自动问答系统中的句子相似度算法的研究

周永梅,陶 红,陈姣姣,张再跃
(江苏科技大学 计算机科学与工程学院,江苏 镇江 212000)

摘 要:文中主要研究了自动问答系统的句子相似度的几种常见算法,基于统计的VSM算法、语义相似度算法、结构的相似度算法,并在此基础上进行改进,提出了一种新的句子相似度算法,提高了自动问答系统的查全率和查准率。主要研究了分词、标注词性和权值、计算词语的相似度,进而计算句子的加权相似度,最后从FAQ中抽取相似度比较高的句子以及答案给用户。最后通过实验进行验证,这种新的句子相似度算法在一定程度上提高了自动问答系统的查全率和查准率,并具有一定的合理性。

关键词:分词;本体;知网;语义相似度;查全率;查准率
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2012)05-0075-04

Study on Sentence Similarity Approach of Automatic Ask & Answer System

ZHOU Yong-mei, TAO Hong, CHEN Jiao-jiao, ZHANG Zai-yue
(School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang 212000, China)

Abstract: Mainly study on several sentence similarity approaches of the automatic ask and answer system, such as VSM algorithm based on statistics, semantic similarity algorithm and structure similarity. Based on this, take weight into account to improve the similarity and put forward the new sentence similarity method, in order to improve the recall and precision of ask & answer system. It mainly focuses on the segmentation, and markes the characteristics and weight of words, removing the stop words, computing the semantic similarity between two words based on the hownet, computing the weighted similarity between two questions, choose the more similar questions and answers from FAQ back to the users at last. Experiment results show the algorithm improves the recall and precision at some extent.

Key words: segmentation; ontology; hownet; semantic similarity; recall; precision

0 引 言

目前,自动问答系统是国内外学者的研究的热点,而句子相似度作为自动问答系统中的核心,也一直是人们研究的重点和热点,目前研究的句子相似度算法共分为两大类:一类是问句与答案的句子相似度计算[1],另一类是问句与问句之间的句子相似度[2~5],文中的研究对象主要是用户提出的问题与常用问题库问题之间的相似度。

●现有的问答系统在计算句子相似度上主要有以下几点欠缺:

(1)没有考虑词的权值对句子相似度的影响,因为自然语言问答系统中,关键词在句子中担当的角色不同,对句子相似度影响也是不同的[1~4];

(2)在计算词语语义相似度采用基于《知网》的词语语义相似度[2],或者采用基于领域本体的词语语义相似度[3],而这两种计算词语语义相似度的方法各有千秋。前者对于专业词汇之间的语义相似度,特别是组合词汇的语义相似度不能准确的计算,因为《知网》没有搜集专业的组合词汇,同样后者对于非专业词汇的相似度的计算也欠缺;

(3)在分词方面只是简单借助于词典或者借助与现有的分词工具,如中科院的ICTCLAS等,前者是能够准确划分专业词汇以及组复合词汇,但经常会出现某些词在词典中没有找到,有的学者把这种词叫做未登记词,导致不能正确的分词,而后者即借助于现有的分词工具不能准确划分专业词汇以及组合词汇。

●针对以上几个方面不足,做了以下改进:

(1)分词方面,在基于词典分词方法上,标注了词语的词性,并引进了权值,对于问句中的未登记词结合现有的分词工具进行分词,并把未登记词添加到词典

收稿日期:2011-09-07;**修回日期**:2011-12-11
基金项目:中国科学院计算技术研究所国家重点实验室开放课题(2009JS095J)
作者简介:周永梅(1985-),女,硕士,研究方向为智能信息处理;张再跃,博士,教授,硕士生导师,研究方向为智能信息处理、数理逻辑。

中,来完善文中的词典,从而提高下一次分词的准确性和效率;

(2)计算词语相似度方面,如果都是专业词语,则相似度采用基于本体的概念相似度方法进行计算,否则采取基于知网的词语相似度的计算方法进行计算;

(3)计算句子相似度方面,研究了几种典型的句子相似度算法:基于统计的VSM算法、语义相似度算法、问句特征结构相似度算法,考虑到权值对句子相似度的影响,引进了分词后的权值,最后把这几种经典的算法综合起来形成新的句子相似度算法。通过实验表明,这种新的句子相似度算法有一定的合理性,与现有的问答系统相比,自动问答系统的查全率与查准率在一定程度上得到了提高。

1 自动问答系统的关键技术

1.1 自动问答系统的模型

文中提出的自动问答系统模型分为三层结构,分别为:用户层、中间层、数据层。其中:

用户层(UI):供用户输入提问的问题,并展示系统返回的答案。

中间层(MI):中间处理层,负责:分词、处理停用词、计算词语相似度、计算句子相似度,返回答案集。

数据层(DI):处理数据库,主要有:专业词库、常用词库、同义词库、停顿词库、领域本体、《知网》本体、FAQ(问题答案集,Frequently Asked Questions)。

问答系统自动答疑的步骤如下:

自动问答系统自动答疑步骤:

步骤1:根据专业词库、常用词库、同义词库对于用户输入的自然语言问句通过正向最大匹配的方法进行分词,对于未登记词先借助于中科院的分词工具后把未登记词添加到词库中;

步骤2:对于分好词的问句根据词的类型,对于属于停用词库进行删除停用词;

步骤3:对于专业词汇采取基于本体的概念相似度方法计算词语语义相似度,对于其他词汇采取基于《知网》本体计算词语语义相似度;

步骤4:根据词语的语义相似度计算句子相似度,分别计算统计的VSM、语义相似度、结构相似度,最后综合起来计算句子相似度;

步骤5:根据计算用户提问的问句与FAQ中问句的句子相似度,并从FAQ库中抽取相似度比较高(大于或等于相似度阈值,且按照相似度大小排序,返回前五个)的问句及其答案作为用户提问问题的答案。

系统的模型运转如图1所示。

1.2 分 词

中文常用的分词方法有三种:基于字符串匹配(或称为词典、词库)的分词方法、基于统计的分词方法和基于理解的分词方法,常用的几种基于字符串匹配方法有:正向最大匹配法、逆向最大匹配法、最少切分法。

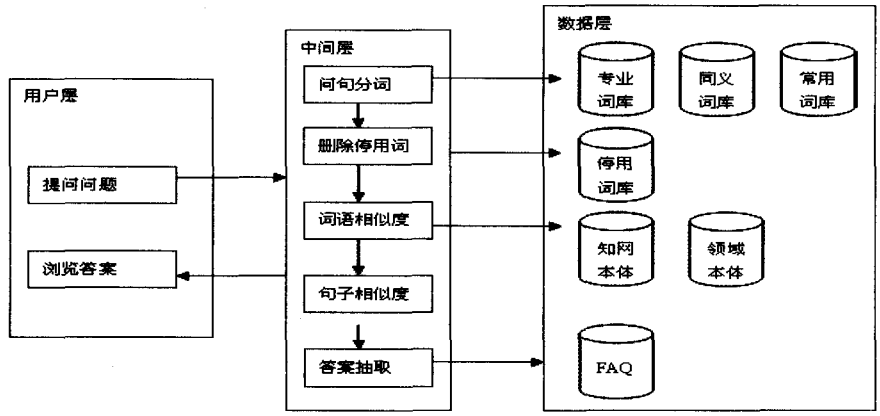

图1 自动问答系统模型

对于自动问答系统来说,问题往往是针对某一具体特定领域的,而不是针对全部自然语言范围内的,问句中的关键词主要是与该领域的概念相关的,而且很多专业关键词是复合词。文中是针对数据结构领域,如"数据结构"、"二叉树"既是专业关键词也是复合词,而"栈"、"树"、"森林"虽然是专业关键词但不是复合词。针对自动问答系统的特点,采用了基于专业词库(主要搜集了所有的专业词库)、同义词库(主要搜集了专业词汇的所有同义词)、停顿词库(主要搜集了语气词、感叹词、字母、数字、助词等等)和常用词库(搜集自然语言的词汇)的字符串正向最大匹配算法。通过分词,可以标注词语的词性,并获取权值。由于搜集的常用词库以及停用词库不全面,导致句子的分词结果不理想,针对这一缺陷,当出现未登记词时,借助于现有的分词工具进行分词,并将未登记词添加到常用词库或者停顿词库中,来不断地完善文中的词库,从而提高下一次分词的准确性和效率。因为停用词不影响句子的相似度(权值为0),所以为了提高自动问答系统的准确性和效率,在分词后,删除停用词。

分词后的权值是根据词汇的词性以及词汇的类型来确定,具体定义为:

$$w=\begin{cases}2 & \text{专业词汇}\\ 1.5 & \text{疑问词}\\ 1 & \text{名词、动词}\\ 0.8 & \text{形容词、副词}\\ 0.5 & \text{其他常用词汇}\\ 0 & \text{停用词汇}\end{cases}$$

1.3 词语语义相似度

（1）文中对于常用词语，采用基于《知网》，参考文献[6]的词语相似度的计算方法，两个汉语词语 W_1 和 W_2，如果 W_1 有 n 个义项（概念）：$S_{11},S_{12},\cdots,S_{1n}$，$W_2$ 有 m 个义项（概念）：$S_{21},S_{22},\cdots,S_{2m}$，规定 W_1 和 W_2 的相似度为各个概念的相似度之最大值，即：

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

（2）文中对于专业词汇采用基于本体的概念相似度，即参考文献[7]的词语的概念相似度，主要从语义重合度、语义距离以及本体结构[7]三方面因素考虑词语的概念相似度，其中结构因素包括概念间的宽度、深度、密度等方面[8]，概念相似度为：

$$\mathrm{Sim}(W_1,W_2)=\sqrt[a]{\mathrm{Sim}(\mathrm{Contactrio})}\times\sqrt[b]{\mathrm{Sim}(\mathrm{Distance})}\times\sqrt[c]{\mathrm{Sim}(\mathrm{Structure})}$$，其中：$a+b+c=1$ (2)

计算词语相似度的算法如下：

> 计算词语相似度的算法：
>
> 步骤1：将问句进行分词，标注词性和权值；
>
> 步骤2：将分词结果消除停用词；
>
> 步骤3：计算词语 w_1,w_2 的语义相似度
>
> (a) 判断 w_1,w_2 是否为同义词，若是则相似度为1，否则转到(b)；
>
> (b) 判断 w_1,w_2 是否都是专业词汇，若是，则采用公式(2)的方法计算词语相似度；否则采用公式(1)的方法计算词语相似度。

在构建领域本体以及在计算专业词汇语义相似度的算法中参考 protégé 的用户手册[8]和 jena 的 API 文档[9]。

1.4 句子相似度

（1）基于统计的 VSM 相似度计算。

FAQ 中的每一个问句都可以用一个 n 维的向量 $T=\langle T_1,T_2,\cdots,T_n\rangle$ 来表示，向量中特征值通过统计方法得到，则问句 T 和 T' 的相似度为[10]：

$$\mathrm{Sim}(T,T')=\cos(T,T')=\frac{\sum_{i=1}^{n}T_i\times T'_i}{\sqrt{(\sum_{i=1}^{n}T_i^2)\times(\sum_{i=1}^{n}T'^2_i)}} \tag{3}$$

公式(3)引入权值后改进变为公式(4)：

$$\mathrm{Sim}_1(T,T')=\cos(T,T')=\frac{\sum_{i=1}^{n}w_iT_i\times w_iT'_i}{\sqrt{(\sum_{i=1}^{n}(w_iT)^2)\times(\sum_{i=1}^{n}(w_iT'_i)^2)}} \tag{4}$$

（2）基于关键词的语义相似度计算。

设 A 和 B 为两问句向量，其中 $A=\{A_1,A_2,\cdots,A_m\}$，$B=\{B_1,B_2,\cdots,B_n\}$，A_m,B_n 分别为 A、B 中的关键词，m、n 表示关键词的个数，$\mathrm{Sim}(A_i,B_k)$ 为 A 中的第 i 个关键词和 B 中第 k 个关键词的相似度，这样就得到一个 $m*n$ 的矩阵，则句子 A 和 B 的相似度为[11]：

$$\mathrm{Sim}(A,B)=(\frac{\sum_{i=1}^{m}a_i}{m}+\frac{\sum_{j=1}^{n}b_j}{n})/2$$，其中 $a_i=\max(\mathrm{Sim}(A_i,B_{1..n}))$，$b_i=\max(\mathrm{Sim}(B_i,A_{1..m}))$ (5)

引入权值后基于关键词的语义相似度公式如下：

$$\mathrm{Sim}_2(A,B)=(\frac{\sum_{i=1}^{m}w_ia_i}{\sum_{i=1}^{m}w_i}+\frac{\sum_{j=1}^{n}w_jb_j}{\sum_{j=1}^{n}w_j})/2 \tag{6}$$

（3）基于问句结构的相似度计算。

问句的结构相似度包括4个方面[12]：词形相似度，即相同关键词的个数；句长相似度，即两条问句的长度；词序相似度，即关键词的顺序；距离相似度，即从相同关键词的距离来标注句子的相似度，各部分的相似度计算方法详见参考文献[4]，问句相似度的计算公式为公式(7)：

$$\mathrm{Sim}_3(A,B)=\lambda_1\mathrm{WordSim}(A,B)+\lambda_2\mathrm{LenSim}(A,B)+\lambda_3\mathrm{OrdSim}(A,B)+\lambda_4\mathrm{DisSim}(A,B)$$ 其中：$\lambda_1+\lambda_2+\lambda_3+\lambda_4=1$ (7)

文中对词形相似度、句长相似度进行改进，引进了词语的权值，分别为公式(8)和公式(9)：

$$\mathrm{WordSim}(A,B)=\frac{2\sum_{m}^{k}w_m}{\sum_{i}^{m}w_i+\sum_{j}^{n}w_j} \tag{8}$$

$$\mathrm{LenSim}(A,B)=1-\frac{|\sum_{i}^{m}w_i-\sum_{j}^{n}w_j|}{\sum_{i}^{m}w_i+\sum_{j}^{n}w_j} \tag{9}$$

（4）基于综合特征的相似度计算。

问句的统计、语义、结构特征从不同的角度反映了问句的相似度，由于问句一般词语数量不多、用户表达不一定规范等方面原因，文中认为应综合考虑问句多个方面的特征计算问句相似度，句子相似度计算公式为公式(10)：

$$\mathrm{Sim}_{\mathrm{sentence}}(A,B)=\alpha_1\mathrm{Sim}_1(A,B)+\alpha_2\mathrm{Sim}_2(A,B)+\alpha_3\mathrm{Sim}_3(A,B)$$，其中 $\alpha_1+\alpha_2+\alpha_3=1$ (10)

2 实验结果

文中的自动问答系统是针对数据结构这门学科进行自动问答，专业词库搜集了有关数据结构这门学科的专业词汇，实验中搜集了100多条问题答案库，涉及数据结构这门学科的每一个章节。实验采取用户自然语言提问问题的方式，系统根据句子的相似度由高到低返回问题以及问题的答案给用户，系统还定义了相

似度阈值(γ),即返回的句子的相似度不低于这个相似度阈值。实验以查全率和查准率来评价自动问答系统的性能,查全率和查准率分别定义为:

查准率 =回答正确问题条数 / 提问问题的条数

查全率 =回答正确问题条数 / 有答案问题条数

文中测试了50条问题,其中10条问题没有答案,测试了两组,分别在γ=0.6和0.8进行测试,并分别与参考文献[2,3]进行了对比,对比的实验结果见表1和表2,其中方法1是参考文献[2]的方法,方法2是参考文献[3]的方法,方法3是文中的方法:

表1　γ=0.6的实验结果

	正确数	错误数	查准率	查全率
方法1	35	5	70%	87.5%
方法2	36	4	72%	90%
方法3	39	2	78%	97.5%

表2　γ=0.8的实验结果

	正确数	错误数	查准率	查全率
方法1	30	10	60%	75%
方法2	33	7	66%	82.5%
方法3	37	3	74%	92.5%

3　结束语

文中主要创新点是提出了一种新的句子相似度计算方法,提高了句子的相似度,进而提高自动问答系统的查全率和查准率。但是文中的知识库采用FAQ的形式,问答时系统先计算用户问句与FAQ中问句相似度,列出一个或多个相似度最高的问句供用户选择,或直接以相似度最高的问句对应的答案作为回答。这种自动问答系统对已有的问题的回答准确率高,缺点是计算问句相似度时以关键词匹配为主,对问句缺乏语义理解,如果FAQ中没有与用户提问相匹配的问句,系统不具备自动回答能力,另外需要事先准备大量问句-答案的集合。所以接下来的工作是改进模型,对于从FAQ库抽取不到的问题,通过问句的模板对问题进行分类,并根据不同的抽取规则自动从本体抽取答案,并自动完善FAQ库。

参考文献:

[1] 崔　桓,蔡东风,苗雪雷.基于网络的中文问答系统及信息抽取算法研究[J].中文信息学报,2005,18(3):24-31.

[2] 秦　兵,刘　挺,汪　洋,等.基于常见问题集的中文问答系统[J].哈尔滨工业大学学报,2003,35(10):1179-1182.

[3] 刘汉兴,刘财兴,林旭东.基于问句相似度的本体问答系统[J].广西师范大学学报(自然科学版),2010,28(3):88-91.

[4] 周法国,杨炳儒.句子相似度计算新方法及在问答系统中的应用[J].计算机工程与应用,2008(1):165-167.

[5] 杨思春,陈家骏.中文自动问答中句子相似度计算研究[J].情报学报,2008(1):35-41.

[6] 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.台北:[出版者不详],2002:59-76.

[7] 李文杰,赵　岩.基于本体结构的概念间语义相似度算法[J].计算科学与工程,2010,36(28):4-6.

[8] Jena Semantic Web Framework[EB/OL].[2009-08-16]. http://jena.sourceforge.net/.

[9] The Proté gé Ontol ogy Editor and Knowledge Acquisition System[EB/OL].[2009-08-16]. http://protege.stanford.edu/.

[10] 宗裕朋.基于本体的中文智能答疑系统研究与实现[D].上海:上海交通大学,2007.

[11] 郭晓燕,张博锋,方爱国,等.智能答疑中问题相关度算法研究及系统实现[J].计算机应用,2005,25(2):449-452.

[12] Gruber T R. A Translation Approach to Portable Ontology Specifications[J]. Knowledge Aequisition, 1993(5):199-220.

(上接第74页)

[6] 吕彦波.基于支持向量机的入侵检测系统研究[D].西安:西安理工大学,2007.

[7] 沈翠华,邓乃扬,肖瑞彦.基于支持向量机的个人信用评估[J].计算机工程与应用,2004,40(23):198-200.

[8] 姜明辉,袁绪川.个人信用评估PSO-SVM模型的构建及应用[J].管理学报,2008,5(4):511-515.

[9] 李建平,徐伟宣,刘京礼,等.消费者信用评估中支持向量机方法研究[J].系统工程,2004,22(10):35-39.

[10] 余乐安,汪寿阳.基于核主元分析的带可变惩罚因子最小二乘模糊支持向量机模型及其在信用分类中的应用[J].系统科学与数学,2009,29(10):1311-1326.

[11] 杨海燕,周永权.一种支持向量机的混合核函数[J].计算机应用,2009,29(z2):232-235.

[12] 毛建军,蔡卫民.个人信用评分模型比较研究[C]//中国系统工程学会第十四届学术年会论文集.出版地不详:出版者不详,2006:358-363.