基于问句类型的问句相似度计算*

田卫东，强继朋
（合肥工业大学 计算机与信息学院，合肥 230009）

摘　要：目前，问句相似度的计算主要借鉴普通陈述句的相似度计算方法。由于普通陈述句的相似性更多反映的是语句间语义上的匹配符合程度，而衡量问句间的相似性则须同时考虑问句及其答案句之间的相似程度，为此，设计了一种新的问句相似度计算方法。该方法不仅利用问句之间的语义和语法特征考察问句之间的匹配程度，还利用问句的问题类型等信息来间接刻画答案句之间的特征形象，从而以获取问句的深层语义信息，以提高问句相似度计算的准确性。实验验证了该方法的有效性。

关键词：问句相似度；FAQ 问答系统；问句类型；问题分类
中图分类号：TP311；TP391　　**文献标志码**：A　　**文章编号**：1001-3695(2014)04-1090-04
doi：10.3969/j.issn.1001-3695.2014.04.032

Questions similarity computation based on question classification

TIAN Wei-dong，QIANG Ji-peng
（*School of Computer & Information*，*Hefei University of Technology*，*Hefei* 230009，*China*）

Abstract：Now，questions similarity computation methods are mainly based on general declarative sentence similarity computation methods，where only semantic similarity between two general declarative sentences is considered. While considering question similarity，both similarity between two questions and similarity between two answers should be considered. So this paper proposed a new question similarity computation method，which took into account not only semantic and syntactic characteristics between questions，but also information such as question type to indirectly get the characteristics of the answer，in order to extract deep semantic information of question，and finally to improve question similarity computation accuracy. Experimental results show that the method can achieve better accuracy.

Key words：questions similarity；FAQ question answering system；types of questions；classfication of questions

0　引言

FAQ 是问答系统的重要组成部分。问答系统通过将常见的问题及其答案存储起来形成常见问题集，来提高类似问题的答案搜索与合成效率。FAQ 在使用上存在问题集的更新和匹配新问题两个主要的问题，而解决这两个问题的关键，则在于问题（或称问句）相似度的准确计算。

现有文献中，问句相似度的计算主要借鉴普通陈述句的相似度计算方法，如语义[1]、语法、句法结构[2]以及多种方法结合[3~5]的方法。但是问句间相似性计算与普通陈述句间相似性计算既有相似地方，也有所区别。这是因为衡量普通陈述句间的相似性仅须考虑两个陈述句在语义上的相似程度，而衡量问句间的相似性则须同时考虑问句及其答案句之间的相似程度。实际上，问句间相似度的细微差别常会导致答案的大相径庭，例如，句法结构完全不同的两个句子也许问的是同一个问题，而语法句法结构完全相同的句子也许问的是完全不同的问题，亦即问句间相似度计算需要考察句子的深层语义信息。对此，现有普通陈述句的相似度计算方法是难以做到的。

结合答案句的特征信息来协助获取问句的深层语义信息，困难在于，在 FAQ 的问句匹配阶段无法获知答案句。为此，本文引入问题分类的相关研究成果[1~5]，利用问题分类的结果，如问题类别、中心词等信息，来间接刻画答案句的特征信息，从而丰富问句相似性计算时的考虑因素。在此基础上，与传统的问句相似度计算方法所考量因素如语法和语义等相结合，设计了一种新的问句相似度计算方法。该方法虽然考虑了问句信息但并未增加多少额外的计算开销，这是因为在问答系统中，问题分类工作本身是需要首先完成的。

本文给出了问句相似度计算的特点，问句相似度的计算方法、步骤和时间复杂度分析，并从不同方面来验证算法的有效性。

1　问句相似度

普通陈述句之间的相似度更多地反映句子间语义上的匹配符合程度，当相似度达到某个设定的阈值时，就认为两个句子相似。而 FAQ 问答系统中的问句相似度计算，不仅要求两个句子的相似度很高，也要求问句之间对应答案的一致性，所以，普通陈述句相似度计算方法没有考虑到问句与答案的对应，如果用于问句相似度计算，必定会影响 FAQ 问答系统的性能。

为了更明显地显示问句相似度计算和普通陈述句相似度计算的不同，给出了几个简单问句的对比，采用了语义相似度计算方法[6]和多种特征结合的相似度计算方法[7]分别计算了问句的相似度值，结果如表 1 所示。

收稿日期：2013-04-24；**修回日期**：2013-06-15　　**基金项目**：国家自然科学基金资助项目(60603068)

作者简介：田卫东(1970-)，男，安徽合肥人，副教授，主要研究方向为人工智能、数据挖掘(twd1024@163.com)；强继朋(1988-)，男，安徽亳州人，博士研究生，主要研究方向为数据挖掘.

表 1　问句相似度计算的例子

序号	问句	语义	多特征
1	清朝一共有哪几位皇帝? 清朝总共有几位皇帝?	0.88	0.86
2	哪个国家的货币最值钱? 哪个国家的货币最不值钱?	0.93	0.94
3	上网对人有什么好处? 上网对人有什么坏处?	0.97	0.92
4	白宫在哪? 白宫的地址是多少?	0.51	0.47

从结果可以看出,前三对问句都有很好的相似度,但都不是相同的问题,都会影响 FAQ 问答系统的性能;第四对问题,虽然有很低的相似度,但却是相同的问题。原有的算法不能很好地计算这些问句,主要是下面几方面的原因:

a)没有考虑两个问句的类型是否一致。第一对问句就是两个不同的问题,一个是问人,另一个是问数字。第四对问句是相同类型的问题,都是关于地址方面的问题,但是没有考虑类型问题,导致相似度很低。

b)没有考虑问句的句型。第二对问句,下面一句比上面一句多一个否定词,导致了结果的截然不同。

c)没有考虑词语之间的褒贬倾向。第三对问句中“好处”和“坏处”分别是褒义词和贬义词,用基于知网的方法计算相似度有很高的相似值[8]。

为此,本文着重考虑问句的类型。为了解决此问题,在计算问句相似度的过程中,引入了问句的类型作为问句相似度计算的一个因素。问题分类的目的就是确定问题的类型,主要是为了增加约束条件,方便信息检索和答案的提取。目前,并没有统一的问题分类体系。为此,很多问答系统都采用了较为复杂的问题分类体系。

由于本文用到的是中文的问题分类,因此将介绍中文问题分类体系。目前一些中文问题研究的问题分类体系都是从 TREC 会议上的英文分类体系直接翻译过来,采用的都是从粗类到细类的层次分类体系。虽然在某些类型上可能造成歧义,总体来说可也得到了较高的分类精度,还可以增加约束条件便于分类。哈尔滨工业大学在 UIUC 分类体系[9]的基础之上,加上中文问题本身的特点,也采用了层次分类的方法制定了中文的分类体系。本文就采用了哈工大的问题分类体系。其中包括 7 大类和 65 个小类。

2　问句相似度计算

2.1　问句相似度计算方法

1)语法方法

语法相似度主要基于词形、词序、句长和距离的综合方法。该方法从词形相似度、词序相似度、句长相似度和距离相似度四个方面进行考虑。

词形相似度反映两个问句中词语在形态上的相似程度,用两个问句中含有的共同词的个数来衡量。用 $\mathrm{wordSim}(A,B)$ 表示问句 A 和 B 的词形相似度,计算式如下:

$$\mathrm{wordSim}(A,B)=2\times\frac{\mathrm{same}(A,B)}{\mathrm{len}(A)+\mathrm{len}(B)} \tag{1}$$

其中:$\mathrm{same}(A,B)$ 表示 A 和 B 中共同词的个数,当一个单词在 A、B 中出现的次数不同时,以出现次数少的计数;$\mathrm{len}(A)$ 和 $\mathrm{len}(B)$ 分别表示 A 和 B 中词的个数。

词序相似度反映两个句子中词语在位置关系上的相似程度。计算式如下:

$$\mathrm{orderSim}(A,B)=\begin{cases}1-\dfrac{\mathrm{ReWord}(A,B)}{|\mathrm{orderOcc}(A,B)|-1} & |\mathrm{orderOcc}(A,B)|>1\\ 1 & |\mathrm{orderOcc}(A,B)|=1\\ 0 & |\mathrm{orderOcc}(A,B)|<1\end{cases} \tag{2}$$

其中:$\mathrm{orderOcc}(A,B)$ 表示在 A、B 中都出现且只出现一次的词;$\mathrm{PFirst}(A,B)$ 表示 $\mathrm{orderOcc}(A,B)$ 中的词在 A 中的位置序号构成的向量;$\mathrm{PSecond}(A,B)$ 表示 $\mathrm{PFirst}(A,B)$ 中的分量按对应词在 B 中的词序排序生成的向量;$\mathrm{reWord}(A,B)$ 表示 $\mathrm{PSecond}(A,B)$ 各相邻分量的逆序数。

句长相似度反映两个问句在长度形态上的相似程度。用 $\mathrm{LenSim}(A,B)$ 表示问句 A 和 B 的句长相似度,计算式如下:

$$\mathrm{lenSim}(A,B)=1-\mathrm{abs}\left|\frac{\mathrm{len}(A)-\mathrm{len}(B)}{\mathrm{len}(A)+\mathrm{len}(B)}\right| \tag{3}$$

其中,abs 表示绝对值。

距离相似度用相同关键词在问句上的距离来衡量句子的相似度。用 $\mathrm{disSim}(A,B)$ 表示问句 A 和 B 的距离相似度,计算式如下:

$$\mathrm{disSim}(A,B)=1-\mathrm{abs}\left|\frac{\mathrm{sameDis}(A)-\mathrm{sameDis}(B)}{\mathrm{dis}(A)+\mathrm{dis}(B)}\right| \tag{4}$$

其中:$\mathrm{sameDis}(A)$ 表示 A 和 B 中相同的词在 A 中的距离,若相同的词重复出现多次,以产生最大距离为准;$\mathrm{dis}(A)$ 表示 A 中非重复词中最左及最右词之间的距离,若词出现多次,以产生最小距离为准。

由以上四部分可以加权得到问句的语法相似度计算式如下:

$$\begin{aligned}\mathrm{syntaxSim}(A,B)=&\alpha\times\mathrm{wordSim}(A,B)+\beta\times\mathrm{orderSim}(A,B)+\\&\gamma\times\mathrm{lenSim}(A,B)+\mu\times\mathrm{disSim}(A,B)\end{aligned} \tag{5}$$

其中,α、β、γ、μ 是常数,且满足 $\alpha+\beta+\gamma+\mu=1$,显示 $\mathrm{sim}(A,B)\in[0,1]$。

在语法相似度中,一般不难理解词形相似度起着决定性的作用,其他三个方面起着次要的作用,因此 α、β、γ、μ 取值时应该有 $\alpha>>\beta,\gamma,\mu$。本文选择的值为 $\alpha=0.7$,$\beta=\gamma=\mu=0.1$。

2)语义方法

问句语义相似度计算还是以词的语义计算为基础。词的计算采用了刘群等人[8]介绍的基于知网的词语的相似度计算。

设两个问句 A 和 B,A 包含的词为 $w_{11},w_{12},\cdots,w_{1n}$,$B$ 包含的词为 $w_{21},w_{22},\cdots,w_{2m}$,则词语 $w_{1i}(1\leqslant i\leqslant n)$ 和 $w_{2j}(1\leqslant j\leqslant m)$ 之间的相似度表示为 $\mathrm{sim}(w_{1i},w_{2j})$。问句 A 和 B 之间的语义相似度可以根据下面公式计算:

$$\begin{aligned}\mathrm{semanticSim}(A,B)=&\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\max\{\mathrm{sim}(w_{1i},w_{2j})\mid 1\leqslant j\leqslant m\}+\right.\\&\left.\frac{1}{m}\sum_{i=1}^{m}\max\{\mathrm{sim}(w_{1i},w_{2j})\mid 1\leqslant i\leqslant n\}\right)\end{aligned} \tag{6}$$

3)问句类型的方法

用 $\mathrm{classSim}(A,B)$ 表示两个问句类别相似度。由于采用的是哈工大的中文问题分类体系,计算过程中也根据了问题的粗类和细类来计算问句的相似度值。目前,问句分类体系采用哈工大的问题分类体系,计算式如式(7)所示。

$$\mathrm{classSim}(A,B)=\begin{cases}1 & A\text{ 和 }B\text{ 同属于一个细类}\\ 0.5 & A\text{ 和 }B\text{ 同属于一个粗类}\\ 0 & \text{其他情况}\end{cases} \tag{7}$$

2.2　问句相似度计算的步骤

给予两个问句 A 和 B,问句相似度的计算式如式(8)所示。在原有的语法和语义方法的基础之上,结合了问句类型的方法。两个问句的相似度计算的伪代码如算法 1 所示,hownet

(w_{1i}, w_{2j})是调用基于知网的词语相似度计算方法[8]。

$$\mathrm{sim}(A,B) = \alpha \times \mathrm{semanticSim}(A,B) + \beta \times \mathrm{syntaxSim}(A,B) + \gamma \times \mathrm{classSim}(A,B) \quad (8)$$

其中,满足 $\alpha+\beta+\gamma=1$,且满足 $\alpha>\beta,\gamma$。

算法 1 问句相似度计算的伪代码

输入:两个问句 A 和 B。
输出:$\mathrm{sim}(A,B)$。

```
1  对 A 和 B 进行分词处理,去除一些停用词,得到 A 包含的词为 w11,w12,…,w1n, B 包含的词为 w21,w22,…,w2m
2  计算 wordSim, orderSim, lenSim 和 disSim,得到 syntaxSim
3  sis←0;
4  semantic←0;
5  for i = 1; i≤n;i ++
       for j = 1; j≤m;j ++
           if sis < hownet(w1i,w2j) sis = hownet(w1i,w2j)
       semantic + = sis;
       sis = 0;
   semanticSim = semantic/n;
   semantic = 0;
   for i = 1; i≤m;i ++
       for j = 1; j≤n;j ++
           if sis < hownet(w2i,w1j) sis = hownet(w2i,w1j)
       semantic + = sis;
   semanticSim + = semantic/m;
   semanticSim / = 2
6  采用问题分类算法对 A 和 B 进行问题分类,得到 classSim
7  根据式(8),可以得到 sim(A,B)
```

假设 Q 为用户输入的问句,corpus 为常见问题库 FAQ。要找出与 Q 最相似的且大于阈值 w 的句子 S。该过程可以用下面公式进行描述:

$$S = \{\operatorname{argmax}_{S \in \mathrm{corpus}} \mathrm{sim}(Q,S) > w\} \quad (9)$$

其中,S 表示 corpus 中任意的句子。

如果 FAQ 问句的数目比较庞大,首先要进行的是问句特征词的扩展和候选问句的选取,这里不作为重点详述。主要介绍从候选问句选取最相似度的问句,步骤如算法 2 所示。

算法 2 FAQ 问答系统的执行伪代码

输入:用户问句为 Q,FAQ 库中的问句 corpus 和相似度阈值 w。
输出:最相似的问句。

```
1  对用户问句进行分词处理,选取关键词
2  对关键词进行扩展
3  FAQ 库候选问句集的查找,得到候选问题集{h1,h2,…,hn}
4  分别计算得到 Q 与 hi 的相似度值 simi(Q,hi)
5  按照相似度值 simi(Q,hi)从高到低进行排序,如果存在大于阈值 w 的问句,返回相似度值最高的问句对应答案给用户,否则返回 null
```

2.3 时间复杂度分析

时间复杂度分别包括语法相似度、语义相似度和问句类型相似度的时间复杂度。

词形相似度的复杂度为 $O(n\times m)$, 其中 n 和 m 分别为两个问句中词的数目。词序相似度的复杂度为 $O(c^2)$,其中 c 为问句中共同词的数目;句长相似度的复杂度为 $O(1)$;距离相似度的复杂度为 $O(c)$。

调用基于语义方法的时间复杂度为 $O(n\times m\times h)$,其中 h 为调用基于知网计算的两个词语相似度的复杂度。

调用基于问题类型的方法,需要的时间复杂度为 $O(2\times q)$,其中 q 为调用的问题分类算法的时间复杂度。

最后,两个问句的时间复杂度为 $O(n\times m + c^2 + n\times m\times h + q)$。

3 实验结果及分析

本文的实验中,使用的第三方平台有知网和 ICTCLAS 3.0 分词系统。

文本信息处理结果的评价标准较多,最常见的是准确率(precision)和查全率。由于 FAQ 自动问答系统的特殊性,在结果分析中可以忽略查全率。本文中方法准确率的估算用下面两个公式:

$$\mathrm{prec}_1 = \frac{1}{n}\sum_{i=1}^{n} a_i$$

$$\mathrm{prec}_2 = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{r_i} \quad (10)$$

其中:n 为测试的问题总数;a_i 为第 i 个问题的值,如果正确为 1,否则为 0;prec_2 是借鉴其他 FAQ 问答系统中的评价指标[12]。对每一个问题,给出大于阈值的答案,且不能超过三个。r_i 为正确答案的位置,若存在答案,都不正确,则 $1/r_i$ 取值为 0;若 FAQ 库中没有答案,系统没给予答案,$1/r_i$ 取值为 1。

实验用的数据集由两个部分组成,即整理的面试常见问题集和哈工大整理的问题集。面试常见问题集是从网上下载的 200 句作为 FAQ 库,并手工标注问题类别;然后找了 10 个学生,每个学生列举出 10 个面试中被问到的问题,总共 100 句作为测试集。哈工大整理的问题集中的训练集 4 966 句作为 FAQ 库,每个问句都已标过类别;然后从这 4 966 句中随机选择 2 000 句,对这些问句进行了一系列的改变。例如,问句中词语进行同义词的替换、改变问法,以及该主题的其他类似的问法等。同时增加哈工大整理的测试集合 1 300 句同时作为测试集。

实验过程中参数的取值,$\alpha=0.6$,$\beta=\gamma=0.2$,阈值 $w=0.65$。对比的算法:语义的方法采用金博等人[6]的算法,语法结合语义采用张琳等人[7]的算法,参数都是参照原论文里的数字。

以问句 Q:“我们去北京的原因是什么?”为例,比较 Q 与 FAQ 库中的三个问题 R_1、R_2 和 R_3,结果如表 2 所示。

表 2 用户问句与 FAQ 库中问句的句子相似度计算结果

FAQ 库中问句	语义	多特征	本文
R_1:为什么我们去北京?	0.728	0.683	0.778
R_2:什么时候我们去北京?	0.836	0.760	0.666
R_3:北京有什么好玩的地方?	0.724	0.651	0.63

从相似度的计算结果可以看出,不论是采用语义的方法,还是采用多特征的方法,都选用了 R_2 作为最后的结果。从人的主观性中可能不难发现,问句 R_1 可能是想要的结果。采用本文的基于问题类型的方法,选择了 R_1 作为最后的结果,主要是由于只有 R_1 与用户问句的问句类项是一样的。

选用了面试常见问题集,得到的实验结果如表 3 所示。

在 prec_1 和 prec_2 上,本文算法的准确率要比其他算法高,说明考虑了问题的类型对准确率的提高有一定的帮助。

表 3 面试常见问题集的实验结果

实验方法	prec_1	prec_2
语义	0.55	0.60
多特征	0.68	0.72
本文	0.79	0.81

选用了哈工大整理的数据集。考虑到 FAQ 库和测试集中的问句都已标注了问句类别,同时为了证明问题分类准确性对问句相似度计算的影响。第一种方法选择目前已有的算法进行问题分类[9],然后用于相似度计算,简称方法 1;第二种方法

直接用标注的类别进行问句相似度计算,简称方法 2。实验结果如表 4 所示。

从结果可以看出,方法 1 和 2 都比已有算法的准确度要高,说明了引入问题类型对问句相似度的计算是有用的。同时由于方法 2 比方法 1 的准确度要高,说明了问题分类的准确性对问句的相似度是有影响的。问题分类的准确度越高,问句相似度的准确度也越高,因此选择分类准确度高的问题分类算法十分必要。

表 4　哈工大数据集的实验结果

实验方法	$prec_1$	$prec_2$
语义	0.52	0.54
多特征	0.65	0.68
方法 1	0.74	0.75
方法 2	0.79	0.80

实验过程中,哈工大测试集中一定比例的问题类型不变,其他问题的类型随机设置为其他问题类型,从而用来验证问题分类算法的准确度对问句相似度计算准确度的影响。实验结果如图 1 所示。从图 1 可知,问句分类的准确度在 0.2 ~ 0.4 的过程中,随着问句分类准确度的提高,问句相似度计算准确度增加得很慢,主要是因为这一段问句分类的准确度都比较低,问句类型对问句相似度计算过程的影响比较小,又由于综合了多特征的方法,而多特征方法占据着主要的权重。随后,问句相似度计算准确度增加得较快,这是因为问句分类正确可以缩小 FAQ 库中其他类型问句的干扰。总之,问句相似度计算的准确度随着问句类型准确度的提高而提高。

研究最终选择答案的数目对不同方法的影响。数据选择改变后的哈工大的数据集,答案数目取值为 1 ~ 6,实验结果如图 2 所示。

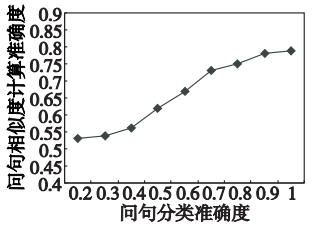

图1　问句分类准确度对问句相似度计算准确度的影响

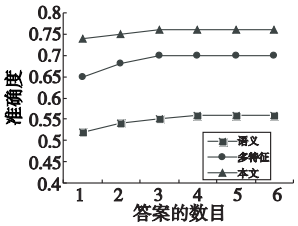

图2　答案的数目对问句相似度计算准确度的影响

从图 2 可以看出,随着答案数目的增加,不同方法问句相似度计算的准确度增加,然后不变。但是,语义的方法和多特征的方法不变后的准确度值与初始值的差值要大于本文方法差值,主要是由于本文的方法考虑到问题类型,缩小了候选问题答案的规模。总之,不论是哪种答案的数目,还是本文算法的准确度最高。

4　结束语

本文通过分析发现,普通陈述句的相似度计算方法不适合用于 FAQ 问答系统中的问句相似度计算,从而引入问句特征的方法,提出了一种新的问句相似度计算方法,实验结果也证明了该方法的有效性。由于该方法受问题分类准确度的影响,考虑结合特定的领域,设计本领域的问题分类算法作为下一步研究的内容。

参考文献:

[1] LIU Qing-lei, GU Xiao-feng, LI Jian-ping. Researches of Chinese sentence similarity based on HowNet[C]//Proc of International Conference on Apperceiving Computing and Intelligence Analysis. 2010: 26-29.

[2] WANG Rong-bo, WANG Xiao-hua, CHI Zhe-ru, *et al*. Chinese sentence similarity measure based on words and structure information [C]//Proc of International Conference on Advanced Language Processing and Web Information Technology. 2008:27-31.

[3] NAN Xuan-guo. The research of sentence similarity computation based on multi-level fusion[C]//Proc of the 7th International Conference on System of Systems Engineering. 2012:617-619.

[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] LI Yu-hua, McLEAN D, BANDAR Z A, *et al*. Sentence similarity based on semantic nets and corpus statistics[J]. IEEE Trans on Knowledge and Data Engineering,2006,18(8):1138-1150.

[6] 金博,史彦军,滕弘飞. 基于语义理解的文本相似度算法[J]. 大连理工大学学报,2005,45(2):291-297.

[7] 张琳,胡杰. FAQ 问答系统句子相似度计算[J]. 郑州大学学报:理科版,2010,42(1):57-61.

[8] 孙景广,蔡东风,吕德新,等. 基于知网的中文问题自动分类[J]. 中文信息学报,2007,21(1):90-95.

[9] 田卫东,高艳影,祖永亮. 基于自学习规则和改进贝叶斯结合的问题分类[J]. 计算机应用研究,2010,27(8):2869-2871.

[8] 刘群,李素建. 基于《知网》的词汇语义相似度计算[C]//第三届汉语词汇语义学研讨会论文集. 2002:8-15.

[9] LI Xin, ROTH D, SMALL K. The role of semantic information in learning question classifiers[C]//Proc of International Joint Conference on Natural Language Processing. 2004:451-458.

[12] 张亮,冯冲,陈肇雄,等. 基于语句相似度计算的 FAQ 自动回复系统设计与实现[J]. 小型微型计算机系统,2006,27(4):720-723.