文章编号:1001 - 9081(2007)09 - 2256 - 02

基于 HowNet 语义相似度的 FAQ 研究

贾可亮[1,2],樊孝忠[1],张 禹[2]
(1. 北京理工大学 计算机学院,北京 100081; 2. 山东经济学院 信息管理学院,济南 250014)
(jiakeliang@ yahoo. com. cn)

摘 要:FAQ 是网站提供在线帮助的主要手段。利用检索机制根据用户提出的问题建立一个候选问句集,利用知网研究了用户问句和候选问句之间的相似度,从中找出最相似的问句,并将相应答案返回给用户。实验表明,该方法提高了问句匹配的准确率。

关键词:知网;Frequently Asked Question(FAQ);句子语义相似度
中图分类号: TP391.1 **文献标志码**:A

Research of FAQ based on the semantic similarities of HowNet

JIA Ke-liang[1,2], FAN Xiao-zhong[1], ZHANG Yu[2]
(1. *School of Computer Science and Technology, Beijing Institute of Technology, Beijing* 100081, *China*;
2. *School of Information Management, Shandong Economic University, Jinan Shandong* 250014, *China*)

Abstract: Frequently Asked Question (FAQ) is the main method to provide online help. A candidate question set was built up according to the user's query by search engine. Semantic similarities of sentences, based on HowNet, were computed between the user's query and the candidate questions. The answer, which is the most similar to the query, was returned to the user. Experiments show that the method has done better in the matching between questions and answers.

Key words: HowNet; Frequently Asked Question (FAQ); sentence semantic similarity

0 引言

常问问题集(FAQ)是把用户提出的常见问题和相关答案组织在一起供用户浏览访问。但随着常问问题集的逐渐积累,问句数量日益增大,逐页浏览式的知识获取途径将越来越难于满足用户的实际需求,继续采用这种解疑方式将浪费用户大量的宝贵时间,面向 FAQ 的自动问答系统是解决该问题的一种有效途径。

自动问答系统允许用户以自然语言进行提问,并返回一个简洁、准确的答案,而不是一些相关网页。与传统的检索系统相比,自动问答系统能够更好地表达用户的检索意图,具有准确、快捷和高效等特点,是目前自然语言处理领域的一个研究热点,是 TREC 会议中最受关注的主题之一[1]。目前,国外已经开发出一些面向英文 FAQ 的问答系统,如 FAQ Finder[2]、FAQshare[3]。在国内,文献[4]采用 TF/IDF 和基于语义距离的方法计算句子的相似度,并用于 FAQ 问句实例的检索,取得了较好的研究成果。

1 系统分析

FAQ 自动问答系统的核心问题是如何快速地将客户所提问题与 FAQ 数据库的问题比较,进而确定与其最相似的问题,如果有,则将对应的答案作为结果回复给客户。从数学的角度看,可以用两个映射表示[5]。$f_1: Q_1 \rightarrow Q_2$, $f_2: Q_2 \rightarrow A_2$。其中,Q_1 为客户提问问题,Q_2 为 FAQ 数据库中的问题集,A_2 为 FAQ 数据库中的答案集。系统结构如图 1 所示。

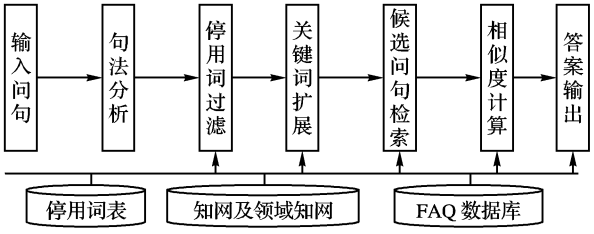

图 1 FAQ 系统结构

词法分析模块主要是对汉语问句进行分词处理,停用词过滤就是去掉客套词(如请问、请问一下)、助词(如的、吗、呢、啊)等对问句意义关系不大但出现频率较高的词。关键词扩展是为后续候选问句检索服务的,主要是进行同义词扩展(如大夫和医生)。

候选问句检索的目的是使后续的相似度计算等较复杂的过程都在候选问题集这个相对较小的范围内进行。文献[4]以问句中的词语为基本单元建立问句的倒排索引,取出 FAQ 问句集与目标问句之间重叠词语数量最大的前 50% 的 FAQ 问句组成候选问题集。这种处理方式存在以下问题:当问题集较大时,取前 50% 的问句仍然过多;且未考虑关键词的扩展问题。

候选问题集就是从大规模问句集中快速取出的一个模糊相关、但相对较小的子集合,因此,该部分的功能可以通过信息检索模块予以实现。这样,一方面可以选择使用成熟稳定的检索系统;另一方面,该模块的功能更改、升级换代也非常容易。

收稿日期:2007 - 03 - 14;**修回日期**:2007 - 05 - 21。 **基金项目**:高等学校博士学科点专项科研基金资助项目(20050007023)。

作者简介:贾可亮(1975 -),男,山东昌乐人,博士研究生,主要研究方向:自然语言理解、汉语自动问答系统、信息抽取; 樊孝忠(1948 -),男,河南许昌人,教授,博士生导师,主要研究方向:自然语言理解、汉语自动问答系统、信息抽取; 张禹(1978 -),男,山东蓬莱人,助教,主要研究方向:自然语言理解、汉语自动问答系统、信息抽取。

确定候选问句集后,下一步的工作就是计算候选问题集中每个问句与输入问句之间的相似度,对应的相似度最大的问句在大于系统指定的阈值时,即为要查找的句子。此时,根据该句子对应的答案 ID,从数据库中自动抽取出有关答案作为输出结果返回给用户。

2　知网及语义相似度计算

知网(Hownet)[6]是一个以汉语和英语的词语所代表的概念为描述对象,以揭示概念与概念之间以及概念所具有的基本属性之间的关系为基本内容的常识知识库。知网对汉语词汇的描述是基于“义原”这一基本概念,义原可以被认为是汉语中最基本的不易于分割的最小语义单位。由于汉语中词在不同的语境中会表达不同的含义,因此 Hownet 把汉语中的词理解为若干个义项的集合。《知网》语义词典中每条记录都是由一个词的一条义项及其描述所组成,即一条记录对应一个词语的一个义项,而每一个义项又是由多个义原来描述的。HowNet 提供了义原分类树,在父节点和子节点之间存在上下位语义关系,因此我们可以利用义原分类树计算两个词之间的语义相似度。

2.1　义原的相似度计算[7]

$$Sim(p_1,p_2) = \frac{2 \times Spd(p_1,p_2)}{Depth(p_1) + Depth(p_2)} \tag{1}$$

其中:p_1,p_2 表示两个义原,$Spd(p_1,p_2)$ 为 p_1,p_2 两个义原的重合度,$Depth(p)$ 为义原在义原树中的深度。

2.2　概念词的相似度计算

知网中实词概念(义项)可以分为 4 个部分:1)第一基本义原描述式,DEF 项中的第一个义原;2)其他基本义原描述式,DEF 项中除第一独立义原以外的所有其他独立义原或具体词;3)关系义原描述式,DEF 项中用“关系义原 = 基本义原”或者“关系义原 = (具体词)”或者“(关系义原 = 具体词)”描述概念的部分;4)符号义原描述式,DEF 项中用“关系符号 基本义原”或者“关系符号(具体词)”描述概念的部分。在此,把两个概念这四部分对应的相似度分别记为 $Sim_1(C_1,C_2)$、$Sim_2(C_1,C_2)$、$Sim_3(C_1,C_2)$、和 $Sim_4(C_1,C_2)$。则概念词的整体相似度为:

$$Sim(C_1,C_2) = \beta_1 Sim_1(C_1,C_2) + \sum_{i=2}^{4} \beta_1\beta_i Sim_i(C_1,C_2) \tag{2}$$

其中,$\beta_i(1 \leqslant i \leqslant 4)$ 并满足:$\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$,$\beta_1 \geqslant \beta_2 \geqslant \beta_3 \geqslant \beta_4 > 0$。

2.3　汉语词的相似度计算

两个汉语词语 W_1 和 W_2,如果 W_1 有 n 个义项(概念):c_{11},c_{12},…,c_{1n},W_2 有 m 个义项:c_{21},c_{22},…,c_{2m}。规定 W_1 和 W_2 的相似度为各义项的相似度之最大值,即:

$$Sim(W_1,W_2) = \max_{i=1\cdots n,j=1\cdots m} Sim(s_{1i},s_{2j}) \tag{3}$$

2.4　汉语句子的相似度计算

两个句子 S_1 和 S_2,S_1 有 n 个词语:w_{11},w_{12},…,w_{1n}。S_2 有 m 个词语:w_{21},w_{22},…,w_{2m}。句子相似度计算方法是分别以两个句子的词集为基础,依次从一个集合中选取一个词分别跟另一个集合中的各个词计算相似度,挑选取得最大相似度的词对,循环直到第一个集合词为空,然后把挑选出来的这些词对的相似度相加,除以第一个集合包含的词数量,最后将分别以两个集合为基础计算的结果平均得到两个句子的相似度,其计算公式如下:

$$Sim(s_1,s_2) = \Big[\sum_{u=1}^{n} \max_{1\leqslant v\leqslant m} Sim(w_{1u},w_{2v})/n + \sum_{v=1}^{m} \max_{1\leqslant u\leqslant n} Sim(w_{1u},w_{2v})/m\Big]/2 \tag{4}$$

3　实验

在为某医院开发的 LYQA 系统中对本文提出的方法进行了验证,共测试 383 个有关于不孕不育方面的问句。考虑到问句检索候选集误差和相似度计算误差,系统在给出最优答案的同时,还给出了前 M 条次优记录,供用户选择,如果最优答案并不真正相关,用户还可以进一步从这些候选答案中查找。

对于回答的正确判断也采用了两种方法:1)判断系统输出的最优解是否相关解? 如是,认为回答正确,否则为回答错误;2)判断系统输出的前 M 条最优解中有无相关解? 如有,认为回答正确,否则为回答错误。本系统中取前 5 条答案。我们采用精确率作为系统的评价标准。其数学公式表示如下:

$$precision = num_{corr}/num_{all} \tag{5}$$

测试结果如表 1。

表 1　FAQ 测试结果

方法	正确回答问句	精确度/%
方法一	306	79.9
方法二	360	94.0

实验结果表明,本文提出的方法取得了较好的效果,最优解的正确率接近 80%,当以前 5 条最优解作为回答正确与否的判断标准时,准确率达到了 94.0%,已经接近或者说可以胜任实际应用的需求。

4　结语

本文提出的 FAQ 系统首先根据用户的提问利用检索系统生成一个候选问题集,然后利用知网语义计算句子相似度,在候选问题集中找到最相似的问句或最相似的前几条问句,进而把答案返给用户。在开发的 LYQA 系统中进行了验证,实验结果表明该方法的精确率已经达到比较理想的效果,已经接近实用水平。

参考文献:

[1] NIST. Text retrieval conference (TREC) homepage[EB/OL]. [2004-06-23]. http://trec.nist.gov/.

[2] HAMMOND K, BURKE R, MARTIN C, *et al*. FAQ finder: a case-based approach to knowledge navigation[C]// Artificial Intelligence for Applications. Los Angeles: [s.n.], 1995: 80-86.

[3] Van LE H, TRENTINI A. FAQ share: a frequently asked questions voting system as a collaboration and evaluation tool in teaching activities[C]// Proceedings of the 14th international conference on software engineering and knowledge engineering. New York: ACM Press, 2002: 557-560.

[4] 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): 1179-1182.

[5] 王继成,潘金贵,张福炎. Web 文本挖掘技术研究[J]. 计算机研究与发展,2000,37(5):514-516.

[6] 董振东, 董强. 知网[EB/OL]. [2003-09-10]. http://www.keenage.com/.

[7] 夏天. 中文信息处理中的相似度计算研究与应用[D]. 北京:北京理工大学,2005:33-35.