综合句法结构及语义相似度的问题推荐技术

段利国 陈俊杰
(太原理工大学计算机科学与技术学院 太原 030024)

摘 要 针对因特网上的大规模问答对资源提出一种新的应用,即在问答系统中加入基于百度知道平台构建的大规模问答对库,通过相似度计算,把库中最相似的问题推荐给用户。实验下载网页10500个,成功提取问答对4687个,运用关键词的TF/IDF、树核函数的句法匹配及问句的语义距离3种方法中的一种、两种和三种进行实验,分别获得79.44%,81.67%和88.33%的准确率。结果表明,综合运用多种方法查找相似问题,效果更好。

关键词 问答系统,信息抽取,问题推荐,语义距离,树核函数
中图法分类号 TP391 **文献标识码** A

Question Recommended Technology of Integrated Sentence Structure and Semantic Similarity

DUAN Li-guo CHEN Jun-jie
(College of Computer Science & Technology, Taiyuan University of Technology, Taiyuan 030024, China)

Abstract A kind of new application was proposed towards large-scale Question Answer(QA) pairs resource in this paper. Large-scale QA pairs library based on BaiDu ZhiDao platform was constructed and joined to QA system firstly. Then the question with the highest similarity in the library was recommended to the user by similarity calculation. We downloaded 10500 Web pages in the experiments and extracted 4687 QA pairs successfully. Results of experimental applications utilizing TF/IDF of keywords, syntax match of tree kernel function, semantic distance of sentences synthetically were given to illustrate the proposed technique. The application of our experiments obtained accurate rate by 79.44%, 81.67% and 88.33% respectively in terms of using 1, 2 or 3 methods abovementioned. The experimental results show that using one more methods synthetically to calculate similarity can acquire more preferable effects.

Keywords Question and answer system, Information extraction, Question recommend, Semantic distance, Tree kernel function

知识在人们的生活和工作中发挥着巨大的作用。随着网络技术的飞速发展,因特网上积累了数量巨大的、包含知识的信息。可是,传统的信息检索工具,如搜索引擎返回给用户的仍然是浩瀚的信息而不是具体的知识。为满足人们日益增长的知识获取需求,支持自然语言提问并能返回准确答案的智能问答系统成为近年来的研究热点。

自从TREC(Text Retrieval Conference)在20世纪90年代增加对问答系统的评测以来,问答系统相关技术得到了很大的发展。长期以来,TREC问答的研究一直主要针对简短的和基于事实的问题,因为这类问题的答案都比较简明。在现实世界中,用户通常会提出更加复杂的问题,他们想获得更长的、更加全面的、包含足够上下文信息的答案。针对这类问题,目前的问答系统还没有好的解决办法。

随着互联网的快速发展,互联网上积累了大规模的问答对,如英文的YahooAnswers(http://answers.yahoo.com),中文的百度知道(http://zhidao.baidu.com)和新浪爱问(http://iask.sina.com.cn)等,这是一种知识共享性服务。如果在问答系统中能够充分利用这些问答对资源来回答用户的提问,将是一件非常有意义的事情。1)支持自然语言问答。在这些网站上,用户可以免费地以自然语言方式发布自己的问题,等待别人的回答,也可以以自然语言方式回答别人提出的问题,实现知识的共享。2)问题不受领域的限制。目前这类网站的日访问量都非常高,仅以百度知道为例,截止到2011年7月25日,百度知道积累了145802766个问题及其答案[1]。随着时间的推移,不计其数的问答对将被收藏到它的数据库中。这些问答对不仅数量多,而且覆盖面广,是真正的开放域问答。3)答案准确率高。网络社区中的问题一般由专业人士回答,而且答案大都经过多位网络用户和问题发布者自己确认,是人类智慧选择的结果,答案准确率非常高。4)可以回答复杂问题。现有的自动问答系统主要针对简短的和基于事实的问题,利用网络问答社区可以回答复杂问题。比如百度知道有数目众多的由具有共同专长、共同兴趣、共同追求的知友组成的知识团队,他们以科学的方法、专业的知识和热忱的态度回答别人的问题。百度知道希望通过团队,加强知道社区平台上用户间的交流,从而互相帮助,共同获取更多知识,同时为百度知道提供更多、更好、更准确的答案[1]。

最近的研究开始关注因特网上的大规模问答服务[2,3]。一些研究开始对这种类型的服务进行特征分析,比如文献

到稿日期:2011-02-01 返修日期:2011-03-15 本文受国家自然科学基金项目(60970059),山西省国际科技合作计划项目(2009081022)资助。
段利国(1970—),男,博士生,副教授,CCF会员,主要研究方向为自然语言处理、汉语问答系统,E-mail:tyutdlg@163.com;陈俊杰(1956—),男,博士,教授,主要研究方向为智能信息处理。

[2]，但是作为相似问题推荐给用户的研究还很少[3]。另一方面，不同的用户有不同的思维方式，因此对于同样一个问题，不同的用户可能有不同的提问方式。受表达能力、表述习惯、文化程度等因素的影响，不同用户能够科学、准确地表达他要问的问题的程度也不相同。文献[4]表明，在使用网络搜索引擎的时候，只有25%的查询能清晰表达用户的意图。用户的问题不准确，会影响系统回答的准确性。

鉴于上述原因，本文在问答系统服务器端配置一个基于关系模式的大规模问答对库，事先从百度知道提取大量问答对存储到问答对库中。用户提问时，先计算用户问题与问答对库中问题的相似度，选取相似度最高的若干个问题推荐给用户。如果用户确认正是他要问的问题，其就将问题对应的答案反馈给用户，如图1虚线所示路径。否则，启动第二策略，即基于Internet实时检索信息并提取答案，如图1实线所示路径。这样，自动问答系统中复杂的信息检索和答案抽取任务，就会转化为在问答对中寻找相似问题，并把对应答案推荐给用户的简单的问题匹配过程。

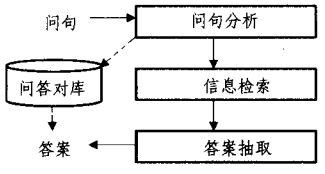

图1　问答系统中获取答案的两种途径

1　大规模问答对库的建立

每个人都有自己的知识体系，这个知识体系可以帮助我们回答一些具体的问题。在问答系统中，问答对库的作用相当于人类的知识体系。问答对的规模和质量对问答系统的性能将产生重要的影响。因此，如何搜集大规模、高质量的问答对，建立问答对库，是开放域问答系统的一项基础性的、重要的研究内容。我们依托百度知道平台建立自己的问答对库，具体分为网页下载、信息提取和建立基于关系模式的问答对库3个阶段。

1.1　基于“百度知道”分类的百度知道网页下载

为了建立问答对库，首先需要收集大规模的常见问题(Frequently Asked Question，FAQ)页面。我们采用的方法是通过网络爬虫来自动下载百度知道页面。百度知道的问答对采用三级分类体系：一级分类包括电脑/网络、生活、医疗健康、体育/运动、电子数码、商业/理财、教育/科学、社会民生、文化/艺术、游戏、娱乐休闲、烦恼、资源共享、地区等14个大类；每个大类下包含若干二级分类；二级分类又包含三级分类若干。表1为大类“文化艺术”的详细分类结构，设定网络爬虫的入口地址时参照了表1分类。

表1　百度知道中大类“文化艺术”的详细分类结构

一级分类	二级分类	三级分类
文化艺术	文学	小说、散文、诗歌、戏剧…
	历史话题	中国、日本、英国、唐朝、三国、清朝、近代、明朝、南北朝、汉代、宋代……
	民俗传统	搬家、风水、习俗、结婚、歇后语、对联、百家姓、方言、生肖……
	地理	中国、亚洲、非洲、北美洲、欧洲…
	器乐/声乐	古筝、琵琶、吉他、钢琴……
	舞蹈	街舞、拉丁舞、芭蕾舞、舞步、背景音乐…
	书画美术	素描、油画、国画、漫画、书法、摄影…

1.2　提取问答对信息

通常的网页正文提取方法是建立HTML文档的DOM(Document Object Model)树，然后递归地遍历DOM树，移除树中的噪音信息，如广告、链接群和非重要节点的信息[5]。这是一种通用的方法，适用于来自不同网站的不同格式的网页的信息提取。

仔细分析百度知道网页的HTML文档，发现每一个百度知道页面都具有相同的结构，具体包含问题、答案、其他回答、问题分类、解决时间等主要信息。对答案进一步分析可知，针对某个问题的回答有4种类型，即精彩回答、最佳答案、推荐答案和其他回答。其中，最佳答案有两种产生方式，一是提问者从众多的回答中选择一个最满意的回答作为最佳答案；二是如果在规定期限内提问者没有选择最佳答案，就通过网友投票的方式来确定最佳答案。精彩回答是知道平台上最优秀的知识，其展现样式最为华丽，在搜索中也会得到更多的展示，从而帮助到更多的人。推荐答案是由高级知道网友推荐的质量较好的回答。显然，这3类问题的回答都依赖人类的智慧，具有很高的正确率。我们在提取信息时可采纳这3类答案，而对于准确率没有保证的其他回答则不予采集。基于这一特点，我们设计了如下简单而高效的信息提取算法。

步骤1　读取网页的HTML文件的内容，赋值给字符串变量S；

步骤2　依据网页标题特征模板〈title〉……_百度知道〈/title〉，从S中提取网页标题，也就是问句本身；

步骤3　如果存在“最佳答案”、“精彩回答”和“推荐答案”等特征标记，则利用这些标记在S中提取相应的答案信息，否则认为是无用的网页，停止信息提取，转步骤5；

步骤4　利用特征“解决时间”提取回答问题的日期和时间；

步骤5　算法结束。

需要说明的是，因为在互联网上存在大量的镜像文件和网页转载现象，所以通过网络爬虫得到的大规模网页集合中存在网页重复现象。一般在网页正文信息提取之前要进行网页去重。但是，考虑到使用上述算法提取信息非常快捷，本文就将比较复杂的网页去重转化为依靠数据库中主键唯一性的约束方法来解决。

1.3　建立基于关系模式的问答对库

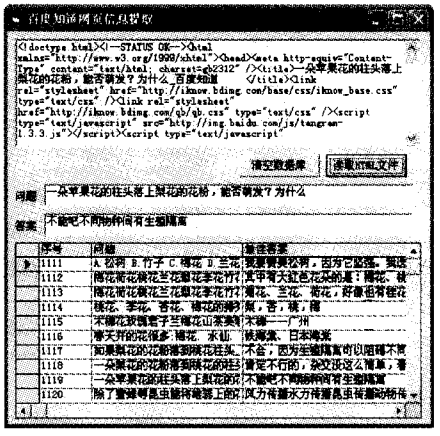

图2　网页信息提取实验界面

本文建立了一个结构为(问题、答案、大类、小类)的问答

对库，并设置问题列为主键。每当提取到新的问答对，都会当作一条新的记录存储到数据库中。存储过程中首先进行主键唯一性检查，如果发现数据库中已经包含同样的问题，则认为是从重复网页中提取到的信息，依据主键唯一的原则，放弃该问答对信息的写入。图 2 是网页信息提取及问答对存储到数据库表中的实验界面。

2 问题推荐相似度计算方法

2.1 基于向量空间模型的 TF/IDF 方法

TF/IDF 方法是一种最基本的度量两个文档之间相似度的方法。

假设问答对库中所有问句包含的词为 $w_1, w_2, \cdots, w_n$，则问答对库中的每一个问题 Q 都可以用一个 n 维向量来表示，$Q=\langle Q_1, Q_2, \cdots, Q_n\rangle$。其中，

$$Q_i = n * \log \frac{M}{m} \tag{1}$$

式中，n 表示关键词 w_i 在问句 Q 中出现的次数，m 为问答对库中含有关键词 w_i 的问句数，M 为问答对库中的问句总数。

用同样的方法可以将用户问句 T 表示为 $T=\langle T_1, T_2, \cdots, T_n\rangle$。这样，问句 Q 和问句 T 之间的相似度就可以用它们向量之间的夹角的余弦值表示：

$$\mathrm{Sim}(Q,T)=\frac{\sum_{i=1}^{n} Q_i \times T_i}{\sqrt{\sum_{i=1}^{n} Q_i^2 \sum_{i=1}^{n} T_i^2}} \tag{2}$$

2.2 基于树核函数的句法匹配方法

在相似问题匹配过程中，纯粹地基于词汇的 TF/IDF 方法有时不能胜任，因为该方法仅仅考虑了关键词的词频因素，没有注意到关键词的位置信息对问句的影响。比如“青蛙吃什么?”和“什么吃青蛙?”两个问句，依据 TF/IDF 方法是完全相似的。实际上，如果用户提问“青蛙吃什么”，系统是不能把“什么吃青蛙”作为相似问题推荐给用户的。

在计算问句相似度时，充分考虑问句的句法结构，往往能够发挥重要作用。树核函数是表现句子句法结构最有效的一种方法[6]。树核函数通过计算两棵解析树之间的相同子树的数量来比较解析树之间的相似度，可以被定义为

$$k(T_1, T_2)=\sum_{n_1 \in N_1, n_2 \in N_2} \Delta(n_1, n_2) \tag{3}$$

式中，n_1 和 n_2 是两个句法树 T_1 和 T_2 中的节点的集合；$\Delta(n_1, n_2)$ 计算以 n_1 和 n_2 为根的共同子树个数，可以通过如下递归算法计算[7]。

步骤 1　如果节点 n_1 和 n_2 的产生规则不一样，$\Delta(n_1, n_2)=0$，否则转步骤 2；

步骤 2　如果节点 n_1 和 n_2 的产生规则一样，而且是叶子节点，$\Delta(n_1, n_2)=1$，否则转步骤 3；

步骤 3　如果节点 n_1 和 n_2 的产生规则一样，而且是叶子节点的前一个节点，$\Delta(n_1, n_2)=\lambda$，否则转步骤 4；

步骤 4　采用如下递归公式计算 $\Delta(n_1, n_2)$：

$$\Delta(n_1, n_2)=\lambda \prod_{k=1}^{nc(n_1)} (1+\Delta(ch(n_1, k), ch(n_2, k))) \tag{4}$$

式中，$nc(n_1)$ 是节点 n_1 的所有子节点数，$ch(n,k)$ 是节点 n 的第 k 个子节点，参数 λ 是一个权重因素，$0<\lambda<1$。

2.3 基于语义的相似度计算方法

在相似问题匹配过程中，纯粹地基于树核函数的句法匹配方法有时也不能胜任，因为该方法没有考虑问句的语义特征。可以借助词语的语义距离来计算两个问句的语义相似度。

知网中存在 Entity, Event, Attribute 等 11 棵义原树。参照文献[8]的做法，在 11 棵义原树中选取 Entity, Event, Attribute, Attribute Value, Quantity, Quantity Value 等 6 棵义原树来计算两个词语之间的语义距离。

定义 1　两个词 w_1, w_2 之间的语义距离 D 定义为这两个词对应的义原在义原树中的最短距离。

如果这两个词中有一个词的义原无法在 6 棵义原树中找到，或者两个词的义原分别处于两棵不同的义原树，则认为这两个词之间的语义距离 $D=\infty$。否则，采用式(5)计算 D：

$$D=|T_1 \cup T_2| - |T_1 \cap T_2| \tag{5}$$

式中，T_1、T_2 分别是两个词所在义原树从树根到该节点语义元素的集合，$T_1 \cup T_2$ 是义原树中从树根到 w_1, w_2 各自语义节点包括的所有义原的集合，$|T_1 \cup T_2|$ 是该集合元素个数，$T_1 \cap T_2$ 表示 w_1, w_2 对应语义树中的相同语义节点集合，$|T_1 \cap T_2|$ 表示公共节点的个数。显然，D 表示义原树中两个词 w_1, w_2 所对应节点之间路径的最短距离。

定义 2　若两个词 w_1, w_2 之间的语义距离为 D，那么 w_1, w_2 之间的相似度可以定义为

$$s(w_1, w_2)=1-D/p \tag{6}$$

因为不同的义原树的深度不同，需通过 p 做归一化处理，其中 p 为义原树中各节点深度之和。因此，两个词相似度取值在 0 到 1 之间[9]。

设用户的问题 T 由词 $T_1, T_2, \cdots, T_n$ 组成，问答对库中的一个问题 Q 由词 $Q_1, Q_2, \cdots, Q_m$ 组成，则词 $T_i(1\leqslant i\leqslant n)$ 和 $Q_j(1\leqslant j\leqslant m)$ 的相似度就可以依据式(7)表示为 $s(T_i, Q_j)$，问句 T 和问句 Q 的相似度可表示为

$$s(T,Q)=\frac{\frac{\sum_{i=1}^{n} t_i}{n}+\frac{\sum_{i=1}^{m} q_j}{m}}{2} \tag{7}$$

式中，$t_i=\max(s(T_i, Q_1), s(T_i, Q_2), \cdots, s(T_i, Q_m))$，$q_i=\max(s(Q_i, T_1), s(Q_i, T_2), \cdots, s(Q_i, T_n))$。

3 实验过程及结果分析

3.1 实验数据

为了尽可能保证数据的平衡性，实验中我们按照表 1 所示的类别下载百度知道网页。共下载网页 10500 个，并成功提取问答对 4687 个，包含了文学、历史话题、民俗、地理、书画等多个方面。提取的全部是具有精彩回答、最佳答案或者推荐答案的问答对。一般认为，如果找到一个相似的问题，就可以直接利用这里的答案。

表 2　提取的问答对在各小类的分布状况

类别	问答对数目
文学	687
历史话题	698
民俗传统	594
地理	634
器乐/声乐	751
舞蹈	599
书画美术	724
总计	4687

由于在一个问题中可能包含多个提问，我们通过问号和疑问词探索法把每个问题切分成单个问句。原因有两个：一是不同的问题可能问的侧重点不同，把它们切分开有利于更好地匹配用户的提问；二是处理器处理短句效果好，长句可能会出现有歧义的句法结构。切分长句后，不同类别得到的问题数如表 2 所列。表 3 给出的是提取的问答对的例子。

表 3　从百度知道网页提取的问答对例子

问题	回答
谁续写的红楼梦？	高鹗
英国会议限制王权的实质是什么	为了资产阶级掌握政权
最经典的二胡曲？	《赛马》，相当经典，有马的声音 《二泉映月》堪称经典
鲁迅的《药》是哪年写的	1919 年 4 月 25 日
清朝入关后的第三位皇帝是谁？	雍正皇帝胤禛

3.2　评测指标

实验的主要目的是评测问题推荐的准确率，准确率通过下式计算：

$$准确率(P)=\frac{正确推荐的问题数(N)}{模拟用户提问的问题总数(M)}\times 100\% \quad (8)$$

3.3　实验方法

为了验证推荐算法的有效性，采用了 3 个不同的相似度计算方法。

方法 1：BoW　即利用基于关键词的 TF/IDF 方法计算问句之间的相似度。

方法 2：BoW＋TK　在方法 1 的基础上加入句法结构，句法结构的匹配采用基于树核函数的句法匹配方法。

方法 3：BoW＋TK＋SEM　在方法 2 中加入语义特征，计算问句的语义相似度。

3.4　实验结果及分析

实验中，我们参照问答对库中的问题，通过人工的方式设计了 180 个问题，模拟用户的提问。设计的问题与问答对库中的问题相比较，形式上尽可能有所区别，而语义上尽可能保持不变。分别采用上述 3 种方法进行测试，实验结果如表 4 所列。方法 2 和方法 3 相对于方法 1 在准确率上的提升幅度如图 3 所示。

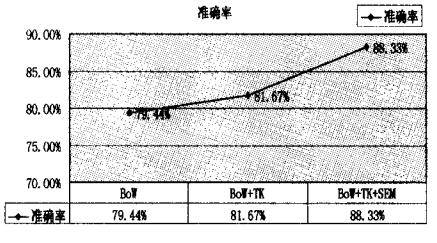

图 3　准确率的提升幅度

表 4　不同方法获得的推荐准确率

推荐方法	测试问题数	准确推荐数	准确率(%)
BoW	180	143	79.44
BoW＋TK	180	147	81.67
BoW＋TK＋SEM	180	159	88.33

分析实验结果可知：

(1)BoW 方法本身的精准度很高(高达 79.44%)，原因可能是实验中设计用户问题时过多地参考了问答对库中的问题，人为地造成二者的高度相似。用户即使输入一些关键词也很可能得到相似问题。

(2)在 BoW 方法的基础上，基于树核函数的句法匹配方法将准确率提升了 2.23%，说明句法树匹配对于相似度计算有明显效果。

(3)当加入语义特征后，准确率在方法 2 的基础上大幅提升了 6.66%。说明在相似度计算过程中要充分考虑句子的语义信息。

分析实验结果可以发现，问题推荐的平均准确率达到 86.14%，似乎有点偏高。原因可能是实验中设计用户问题时过多地参考了问答对库中的问题，人为地造成二者的高度相似。尽管如此，我们仍然能够预见到问题推荐技术在问答系统中的巨大作用。

结束语　问题推荐技术可以弥补用户因表达能力不足而造成的表达意图不明确的缺点，帮助用户构造更准确的提问方式，节省信息检索和答案抽取的时间。此外，通过问题推荐这种交互方式，既可以改善为用户服务的效果，又可以提高问答系统的性能。本文针对因特网上的大规模问答对资源提出一种新的应用，将其作为相似问题推荐给问答系统的用户，并给出了一种应用模式和具体的相似度计算方法。综合使用句法结构和语义相似度技术，可以在问答系统中得到更高的推荐准确率。

当然，问题推荐技术还有很多地方有待进一步探讨和研究。如何充分利用因特网上的问答对为智能问答系统服务，以及探索脱离问答对库的开放域问题推荐算法，将是我们下一步研究的重点。

参 考 文 献

[1]　http://zhidao.baidu.com/,2011-07-25

[2]　Bian J,Liu Y,Agichtein E,et al. Finding the right facts in the crowd:factoid question answering over social media[C]//WWW'08. ACM,2008:467-476

[3]　Wang Kai,Ming Zhao-yan,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

[4]　Strohmaier M, Kröll M, Rner C K. Intentional Query Suggestion:Making User Goals More Explicit During Search [C]// Proceedings of the 2009 Workshop on Web Search Click Data. ACM,2009:68-74

[5]　刘挺，秦兵，张宇，等. 信息检索系统导论[M]. 北京：机械工业出版社，2008

[6]　庄成龙，钱龙华，周国栋. 基于树核函数的实体语义关系抽取方法研究[J]. 中文信息学报，2009,23(1):3

[7]　陈九昌，孔芳，朱巧明，等. 基于树核函数的"it"待消解项识别研究[J]. 中文信息学报，2010,24(5):24

[8]　秦兵，刘挺，王洋，等. 基于常问问题集的中文问答系统研究[J]. 哈尔滨工业大学学报，2003,35(10):1179

[9]　刘宝艳，林鸿飞，赵晶. 基于改进编辑距离和依存文法的汉语句子相似度计算[J]. 计算机应用与软件，2008,25(7):3