

基于常见问题库的多搜索引擎自动问答系统

王慧芝¹, 安玉朋²

(1. 天津大学 电信学院, 天津 300072; 2. 天津市武清区成人教育中心, 天津 301700)

摘要:与传统的搜索引擎相比, 自动问答系统支持自然语言提问, 返回给用户一个简短而准确的答案, 是自然语言处理领域的一个研究热点。文中介绍了一种基于常见问题库的多搜索引擎自动问答系统, 它利用常见问题库和两大搜索引擎, 快速准确地回答用户的问题, 更加智能化地满足用户的检索需求。

关键词:多搜索引擎; 问答系统; 常见问题集; 句子相似度

中图分类号: TP182

文献标识码: A

文章编号: 1673-629X(2006)11-0136-04

Multiple Search - Engine Question - Answering System With Frequently - Asked Question

WANG Hui-zhi¹, AN Yu-peng²

(1. School of Electronic & Information Engineering, Tianjin University, Tianjin 300072, China;

2. Tianjin Wuqing Adult Education Center, Tianjin 301700, China)

Abstract: Comparing with conventional search engine, question - answering (QA) system enables users to ask a question in natural languages and gives users concise and accurate answers. It is a hot research field in the field of natural language processing (NLP). Introduces a multiple search - engine QA system with frequently - asked question (FAQ). In this QA system, FAQ and two search engines are used to answer users' questions fast and accurately. It can satisfy users' search request more intelligent.

Key words: multiple search - engine; question answering system; FAQ; sentence similarity

0 引言

随着互联网的迅速发展和广泛普及, 网上信息呈现爆炸性的增长。人们总希望通过网络信息检索迅速快捷地找到自己所需要的信息, 而这时传统的搜索引擎的弊端就逐渐显露出来。例如: 在百度中输入一个关键词, 常常能找到成百上千以至于上万个网页, 必须逐一阅读这些网页才能找到真正的答案, 并且网上还有哪些相关的网页没有被检索出来, 也无从知道。这就是人们经常所说的“rich data, poor information”。因此, 传统的搜索引擎已不能满足人们的需求, 新的信息检索方式和搜索引擎呼之欲出, 开放域问答式信息检索就是这样新型快捷的检索方式。

早在 20 世纪 60 年代人工智能研究刚开始的时候, 人们就提出了让计算机用自然语言来回答人们的问题, 这就是指自动问答系统。但是, 在初始阶段, 所有的实验都是在非常受限的领域上进行的, 所以自动问答一直被限制在特殊领域的专家系统。最近几年, 随着网络和信息技术的快速发展, 人们想更快地获取信息的愿望, 也使研究学者重新将眼光转到开放域自动问答技术的研究上。

1 系统概况

该自动问答系统包括 4 个部分: FAQ (常见问题库)、问题理解、信息检索和答案抽取。如图 1 所示。

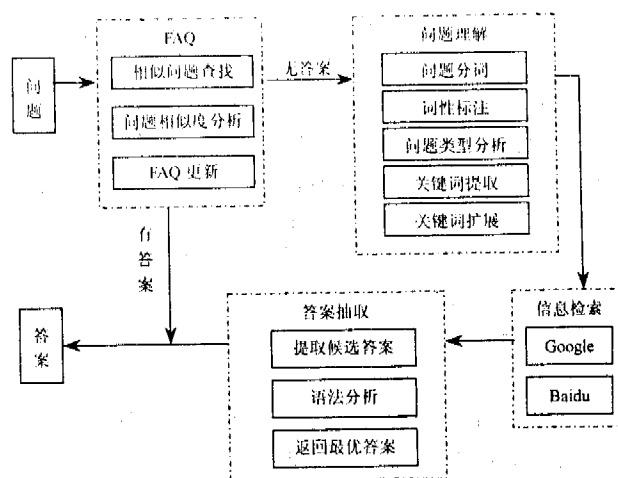


图1 自动问答系统总体框架

对于用户输入的问题, 首先在 FAQ 库 (Frequently - Asked Question) 中搜索, 看看有没有相同的问题。如果有, 就可以把 FAQ 库中这个问题对应的答案直接返回给用户。这样, 对于用户经常问的问题, 问答系统就可以快速准确地给出答案, 而不需要进行语意理解、网络搜索等处理, 大大提高了系统的效率。

收稿日期: 2006-02-27

作者简介: 王慧芝 (1975-), 女, 天津人, 讲师, 硕士, 研究方向为现代网络技术的相关知识。

如果在 FAQ 库中找不到用户输入的问题,就需要我们通过网络搜索来寻找答案。首先要进行问题理解,问题理解是要把用户用自然语言提出的问题转化为符合计算机查询的一系列关键词组合。然后将通过问题理解得到的关键词组合提交给信息检索模块来查找问题的答案,通过检索返回一些相关网页。最后,对于信息检索模块返回的相关网页,答案抽取模块进行语法分析等工作,从而得到问题简短准确的答案返回给用户。

2 主要技术实现

2.1 FAQ 模块

常问问题库把用户经常问的问题、答案和问题查询次数保存起来。对用户提问的问题先在 FAQ 库中搜索,看看有没有相同的问题。如果有,就可以直接把 FAQ 库中这个问题的答案返回给用户。这样,对于用户常问的问题,问答系统就可以很快给出答案,而不需要经过复杂的处理,还能保证答案的正确。所以有了 FAQ 库之后,既能提高问答系统的效率,又能提高准确性。

2.1.1 相似问题集的查找

这一步骤的目的是要从常问问题库(FAQ)中找出若干个候选的问题组成候选问题集,以缩小查找的范围,使后续的相似度分析等较复杂的处理过程都在候选问题集这个相对较小的范围内进行。

2.1.2 问题相似度分析

候选问题集确定后,下一步是从这个集合中找出和用户输入的问句(这里称为目标问句)最相似的问句。文中所用的方法是计算候选问题集中每个问句和目标问句之间的相似度,对应的相似度最大的问句就是要找的句子。计算相似度的方法有很多,一种是基于向量空间模型的 TFIDF 方法,另一种是基于语义^[1]的方法。

2.1.3 FAQ 库的更新

计算出用户所输入的目标问句和候选问题集中每个问句的相似度,如果所有这些计算出来的相似度的最大值大于一定的阈值^[2],那么就认为最大的相似度所对应的问句和用户的目标问句问的是同一个问题。可以直接将这个问句对应的答案输出给用户。

如果最大相似度的值小于阈值,就认为 FAQ 库中没有用户所问的问题,那么必须利用其他的方法(如信息检索、答案抽取等)来找出答案。如果能够找到答案,就可以将用户所问的这个问题和对应的答案加入 FAQ 库。

2.2 问题理解模块

问题理解模块主要包括问题分词、词性标注、问题类型分析、关键词提取和关键词扩展 5 个部分。

2.2.1 问题分词和词性标注

词语是信息表达的最小单位,而汉语不同于西方语言,其句子的词语间没有分隔符(空格),因此需要进行词语切分。汉语词语切分中存在切分歧异,例如句子“当好人大代表”可切分为“当好/人大/代表”,也可能被错误地

切分为“当/好人/大代表”。因而需要利用各种上下文知识解决词语切分歧异。自 20 世纪 80 年代研究汉语自动分词以来,已经提出了多种分词方法,如正向最大匹配、逆向最大匹配、有穷多层次列举、邻接约束、联想-回溯、词频统计、专家系统、神经网络等方法。不同的分词方法模拟了人类分词行为的不同侧面,取得了不同的成效。这里用的是改进的联想-回溯分词算法(Association-Backtracking Word Segmentation)^[1]。该算法不过分依赖词表,而是较多地利用了汉语本身的知识,提出一些处理歧义结构的实用分词规则,采用切分标志法和有穷多次列举的方法,以提高分词精度、速度和分词正确率。

联想-回溯分词算法的结构框图如图 2 所示。

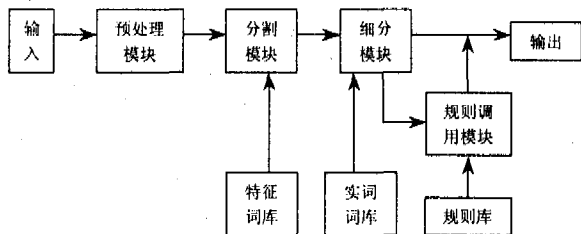


图 2 联想-回溯分词算法的结构框图

其基本思想为:首先将待切分的汉字符串序列依特征词词库分割为若干子串,每个子串或成为词或为词群(几个词组合而成的线性序列);然后利用实词库和规则库再将词群细分为词。

预处理模块将用户提出的问题依照各种形态标志(主要是标点符号和数字等非汉字字符)分解成独立的、可直接处理的子串序列;分割模块对预处理后的问题进行扫描,它以特征词库中的词作为词切分标志、依靠联想规则将一个子串分割为多个更小的子串;细分模块依据实词库内容将从分割模块得到的子串切分为词;规则调用模块调用相应的规则处理歧义组合结构,或调用通用规则切分出类型词。

词性标注是在给定句子中判定每个词的语法范畴,确定其词性并加以标注的过程^[3]。在词语切分的基础上,利用基于规则和基于 Markov(马尔科夫链)模型的概率统计的方法进行词性标注。基于 Markov 链随机过程的 n 元法(n -grams)统计分析方法,已经成功地应用于词性标注和语音识别等领域中,能达到较高的精度。利用 n 元法,在已经切分好的词后面用特定的符号标明这个词的词性。例如:/NG 表示普通名词,/NP 表示专有名词,/UT 表示时态助词,/NPFF 表示姓等。

2.2.2 问题类型分析

通过对大量问题的统计分析发现,用户提出的问题大概可以分为若干种类型,如询问人(谁、什么人)、询问时间(何时、哪年)、询问地点和位置(哪里、什么地方)、询问数量(多少)、询问原因(为什么、为何)、询问定义(什么是)、其他等。对于不同类型的问题,答案的组成方式也不同。

问题分类主要依靠问题中的疑问词。在英语中,疑问词数目有限且在句中的位置固定,如以 wh 开头和 how 开

始的句子。而对于汉语,问句提问的方式就太多了。如果问句中出现疑问词“哪里”或“哪儿”,可以很容易地判断出问题类型为“询问地点”;可是如果问句中出现“什么”,我们就不能仅靠疑问词来判断出问题类型了,此时就需要通过问句中的另外一个词(疑问修饰词)来判断问题类型。

可根据事先规定好的问题类别进行分类,对于不能判断类型的疑问词,按照一定的规则提取疑问修饰词,确定问题类型。

2.2.3 关键词提取和关键词扩展

问题关键词的提取直接影响到后面的检索的结果,并不是问题中所有的词都可以作为关键词,有许多语气词、助词等对问题的解答没有任何的帮助,应直接过滤掉,例如:啊、把、不但、别的等。一般来说,关键词主要有名词、动词、形容词、限定性副词等组成,但在实际应用中为了提高检索的精度,可以把除了疑问词以外的大部分词作为关键词。关键词可能分为两种:一般性关键词、“必须含有”性关键词。所谓“必须含有”的关键词指的是这些关键词必须在答案中含有,而一般性关键词可以不被答案所包含。关键词被赋予不同的权重,在检索答案时这些权重用来计算句子的权重,通常名词、具有限定性副词、时间有比较高的权重。

在检索问题答案时,对关键词进行适当的扩展可以提高系统的召回率。如果不对关键词进行任何扩展,有可能会造成检索的失败;但如果扩展不适当就会极大地降低检索的准确率。这里,在两方面进行关键词扩展:同义词扩展和答案特征词扩展。同义词扩展指将关键词的同义词也作为关键词,主要包括名词和动词的同义词的扩展;答案特征词扩展指对于某些类型的问题,所对应的答案中经常会出现某种共同特征的词,将这些词也做相应的扩展。例如询问数量(多少)时,可以扩展一些表示数量的单位,这样可以提高系统查询的正确率。

2.3 信息检索模块

当在 FAQ 库中找不到答案时,只有依赖整个网络作为知识库,在网络中检索问题的答案。因此搜索引擎的选择决定了获得知识的范围和方法。

百度搜索引擎是目前世界上数据更新时间最快、中文信息量最大的中文搜索引擎,以其优秀的中文信息检索与传递技术被公认为是众多搜索引擎中的佼佼者。百度搜索引擎由 4 部分组成:蜘蛛程序、监控程序、索引数据库、检索程序,检索范围涵盖了中国大陆、香港、台湾、澳门以及新加坡等华语地区和北美、欧洲的部分站点。它使用了高性能的调度算法使得搜索器能在极短的时间内收集到最大数量的互联网信息。百度以关键词搜索为核心,支持布尔逻辑检索、限定范围检索等基本搜索方法,支持动态网页的检索。

Google 中文搜索引擎是收集亚洲网站最多的搜索引擎之一,检索速度极快,检索命中率。它采用了完善的文本对应技术和先进的基于 PAGERANK 和 HITS 算法

的排序技术。当用户输入关键字搜索时,Google 不仅会去搜索包含关键字的网页,同时还会搜索和这些网页具有高相关性的网页,按照关键字的接近度区分搜索结果的优先次序。筛选与关键字较为接近的结果。并且,在显示的结果中,Google 将检索结果按相关性从大到小排序,只摘录包含用户查询字符串的内容作为网页简介,且查询字符串醒目地高亮显示。这使用户尽可能地不受其他无关结果的烦扰,从而节省了查阅时间,同时也大大提高了查询结果的精确度。与其他中文搜索引擎相比,Google 是目前唯一支持中文图片搜索的搜索引擎,但美中不足的是,其数据的更新速度无法进一步提高,这在一定程度上影响了用户对信息的时效需求^[4]。

所以在这里,使用百度和 Google 来进行问题答案的检索,可以提高检索率和召回率。

2.4 答案抽取模块

信息检索模块返回的是一些网页,而问答系统所要返回的应该是简短的答案,因此需要答案抽取模块将正确答案从网页中提取出来,答案的形式应该是词语、句子或者段落。答案抽取模块主要包括 3 个部分:提取候选答案、语法分析和返回最优答案。

答案抽取的基本过程为:首先将搜索到的大量网页处理并去重后得到纯文本,从这些相关文档中搜索出可能包含候选答案的句组(summary);对搜索出的句组进行评估打分,打分时要考虑文本的排序位置,选出最有可能包含答案的前若干组;根据问题类型分析确定答案的类别和对句组的评分情况,从这些句组里抽取最佳的答案返回给用户。

2.4.1 以句子作为答案

为了处理的方便,很多的问答系统返回的是句子作为答案。在这种系统中,答案的抽取步骤如下:

- (1) 把检索出来的文档分成句子;
- (2) 按照一定的算法,给每个句子打分;
- (3) 对句子按照分值进行排序;
- (4) 根据问题的类型对候选答案重新排序。

经过重新排序后,排在最前面的那个句子就是问答系统返回的最终答案。

2.4.2 以词或短语作为答案

如果以句子作为答案,处理起来相对简单一些。但是,对于那些问时间地点的问题,其答案就比较简短,而用不着一句话。比如,对于问题:“中华人民共和国是什么时候成立的?”可能检索出这样的一句话:“自从 1949 年 10 月 1 日中华人民共和国成立以来至 1994 年底止,我国已经同世界上的约 160 个国家建立了外交关系,而且还同更多的国家和地区发展了经济贸易关系和文化往来。”从这个例子可以看出,所要的答案只是这句话中的一小部分,如果能把这整句话作为答案都提交给用户的话,显然冗余信息太多。所以有些问答系统希望直接把包含答案的那段话抽取出来。

2.4.3 以文摘作为答案

对于有些问题,简短的一个短语或者一句话很难说清楚,比如对于问题“火烧圆明园是怎么回事?”。像这种问题,在互联网上有许多相关的报道,如果把这些相关报道都交给用户的话,那么用户将要花很多时间来阅读。如果能把这些相关报道做成一个简短的文摘,让用户只要看文摘就能知道整个事件的前因后果,那么将会为用户带来很大的方便。这就需要用到多文档自动文摘技术。多文档自动文摘模块把信息检索模块检索出来的相关文档做成文摘,再把这个文摘作为答案返回给用户。^[5]

在答案抽取过程中,对句组评估打分时需要设定一种打分规则,计算问题和答案之间的相似度。对句组打分可以依据以下条件:

- 1) 是否是期待的答案类型,例如:用户询问时间时,打分的句组中不含时间,则表示此句组肯定不是答案。
- 2) 含有匹配关键词和扩展关键词的个数;
- 3) 含有匹配关键词和扩展关键词之间的最大距离。

3 结 论

中文自动问答系统可以说是一种新型的中文智能搜索引擎,用户既能用自然语言句子提问,又能为用户直接返回所需的答案,而不是相关的网页。所以,问答系统能

更好地满足用户的检索需求,能更快地找出用户所需的答案。对于问答系统,用户不需要把自己的问题分解成关键字,用户可以把整个问题直接交给问答系统。问答系统就像一个知识渊博的专家,可以快速准确地回答任何问题。比如,用户提交一个问题“上海的简称是什么?”问答系统将会直接给出答案“上海的简称是沪”。可以看出,问答系统要比传统的搜索引擎方便、快捷、高效。自动问答系统还可应用在远程教育、企业客户咨询等方面。广阔的应用前景正推动着自动问答技术的快速发展,相信在不久的将来问答系统将会取得重大的突破并且得到广泛的应用。

参考文献:

- [1] 秦 兵,刘 挺,基于常问问题集的中文问答系统研究[J].哈工大学报,2003,35(10):1179-1182.
- [2] 王 洋,秦 兵,郑实福.句子相似度计算在 FAQ 中的应用[EB/OL].2002-10-30.中国语言文字网.
- [3] 刘开瑛.中文文本自动分词和标注[M].北京:商务印书馆,2000:162-199,231-246.
- [4] 薛万新.常用中文搜索引擎的特征分析[J].科技情报开发与经济,2004,14(7):209-210.
- [5] 郑实福,秦 兵,刘 挺,等.中文自动问答系统综述[J].中文信息学报,2002,6(16):46-52.

(上接第 135 页)

使其缺少强有力的支撑体系。这些都使得 WSMO 在与 OWL-S 的比较中处于劣势^[7]。

4 结论与展望

在确定了语义 Web 服务领域的服务描述框架的基础上,可以进行关于语义 Web 服务发现的研究。目前在语义 Web 服务发现领域的研究也可以根据 OWL-S 和 WSMO 的体系大体分为两类。在 WSMO 领域的研究,服务与一个包含语义的服务描述绑定,通过用户选择目标组件本体来调用 Web 服务。在 OWL-S 领域,因为 OWL-S 框架还在完善过程中,尚没有完整地提出语义 Web 服务发现框架。鉴于 OWL-S 框架与 WSDL 与 UDDI 的紧密联系,所以目前的研究主要是如何建立 OWL-S 与 UDDI 和 WSDL 的映射关系^[8],对目前已经存在且标准化较高的 UDDI 注册标准进行一些改进,就可以完成基于语义的服务发现,而不需要像 WSMO 完全重建一套框架。可以较为轻松地实现现有 Web 服务到语义 Web 服务的过渡。

由于 OWL-S 对 Web 服务领域标准和语义 Web 领域标准的兼容性较好和开放灵活的定义方式使得其逐渐成为语义 Web 服务描述框架的事实标准,并且以其为基础的发现、组合和调用的研究也正在进行中。

参考文献:

- [1] 梁晓路,梁宇奇.Web Services 技术构架和应用[M].北京:电子工业出版社,2003.
- [2] Paolucci M, Kawamura T, Payne T R, et al. Semantic Matching of Web Services Capabilities[C]//1st International Semantic Web Conference. Italy:[s. n.],2002.
- [3] HP Labs. HP homepage[EB/OL].2005. <http://www.hpl.hp.com/semweb/swws.htm>.
- [4] W3C. Web Service Modeling Ontology (WSMO) Submission[EB/OL].2005-06. <http://www.w3.org/Submission/2005/06/>.
- [5] W3C. OWL Web Ontology Language for Services[EB/OL].2004-07. <http://www.w3.org/Submission/2004/07/>.
- [6] W3C. Team Comment on the OWL-S Submission[EB/OL].2004-07. <http://www.w3.org/Submission/2004/07/Comment>.
- [7] W3C. Team Comment on WSMO[EB/OL].2005-06. <http://www.w3.org/Submission/2005/06/Comment>.
- [8] Paolucci M, Kawamura T, Payne T R, et al. Importing the Semantic Web in UDDI[C]//1st International Semantic Web Conference. Italy:[s. n.],2002.

热烈祝贺中国计算机事业创建五十周年庆典大会召开!