为相关数据集中的数据。另外的 30 个类别作为非相关类别，即这些类别中的问句作为非相关数据集中的数据。之所以前 30 个类别为相关类别，是因为本文的查询由前 30 个类别提供，共计 143 个查询，这些查询由不同的长度、句式构成，以保证查询的公平合理性。最后的实验数据集的相关信息如表 3.1 和 3.2 所示。

表 3.1 相关语料集得统计信息情况

Tab. 3.1 Statistical information of relevant corpus

问题数	答案数	最佳答案数量	类别数	问题平均答案数
165, 847	1, 506, 186	165, 911	30	9. 08

表 3.2 非相关语料集的统计信息情况

Tab. 3.2 Statistical information of irrelevant corpus

问题数	答案数	最佳答案数量	类别数	问题平均答案数
35,151	342,255	35,164	30	9.73

在我们的实验中由于选取的语料都是已解决的问题答案对，即每个问题都有最佳的答案，并且问题所对应的最佳答案一般都是由提问这个问题的用户经过对比其他答案和深思熟虑后选出来的，或者是由社区问答系统中的其他有过类似问题的用户投票选出来的，因此我们可以认为问题对应的最佳答案代表了该问题的最准确信息。并基于这样的假设：我们根据有共同最佳答案的问句是最相似的来判定两个句子是否相似，因此分别选出这 143 个查询所对应有共同最佳答案的问句作为标准答案集，每个查询都对应 5 到 6 个这样的对应问句。

接着我们从前 30 个相关类别中去除这 143 个查询问句以及标准答案问句，重新分别对相关数据集中的 30 个类别，每个类别随机选择 500 个问句，共选取 15000 个句子作为噪音句，同时把这 143 个查询对应的标准答案集和这 15000 个噪音句融合在一起，并在此数据集上进行实验，另外采用相同的方法从另外的非相关类中选取 15000 个问句作为噪音句，按上面的方法组成数据集，并在此数据集上进行同样的实验。

3.3.2 评价指标

根据社区问答系统中检索任务的特殊性，在本章中我们选取两种评价指标 P@N 和 MAP(Mean Average of Precision)对所采用的检索模型进行评价。P@N 和 MAP 的计算方法如下：

（1）P@N ：*P@N* 表示检索出来的前 *N* 个候选问句中，相关问句所占的比例。计算公式为公式（3.13）：

$$P@N=\frac{\sum_{i=1}^{n} rel(i)}{N} \tag{3.13}$$

其中 $rel(i)$ 表示第 i 个候选查询问句和当前查询问句是否相关，取值为{0，1}。0 表示不相关，1 表示相关。

（2）MAP：MAP 主要是指单个查询的平均准确率，即每个查询返回结果准确率的平均值。计算公式为公式（3.14）：

$$MAP=\frac{1}{Q_r}\sum_{q\in Q_r}\frac{\sum_{r=1}^{n} p(r)\times rel(r)}{|R_q|} \tag{3.14}$$

其中 Q_r 表示要进行查询的问句集合，R_q 表示和当前查询相关的问句，r 值表示在检索出来的排序列表中的排列的位置，N 是检索的问句的总个数，$rel(r)$ 表示第 r 个候选查询问句和当前查询问句是否相关，取值为{0，1}，0 表示不相关，1 表示相关。$P(r)$ 表示前 r 个检索的问句中，相关问句所占的比例。

3.3.3 实验结果与分析

为了能够更加公平合理的评价本章提出检索模型的性能，在这里采用学术界广泛认同的 P@N 和 MAP 两个指标对实验结果进行评价。同时为了显示本章提出的模型较之前的模型有一定的优势，在这里采用了 7 种不同的检索模型进行对比实验。表 3.3 中是 7 种不同方法的描述，其中（1）和（2）是文献[13]提出的方法（BaseLine），

表 3.3 实验方法和表述

Tab. 3.3 Experimental methods and description

方法名称	方法描述
（1）SM[13]	统计模型
（2）SM+SEM[13]	在 SM 基础上上引入语义信息（基于 WordNet）
（3）ISM	改进的统计模型
（4）ISM+SEM	在模型（3)基础上引入语义信息(基于 WordNet）
（5）ISM+QTFD	在模型（3）的基础上引入主题和焦点信息
（6）ISM+QTFD+SEM	在模型（5）的基础上引入语义信息
（7）ISM+QTFD+SEM+BAIM(our)	在模型（6）的基础上引入答案信息(本文方法)

采用相关类中的实验结果如表 3.4

表 3.4 在相关语料集上 7 种模型的 MAP 和 P@3 值

Tab. 3.4 MAP and P@3 performance of the seven models on relevant corpus

模型	MAP	P@3
SM[13]	0.5526	0.5642
ISM[13]	0.5963	0.6039
SM+SEM	0.6894	0.7236
ISM+SEM	0.7396	0.7647
ISM+QTFD	0.7156	0.7403
ISM+QTFD+SEM	0.7427	0.7864
ISM+QTFD+SEM+BAIM (our)	0.7571	0.8228

从表 3.4 结果可以看出：

（1）实验结果显示，在采用改进后的统计模型（ISM）比文献[12]中提出的统计模型（SM）在 MAP 和 P@3 上都有了小幅的提高，这表明在问句相似性计算过程中遵循语言学的知识，即对句子中的名词和动词相对于其他词性的词赋予较高的权重，更加有利于句子的相似性计算，同时进行同义词扩展也有利于两个句子之间相似度的比较。弥补了原来模型的不足。

（2）在改进后的统计模型（ISM）中引入语义信息即（ISM+SEM）要比 ISM 引入 QTFD 信息即（ISM+QTFD）在 MAP 上高出 0.024，而在 P@3 上也高出 0.0244。这表明在相关类别的问句语料集中语义特征起到主要作用，而词本身特征起到次要作用。这主要是因为在相同类别中所有问句都是关于一个类别的，问句中的主题词即关键词基本相同，所以词本身特征很难区分句子间的不同。这时候语义信息将起到绝对作用。

（3）在改进后的统计模型（ISM）中同时引入语义信息和 QTFD 信息即（ISM+QTFD+SEM）相对于两者单独引入（ISM+SEM 和 ISM+QTFD），不管是在 MAP 还是在 P@3 上都有相应的提高，说明综合运用词信息，语义信息和统计信息可以得到较好的结果。

（4）在 ISM+QTFD+SEM 中引入 BAIM 信息即引入答案信息相对于 ISM+QTFD+SEM 在 MAP 上提高 1.9%，在 P@3 上提高 4.6%，说明综合利用词信息，语义信息，统计信息和答案信息可以更加充分的挖掘句子之间的信息，从而更有利于句子之间的相似度计算。同时也证明我们提出的方法是可行的。

采用非相关类中的实验结果如表 3.5。

表 3.5 在非相关语料集上 7 种模型的 MAP 和 P@3 值

Tab. 3.5 MAP and P@3 performance of the seven models on irrelevant corpus

模型	MAP	P@3
SM[13]	0.6435	0.6632
ISM[13]	0.7091	0.6874
SM+SEM	0.7325	0.7535
ISM+SEM	0.7612	0.7763
ISM+QTFD	0.8162	0.8243
ISM+QTFD+SEM	0.8268	0.8362
ISM+QTFD+SEM+BAIM (our)	0.8450	0.8787

从表 3.5 和表 3.4 的对比可以看出：

（1）不管是在 MAP 值，还是在 P@3 值，整体上在非相关领域数据集上的实验结果都要比在相关领域数据集上的实验结果好。这主要因为在相关领域数据集中，噪音句和标准答案句之间相似性较高，不管是在语义特征还是词本身特征都存在较大的干扰；而在非相关领域数据集中，噪音句和标准答案句分别属于不同的领域，因此它们之间的相似度较低，存在的干扰度较小，所以整体上非相关领域数据集上的结果要好些。

（2）在改进后的统计模型（ISM 中引入语义模型（ISM+SEM）要比 ISM 引入 QTFD 信息（ISM+QTFD）在 MAP 上降低了 0.055，而在 P@3 上也降低了 0.048。说明在非相关类中语义特征起到次要作用，而词本身特征起到主要作用。这主要是因为在非相关类别中问句之间分别属于不同的类别，它们的关键词和主题词存在很大的差异，所以词本身的特征在相似度计算过程中其绝对的作用。

（3）从表 3.4 中 ISM+SEM 与 ISM+QTFD 结果比较和表 3.55 中 ISM+SEM 与 ISM+QTFD 结果比较可以看出，在相关领域中，语义特征起到了主要作用，因为相关领

域在同一个类别下的同义词较多，基于语义特征的方法正好解决此问题；而在非相关领域中，词本身特征起到了主要作用，因为非相关领域同义词出现的概率较少，而且非相关领域歧义情况也出现得较少。

（4）从表 3.5 中可以看出本章提出的多种特征融合的方法（ISM+QTFD+SEM+BAIM）在相关和非相关领域都得到了较高的评价结果，由此表明该方法是一个可行的通用方法。

3.4 本章小结

本章提出了一种基于特征融合的问句相似度计算方法，首先从如何理解问句出发对问句进行了详细的分析，根据问句自身的特点，建模时较全面地运用了语言学知识特点，充分挖掘了句子本身的结构信息；同时根据社区问答系统存在大量问答对的特点，把问句对应的答案信息引入到句子相似度计算中。从多角度的对比实验结果表明，将词序特征、语义特征、统计特征和答案特征等进行适当融合，用于问句的相似度计算，如实验部分所述，显示了很好的性能，从而也表明本章提出的方法能够有效解决社区问答系统中句子匹配的问题。

4 融合问句类别信息和答案类别信息的检索模型

本章主要通过以下几个方面来介绍所提出的融合问句类别信息和答案类别信息的检索模型：方法思想的起源，方法用到的三个模型（语言模型、基于问句类别信息平滑的语言模型、基于问句答案类别信息平滑的语言模型），并通过实验设计和对比结果分析来论证方法的有效性。

4.1 引言

在社区问答系统问句检索这个领域，已经涌现出了一些利用问句类别信息来提高问句检索性能的方法[23,24,37]。Cao 等人[23]利用分类器去计算每个查询属于不同类别的概率，然后对语言模型进行平滑，来提高问句检索的性能。它检索性能的提升是依赖于分类器的准确率。然而，在社区问答系统中的分类问题是一个大规模的层次分类问题，并且由于分类误差造成的检索性能下降是很难改善的[24]。Cai 等人[37]把类别信息融合到基于翻译的语言模型中，来提高检索的性能，但忽略了检索的效率。

针对上述问题，本章提出的方法利用用户已经选定的类别，结合问句所对应的答案类别信息来提高问句检索的效率和性能。该方法分别利用问句的类别信息和问句答案类别信息对传统的语言模型进行平滑，然后线性加权两种平滑结果，来提高社区问答系统中问句检索性能。那么为什么在问句检索过程中使用问句的类别信息就能提高问句检索的性能和效率。

首先性能方面，我们可以看这样一个例子，查询(q): "Can you recommend me some Chinese food?"，这个问题主要是想得到一些关于中国美食方面的信息，并不是其他国家的美食信息。而下面这个问题，(d) "Can you recommend me some Japanese food?"的答案明显不是上一个问题的答案，虽然这两个问句在句法结构上十分相似。对于一般的检索方法，在检索关于中国美食问题时，是很难过滤掉其他国家美食的问题，如果采用基于类别信息，可以很容易过滤掉不相关的类别信息，从而提高模型的检索性能。

其次效率方面，还举上面的例子，一个问句所在类别是关于中国美食的，那么在这个类别下面的所有问句都是关于中国美食的。因此对于用户提出的一个关于中国美食的问题，我们可以优先的在这个类别下检索，从而避免在不相关类别下检索，这样就会提高问句的检索效率。

4.2 检索模型概述

4.2.1 算法思想

首先从 Yahoo! Answers 网站抽取“问答对”语料集，然后建立问答对之间的索引。对于任意的一个查询 Q，分别利用问句自身的类别信息和问句所对应答案的类别信息对语言模型进行平滑，得到两个相似得分，最后线性加权两个相似得分得到最终结果。根据最终结果，从相应的答案索引中提取出最佳答案返回给用户。算法流程如图 4.1 所示。

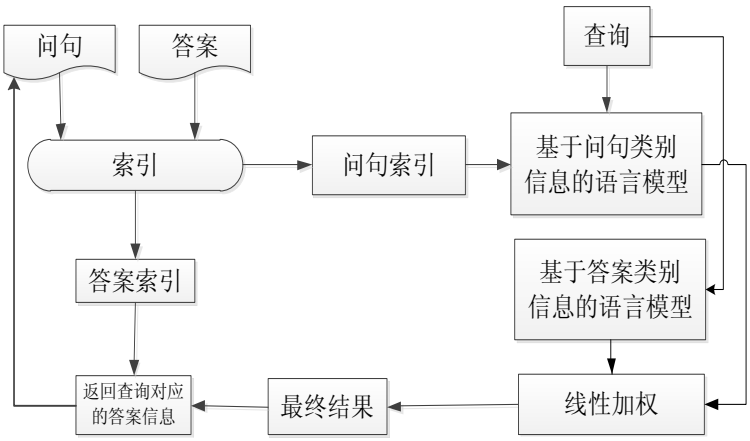

图 4.1 算法流程图

Fig. 4.1 Overview of algorithm

4.2.2 语言模型

语言模型是信息检索中比较成熟的模型之一，它在过去很多信息检索任务中都表现出了良好的性能[28,38,39]，同时也在问句检索任务中显示出了良好的性能[40]。它的基本思想是根据文档语言模型生成查询文本的概率来对文档经行排序。即对于一个查询 q 和一个文档 d，采用 Jelinek-Mercer 平滑方法[28]的语言模型公式如下：

$$
\begin{aligned}
P(q \mid d) &= \prod_{w \in q} P(w \mid d) \\
P(w \mid d) &= (1-\alpha) P_{ml}(w \mid d) + \alpha P_{ml}(w \mid Col) \\
P_{ml}(w \mid d) &= \frac{tf(w,d)}{\sum_{w' \in d} tf(w',d)} \\
P_{ml}(w \mid Col) &= \frac{tf(w,Col)}{\sum_{w' \in C} tf(w',Col)}
\end{aligned}
\tag{4.1}
$$

其中w是一个查询q中的单词，α是平滑参数，$p_{ml}(w \mid d)$是对于一个单词w在文档d中的最大似然估计，$tf(w \mid d)$是对于单词w在文档d中的频率。$tf(w \mid Coll)$是单词w在整个文档集合$Coll$中的频率。$P_{ml}(w \mid Coll)$是单词w在文档集合$Coll$中的最大似然估计。

4.2.3 基于问句类别信息平滑的语言模型

在社区问答系统中，例如 Yahoo! Answers 中每个用户在上面提问问题的时候，都要事先给问题选定一个类别，以有利于系统能更加快速准确的为用户提供相应的答案，因此社区问答系统中的每一个问句都有一个相应的类别。本章之所以采用问句的类别信息对语言模型进行平滑，主要是基于以下两点思想：

（1）在相同问句类别下进行检索过程中，出现词频较高的词语，对问句检索起到次要的作用，因为在相同问句类别内所有的问句都是关于同一个类别的，而词频较高的词，基本都是和类别相关的，所以在同一个问句类别内，词频较高的词，对于问句检索起到次要作用。

（2）在不同的问句类别下进行检索过程中，出现词频较高的词语在问句检索中起到比较重要的作用，因为在不同的问句类别下的问句都是关于不同问句类别的，所以问句类别中词频较高的词语更加有利于区分不同的问句类别，因此对于在不同问句类别下的问句检索，问句类别中词频较高的词语起到比较重要的作用。

基于以上的两点思想，对于用户的问句查询q和相应的候选问句d，我们用$P(q \mid d)$表示候选问句d产生查询问句q的概率，$p(w \mid d)$表示检索模型用来衡量候选问句 d 和查询问句 q 的相关度。为了计算$P(q \mid d)$，我们需要估计语言模型$p(w \mid d)$，因此，修改公式（4.1）变为公式（4.2）如下所示：

$$\begin{aligned} P(q \mid d) &= \prod_{w \in q} P(w \mid d) \\ P(w \mid d) &= (1-\alpha) P_{ml}(w \mid d) \\ &+ a[(1-\beta) P_{ml}(w \mid Cat(d)) + \beta P_{ml}(w \mid Coll)] \end{aligned} \tag{4.2}$$

其中w表示查询问句q中的单词，$Cat(d)$表示查询问句q的问句所在类别信息，α和β是不同的平滑参数。$p_{ml}(w \mid d)$是采用公式（4.1）计算出的单词w在候选文档d中的最大似然估计。$p_{ml}(w \mid Coll) = \frac{tf(w, Coll)}{\sum_{w' \in Coll} tf(w', Coll)}$表示单词$w$在整个问句集合$Coll$中的

最大似然估计。$P_{ml}(w \mid Cat(d)) = \frac{tf(w, Cat(d))}{\sum_{w' \in Cat(d)} tf(w', Cat(d))}$ 表示 w 在问句类别 $Cat(d)$ 中的最大似然估计。把使用公式（4.2）计算候选问句 d 和查询问句 q 最大似然估计最为最后的相似得分，在这里把这种方法记为 LM+C。

接下来我们将展示在社区问答系统问句检索过程中，对于具体问句类别中的高频词语，在不同问句类别下对于问句的检索起更重要的作用。在这里我们定义 P_{cs} 表示在问句类别集合 $Cat(d)$ 中出现的单词集合，同样的 P_{cu} 表示在问句类别集合 $Cat(d)$ 中不出现的单词集合。因此对于一个文档 d 生成查询 q 的概率可使用如下公式（4.3）计算

$$\begin{aligned} \log P(q \mid d) &= \sum_{w \in q} \log P(w \mid d) \\ &= \sum_{w \in q}^{tf(w,Cat(d))>0} \log P_{cs}(w \mid d) + \sum_{w \in q}^{tf(w,Cat(d))=0} \log P_{cu}(w \mid d) \\ &= \sum_{w \in q}^{tf(w,Cat(d))>0} \log \frac{P_{cs}(w \mid d)}{P_{us}(w \mid d)} + \sum_{w \in q} \log P_{cu}(w \mid d) \end{aligned} \tag{4.3}$$

根据基于问句类别信息平滑模型，即公式（4.2），我们可以得到公式（4.4）

$$\begin{aligned} P_{cs}(w \mid d) &= (1-\alpha) P_{ml}(w \mid d) + \\ &\alpha[(1-\beta) P_{ml}(w \mid Cat(d)) \\ &+ \beta P_{ml}(w \mid Coll)] \\ P_{cu}(w \mid d) &= \alpha\beta P_{ml}(w \mid Coll) \end{aligned} \tag{4.4}$$

根据公式（4.3）和公式（4.3）我们可以得到公式（4.5）

$$\begin{aligned} \log P(q \mid d) = &\sum_{w \in q}^{tf(w,Cat(d))>0} \log\left(\frac{(1-\alpha) P_{ml}(w \mid d) + \alpha(1-\beta) P_{ml}(w \mid Cat(d))}{\alpha\beta P_{ml}(w \mid Coll)} + 1\right) \\ &+ \sum_{w \in q} \log \alpha\beta P_{ml}(w \mid Coll) \end{aligned} \tag{4.5}$$

从公式（4.5）的右边第二部分可以看出，它是独立于文档 d 的，因此它对我们的排序得分可以忽略不计。从右边第一部分可以看出，对于在不同问句类别下的问句集合，$P_{ml}(w \mid Cat(d))$ 越大，则 $P(q \mid d)$ 越大。在一个问句中越多的单词出现在一个问句类别下，那么它在同一个问句类别中的得分就越高，因此表明，采用问句类别信息平滑的语言模型对不同的问句在不同的问句类别下起到关键的作用。

接下来我们讨论在同一个问句类别中出现频率高的词语，在同一问句类别下起次要作用。我们定义 $P_s(w \mid d)$ 表示单词 w 在问句 d 中出现的单词的概率，即 $tf(w,d) > 0$ 的

情况，$P_u(w \mid d)$表示单词w在问句d中不出现的概率，即$tf(w,d)=0$)的情况，因此一个问句d生成问句q的概率可以表示为如下公式（4.6）

$$\begin{aligned}\log P(q \mid d) &= \sum_{w \in q} \log P(w \mid d) \\ &= \sum_{w \in q}^{tf(w,d)>0} \log P_s(w \mid d) + \sum_{w \in q}^{tf(w,d)=0} \log P_u(w \mid d) \\ &= \sum_{w \in q}^{tf(w,d)>0} \log \frac{P_s(w \mid d)}{P_u(w \mid d)} + \sum_{w \in q} \log P_u(w \mid d)\end{aligned} \tag{4.6}$$

根据主题类别信息平滑模型即公式（4.2）我们可以得到公式（4.7）

$$\begin{aligned}P_s(w \mid d) &= (1-\alpha)P_{ml}(w \mid d) + \alpha[(1-\beta)P_{ml}(w \mid Cat(d)) + \beta P_{ml}(w \mid C)] \\ P_u(w \mid d) &= \alpha(1-\beta)P_{ml}(w \mid Cat(d)) + \beta P_{ml}(w \mid C)\end{aligned} \tag{4.7}$$

把公式（4.7）带入公式（4.6）中可以得到公式（4.8）

$$\begin{aligned}Log(q \mid d) = &\sum_{w \in q}^{tf(w,d)>0} \log\left(\frac{(1-\alpha)P_{ml}(w \mid d)}{\alpha[(1-\beta)P_{ml}(w \mid Cat(d)) + \beta P_{ml}(w \mid C)]} + 1\right) \\ &+ \sum_{w \in q} \log(\alpha[(1-\beta)P_{ml}(w \mid Cat(d)) + \beta P_{ml}(w \mid C)])\end{aligned} \tag{4.8}$$

从公式（4.8）中我们可以看出，对于任意的查询q和候选问句d，公式右边的第二部分都是一样的，因此它对最后的排序得分没有影响，在此可以忽略。但是对于公式右边的第一部分，它表示单词w在$Cat(d)$中最大似然估计的倒数，即$P_{ml}(w \mid Cat(d))$的倒数。因此可以看出这里的基于问句类别信息平滑和我们熟悉的 IDF 起到的作用是一样的。在一个具体问句类别中出现次数越多的单词，在同一类别检索中起到的作用越次要。

综上所述可以看出，对于查询q在不同问句类别下检索过程中，基于问句类别信息的语言模型可以得到更好的结果，但是在相同问句类别下的检索只能和 IDF 起到相同的效果。因此可以看出，基于问句类别信息的语言模型更容易区分出不同类别下的问句。

4.2.4 基于答案类别信息平滑的语言模型

在社区问答系统中之所以问句的答案及答案的类别信息对问句检索有帮助。是因为，一方面，在社区问答系统中提问的人都是来自现实生活中的人，因此提问的问题基本都反映人们的真实生活，并且回答问题的人也是来自于现实生活中的人，而且一般这些人都是对某一领域有一定研究的人。因此他们提供的答案一般比较详尽，更能表达出问题的本意。另一方面，通常情况下用户在提问的问题时，一般对于所提问问题的领域都不怎么熟悉，所以在提问问题时有时候表达不是很清楚，通常就造成提问的问题一般比较短，但相对于回答问题的人一般都是对某一领域有一定了解的人，因此回答的答案

通常比较详细。所以说问题所对应的答案信息往往比问题含有更加丰富的信息。因此采用问题答案的类别信息对语言模型进行平滑，能得到更好的效果。那么该怎样把答案信息融入到我们的模型中，因为我们实验中所使用的语料集中的每个问句都有唯一相对应的最佳答案，因此可以假设问句所在的类别决定了答案所在的类别。改写上面的公式（4.8），在这里把公式（4.8）中的问句类别改成相应的答案类别，修改后的公式为如下公式（4.9）

$$Log(q \mid ans) = \sum_{w \in q}^{tf(w,ans)>0} \log\left(\frac{(1-\alpha)P_{ml}(w \mid ans)}{\alpha[(1-\beta)P_{ml}(w \mid C(ans)) + \beta P_{ml}(w \mid Ca)]} + 1\right) + \sum_{w \in q} \log(\alpha[(1-\beta)P_{ml}(w \mid C(ans)) + \beta P_{ml}(w \mid Ca)]) \quad (4.9)$$

其中ans表示问句所对应的答案信息，$C(ans)$表示应问句答案所在的类别信息，Ca表示语料集中的所有问句的答案集合。

4.2.5 融合问句类别信息和答案类别信息平滑的语言模型

由 4.2.3 节和 4.2.4 节介绍可知，在社区问答系统中，对于问句检索任务，除了问句本身所能提供的信息外，问句所对应的答案信息，对问句的检索任务也是十分有帮助的。如何有效的利用这两者信息来提高问句检索的性能，将成为本章研究的重点。在这里我们采用线性加权的方式结合两者信息，之所以采用线性加权的方式，是因为这种方式简单直观，并且效果显著，参数容易控制和调节。融合两者信息后的公式为如下公式(4.10)：

$$Score(q,d) = (1-\alpha)\log(q, C(d)) + \alpha \log(q, ans(d)) \quad (4.10)$$

在这里$Score(q,d)$为最后的相似得分，α是一个线性参数，α越大说明问句对应的答案信息越重要，反之则说明问句本身的信息越重要。$\log(q, C(d))$是根据公式（4.8）得到的得分，$\log(q, ans(d))$是根据公式（4.9）得到的得分。

4.3 实验设计

4.3.1 实验数据

本次实验语料是取自 Yahoo! Answers 的一个数据集。由于数据集，存在一定的数据稀疏性，即有些类别的问句数量比较多，有些类别的问句数量非常少。为了保持数据集的平衡性，我们选取相对比较热门的问句类别。首先根据问句类别中问句的数量对每个问句类别进行排序，在这里选取问句类别中问句数量最多的 60 个类别。在我们所选取的问句类别中，每个类别的问句数量均在 1000 个以上。实验的语料信息如表 4.1：

表 4.1　语料集得统计信息情况

Tab. 4.1　Statistical information of corpus

问题数	答案数	最佳答案数量	类别数	问题平均答案数
200,998	1,848,441	201,075	60	9.405

在我们的实验中由于选取的语料都是已解决的问题答案对，即每个问题都有最佳的答案，并且问题所对应的最佳答案一般都是由提问这个问题的用户经过对比其他答案和深思熟虑后选出来的，或者是由社区问答系统中的其他有过类似问题的用户投票选出来的，因此我们可以认为问题对应的最佳答案代表了该问题的最准确信息。并基于这样的假设：我们根据有共同最佳答案的问句是最相似的来判定两个句子是否相似，因此分别选出这 143 个查询所对应有共同最佳答案的问句作为标准答案集，每个查询都对应 5 到 6 个这样的对应问句。

从 60 个类别的问句中去除这 143 个查询问句以及标准答案问句，重新对每个类别随机选择 500 个问句,共选取 30000 个句子作为噪音句，同时把这 143 个查询对应的标准答案集和这 30000 个噪音句融合在一起，并在此数据集上进行实验。

4.3.2　参数选择

在本次的实验中需要两个平滑参数，表 4.2 展示出的结果是在一个小的语料集上的结果，这个语料包含 30 个查询和 1000 个句子，我们使用 LM+C 模型采用 MAP 值作为评价指标，在这个小语料集上对参数进行选择，最后显示在 α=0.2, β=0.2 的情况下去的最好的结果。

表 4.2　参数选择（MAP）

Tab. 4.2　Parameter selection(MAP)

β \ α	0.1	0.2	0.3
0.1	0.8463	0.8534	0.8375
0.2	0.8575	0.8643	0.8524
0.3	0.8478	0.8545	0.8482

4.3.3　实验结果与分析

为了能够更加公平合理的评价本章提出检索模型的性能，在这里采用学术界广泛认同的 P@N 和 MAP 两个指标对实验结果进行评价。同时为了显示本章提出的模型较之

前的模型有一定的优势，在这里采用了 6 种不同的检索模型进行对比实验。表 4.3 中是 6 中不同方法的描述，其中（1）、（2）、（4）、（5）分别是文献[28]、[23]、[29]、[41]提出的方法（Base Line），

表 4.3　实验方法和表述

Tab. 4.3　Experimental methods and description

方法名称	方法描述
（1）LM[28]	语言模型（采用 Jelinek_Mercer 平滑）
（2）LM+C[23]	基于问句类别信息的语言模型
（3）LM+A	基于答案类别信息的语言模型
（4）TRLM[29]	基于翻译的语言模型
（5）CLM[41]	集成类别信息的语言模型
（6）LM+C+A	融合问句类别信息和问句答案信息的语言模型（本文方法）

在语料集中的实验结果如表 4.4

表 4.4　在语料集上 6 种模型的 MAP 和 P@1 值

Tab. 4.4　MAP and P@1 performance of the six models on relevant corpus

模型	MAP	P@1
LM[28]	0.8463	0.7587
LM+C[23]	0.8504	0.7867
LM+A	0.8662	0.7943
TRLM[29]	0.8678	0.7968
CLM[41]	0.8756	0.8075
LM+C+A	0.8784	0.8125

从表 4.4 结果可以看出：

（1）在本章的实验中采用基于问句类别信息的语言模型（LM+C）比传统的语言模型（LM）在 MAP 和 P@1 上都有一定的提高，这主要是因为采用类别信息的语言模型除了拥有传统语言模型的优势外，还考虑了句子本身的类别信息。同时也证明了问句的类别信息在提高问句检索性能上是有帮助的。(Row2 VS Row1)

（2）采用基于问句答案类别信息的语言模型(LM+A)比采用基于问句类别信息的语言模型(LM+C)在 MAP 提高了 0.0218，在 P@1 上提高了 0.0105。这表明问句的答案信息比问句本身的信息更能反映出问题的实质性内容，也能更好的表达用户的真实意图。同时也证明了问句的答案类别信息在问句检索中具有更强的区分度。(Row3 VS Row2)

（3）采用基于翻译的语言模型(TRLM)要比采用基于问句类别信息的语言模型(LM+C)和采用基于问句答案类别信息的语言模型(LM+A)在 MAP 和 P@1 上都有所提高，这是因为采用基于翻译模型的语言模型除了能更加全面的考虑问句本身的信息外，又融入了一些外部的信息来对问句进行补充，所以检索效果要好一些。(Row4 VS Row3，Row2)

（4）集成类别信息的语言模型（CLM）要比采用基于翻译的语言模型(TRLM) 在 MAP 和 P@1 上都有所提高，主要是因为 CLM 对问句的类别信息进行了高度的集成，从分类类别上更加的精准化，同时考虑了更多的语言信息。

（5）采用融合问句类别信息和问句答案类别信息的方法（LM+C+A）,比以上所有的模型不论是 MAP 值，还是 P@1 值都有所提高，可以看出在考虑问句自身类别信息的情况下，再把问句所对应的答案类别信息融合进来，既用到了问句本身的信息又用到了问句所对应的答案信息，从而能更好的表达句子的本身含义，更加有利于问句的检索。从中也证明本章提出的模型是一个有效的检索模型。

4.4 本章小结

本文主要探讨了社区问答系统中的问句检索模型，提出一种融合问句类别信息和问句答案类别信息的问句检索模型。首先从如何理解问句出发，对问句进行了详细的分析，并考虑了社区问答系统自身的特点，即每个问句都有相应的类别信息。同时也考虑了问句所对应的答案信息，利用问句自身的类别信息和问句所对应答案类别信息分别对语言模型进行平滑，最后线性加权两个平滑结果最为最后的相似度等分。多角度的对比实验结果表明，将问句类别信息和问句答案及其类别信息适当融合，用于问句的检索，能够有效解决社区问答系统中句子检索的问题。如实验部分所述，显示了很好的性能。

5 融合问句主题信息和答案主题信息的检索模型

本章主要从以下几个方面来介绍所提出的融合问句主题信息和答案主题信息的检索模型：方法思想的起源，方法用到的四个模型（LDA 模型、语言模型、基于问句主题信息平滑的语言模型、基于问句对应答案主题信息平滑的语言模型），并通过实验设计和对比结果分析来论证方法的有效性。

5.1 引言

随着社区问答系统的兴起，很多学者都把 LDA 模型应用到了社区问答系统中的各个研究领域[42-44]。在问句检索领域，Celikyilmaz 等人[45]把 LDA 应用到了问句检索领域，不过他们只是单纯的利用 LDA 用于问句的相似度计算，同样文献[46]也是利用 LDA 对问句进行相似度计算。而本章提出的方法，主要是利用 LDA 进行主题信息提取。然后利用提取后的主题信息对语言模型进行平滑。

本章提出的融合问句主题信息和答案主题信息的检索模型，首先主要是利用问句本身的主题信息和问句所对应答案的主题信息，分别对传统的语言模型进行平滑，然后线性加权两种平滑结果，来提高社区问答系统中问句的检索性能。相对于 4.2.3 节提出的基于问句类别信息的语言模型有一定的相似之处，但也有区别。

相似之处在于，基于问句类别信息平滑的语言模型中是利用问句的类别信息来过滤掉不相关的类别问句，从而提高问句检索的效率和性能，而基于问句主题信息平滑的语言模型是利用问句的主题信息来过滤掉不相关主题的问句，来提高问句检索的效率和性能。

不同之处在于，主题信息相对于类别信息来说，主题信息范围更广泛。比如对于“计算机”这个主题，下面可以包含“硬件”和“软件”两个类别。所以采用主题信息相当于对相似类别之间进行合并，因此利用问句主题信息可以减轻由于误分类对检索性能的影响，从而表现出更好的性能。

5.2 检索模型概述

5.2.1 算法思想

首先从 Yahoo! Answers 网站抽取“问答对”语料集，然后建立问答对之间的索引。然后使用 LDA 主题模型分别对问句集合和答案集合进行主题信息提取。对于任意的一个查询 Q，利用提取出来的问句主题信息和答案主题信息分别对语言模型进行平滑，得

到两个相似得分，最后进行线性加权两个相似得分得到最终结果。根据最终结果，从相应的答案索引中提取出最佳答案返回给用户。算法的流程图如图 5.1 所示。

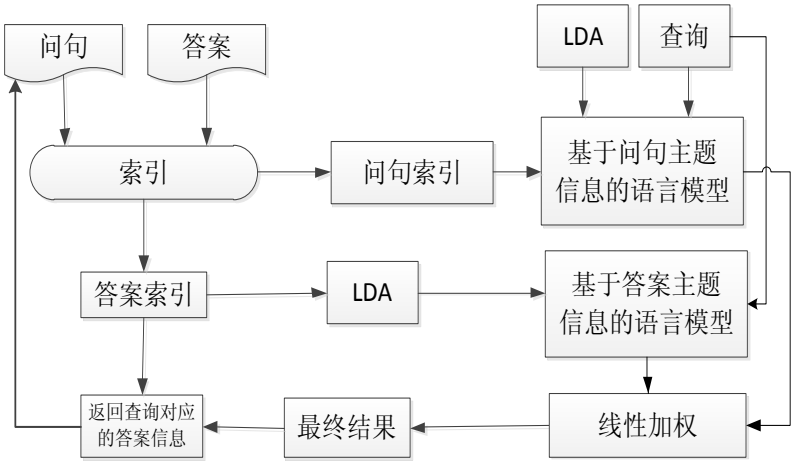

图 5.1　算法流程图

Fig. 5.1　Overview of algorithm

5.2.2　LDA（Latent Dirichlet Allocation）主题模型

在介绍 LDA[47]主题模型之前，先介绍一下主题模型[48]，主题模型是一种概率模型。它在文本挖掘领域应用十分广泛。在传统判断两个文档相似性的方法中，通常是查看两个文档共同出现的词语的多少来衡量两个文档的相似程度。但这种方法明显有一个缺陷，即没有考虑到文字背后隐藏更深的一些信息，比如文字背后的语义关联信息等等。即可能存在这样的两个文档，它们几乎没有共同出现的词语，但这两个文档却是相似的。而在社区问答系统的问句检索中，出现的情况更为复杂，不仅会出现两个句子中没有任何相同的词汇，但两个句子是相似的。而且还会出现两个句子中出现的单词完全一样但表达的意义完全不相同。比如“How can I lose weight in a few months？”和“Are there any ways of losing pound in a short period？”都是关于寻求减肥方法的问题，但是他们几乎不含有任何相同的词汇，并且具有不同的句法特征，但这两个句子的意思是一样的，传统的利用词频的方法很难解决上述的问题。又比如“How many London to New York flights are there in a day？”和“How many New York to London flights are there in a day？”都是询问一天飞机的航班情况，并且所有的单词都一样，不过一个是询问从伦敦到纽约，另一个是询问从纽约到伦敦，两个句子中的特征词完全一样，但意思完全不同。因此对于问句这种短文本之间的相似度计算，考虑背后的语义信息是必不可少的。而对语义挖

掘目前最好的模型就是主题模型，LDA 就是其中一种比较有效的模型之一。Yohan 等人[49]针对特定领域对 LDA 进行改进，使其表现了更好的性能。

LDA 模型是目前应用比较广泛的主题模型之一。近期不少学者，针对不用的研究领域，对 LDA 模型提出了自己的改进算法[50-52]。LDA 模型主要包含词，文档和主题三层结构。其中 Dirichlet 分布主要是指文档到主题的分布，而多项式分布主主要是指主题到词的分布。如图 5.2 展示了 LDA 图模型表示：

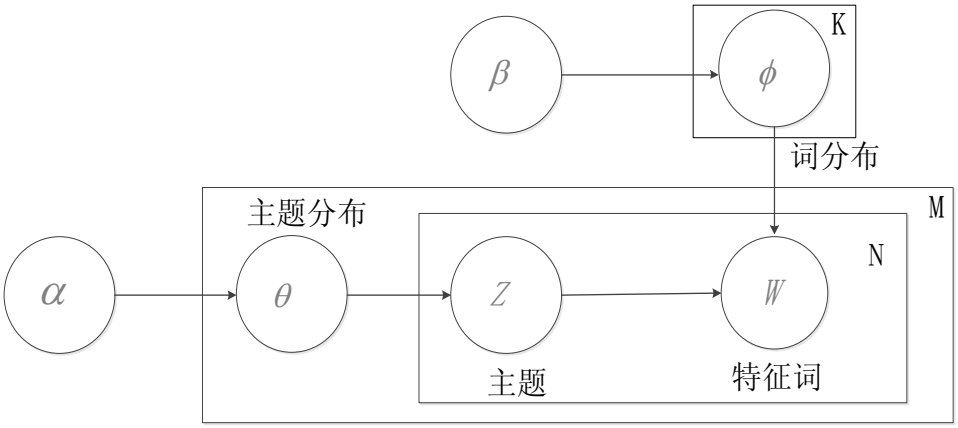

图 5.2　LDA 图模型

Fig. 5.2　Graphical model of LDA

在社区问答系统中，对语料库中的每个问题，LDA 定义了如下的生成过程[53]：

（1）对于每个主题 K，由 $Dirichlet(\beta)$ 分布得到该主题上的一个词多项式分布

（2）对于每个问题 d，由 $Dirichlet(\alpha)$ 分布得到该问题上的一个主题多项式分布

（3）对于每个问题中的每一个单词 w_i：

①从主题多项式 θ_d 采样得到主题 K。

②从该主题上的词多项分布 $\phi^{(t)}$ 采样得到单词 w_i。

由上面的生成过程可以看出，在社区问答系统中利用 LDA 模型生成问句中词的过程是一个概率抽样过程，分别由参数 (α,β) 来确定问题集合中词、问句和主题这三层结构的主题模型，其中参数 α 控制了问题集合中对于不同主题的相对出现的概率，参数 β 主要是调节对于所有主题中不同词的出现概率分布。

在本章中，主要是利用 LDA 主题模型，首先分别对问句集合和问句所对应的答案集合进行主题信息提取，然后分别计算出每个问句属于某个主题的概率和每个答案属于某个主题的概率，最后在后面 5.2.3 节和 5.2.4 节中使用此概率对语言模型平滑。

5.2.2 语言模型

语言模型是信息检索中比较成熟的模型之一，它在过去很多信息检索任务中都表现出了良好的性能[28,38,39]，同时也在问句检索任务中显示出了良好的性能[17]。它的基本思想是根据文档语言模型生成查询文本的概率来对文档经行排序。即对于一个查询 q 和一个文档 d，采用 Jelinek-Mercer 平滑方法[28]的语言模型公式如下：

$$
\begin{aligned}
P(q \mid d) &= \prod_{w \in q} P(w \mid d) \\
P(w \mid d) &= (1-\alpha) P_{ml}(w \mid d) + \alpha P_{ml}(w \mid Coll) \\
P_{ml}(w \mid d) &= \frac{tf(w,d)}{\sum_{w' \in d} tf(w',d)} \\
P_{ml}(w \mid Coll) &= \frac{tf(w,Coll)}{\sum_{w' \in C} tf(w',Coll)}
\end{aligned}
\tag{5.1}
$$

其中 w 是一个查询 q 中的单词，α 是平滑参数，$p_{ml}(w \mid d)$ 是对于一个单词 w 在文档 d 中的最大似然估计，$tf(w \mid d)$ 是对于单词 w 在文档 d 中的频率。$tf(w \mid Coll)$ 是单词 w 在整个文档集合 $Coll$ 中的频率。$P_{ml}(w \mid Coll)$ 是单词 w 在文档集合 $Coll$ 中的最大似然估计。

5.2.3 基于问句主题信息平滑的语言模型

在社区问答系统中，例如 Yahoo! Answers 中每个用户在上面提问问题的时候，都要事先给问题选定一个类别，以有利于系统能更加快速准确的为用户提供相应的答案，因此社区问答系统中的每一个问句都有一个相应的类别。但由于用户在问答系统中提问问题的时候可能不太清楚自己所提问问题的具体类别，而选择错误的类别，针对这个问题，本章之所以采用问句的主题信息对语言模型进行平滑，主要是基于以下三点思想：

（1）对于任意一个查询 q，在同一个主题信息中进行检索时，对于出现词频较高的词语，对问句检索起到次要的作用，因为在同一个主题信息下所有的问句都是关于同一个主题的，而一般词频较高的词，基本都是和主题息息相关的，因此词频较高的词语对问句之间的区分度不是很大，所以在同一个主题信息内，对于词频较高的词，对于问句检索起到次要作用。

（2）对于任意一个查询q，在不同的主题信息中进行检索时，对于出现词频较高的词语，在问句检索中起到比较重要的作用。这主要是因为在不同的主题信息中的问句都是关于不同主题的，而一般词频较高的词，基本都是和主题息息相关的，所以在不同的主题中，词频较高的词语更加有利于区分不同的问句，因此对于在不同主题信息中的问句检索，问句中词频较高的词语起到比较重要的作用。

（3）由以上两点可以看出，本节提出的基于问句主题信息平滑的语言模型，相对于 4.2.3 节提出的基于问句类别信息的语言模型有一定的相似之处。但也有区别，本节提出的模型相对于 4.2.3 节提出的模型不同之处在于，基于问句类别信息平滑的语言模型取的好的结果的前提是问句分类的准确性，而在社区问答系统中，用户在提问问题的时候可能不太清楚自己所提问问题的具体类别，而选择错误的类别，这将会影响基于问句类别平滑的语言模型的性能。而基于问句主题信息平滑的语言模型正好弥补了这个不足，它是基于主题信息的，主题信息是隐藏在问句集合背后的信息，一个主题信息可能包含几个相似的类别。比如一个关于“计算机”的主题，可能包含“硬件”和“软件”两个类别。对于一个用户在社区问答系统中提问关于计算机问题的时候，可能不太清楚是硬件问题还是软件问题，而选择错误的类别提问，这就会给基于问句类别信息平滑的语言模型造成误差，但基于问句主题信息类别平滑的语言模型可以很好的解决这个问题，因为它们有相同的主题。

基于以上的三点思想，对于用户的问句查询q和相应的候选问句d，我们用$P(q \mid d)$表示候选问句d产生查询问句q的概率，$p(w \mid d)$表示检索模型用来衡量候选问句 d 和查询问句 q 的相关度。为了计算$P(q \mid d)$，我们需要估计语言模型$p(w \mid d)$，因此，修改公式（5.1）变为公式（5.2）如下所示：

$$\begin{aligned} P(q \mid d) &= \prod_{w \in q} P(w \mid d) \\ P(w \mid d) &= (1-\alpha) P_{ml}(w \mid d) \\ &+ a[(1-\beta) P_{ml}(w \mid Topic(d)) + \beta P_{ml}(w \mid Coll)] \end{aligned} \tag{5.2}$$

其中w表示查询问句q中的单词，$Topic(d)$表示查询问句q的问句所在主题，α和β是不同的平滑参数。$p_{ml}(w \mid d)$是采用公式（4.1）计算出的单词w在候选文档d中的最大似然估计。$p_{ml}(w \mid Coll) = \frac{tf(w, Coll)}{\sum_{w' \in Coll} tf(w', Coll)}$表示单词$w$在整个问句集合$Coll$中的最大似然估计。$P_{ml}(w \mid Topic(d)) = \frac{tf(w, Topic(d))}{\sum_{w' \in Topic(d)} tf(w', Topic(d))}$表示$w$在问句类别$Topic(d)$中的

最大似然估计。把使用公式（5.2）计算候选问句 d 和查询问句 q 最大似然估计最为最后的相似得分，在这里把这种方法记为 LM+T。

接下来我们将展示在社区问答系统问句检索过程中，对于具体主题中的高频词语，在不同问句主题中对于问句的检索起更重要的作用。在这里我们定义P_{cs}表示在问句主题集合$Topic(d)$中出现的单词集合，同样的P_{cu}表示在问句主题集合$Topic(d)$中不出现的单词集合。因此对于一个文档d生成查询q的概率可使用如下公式（5.3）计算

$$\begin{aligned}\log P(q \mid d) &= \sum_{w \in q} \log P(w \mid d) \\ &= \sum_{w \in q}^{tf(w,Topic(d))>0} \log P_{cs}(w \mid d) + \sum_{w \in q}^{tf(w,Topic(d))=0} \log P_{cu}(w \mid d) \\ &= \sum_{w \in q}^{tf(w,Topic(d))>0} \log \frac{P_{cs}(w \mid d)}{P_{us}(w \mid d)} + \sum_{w \in q} \log P_{cu}(w \mid d)\end{aligned} \tag{5.3}$$

根据基于问句类别信息平滑模型，即公式（5.2），我们可以得到公式（5.4）

$$\begin{aligned}P_{cs}(w \mid d) &= (1-\alpha)P_{ml}(w \mid d) + \\ &\alpha[(1-\beta)P_{ml}(w \mid Topic(d)) \\ &+ \beta P_{ml}(w \mid Coll)] \\ P_{cu}(w \mid d) &= \alpha\beta P_{ml}(w \mid Coll)\end{aligned} \tag{5.4}$$

根据公式（4.3）和公式（4.3）我们可以得到公式（4.5）

$$\begin{aligned}\log P(q \mid d) = &\sum_{w \in q}^{tf(w,Topic(d))>0} \log\left(\frac{(1-\alpha)P_{ml}(w \mid d) + \alpha(1-\beta)P_{ml}(w \mid Topic(d))}{\alpha\beta P_{ml}(w \mid Coll)} + 1\right) \\ &+ \sum_{w \in q} \log \alpha\beta P_{ml}(w \mid Coll)\end{aligned} \tag{5.5}$$

从公式（5.5）的右边第二部分，即$\sum_{w \in q} \log \alpha\beta P_{ml}(w \mid Coll)$可以看出，它是独立于文档$d$的，因此它对我们的排序得分可以忽略不计。从公式（5.5）右边第一部分可以看出，对于在不同问句主题下的问句集合，$P_{ml}(w \mid Topic(d))$越大，则$P(q \mid d)$越大。所以对于一个查询q，如果q中的单词在一个主题中出现的越多，那么它和这个主题集合中的问句的相似度得分就越高。因此表明，采用基于主题信息平滑的语言模型能够更好的区分在不同主题下的问句。

接下来我们讨论在同一个主题信息中出现频率较高的词语，在同一主题中起次要作用。我们定义$P_s(w \mid d)$表示单词w在问句d中出现的单词的概率，即$tf(w,d) > 0$的情

况，$P_u(w \mid d)$表示单词w在问句d中不出现的概率，即$tf(w,d)=0)$的情况，因此一个问句d生成问句q的概率可以表示为如下公式（5.6）

$$\begin{aligned}\log P(q \mid d) &= \sum_{w \in q} \log P(w \mid d) \\ &= \sum_{w \in q}^{tf(w,d)>0} \log P_s(w \mid d) + \sum_{w \in q}^{tf(w,d)=0} \log P_u(w \mid d) \\ &= \sum_{w \in q}^{tf(w,d)>0} \log \frac{P_s(w \mid d)}{P_u(w \mid d)} + \sum_{w \in q} \log P_u(w \mid d)\end{aligned} \tag{5.6}$$

根据基于主题信息平滑模型即公式（5.2）我们可以得到公式（5.7）

$$\begin{aligned}P_s(w \mid d) &= (1-\alpha)P_{ml}(w \mid d) + \alpha[(1-\beta)P_{ml}(w \mid Topic(d)) + \beta P_{ml}(w \mid Coll)] \\ P_u(w \mid d) &= \alpha(1-\beta)P_{ml}(w \mid Topic(d)) + \beta P_{ml}(w \mid Coll)\end{aligned} \tag{5.7}$$

把公式（5.7）带入公式（5.6）中可以得到公式（5.8）

$$\begin{aligned}Log(q \mid d) = &\sum_{w \in q}^{tf(w,d)>0} \log\left(\frac{(1-\alpha)P_{ml}(w \mid d)}{\alpha[(1-\beta)P_{ml}(w \mid Topic(d)) + \beta P_{ml}(w \mid Coll)]} + 1\right) \\ &+ \sum_{w \in q} \log(\alpha[(1-\beta)P_{ml}(w \mid Topic(d)) + \beta P_{ml}(w \mid Coll)])\end{aligned} \tag{5.8}$$

从公式（5.8）中我们可以看出，对于任意的查询q和候选问句d，公式右边的第二部分都是一样的，因此它对最后的排序得分没有影响，在此可以忽略。但是对于公式右边的第一部分，它表示单词w在$Topic(d)$中最大似然估计的倒数，即$P_{ml}(w \mid Topic(d))$的倒数。因此可以看出这里的基于问句主题信息平滑和我们熟悉的 IDF 起到的作用是一样的。在一个具体问句主题中出现次数越多的单词，在同一主题检索中起到的作用越次要。

综上所述可以看出，对于查询q在不同问句主题下检索过程中，基于问句主题信息的语言模型可以得到更好的结果，但是在相同问句主题下的检索只能和 IDF 起到相同的效果。因此可以看出，基于问句主题信息的语言模型更容易区分出不同主题下的问句。

5.2.4 基于答案主题信息平滑的语言模型

在 4.2.4 节已经介绍了在社区问答系统中，问句所对答案信息在问句检索中的作用。在这里我们将要介绍问句答案主题信息在问句检索中的作用，由于答案信息相对于问句本身的信息具有更丰富的信息，因此利用 LDA 对答案集合进行主题信息提取时，主题信息会更加的明确，主题更容易确定。在 5.2.3 节详细介绍了问句主题信息在问句检索中的作用，以此类推，由于答案信息相对于问句信息具有更加丰富的信息，因此采用基于答案主题信息对语言模型进行平滑，将得到更好的效果。那么如何把答案主题信息融

入到我们的模型中，在这里我们改写上面的公式（5.8），在这里把公式（5.8）中的问句主题信息改成相应的答案主题信息，修改后的公式为如下公式（5.9）

$$Log(q \mid ans) = \sum_{w \in q}^{tf(w,ans)>0} \log\left(\frac{(1-\alpha)P_{ml}(w \mid ans)}{\alpha[(1-\beta)P_{ml}(w \mid Topic(ans)) + \beta P_{ml}(w \mid anscoll)]} + 1\right) + \sum_{w \in q} \log(\alpha[(1-\beta)P_{ml}(w \mid Topic(ans)) + \beta P_{ml}(w \mid anscoll)]) \quad (5.9)$$

其中ans表示问句所对应的答案信息，$Topic(ans)$表示应问句答案所在的主题信息，$anscoll$表示语料集中的所有问句的答案集合。

5.2.5 融合问句类别信息和答案类别信息平滑的语言模型

由5.2.3节和5.2.4节介绍可知，在社区问答系统中，对于问句检索任务，除了问句本身所能提供的主题信息外，问句所对应的答案信息提供的主题信息，对问句的检索任务也是十分有帮助的。如何有效的利用这两者信息来提高问句检索的性能，将成为本章研究的重点。在这里我们采用线性加权的方式结合两者信息，之所以采用线性加权的方式，是因为这种方式简单直观，并且效果显著，参数容易控制和调节。融合两者信息后的公式为如下公式（5.10）：

$$Score(q,d) = (1-\alpha)\log(q, Topipc(d)) + \alpha \log(q, ans(d)) \quad (5.10)$$

在这里$Score(q,d)$为最后的相似得分，α是一个线性参数，α越大说明问句对应的答案信息越重要，反之则说明问句本身的信息越重要。$\log(q, Topic(d))$是根据公式（5.8）得到的得分，$\log(q, ans(d))$是根据公式（5.9）得到的得分。

5.3 实验设计

5.3.1 实验数据

本次实验语料是取自Yahoo! Answers的一个数据集。由于数据集存在一定的数据稀疏性，即有些类别的问句数量比较多，有些类别的问句数量非常少。为了保持数据集的平衡性，我们选取相对比较热门的问句类别。首先根据问句类别中问句的数量对每个问句类别进行排序，在这里选取问句类别中问句数量最多的60个类别。在我们所选取的问句类别中，每个类别的问句数量均在1000个以上。实验的语料信息如表5.1：

表5.1 语料集得统计信息情况

Tab. 5.1 Statistical information of corpus

问题数	答案数	最佳答案数量	类别数	问题平均答案数
200,998	1,848,441	201,075	60	9.405

在我们的实验中由于选取的语料都是已解决的问题答案对，即每个问题都有最佳的答案，并且问题所对应的最佳答案一般都是由提问这个问题的用户经过对比其他答案和深思熟虑后选出来的，或者是由社区问答系统中的其他有过类似问题的用户投票选出来的，因此我们可以认为问题对应的最佳答案代表了该问题的最准确信息。并基于这样的假设：我们根据有共同最佳答案的问句是最相似的来判定两个句子是否相似，因此分别选出这 143 个查询所对应有共同最佳答案的问句作为标准答案集，每个查询都对应 5 到 6 个这样的对应问句。

从 60 个类别的问句中去除这 143 个查询问句以及标准答案问句，重新对每个类别随机选择 500 个问句,共选取 30000 个句子作为噪音句,同时把这 143 个查询对应的标准答案集和这 30000 个噪音句融合在一起，并在此数据集上进行实验。

5.3.2 参数选择

在本次的实验中需要两个平滑参数，表 5.2 展示出的结果是在一个小的语料集上的结果，这个语料包含 30 个查询和 1000 个句子，我们使用 LM+T 模型采用 MAP 值作为评价指标，在这个小语料集上对参数进行选择，最后显示在α=0.2，β=0.2 的情况下去的最好的结果。

表 5.2 参数选择（MAP）

Tab. 5.2 Parameter selection(MAP)

β \ α	0.1	0.2	0.3
0.1	0.8572	0.8623	0.8564
0.2	0.8643	0.8714	0.8524
0.3	0.8576	0.8619	0.8436

5.3.3 实验结果与分析

在本次实验中，为了能够更加公平合理的评价本章提出检索模型的性能，在这里采用学术界广泛认同的 P@N 和 MAP 两个指标对实验结果进行评价。同时为了显示本章提出的模型较之前的模型在性能上有一定的优势，在这里采用了 7 种不同的检索模型进行对比实验。并在实验结束后，对实验结果进行详细的对比分析。表 5.3 中是 7 中不同方法的描述，其中（1）、（2）、（4）、（5）分别是文献[28]、[23]、[29]、 [41]提出的方法（Base Line），

表 5.3　实验方法和表述

Tab. 5.3　Experimental methods and description

方法名称	方法描述
（1）LM	语言模型（采用 Jelinek_Mercer 平滑）(Base Line)
（2）LM+C	基于问句类别信息的语言模型 (Base Line)
（3）LM+A	基于答案类别信息的语言模型
（4）TRLM	基于翻译的语言模型 (Base Line)
（5）CLM	集成类别信息的语言模型 (Base Line)
（6）LM+T	基于问句主题信息的语言模型
（7）LM+T+A	融合问句主题信息和问句答案主题的语言模型（本章中的方法）

在语料集中的实验结果如表 5.4

表 5.4　在语料集上 7 种模型的 MAP 和 P@1 值

Tab. 5.4　MAP and P@1 performance of the seven models on relevant corpus

模型	MAP	P@1
LM	0.8463	0.7587
LM+C	0.8504	0.7867
LM+A	0.8662	0.7943
TRLM	0.8678	0.7968
CLM	0.8756	0.8075
LM+T	0.8722	0.7972
LM+T+A	0.8873	0.8164

从表 5.4 结果可以看出：

（1）在本章的实验中采用基于问句类别信息的语言模型（LM+C）比传统的语言模型（LM）在 MAP 和 P@1 上都有一定的提高，这主要是因为采用类别信息的语言模型除了拥有传统语言模型的优势外，还考虑了句子本身的类别信息。同时也证明了问句的类别信息在提高问句检索性能上是有帮助的。(Row2 VS Row1)

（2）采用基于问句答案类别信息的语言模型(LM+A)比采用基于问句类别信息的语言模型(LM+C)在 MAP 提高了 0.0218，在 P@1 上提高了 0.0105。这表明问句的答案信息比问句本身的信息更能反映出问题的实质性内容，也能更好的表达用户的真实意图。同时也证明了问句的答案类别信息在问句检索中具有更强的区分度。(Row3 VS Row2)

（3）采用基于翻译的语言模型(TRLM)要比采用基于问句类别信息的语言模型(LM+C)和采用基于问句答案类别信息的语言模型(LM+A)在 MAP 和 P@1 上都有所提高，这是因为采用基于翻译模型的语言模型除了能更加全面的考虑问句本身的信息外，

又融入了一些外部的信息来对问句进行补充，所以检索效果要好一些。(Row4 VS Row3，Row2)

（4）集成类别信息的语言模型（CLM）要比采用基于翻译的语言模型(TRLM) 在MAP 和 P@1 上都有所提高，主要是因为 CLM 对问句的类别信息进行了高度的集成，从分类类别上更加的精准化，同时考虑了更多的语言信息。（Row5 VS Row4）

（5）采用基于问句主题信息平滑的语言模型（LM+T）相对于采用基于翻译的语言模型（TRLM）、基于问句答案类别的语言模型（LM+A）、基于问句类别信息的语言模型（LM+C）在 MAP 和 P@1 上都有所提高，表明采用宏观的主题信息能更好的区分问句。同时，采用 LDA 对问句进行主题信息提取后，每个问句都属于相应的主题类别，主题类别的确定，相对于原始的采用基于类别的模型，解决了问句原始误分类的问题。所以效果要好。但同时比 CLM 在 MAP 和 P@1 要低，主要是因为 CLM 精确的分类器，导致原始基于分类的误分率降低，同时 CLM 采用了更好的相似度计算策略。

（6）采用融合问句主题信息和问句答案主题信息的方法（LM+T+A）,比以上所有的模型不论是 MAP 值，还是 P@1 值都有所提高，可以看出在考虑问句自身主题信息的情况下，再把问句所对应的答案主题信息融合进来，既用到了问句本身的信息又用到了问句所对应的答案信息，从而能更好的表达句子的本身含义，更加有利于问句的检索。从中也证明本章提出的方法是一个有效的检索方法。

5.4 本章小结

本章主要探讨了社区问答系统中的问句检索模型，提出一种融合问句主题信息和问句答案主题信息问句检索模型。从如何理解问句出发，对问句进行了详细的分析，并考虑了社区问答系统自身的特点，即每个问句都有相应的最佳答案。因此该模型首先利用 LDA 分别对问句集合和问句对应的答案集合进行主题信息抽取，然后利用抽取出来的问句主题信息和答案主题信息分别对语言模型进行平滑，最后线性加权两个平滑结果得到最后的相似度等分。多角度的对比实验结果表明，将问句主题信息和问句答案及其主题信息适当融合，用于问句的检索，能够有效解决社区问答系统中句子检索的问题。如实验部分所述，显示了很好的性能。

结　论

随着 web2.0 技术的发展，社区问答系统的发展如雨后春笋。由于社区问答系统不仅能够对专业的问题进行有效的检索，而且还能给用户带来交互式的体验。因此社区问答系统吸引了大量的用户参与其中。在社区问答系统中，人们可以自由的提出自己所关心或关注的问题，并且会得到一些领域专家用户的回答，从而可以更好的解决自己提出的问题。由于任何人都可以在社区问答系统中提出自己的问题和回答别人提出的问题，因此 Yahoo! Answers 等社区问答系统建立几年来已经积累了大量的问答对。问句检索的研究就是为了能够有效地利用这些历史的问答对信息，快速找到与用户关心问题相同或相近的原有问题，缩短用户得到想要的答案的等待时间的。因此本文所研究的问题是社区问答系统中的问句检索技术，主要的工作内容为如下几个方面。

（1）针对传统问句检索模型，缺少语义信息的不足。本文提出了基于特征融合的问句相似度计算方法，该方法主要利用问句本身的词序特征、统计特征、语义特征和问句对应答案的答案特征相结合来解决问句检索问题，在 Yahoo！Answers 数据集上的结果显示，该方法具有一定的有效性。

（2）在对社区问答系统的研究中发现，问句的类别信息在问句检索中有着举足轻重的作用。随后本文提出了融合问句类别信息和答案类别信息的问句检索模型，该模型主要是分别利用了问句本身的类别信息和问句对应的答案类别信息对语言模型进行平滑，最后线性结合两个平滑的结果，对问句进行检索。在 Yahoo！Answers 数据集上的实验结果，表明了该方法在问句检索中的有效性。

（3）在研究问句类别对问句检索有效性的过程中，发现误分类问题对于问句检索的性能有十分重要的影响。同时发现问句背后隐藏的主题信息对于问句检索有着十分重要的意义，因此本文最后提出融合问句的主题信息和问句对应答案主题信息的问句检索模型，该模型首先使用 LDA 分别提取问句集合中的主题信息和问句对应答案集合的主题信息，然后利用问句的主题信息和问句所对应答案的主题信息分别对语言模型进行平滑，最后线性结合两个平滑结果，对问句进行检索。在 Yahoo！Answers 数据集上的实验结果，表明了该方法在问句检索中的有效性。

下一步工作主要是对实验中存在的一些问题进行改进，主要包括：

（1）本文提出的基于特征融合的问句相似度计算方法，虽然能够有效的弥补其他模型中缺少语义信息的不足，但是在利用 WordNet 来提取语义信息过程中，速度十分缓慢。所以未来的工作将是在保证问句检索性能的前提下，提高问句的检索效率。

（2）融合问句类别信息和答案类别信息的问句检索模型，由于利用了问句类别信息过滤掉了不相关的问句类别，因此在问句检索效率和性能上都用明显的提高。但是这种提高的前提是问句分类的准确率，然而社区问答系统中的分类问题是一个大规模的层次分类问题。目前还没有出现特别有效的方法，因此未来的工作重点可以放在研究大规模的层次分类问题。

（3）融合问句的主题信息和问句对应答案主题信息的问句检索模型，由于利用主题信息合并了相近的问句类别，所以相对于基于类别的问句检索，在性能上有一定的提高，但由于增加了相近的类别，因此检索的效率有一定的下降。另外 LDA 在提取主题信息过程中，也会存在一定的误差。因此未来的工作，可以把重点放在如何在短文本中有效的提取主题信息。

面对传统搜索引擎暴露出来的诸如不能对于专业的问题进行有效的检索、无法给用户带来交互式的体验等问题。社区问答系统的出现，在一定程度上弥补了上述不足。由于社区问答系统中有着庞大的用户群，和大量的历史数据，吸引了大量的学者投身到社区问答系统的研究中。因此社区问答系统的研究，在未来仍然是热门的研究之一。

参 考 文 献

[1] http://www.elickz.com/stats/.全球互联网状况统计

[2] 张亮. 面向开放域的中文问答系统问句处理相关技术研究[D].南京:南京理工大学,2006.

[3] DUAN H, CAO Y, LIN C Y, et al. Searching Questions by Identifying Question Topic and Question Focus[C]//ACL. 2008: 156-164.

[4] LIU D R, CHEN Y H, KAO W C, et al. Integrating expert profile, reputation and link analysis for expert finding in question-answering websites[J]. Information Processing & Management, 2013, 49(1): 312-329

[5] TOBA H, MING Z Y, ADRIANI M, et al. Discovering high quality answers in community question answering archives using a hierarchy of classifiers[J]. Information Sciences, 2014, 261: 101-115.

[6] SHTOK A, DROR G, MAAREK Y, et al. Learning from the past: answering new questions with past answers[C]//Proceedings of the 21st international conference on World Wide Web. ACM, 2012: 759-768.

[7] RIAHI F, ZOLAKTAF Z, SHAFIEI M, et al. Finding expert users in community question answering[C]//Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012: 791-798.

[8] CHEN L, ZHANG D, MARK L. Understanding user intent in community question answering[C]//Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012: 823-828.

[9] BURKE R D, HAMMOND K J, KULYUKIN V, et al. Question answering from frequently asked question files: Experiences with the faq finder system[J]. AI Magazine, 1997, 18(2): 57.

[10] BERGER A, CARUANA R, COHN D, et al. Bridging the lexical chasm: statistical approaches to answer-finding[C]//Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2000: 192-199.

[11] JIJKOUN V, DE RIJKE M. Retrieving answers from frequently asked questions pages on the web[C]//Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005: 76-83.

[12] RIEZLER S, VASSERMAN A, TSOCHANTARIDIS I, et al. Statistical machine translation for query expansion in answer retrieval[C]//ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. 2007, 45(1): 464.

[13] SONG W, FENG M, GU N, et al. Question similarity calculation for FAQ answering[C]//Semantics, Knowledge and Grid, Third International Conference on. IEEE, 2007: 298-301.

[14] 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. ACM, 2005: 617-618.

[15] 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. ACM, 2005: 84-90.

[16] LEE J T, KIM S B, SONG Y I, et al. Bridging lexical gaps between queries and questions on large online Q&A collections with compact translation models[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008: 410-418.

[17] BERNHARD D, GUREVYCH I. Combining lexical semantic resources with question & answer archives for translation-based answer finding[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, 2009: 728-736.

[18] SURDEANU M, CIARAMITA M, ZARAGOZA H. Learning to Rank Answers on Large Online QA Collections[C]//ACL. 2008: 719-727.

[19] WANG B, WANG X, SUN C, et al. Modeling semantic relevance for question-answer pairs in web social communities[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 1230-1238.

[20] COLLINS M, DUFFY N. Convolution kernels for natural language[C]//NIPS. 2001, 2001: 625-632.

[21] WANG K, MING Z, 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. ACM, 2009: 187-194.

[22] XUE G R, XING D, YANG Q, et al. Deep classification in large-scale text hierarchies[C]//Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2008: 619-626.

[23] CAO X, CONG G, CUI B, et al. The use of categorization information in language models for question retrieval[C]//Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009: 265-274.

[24] CAO X, CONG G, CUI B, et al. A generalized framework of exploring category information for question retrieval in community question answer archives[C]//Proceedings of the 19th international conference on World wide web. ACM, 2010: 201-210.

[25] 许展乐, 张琳. 中文问答系统中句子相似度计算方法研究[J]. 现代计算机 (专业版), 2010, 5: 010.

[26] 周舫, 郑逢斌. 汉语句子相似度计算方法及其应用的研究 [D]. 开封：河南大学, 2005.
[27] ROBERTSON S E, WALKER S, JONES S, et al. Okapi at TREC-3[J]. NIST SPECIAL PUBLICATION SP, 1995: 109-109.
[28] ZHAI C, LAFFERTY J. A study of smoothing methods for language models applied to ad hoc information retrieval[C]//Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001: 334-342.
[29] XUE X, 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. ACM, 2008: 475-482.
[30] ZHOU G, CAI L, ZHAO J, et al. Phrase-based translation model for question retrieval in community question answer archives[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 653-662.
[31] SINGH A. Entity based q&a retrieval[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012: 1266-1277.
[32] 吕学强, 任飞亮, 黄志丹, 等. 句子相似模型和最相似句子查找算法[J]. 东北大学学报: 自然科学版, 2003, 24(6): 531-534.
[33] ZHANG Y, WANG X, WANG X, et al. Diversifying Question Recommendations in Community-Based Question Answering[C]//Neural Information Processing. Springer Berlin Heidelberg, 2011: 177-186.
[34] WANG J, LI Z, HU B. A context approach to measuring similarity between questions in the community-based QA services[C]//Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on. IEEE, 2010, 5: 2408-2411.
[35] XU J, CROFT W B. Query expansion using local and global document analysis[C]//Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1996: 4-11.
[36] WANG X, ZHAI C X. Mining term association patterns from search logs for effective query reformulation[C]//Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008: 479-488.
[37] CAI L, ZHOU G, LIU K, et al. Learning the Latent Topics for Question Retrieval in Community QA[C]//IJCNLP. 2011, 11: 273-281.
[38] PONTE J M, CROFT W B. A language modeling approach to information retrieval[C]//Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998: 275-281.

[39] PETKOVA D, CROFT W B. Hierarchical language models for expert finding in enterprise corpora[J]. International Journal on Artificial Intelligence Tools, 2008, 17(01): 5-18.

[40] SORICUT R, BRILL E. Automatic Question Answering: Beyond the Factoid[C]//HLT-NAACL. 2004: 57-64.

[41] JI Z, XU F, WANG B. A category-integrated language model for question retrieval in community question answering[M]//Information Retrieval Technology. Springer Berlin Heidelberg, 2012: 14-25.

[42] WAN Q, HUANG S, WEI M. Research on Pretreatment of Questions Based on Large-scale Real Questions Set[J]. Journal of Networks, 2013, 8(8).

[43] ZHANG Z, LI Q, ZENG D, et al. Extracting evolutionary communities in community question answering[J]. Journal of the Association for Information Science and Technology, 2014.

[44] SUN Y, WANG X, WANG X, et al. Ensemble similarity measure for community-based question answer[J]. The Journal of China Universities of Posts and Telecommunications, 2014, 21(1): 116-121.

[45] CELIKYILMAZ A, HAKKANI-TUR D, TUR G. LDA based similarity modeling for question answering[C]//Proceedings of the NAACL HLT 2010 Workshop on Semantic Search. Association for Computational Linguistics, 2010: 1-9.

[46] 熊大平, 王健, 林鸿飞. 一种基于 LDA 的社区问答问句相似度计算方法[J]. 中文信息学报, 2012, 26(5): 40-45.

[47] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. the Journal of machine Learning research, 2003, 3: 993-1022.

[48] 许景楠. 基于评论和评分的个性化推荐算法研究[D]. 杭州：浙江大学，2013

[49] JO Y, OH A H. Aspect and sentiment unification model for online review analysis[C]//Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011: 815-824.

[50] 王萍. 基于概率主题模型的文献知识挖掘[J]. 情报学报, 2011, 30(6): 583-590.

[51] 单斌, 李芳. 基于 LDA 话题演化研究方法综述[J]. 中文信息学报, 2010, 24(6): 43-49.

[52] 杨潇, 马军, 杨同峰, 等. 主题模型 LDA 的多文档自动文摘[J]. 智能系统学报, 2010, 5(2): 169-176.

[53] 熊大平. 社区问答中问句相似度计算和分类技术的研究[D]. 大连：大连理工大学，2013.

攻读硕士学位期间发表学术论文情况

1　基于特征融合的社区问答问句相似度计算. **杨海天**，王健，林鸿飞. 江西师范大学学报（自然科学版），2013 年，37（2）：125-129. 主办单位：江西师范大学。（本硕士学位论文第一、二、三、四、五章）

2　一种基于主题类别信息问句检索的新方法. **杨海天**，王健，林鸿飞. 计算机应用与软件（已录用）. 主办单位：上海市计算技术研究所、上海计算机软件技术开发中心。（本硕士学位论文第四、五章）

备注：国家自然科学基金资助项目(60973068)、高等学校博士学科点专项科研基金资助项目 (20090041110002)、辽宁省自然科学基金 (201202031)

致　　谢

在论文即将完成之际，回顾三年的研究生生活，感慨万千。研究生期间的三年学习生活，不仅使我在专业基础理论方面的知识更加扎实，而且在为人做事方面也有很大的进步，学会的懂得分享，在分享中享受快乐。在此我心怀感激。

首先，在此我要感谢我敬爱的导师王健老师，王老师为人和蔼、治学严谨、责任心强、乐观的精神、认真务实的态度，这些高尚的品质都深深的影响着我，让我受益匪浅。王老师不仅在学术上对我进行督促和指导，在生活上也对我关怀备至。使我在这座陌生的城市有了家的感觉。在小论文发表及毕业论文书写期间，从研究方向的确定到实验中出现的各种问题，王老师都对我细心的指导，遇到问题给我及时的解答。使我从一个学术的门外汉，快速的融入到学术的大家庭之中。她严谨的工作态度、渊博的知识和对科研的执着，将是我以后工作、生活和学习上的榜样。同时也要感谢实验室的林鸿飞老师和杨志豪老师，两位老师治学严谨的态度、敬业的精神深深的影响着我。

其次，要感谢一直以来给予我最直接帮助的熊大平师兄、徐谦师兄以及教研室其他师兄师姐。感谢大平师兄在我确定方向和实验中给我的细心指导，每当遇到实验中的问题，大平师兄都不难其烦的给我讲解，使我受益良多。感谢徐谦师兄给我营造了一个好学习生活氛围，每一次的春游，每一次的聚餐，都给我留下了美好的回忆。感谢李瑞敏师姐、任克江师兄在学术论文方面给予了我很多的指导，每一次的组会都市我受益良多。在此感谢以上已经毕业的师兄师姐，希望他们工作顺心，事业有成。同时也感谢文本组的现有师弟师妹们，有了你们才使得我的研究生生活更加充实，希望你们在今后的研究生生活中更加的充实精彩。感谢我同届的郭青、闫俊、任巨伟、徐博、刘晓霞、李浩瑞、李宗耀、魏现辉、于凤英、何文译等一起陪我走过了研究生三年生活，希望你们以后事业有成，身体健康。

最后衷心的感谢我的父亲，感谢您对我多年的养育和支持，使我能够安于学业。感谢我的妹妹对我的理解和支持，感谢所有爱我的人和我爱的人。

大连理工大学学位论文版权使用授权书

本人完全了解学校有关学位论文知识产权的规定，在校攻读学位期间论文工作的知识产权属于大连理工大学，允许论文被查阅和借阅。学校有权保留论文并向国家有关部门或机构送交论文的复印件和电子版，可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印、或扫描等复制手段保存和汇编本学位论文。

学位论文题目：_____

作 者 签 名：_____ 日期：_____年____月____日

导 师 签 名：_____ 日期：_____年____月____日