分布式自动答疑系统①

李园伟，宁可为，王 炜

(新疆师范大学 教育科学学院，乌鲁木齐 830054)

摘 要：作为远程教育中的重要组成部分，自动答疑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。在 HADOOP 框架下，采用改进的编辑距离算法对汉语句子的相似度进行计算能够使自动答疑系统更加快捷以及智能化。

关键词：分布式技术；语句相似度；答疑系统；远程教育

Automatic Question-Answering System Based on Distributed Technology

LI Yuan-Wei, NING Ke-Wei, WANG Wei

(College of Educational Science, Xinjiang Normal University, Urumqi 830054, China)

Abstract: As an important component of distance education, automatic question-answering system allows users to ask questions in natural language, it can return a concise and accurate answers.Based on Hadoop framework, using an improved edited distance algorithm to calculate the similarity of Chinese sentences can make the system more efficient and more intelligent .

Key words: hadoop; sentence similarity; automatic question-answering system; distance education

1 引言

随着终生学习理念的确立，远程教育扮演着越来越重要的角色，而线上答疑是远程教育过程中重要的一环。面向 FAQ 的答疑系统可以在保持 FAQ 原有功能的前提下，给用户提供一种更为方便、快捷的答疑途径。目前在国内的研究中，秦兵、刘挺等采用基于语义的方法计算句子的相似度，这需要很大规模的 FAQ 库;樊康新采用倒排索引技术加速 FAQ 库的检索速度，但是没有涉及到分布式处理技术，句子相似度的计算方法也存在不足。本文中采用了改进编辑距离算法进行汉语句子的相似度计算，并利用 HADOOP 框架，实现了一个面向远程教育的分布式自动答疑系统，能够很好地提高学习效率及资源的利用率。

2 系统设计概述

系统接收用户提交的自然语言描述式后,首先对问题进行分词，找出其中的关键词组并扩展，确定问题类型并定位到相应的检索服务器；然后根据关键序列先在与类型相对应的 FAQ 库中进行模糊检索，根据返回的问题集的大小，确定是否使用并行算法进行相似度计算；最后通过相似度算法检测是否有相同的问题,若有，则直接把 FAQ 库中这个问题的答案返回。当用户认为问答一致时，答疑过程结束，当用户认为回答的问题偏差很大时，可以将这样的问题提交给系统，让教师对答案进行人工修正后再入 FAQ 库。随着系统使用时间的增加，问题库将因教师不断的回答而自动扩大，系统将变得越来越实用和高效。答疑系统工作流程如图 1 所示。

3 系统实现

本答疑系统主要包括：知识库存储、问题理解、知识库索引模块和知识检索模块四个部分。其中问题

① 基金项目:新疆师范大学研究生科技创新项目基金(20101105);国家社科基金“十一五”规划 2010 年度教育学一般课题(BCA100025)
收稿时间:2011-10-09;收到修改稿时间:2011-11-10

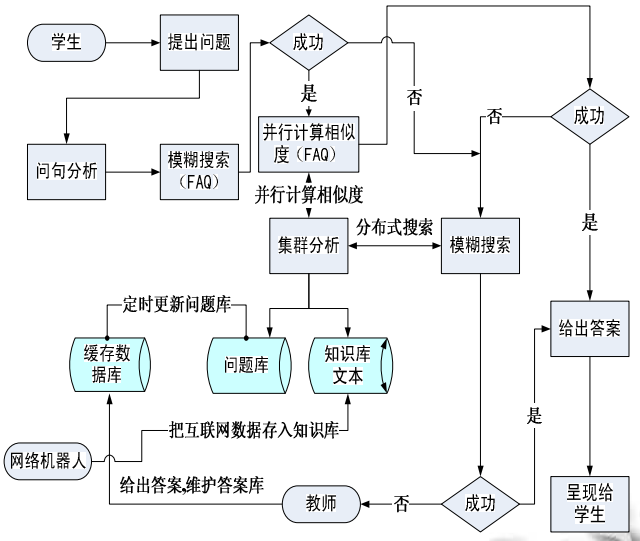

图 1 答疑系统工作流程图

理解模块是对问句进行预处理操作，包括问题的分类、关键词的提取和关键词扩展；知识库索引模块是通过 Lucene 构建索引；知识检索模块则通过分词分解问句检索出相关问题再进行相似度计算。

3.1 知识库的存储

答疑的知识库的形成主要由各科老师、管理员长期添加形成的。简单的单机关系型数据库无论从速度上还是存储容量上都满足不了需求。本文利用倒排索引结构对 FAQ 库进行索引然后直接存入分布式存储系统中。另外在本答疑系统中，问题和答案是以记录的形式存储在数据库表中，通过 Lucene 生成索引时只需对问题进行索引，而答案则直接通过 HDFS 进行分布式存储 1。这种方法生成的索引所占空间极小，因而有利于快速检索。知识库的结构如图 2 所示：

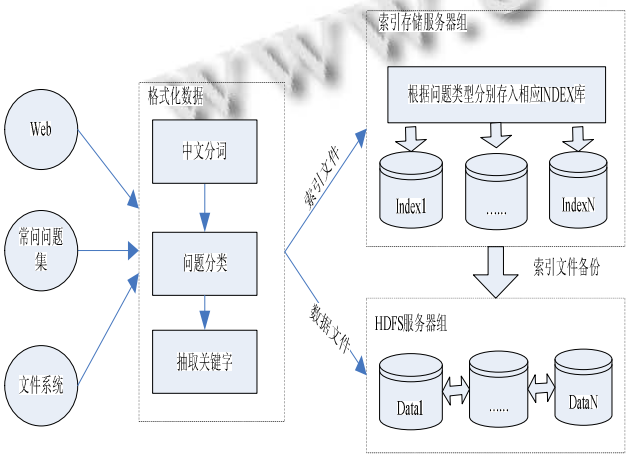

图 2 知识库分布式存储结构图

3.2 问题的理解

问题理解是答疑系统进行用户答疑的第一个步骤，其核心是正确分析理解用户提问意途，分析出用户所提问题属于哪类问题，问题中重点词汇有哪些，这样一方面可以缩小问题的匹配范围，另一方面可以提高匹配的准确率。

3.2.1 问句分类算法

在疑问句中，如果有疑问词，可以立刻判断出问题的类型，例如：“为什么”等。如果问句中的疑问词是通用疑问词，则通过一定规则找出问句中的疑问修饰词，例如“什么”，就需要根据其后面的名词来判断问题类型。由于汉语的灵活性，通用疑问词和疑问修饰词之间的位置关系是不确定的，因此，需借助句子中的其它的词所提供的信息来确定疑问修饰词。汉语中大部分简单问句有以下几个特征：

1）问句中只含有一个动词和一个疑问词。

2）如果通用疑问词之后不含有名词，那么疑问词之前的最后一个名词是疑问修饰词。

3）如果通用疑问词之后含有名词，又分为两种情况：

a.疑问词之后含有动词，且动词和通用疑问词之间不存在名词，那么整个问句中的最后一个名词是疑问修饰词。

b.疑问词之后不含有动词，或动词和通用疑问词之间含有名次，那么疑问词后面的第一个名词是疑问修饰词。

根据上面的几个特征，我们对各个疑问词进行了分析，并赋予了一定的编号，以便于以后的分析。

3.2.2 关键词组的形成

在问句中含有一些关键词代表了问句的主体含义，除了关键词，问句通常还包含一些没有太大检索意义的词，如“别的、并且”，其表义值非常低，通过构造一张停用词表（stop list），凡是这张表中出现的词都将被过滤掉。

在对问题库中的问题进行比较和匹配时，某些词常常不是原来问题的关键词，而是这些词的同义扩展。例如：“美国的首都在哪里？”与“美国的首府在哪里？”，两个问题中的关键词分别是“首都”和“首府”，如果没有关键词扩展，就会大大降低匹配和搜索的召回率。我们利用了同义词林对关键词进行了同义扩展。例如：中国接壤哪些国家？根据关键字抽取及扩展得

出：中国/接壤/接近/濒临/逼近/挨近/靠近/哪些/国家。

给同义词增加一个权值属性，同义词权值在系统运行初始化为 0，随后根据用户使用答疑系统过程中提问的次数而适当增加其权值。使用这种反馈方法的原因一方面是为了系统在运行过程中通过机器学习方法学习用户使用答疑系统习惯，增大或者减小 N 的取值，使得系统在使用过程中变得“智能”，能够越来越准确的为用户提供一定数目的准确的结果，同时在这个过程中答疑系统的语义处理模块得到训练，提高了答疑的准确度和效率。

3.3 语句相似度计算

3.3.1 词语相似度计算

词语相似度是语句相似度计算的基础。在《知网》中每个词汇可表达为若干个“概念”。“义原”是用于描述一个“概念”的最小意义单位[2]。本文采用中科院计算机研究所的“基于《知网》的词汇语义相似度计算方法”，来计算待比较两个语句中词语的相似度，即把两个词语之间的相似度归结到两个概念之间的相似度。《知网》将词语分为已登录概念词语和未登录概念词语。

3.3.1.1 登录概念词语的相似度计算

概念词分为实词和虚词两类，由于实词和虚词差别很大，本文约定实词和虚词的相似度置为 0。其中实词义项相似度计算分成以下四部分：

第一独立义原描述式：将两个概念的第一位置的相似度记为 d1=Sim1(S1,S2)

其他独立义原描述式：语义表达式中除第一独立义原以外的所有其他独立义原描述式的相似度记为 d2=Sim2(S1,S2)。

关系义原描述式：语义表达式中所有的用关系义原描述式的相似度记为 d3=Sim3(S1,S2)。

符号义原描述式：语义表达式中所有的用符号义原描述式的相似度记为 d4=Sim4(S1,S2)。

定义 1. 设两个实词概念 G1 和 G2，G1 有 4 个义原描述式：d11,d12,d13,d14,G2 有 4 个义原描述式:d21,d22,d23,d24，则概念 G1 和 G2 的相似度记为：

$$Sim(G_1,G_2)=\beta_1 Sim(G_1,G_2)+\sum_{i=2}^{4}\beta_1\beta_i Sim(G_1,G_2)$$

其中，βi(1≤i≤4)是可以调节的参数，各部分的重要程度通过βi 进行限定，且有：β1+β2+β3+β4=1，β1≥β2>β3≥β4>0。

定义 2. 设两个汉语词语 W1 和 W2，W1 有 n 个概念：g11,g12,…，g1n，W2 有 m 个概念：g21,g22,…，g2m，则 W1 和 W2 的相似度为各概念的相似度之最大值，即：

$$Sim(W_1,W_2)=\max_{i=1\ldots n,\ j=1\ldots m} Sim(g_{1i},g_{2j})$$

3.3.1.2 未登录概念词语的相似度计算

1）未登录概念词的切分及语义确定

利用逆向最大匹配法将未登录概念词切分成多个概念，如:“长沙市”=“长沙”＋“市”。假设基本组合概念 BCC 由原子概念 PC1 和 PC2 组成，即 BCC=PC1+PC2，并且 PC1 和 PC2 的语义表达式分别为 Def(PC1)和 Def(PC2)，则 BCC 的语义表达式表示为：

$$Def(BCC)=Def(PC2)\cup Def(PC1) \quad (1)$$

同时，根据“重心后移”原则[3]，令 PC2 的第一基本义原作为 BCC 的第一基本义原。令 PS(C)表示概念 C 的第一基本义原，则有 PS(BCC)=PS(PC2)。

2）组合概念的相似度计算

组合概念相似度计算又可分为两种：组合概念(未登录词)和原子概念(登录词)的相似度计算；组合概念与组合概念的相似度计算。

对于第一种情况，可以把参与运算的原子概念作为组合概念的参照概念，求解组合概念的语义表达式，进而计算两个语义集合的相似度。对第二种情况本文采取了一种简化策略：首先根据公式（1）计算两个组合概念各自的语义表达式，然后相互以对方为参照概念修正自己的语义关系，最后以修正后的语义集合进行相似度计算。

3.3.2 语句相似度计算

对于两个汉语相似度计算最终是一组词语的相似计算，首先通过分词程序将句子简化为一组关键词语，其中只包含名词、动词、形容词、限定性副词等同语义关系密切的词语，去掉会引入噪声数据的虚词，然后计算两个关键词组成的集合之间的相似度。

定义 3. 设两个语句 A 和 B，语句 A 由词语 x1,x2,…,xm 组成，语句 B 由词语 y1, y2,…, y3 组成，则语句 A 与 B 的相似为两组词语间的语义相似度，记为 S。语义相似度 S 的计算详见文献[4],在此不赘述。

4 实验结果与分析

在局域网的环境下，本系统利用基于 Hadoop 框架下的 Map/Reduce 编程模型,采用 master/slave 结构，使用了由 6 台 PC 机搭建的服务器集群，其中一个做为管理任务的 JobTracker 主服务器，其他作为 TaskTracker 的从服务器[5]。MySQL 安装在 192.168.1.1 名称节点上。系统从性能及返回结果准确度两方面进行分析与测试如下：

4.1 系统性能测试

本文通过前台输入检索信息“日本的首都在哪里”的自然语言表达式,对 Hadoop+Lucene 检索和数据库模糊查询的查询速度进行测试,结果如表 1 及图 3 所示：

表 1 查询时间对比表

问题数目	SQL 查询	Hadoop+lucene 查询
5,990	0.068s	0.078s
17,794	0.156s	0.109s
27,647	0.316s	0.109s
65,012	0.547s	0.110s
110,768	0.719s	0.110s
278,540	1.180s	0.141s

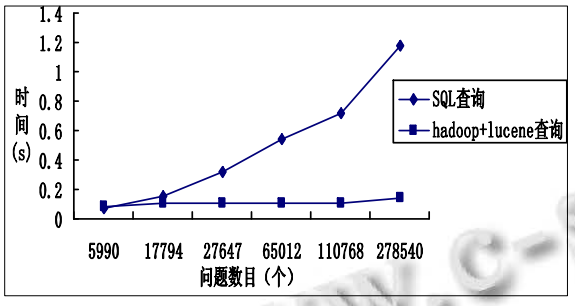

图 3 SQL 与 Hadoop+lucene 查询对比

在上述实验结果中可以看出,在问题数量非常小及问句字段含关键字少的情况下，SQL 查询及 Hadoop+Lucene 查询的所耗的时间量基本相同，随之问题数量的增加 SQL 查询的时间明显高于 Hadoop+Lucene 索引查询，SQL 查询时间几乎成倍增长，而 Hadoop+Lucene 索引查询时间以线性方式缓慢增加。另外当以 Hadoop+Lucene 进行检索时，其中一台设备宕机的情况下，仍然能够很好的运行，SQL 查询单机查询则不能。因此当系统长时间运行，问题库数据将达到 TB 级，无论从速度上还是安全性能来讲，SQL 根本满足不了需求，将 Hadoop+Lucene 用在海量数据检索，相对模糊查询来说，速度优势是很明显的。

4.2 答疑返回结果测试

由于本系统综合测试只涉及问题的数量以及问题的相似度，并不涉及问题的内容，因此现有问题数据主要来源百度知道互动平台，共收集问题 200 个，我们采用人工提问方式对 200 个问题进行集中测试，判断系统匹配的准确程度，系统按照相似度数值由大到小的顺序返回前 M 个候选问题-答案对。实验中采用刘群[2]等人提出的基于语义词典的方法与本文实现算法进行了对比，给出了相应的结果。实验中的参数 M=4，即取前 4 个最优解；Sim=0.60，即相似度阈值最小为 0.60。

本研究所采用结果评价指标为准确率(Precision)、召回率(Recall)和宏平均 F-值。其中准确率与召回率分别做如下定义：

$$\text{准确率 P} = \frac{\text{答对的问句数量}}{\text{识别的总问句数量}}$$

$$\text{准确率 P} = \frac{\text{答对的问句数量}}{\text{人工对测试数据识别结果数量}}$$

准确率和召回率从两个方面反映了相似问句的识别结果，它们之间相互抑制，因此二者应当综合考虑。宏平均 F-值就是一个综合指标，公式定义如下：

$$F_\beta(R,P) = \frac{(\beta^2+1)\times PR}{\beta^2 P + R}$$

其中，参数 β 用来为准确率（P）和召回率（R）赋予不同的权重，本文中令 $\beta=1$，即准确率和召回率的权重相同。实验结果如表 2 及表 3 所示：

表 2 相似度测试部分实验结果

用户问题	番茄是什么味道?	计算机是什么时候产生的?	为什么国民大革命失败了?
问题库中相近（同）问题 1	蕃茄的口味如何?	计算机的何时诞生?	国民大革命失败的原因是什么?
相似度	0.738	0.671	0.65
问题库中相近（同）问题 2	西红柿是水果吗?	计算机哪一年出现的?	国民大革命失败的标志是什么?
相似度	0.510	0.642	0.61

(下转第 64 页)

裕度应大于 6dB。所以从频域上分析系统稳定，而且满足相应的相位裕度和幅值裕度要求。从时域上分析达到稳定的时间约为 185ms，上升时间为 88ms，并且稳态误差为 1.5%，故能满足系统的控制要求。

参考文献

1 Astrom KJ, Hagglund T. PID Controllers: Theory, Design and Tuning. Research Triangle Park: Instrument Society of American. 1995.

2 Conway J, Watts S. A software Engineering Approach to LABVIEW. New Jersey: Prentice Hall PTR. 2003.

3 Mokhtari M, Marie M. Engineering Applications of MATLAB 5.3 and SIMULINK 3. Beijing: Publishing House of Electronics Industry. 2002.

4 孙振华.INSTRON-1343 电液伺服疲劳试验机计算机控制及数据处理系统.实验技术与管理,1999,16(2):34–35.

5 许贤良,王传礼.液压传动系统.北京:国防工业出版社,2008.

6 厉虹,杨黎明,艾红.伺服技术.北京:国防工业出版社,2008.

7 吴麒,王诗宓.自动控制原理.第 2 版.北京:清华大学出版社,2006.337–344.

8 梅晓榕.自动控制原理.第 2 版.北京:科学出版社,2007.144–146.

9 吴新余,周井泉,沈元隆.信号与系统—时域、频域分析及 MatLab 软件的应用.北京:电子工业出版社,1999.

10 Alan V. Oppenheim, Alan S. Willsky, S. Hamid, et al. Signals and Systems. Prentice Hall, 2002.

11 朱骥北.机械控制工程基础.北京:机械工业出版社,2006.

(上接第 25 页)

问题库中相近（同）问题 2	怎样做西红柿鸡蛋汤?	计算机什么时期发展最快?	国民革命失败的原因是什么?
相似度	0.500	0.528	0.341
问题中距较远（同）问题 3	西红柿对身体好吗?	第一台计算机问世于何时?	怎么理解中共在国民革命的作用?
相似度	0.381	0.400	0.231

表 3 相似度计算方法实验结果对比

方法	准确率	召回率	F-指标值
基于刘群等人的语义计算方法	78.82	77.75	78.28
本文的计算方法	88.21	85.54	86.85

从上述结果我们可以看出，在基于语义的相似度计算中，大部分问题的相似度计算在趋势上符合人们的直观感觉，因此以 M 条最优解作为回答正确与否的判断标准时，准确率超过了 85%，所以从整体上看，系统根据可扩展语义的相似度计算来进行问题检索可以胜任实际应用的需求。

参考文献

1 王学松.Lucene+Nutch 搜索引擎开发.北京:人民邮电出版社,2008.

2 刘群,李素建.基于《知网》的词汇语义相似度计算.计算语言学及中文信息处理,2002,7:59–76.

3 夏天.汉语词语语义相似度计算研究.计算机工程,2007, (3):191–194.

4 宁可为,王炜,李园伟.基于 Hadoop 句群相似度计算研究.计算机系统应用,2010,19(12):59–63.

5 White T. Hadoop: The Definitive Guide. 北京:清华大学出版社,2010.166–317.