基于问句相似度的中文FAQ问答系统研究

叶正，林鸿飞，杨志豪

（大连理工大学计算机科学与工程系，大连 116024）

摘 要： 常见问题（FAQ）问答系统是一种在已有的“问题—答案”对集合中找到与用户提问相匹配的问句,并将其对应的答案返回给用户的问答式检索系统。其关键问题是用户提出问句与 FAQ 库中问句进行相似度计算。本文通过对常见问句特点的研究，给出一种基于分解的向量空间模型和语义概念的问句相似度计算方法。实验表明，与传统的基于向量空间模型的 TF-DF 问句相似度计算方法相比，可以提高问句匹配的精度。

关键词： 问句相似度；语义相似度；常见问题集；向量空间模型

FAQ QA system based on sentence similarity

YeZheng,Lin Hongfei,Yang Zhihao

(Department of Computer Science and Engineering, Dalian University of Technology , Dalian 116024)

Abstract: The key question of FAQ QA system is to calculate the similarity between the user questions and FAQ pairs. This paper presented a question similarity computation approach based on splited vector space model and semantic concept. The experiment indicates that precision of question match can be improved compared to traditional question similarity computation based on TF-IDF computation of vector space model.

Keywords: Sentence similarity; semantic similarity; FAQ; vector space model

1 引言

FAQ问答系统是一种在已有的“问题—答案”对集合中找到与用户提问相匹配的问句，并将其对应的答案返回给用户的问答式检索系统[1]。目前国内外关于FAQ问答系统中问句相似度的计算方法研究，主要有基于统计的方法和基于语义概念的方法。国外Robin D.Burke等人利用WordNet计算问句的语义相似度，并与基于TF-IDF的句子相似度的方法相融合[2]；国内秦兵等人利用知网（HowNet），采用计算句子的语义相似度的方法来找出匹配的问句[3]。

本文针对问句本身的特点，给出一种基于分解的向量空间模型和语义概念的问句相似度计算方法，并详述了其实现过程。

基金资助：国家自然科学基金（60373095）

作者简介：叶正（1981），男，湖北，硕士研究生 E-mail：jeafyezheng@163.com. 林鸿飞（1962），男，辽宁，教授，博士，E-mail：hflin@dlut.edu.cn。杨志豪，男，博士，E-mail：yangzh@dlut.edu.cn

2 系统的设计与实现

该系统主要包括4个部分：问句预处理模块、问句的索引和检索模块、问句相似度计算模块和FAQ库的更新模块。为了便于描述，首先给出本文用到的术语和系统的结构图，如图1所示。

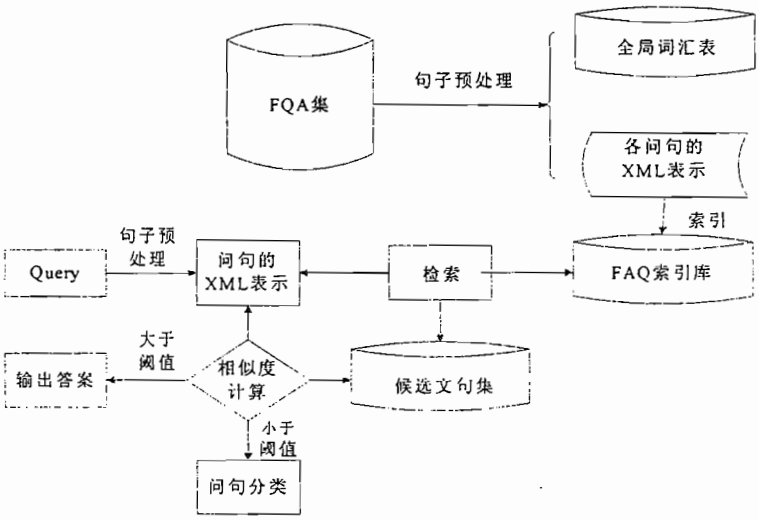

图1：系统结构图

Fig.1 The system structure chart

2.1 问句的预处理

问句的预处理主要包括：分词、词性标注、去停用词和问句标注。其中问句标注主要完成主题词、问点、疑问词以及问句对应的答案进行标注，并以通用的 XML 格式存储，可以把一个问句向量分解成三个分向量。问句标注算法如下[4]：

①提取<topicword 主题词>标记内容

主题词是指该问句中疑问的对象。由于系统主要是应用于校内，所以问句主要是针对校内对象提问的，因此主题词是校内的一些概念主体。

②提取<focus 问点>标记内容

问点是指问句的疑问焦点，主要包括询问各种属性。如定义 define、功用 function、特点 speciality、分类 subclass、外形 form 等，可根据一些特殊词词组合标识。

③提取<questionword 疑问词>标记内容

事先从大量问句中找到高频的疑问词集合，以之为参考比较每个问句中的疑问词。

2.2 问句的索引和检索

FAQ 集中问句经过预处理后，下一步就是对标注后的 FAQ 集建立倒排索引，一方面可以提高系统的检索效率，另一方面可以检索到一个小的候选问句集。

2.3 问句相似度算法

本文利用《HIT－IRLab 同义词词林（扩展版）》来计算词与词间的相似度，得到词与词间相似度后，就可以计算出每两个分向量间的相似度，最后线性加权就能得出问句间的相似度。本论文中只找出相似度超过某一给定阈值的前 5 个问句。

2.3.1 词与词的语义相似度的计算

在《HIT－IRLab同义词词林（扩展版）》中，将词义分为大、中、小类描述了一个由上到下，并将所收录的

词按词义分门别类组织在其中，其编码的方法说明见表1。

表1 词语编码表

Tab.1 the coding table of words

编码位	1	2	3	4	5	6	7	8
符号举例	D	a	1	5	B	0	2	=\#\@
符号性质	大类	中类	小类		词群	原子词群		
级别	第1级	第2级	第3级		第4级	第5级		

本文中词与词间相似度计算方法如下：

c) 如果两个词的前七个编码位相同且第八位都为“＝”或“@”，则相似度为 1

d) 如果两个词的前七个编码位相同且第八位都为“＃”，则相似度为 1/2

e) 如果两个词的前 i－1 个编码位都相同且第 i 个编码位不同，则相似度为 $1/(10-i)$

2.3.2 问句相似度计算方法

设 A 和 B 为两问句向量，分解后的分向量分别为 A_1， A_2， A_3 和 B_1， B_2， B_3，其中 $A_i=\{a_{i1},a_{i2},\mathrm{L}\ a_{im}\}$，$B_i=\{b_{i1},b_{i2},\mathrm{L}\ ,b_{in}\}$，$a_{il}$，$b_{ik}$ 分别为 A_i，B_i 中关键词，i=1,2,3，m，n 分别为 A_i，B_i 中关键词的个数。记 $s(a_{il},b_{ik})$ 为 A_i 中的 l 个关键词和 B_i 中第 k 个关键词的相似度。

首先计算 A_i 和 B_i 任意两个关键词的相似度就可以得到关键词相似度矩阵：

$$W_{A_iB_j}=\begin{bmatrix} s(a_{i1},b_{i1}),\mathrm{L}\ ,s(a_{i1},b_{in}) \\ \mathrm{M} \\ s(a_{im},b_{i1}),\mathrm{L}\ ,s(a_{im},b_{in}) \end{bmatrix}$$

则分向量 A_i 和 B_i 间的语义相似度为：

$$S(A_i,B_i)=\frac{\sum_{l=1}^{m}\max(s(a_{il},b_{i1}),\mathrm{L}\ ,s(a_{il},b_{in}))}{m}$$

得到每个分向量间的相似度后，则问句 A 和 B 间的相似度为：

$$Sim(A,B)=\sum_{i=1}^{3}\delta_i S(A_i,B_i)$$

其中，$\delta_i(1<i<3)$ 表示各分向量的重要程度，且 $\delta_1+\delta_2+\delta_3=1$

2.3.3 FAQ 库的更新模块

如果系统判定用户提出的问句在 FAQ 库中没有匹配的问句时，则系统会对用户问句进行分类，发给相应管理人员回答，并将用户的问题和对应的答案加入到 FAQ 库中。

3 实验结果与分析

为了衡量系统的检索质量，本文使用的是文献[2]中给出的评价方法，文献[2]中指出，相对传统的招回率和精确率，其评价方法能更好的反应 FAQ 问答系统的性能。文献[2]修改了计算招回率的方法，其计算公式为：

$$recall=\frac{c}{n}$$

其中 $recall$ 表示系统的招回率，n 为用户问句个数，c 为系统返回正确答案的问句个数。

对于系统正确率，文献[2]没有使用传统的准确率的概念，而使用的是不匹配率，其计算公式为：

$$rejection = \frac{c}{qn}$$

其中 $rejection$ 表示系统的不匹配率，qn 为用户问句的个数（所用用户问句的正确答案都不在 FAQ 集中），c 为系统判定 FAQ 集中没有正确答案的个数。

实验过程中，收集了学校各部门留言版上的 300 个真实问句，其中 200 个是是包含答案的，剩下的 100 个没有作答，另外还人工构造了 100 个语义相近的问句，分别用来测试招回率和不匹配率。根据以上评价方法，本文选取两组不同问句相似度阈值 θ，实验结果如表 2 所示。

表 2 实验结果

Tab.2 The result of experimentation

问句相似度计算方法	recall		rejection	
	$\theta = 0.6$	$\theta = 0.8$	$\theta = 0.6$	$\theta = 0.8$
基于 TF-IDF 方法	75%	68%	71%	78%
基于分解向量空间模型和语义的方法	81%	73%	78%	85%

从表 2 中，可以看到基于分解的向量空间模型和语义的方法能取得较好的效果。提高问句相似度阈值后，虽然不匹配率能得到提高，但是招回率会迅速降低。

4 结束语

本文详细讲述了 FAQ 问答系统的一种设计方法，并给出了一种基于分解的向量空间模型和语义概念的问句相似度计算方法。目前我们的研究主要集中在如何提高系统的招回率，然而，判断出某个问题的答案不在 FAQ 集中也很重要，因为这样可以便于下一步的处理。今后研究重点将放在如何提高问句的理解深度上，以提高系统的不匹配率。

参考文献：

[1] 吴友政,赵军,段湘煜,徐波 问答式检索技术及评测研究综述 中文信息学报， 2004 年，19（3）：1-13

[2] BURKE R D, HAMMOND K J ,KULYUKIN V ,etal. Question answering from frequently asked question files:experiences with the FAQ finder system p[J].AI Magazine ,1997, 18:57-66

[3] 秦兵，刘挺等。 基于常见问题集得中文问答系统[J] 哈尔滨工业大学学报，2003，35：1179－1182

[4] 余正涛，樊孝忠等 基于自然语言理解的受限领域自动应答系统 计算机工程，2004，30（18）：35-37

[5] Vlentin Jijkoun,Maarten de Pijke .Retrieving Answers from Frequently Asked Questions Pages on the Web Proceedings of the 14th ACM international conference on Information and knowledge management ,2005, 76-83

[6] 张宇;刘挺;高立琦等;基于常问问题集的在线客服实验研究 全国第八届计算语言学联合学术会议（JSCL-2005）论文集 , 2005 年

[7] 车万翔; 刘挺; 秦兵; 等，基于改进编辑距离的中文相似句子检索 高技术通讯，2004，（7）：15-19