●苏小虎，杨思春（安徽工业大学 计算机学院，安徽 马鞍山 243002）

基于改进 VSM 的中文问答系统研究*

摘 要：针对向量空间模型中的权重计算公式仅考虑词语项在文档中的相关频数，提出词语项本身的领域权重概念，改进了向量空间模型的权重计算。同时结合关键词距离和关键词顺序信息，实现了句子相似度计算，以特定课程的 FAQ 库检索作 S@n 测试对比，结果表明改进后的相似度模型提高了 S@n 值。

关键词：向量空间模型；领域权重；中文问答系统

Abstract: In view of the fact that only the term's relevant frequency in documents is considered in the weight calculation formula of Vector Space Model (VSM), a concept of term's domain weight is put forward to improve the weight calculation of VSM. Further more, the keywords' distance and sequence are combined to realize the similarity calculation of the sentence. By conducting the S@n test with the special course's FAQ database, the results show that the S@n value is increased by the improved similarity model.

Keywords: VSM; domain weight; Chinese Question Answering System

在每年一度的文本信息检索（TREC）会议上，自动问答（Question Answering Track）一直是最受关注的主题之一[1]。当前的中文问答系统主要有针对开放域和针对特定域两大类。开放域的问答系统如南京理工大学张亮博士研究的中文问答系统[2-3]等，特定域的问答系统如北京师范大学 Vclass 教学平台中的 Askme 答疑部件[4]等。因开放域的中文问答系统经常不能返回令人满意的简洁答案，所以大多研究具有实际效用的特定领域，如税务业务咨询[5]等。本文在对向量空间模型（VSM）中权重计算方法的改进中引入领域权重概念，并辅以关键词距离和关键词顺序信息，研究基于特定领域 FAQ 的中文问答系统。

1 基于特定领域 FAQ 的中文问答系统设计

1.1 系统主体结构图

整个系统主体分为三大模块：①问句分析模块；②FAQ 库检索模块；③答案处理模块，如图 1 所示。问句分析模块针对用户问句作处理，主要包括停用词过滤、问句类型分析判断、分词等，目的是获取用户问句中的主要提问信息，即“理解”用户的提问意图；FAQ 库检索模块主要是根据“理解”的用户意图检索 FAQ 库，也即计算用户问句和 FAQ 库中各问题的相似度，从中找出大于指定相似度值（阈值）的相关问题，构成候选问题集；答案处理模块对候选问题集分析处理，再把处理结果反馈给用户，有答案的显示答案，无答案的给出相应提示信息。

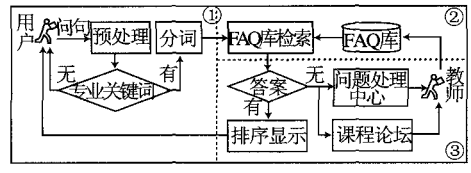

图 1 系统框架

1.2 FAQ 库设计

针对 VB 课程辅导答疑，FAQ 一部分来自任课教师平时的辅导答疑积累，另一部分来自网上 VB 论坛和学生留言，共 1 765 条，通过人工整理分成标准控件、常用函数、错误信息等 25 个类别。

FAQ 库材料暂以文件和文件夹的形式存储。其中问题以文件形式存储，文件名是 FAQ 的问题，内容就是答案；问题的类别以文件夹形式存储，再把相关课程的所有内容放在相应的课程文件夹中。如 VB 课程的 FAQ 放在以“VB 课程”为名的文件夹中，而与标准控件相关的问题文件全放在以“标准控件”为名的问题类别文件夹中，见图 2。以文件存储 FAQ 库，可方便使用 doc，htm/html 等多种文件格式，答案不再是单纯的文本形式，可以是声、图、文并茂，使答案更直观，更易理解。

1.3 问句分析

问答系统需要对用户提交给它的自然语言问句进行分析处理，包括问句预处理和分词，具体流程为：

1）问句预处理。首先，对问句进行关键词过滤，不

* 本文为安徽省高校省级自然科学基金项目研究成果之一，项目编号：KJ2007B245。

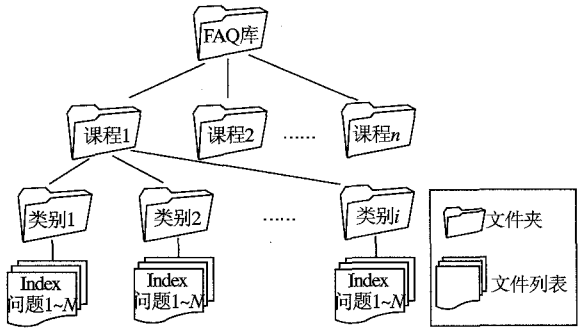

图2 FAQ 库材料存储组织

含任何关键词的问句不再作任何后续处理，直接反馈给用户提示信息，从而避免对无关问句做大量的计算工作，节约了系统开销。然后根据问句类型关键词判定问句所属类型[6-7]，既进一步明确用户提问意向，又缩小了检索范围，提高了检索速度。最后依据停用词词典删除停用词、标点符号，剔除无关信息。

2）分词。先依据本课程所涉及的中、英文领域关键词词典和普通词典进行分词，并对领域关键词和非领域关键词[7-8]（指名、动、形、限定性副词）作标记。采用双向最大匹配算法，对于两次分词结果一致的，任取其一作为分词结果；两次分词结果不一致的，在不一致的两个词中取关键词；如果两个都是关键词，则两个都保留且均参与后面的相似度计算。再依据同义词词典进行同义词改写和扩展，如“Form”与“窗体”等。领域同义词是依据相关领域收集整理，问句经过分词后，每一项结果如 W_i（pos，synonym，skey，key，tf，nk，log，wik，loc）形式。其中，pos 表示词性，synonym 表示同义词，skey 表示领域关键词，key 表示关键词，此 4 项在分词阶段获得相应的值。tf 表示文档内频数，nk 表示文档频数，wik 表示词的权重，log 表示 log（N/n_k），loc 表示 W_i 在本句子中的相对位置，此 5 项在后期需要计算相似度时才处理。本文中领域关键词指特定领域中的专业术语，如 VB 课程中“窗体”（Form）等名词；非领域关键词指一般的名、动、形、限定性副词等，文中未特别指名的关键词则包含两者。

1.4 FAQ 库检索

FAQ 库检索，就是将用户问句与 FAQ 库中的问题逐一进行匹配。在目前的各种检索模型中，最常用的是布尔模型和向量空间模型. 布尔模型虽然实现简单，搜索速度快，但查准率低。本文采用基于向量空间模型的句子相似度计算来实现问句与 FAQ 问题的匹配。鉴于相似度计算的时间复杂度比较大，具体处理时，分两步进行，第一步基于领域关键词在各问题类别中的频度确定的问句所属问题类别和问题预处理阶段获得的问句类型进行初步匹配；第二步基于句子相似度的计算，这样能避免对两个毫不相干的句子也进行复杂耗时的相似度计算工作，从而节约了大量的时间和系统开销。具体 FAQ 库检索流程见图 3。

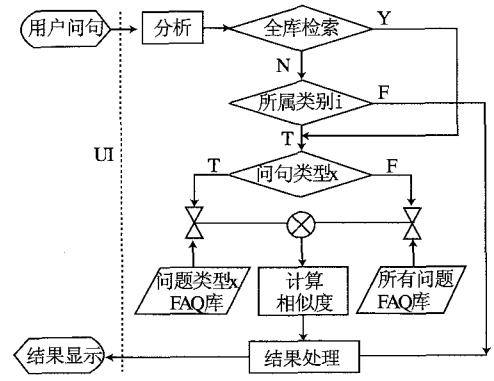

图3 FAQ 库检索模型

1.4.1 初步匹配 具体处理过程的算法描述如下：

1）是否全库检索，否则转 3）。

2）确定用户问句属于哪个问题类别。

3）条件 Flag = 1 吗？否则直接反馈给用户相应的提示信息，结束本次提问。

4）确定问句类型：①问句无类型，User_ Q 与所有 Flag = 1 的 FAQi 进行匹配；②问句有类型，User_ Q 与 Flag = 1 且 User_ Q_ type = FAQi_ type 的 FAQi 进行匹配。

说明：Flag = 1 表示问句与 FAQi 中至少有一个关键词相同或同义；User_ Q 表示用户问句，User_ Q_ type 表示用户问句类型；FAQi 表示 FAQ 库中的第 i 条记录（即第 i 个文件的文件名），FAQi_ type 表示 FAQi 的问句类型。

1.4.2 相似度计算 笔者在综合权衡几种相似度算法的时间复杂度和精确度的基础上，采用了算法时间复杂度和精确度均相对适中的向量空间模型，但对其中的权重作了改进。还引用了关键词距离、关键词顺序两项辅助的句子相似度计算方法，进一步提高计算精度。总的相似度计算分 3 个部分计算，最后综合 3 个部分取得最终的相似度。

1）传统向量空间模型。向量空间模型[9]中，文档是由相互独立的项（指基本语言单位）T_1，T_2，…，T_m 构成，令 $D = (D_1, D_2, \cdots, D_n)$ 表示由 m 个项构成的 n 个文档的文档集合。其中 $D_k = (W_{1k}, W_{2k}, \cdots, W_{mk})^T$ 就是文档向量。那么，两个文档 D_1、D_2 之间的相似度就可用公式（1）来计算。

$$\text{simVSM}(D_1, D_2) = \frac{\sum_{k=1}^{n} W_{1k} \cdot W_{2k}}{\sqrt{\sum_{k=1}^{n} W_{1k}^2 \cdot \sum_{k=1}^{n} W_{2k}^2}} \quad (1)$$

其中，W_{ik} 表示项 T_i 发生在文档 D_k 中的权重，计算方法主要根据 tf-idf 公式，目前效果较好的权重评价函数是：

$$W_{ik}' = \text{tf}_{ik} \cdot \text{idf}_k \quad (2)$$

公式（2）中，tf_{ik}表示项T_i在文档D_k中的频数，即文档内频数，idf_k表示项的反比文档频数，$idf_k = \log(N/n_k)$，其中N表示文档集中的文档总数，n_k表示文档集中包含项T_i的文档频数。由于公式（2）没有考虑文档的长度因素，因此越长的文档越有可能被检索到，为了避免这种情况的发生，对公式（2）归一化调整后得公式（3）。

$$W_{ik}'' = \frac{tf_{ik} \cdot \log(N/n_k)}{\sqrt{\sum_{k=1}^{n} tf_{ik}^2 \cdot [\log(N/n_k)]^2}} \quad (3)$$

2）向量空间模型的权重改进。公式（3）中的W_{ik}''，反映了项在文档中的重要程度，它是依据项在文档中的相关频数计算而得，权且称之为项的频率权重。为了进一步体现不同项在文档中的重要程度，在原有频率权重W_{ik}''的基础上再考虑各项本身的重要程度，权且称之为项的领域权重，记为W_i。如在特定领域文档中，其领域关键词项肯定比非领域关键词项重要，比一般的普通词项更重要，即更能体现句子的中心意思。基于此思想引入项的领域权重W_i，对公式（3）中项的频率权重W_{ik}''的值作进一步的修正，即得公式（4）：

$$W_{ik} = \alpha_1 \cdot W_{ik}'' + \alpha_2 \cdot W_i \quad (4)$$

其中：$\alpha_1 + \alpha_2 = 1$，这里取$\alpha_1 = 0.8$、$\alpha_2 = 0.2$。W_i的值根据项本身的情况来定：如果项是领域关键词，则取$W_i = 1$；如果项是非领域关键词，则取$W_i = 0.8$；如果项是其他一般的普通词项，则取$W_i = 0.6$。

3）两项辅助的句子相似度计算方法。文献［10］中在基于关键词相似度的基础上，利用关键词距离、顺序信息来修正句子相似度的计算结果。在向量空间模型中，为了计算的方便，没有考虑项与项之间的距离和项的先后次序关系，所以也可引入此两项辅助的句子相似度计算方法，进一步修正向量空间模型的计算结果。

·基于关键词距离的相似度。关键词间距离反映了句子的凝聚程度。假设有句子D_1、D_2，如果两句中相同关键词之间的距离越小，则两者越相似。参见公式（5）。

$$SimDis(D_1, D_2) = 1 - \frac{Dis(D_1)}{Dis(D_1) + Dis(D_2)} \quad (5)$$

其中，$Dis(D_1)$表示问句中非重复关键词中最左及最右关键词之间的距离，$Dis(D_2)$表示答案中与问句相同的最左及最右关键词之间距离。如果同一关键词在句子中出现多次则以产生最小距离的关键词为准。

·基于关键词顺序的相似度。关键词顺序反映了它们之间的先后次序关系。假设有句子D_1、D_2，如果两句中关键词顺序越接近，则两者越相似。参见公式（6）。

$$SimOrd(D_1, D_2) = 1 - \frac{Rev(D_1, D_2)}{MaxRev(D_1, D_2)} \quad (6)$$

其中，$MaxRev(D_1, D_2)$表示数量为$KeyWC(D_1)$的自然数序列的最大逆序数；$Seq(D_1)$表示以答案中关键词对应在问句中的位置构成的自然数序列；$Rev(D_1, D_2)$表示$Seq(D_1)$的逆序数。

4）最终相似度。

$$Sim(D_1, D_2) = \beta_1 \cdot SimVsm(D_1, D_2) + \beta_2 \cdot SimDis(D_1, D_2) + \beta_3 \cdot SimOrd(D_1, D_2) \quad (7)$$

这里的参数值依据实验（限于篇幅相关实验数据略）初步调整为：$\beta_1 = 0.8$，$\beta_2 = 0.1$，$\beta_3 = 0.1$。

具体计算相似度时，先统计计算问句的分词结果W_i（pos，synonym，key，skey，tf，nk，log，wik，loc）中的相应词的tf，nk，log，wik，loc值；再对需要处理的FAQ_i（FAQ_i表示FAQ库中的第i个文件的文件名）进行和问句相同的问题预处理，并对FAQ_i的分词结果faqwi（pos，synonym，key，skey，tf，nk，log，wik，loc）中的各项统计计算赋值；最后代入公式（7）计算二者的相似度。设定相似度的阈值（本系统设定的阈值为0.6），把高于阈值的FAQ_i加入候选答案集。

2 答案处理

答案处理主要包括3个方面：

1）对检索到的候选答案集处理。如果候选答案集为空，表示用户输入的问题暂时没有相关答案，反馈给用户相应“暂无答案”的提示，并提醒用户可把“暂无答案”问题提交到相关课程“论坛”、留言板或发Email给相关老师，最后转入后台对“无答案问题”作处理；如果答案集只有一项，直接显示问题答案（即FAQ库中相应文件的内容）；如果答案集有多项则按相似度大小排序，构成相关问题列表显示，把最相似的问题答案反馈给用户。

2）显示问题答案。目前FAQ库的问题文件类型有3种：txt，doc，htm/html。显示答案时，先判断问题文件类型。对txt格式的问题文件需要读出其内容并显示；对doc，htm/html格式的文件，则只需访问指定位置的以“问题”为主名、以“doc或htm/html”为扩展名的文件并按指定方式直接显示文件。对于一些问题虽然有了答案，但用户如果对答案不够满意，也可选择提交到有关课程的论坛中讨论，也可选择留言或通过Email提交给相关老师。

3）无答案问题处理。对暂时没有答案的问题经领域关键词过滤后区别对待。①确实与本课程相关的问题，一律自动提交到“问题处理中心”文件夹（即在该文件夹中创建以用户“问句”为名称的空文件），由管理员集中，转送相关老师作答后，再补充到FAQ库中。②对于那些与课程本身毫无关系的问题，如“作业是什么，何时交?”、“什么时候考试?”等本可以直接废弃，但考虑这

些问题还有点意义，而且有时也有必要作特别说明，且问得比较多的问题，单独收集到“相关问题. txt”文件中。③对于那些毫无意义的问题，如“你下午去打球吗?”，则一个不留直接废弃（目前也没真正废弃，而是保存在“废弃问题. txt”文件中）。本系统中对②和③的区分是通过一本特殊的小词典，该小词典中包含了如“考试”、“考试重点”等词。

3 实验及分析

为了评测对 VSM 改进方法的实际效果，构造了三组测试集。第一组是从 Web 上 VB 相关论坛中搜集到的问句，共 262 条；第二组是从收集的用户问句中随机抽取了 317 条；第三组是直接从 FAQ 库中抽取了 124 条，再加以人工改造。评测时为了便于程序自动统计和人工核对，采用类似 TREC 的 S@ n（success at n）[6,11] 方法，即正确答案在前 n 个结果中的比例。对错判的评价主要考虑匹配到问句是否为正确答案，系统返回认为正确而实际错误，则认为是发生错误；对于系统没有发现正确的匹配问句不在考察范围内，取 S@1，评测结果如表 1 所示。

表 1 测试实验结果

计算模型	测试集	答对题数	无答案数	S@1
①VSM	第一组	35	167	0.36
	第二组	117	114	0.57
	第三组	101	0	0.81
②改进权重的 VSM	第一组	36	167	0.37
	第二组	120	114	0.59
	第三组	111	0	0.89
③改进权重的 VSM + 两项辅助方法	第一组	36	167	0.37
	第二组	123	114	0.60
	第三组	113	0	0.91

实验结果分析：

1）三组数据的比较。第一组数据的 S@1 值偏低。本 FAQ 库目前的服务对象考虑到是在校学生，构建库时参照了 VB 教材的教学大纲。而第一组数据来自 Web 论坛，其中的问题绝大多数来自一些工程技术人员或 VB 编程爱好者，问题一般较深入，超出了教材大纲的范围。所以，既出现了 63.7% 的问题未匹配上，又有约 60% 的问题反馈答案错误。第二组数据虽然来自学生用户，但 S@1 值也不够理想，无答案题数也接近 40%，经仔细查看发现，一些问题确实与课程无关，如“哪里有考试大纲?”。第三组数据的 S@1 值较好，但还是有近 10% 的问题的答案出错。经查证发现，是因对这些问句的所属问题类别判断出错，说明单凭领域关键词在各问题类别中的频率来判断问句的所属问题类别还不够准确，还有待更好的解决办法，暂时增加了“全库检索”的选择项。

2）三种相似度计算方法的比较。三组测试集的 S@1 值在相似度计算方法①到②、②到③的改进中，均有一定程度的提升，虽然不是很明显，但足以说明对传统 VSM 的改进还是有一定成效的，随着 FAQ 库规模的扩大，这种成效将会更明显。

4 结束语

本文介绍了一个基于特定领域的问答系统，以特定领域 FAQ 为知识来源，采用改进的 VSM 模型和两项辅助句子相似度的计算方法来抽取答案。该系统的显著特点是：①FAQ 库的组织采用了简单的文件夹和文本文件形式，使得 FAQ 库的增加和移植更方便。②句子相似度的计算方法在传统 VSM 基础上作了较大改进。从实验的结果来看，这种改进是有成效的，下一步工作应考虑改进问题自动分类算法、进一步引入句法语义层次[12] 的检索模型和支持更多的文件格式。□

参考文献

[1] 秦兵，刘挺，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（10）

[2] 张亮，黄河燕，胡春玲. 中文问答系统模型研究[J]. 情报学报，2006，25（2）：197-201

[3] 张亮. 面向开放域的中文问答系统问句处理相关技术研究[D]. 南京：南京理工大学，2006

[4] 柳泉波，黄荣怀，何克抗. 智能答疑系统的设计与实现[J]. 中国远程教育，2000（8）

[5] 陈义，胡志宇，曾玮，等. 税务业务咨询问答系统[J]. 计算机应用与软件，2007（2）

[6] 闫宏飞，陈翀. 词汇与中心词的距离信息对问句相似度匹配的影响[J]. 清华大学学报：自然科学版，2005，45（S1）

[7] 吴栋，滕育平. 中文信息检索引擎中的分词与检索技术[J]. 计算机应用，2004（7）

[8] 张晓辉，何玉廉，刘光然. 智能答疑系统中搜索技术的研究[J]. 微计算机应用，2006（5）

[9] Salton G. The SMART retrieval system- experiments in automatic document processing[M]. Englewood Cliffs, NJ: Prentice-Hall, Inc, 1971

[10] 王宇，战学刚，蔡建山. 基于网络的中文问答系统的研究[J]. 计算机工程与应用，2006（7）

[11] Hawking D, Craswell N. Very large scale retrieval and Web search[M]. [S. l.]: The MIT Press, 2005

[12] 刘亚军，徐易. 一种基于加权语义相似度模型的自动问答系统[J]. 东南大学学报：自然科学版，2004（9）

作者简介：苏小虎，男，1974 年生，硕士，讲师。

杨思春，男，1970 年生，硕士，副教授。

收稿日期：2008-02-25