基于加权潜在语义分析的答案抽取①

陈永平[1]，杨思春[2]，苏　新[1]，毛万胜[1]

[1](马鞍山职业技术学院 计算机系，马鞍山 243000)
[2](安徽工业大学 计算机学院，马鞍山 243002)

摘　要：问答系统应该能够用准确、简洁的语言回答用户用自然语言提出的问题,其关键和核心实现技术是答案抽取。结合关键词在用户问句和返回文档中的权重，通过潜在语义分析技术实现了中文问答系统中的答案抽取。实验结果表明，加权 LSA 的 MRR 值要明显优于未加权 LSA 和空间向量模型的 MRR 值，实际用于回答用户提出的问题具有较好的效果。

关键词：问答系统，答案抽取，潜在语义分析，空间向量模型

Answer Extraction Based on Weighted Latent Semantic Analysis

CHEN Yong-Ping[1], YANG Si-Chun[2], SU Xin[1], MAO Wan-Sheng[1]

[1](Department of Computer, Ma'anshan Vocational Technology College, Ma'anshan 243000, China)
[2](School of Computer, Anhui University of Technology, Ma'anshan 243002, China)

Abstract: Question answering system returns precise and concise answers for user questions in natural language, and its core technology is answer extraction. Based on weight importance of different keywords in user' questions and returned documents, a method for computing keyword weight is proposed. In the meantime, the weighted Latent Semantic Analysis technique is also introduced in this process. Experimental results show that the MRR of the proposed method is better than that of Vector Space Model, and gets a more satisfactory performance.

Key words: answer question system; answer extracting; latent semantic analysis(LSA); vector space model(VSM)

1　引言

自动问答是当前自然语言处理的研究热点[1]，它包括三个模块：问题分析、文档检索和答案抽取。问题分析模块对问题进行分类、关键词提取和关键词扩展；文档检索模块将问题分析的处理结果作为查询提交给搜索引擎，以获取与问题相关的网页；答案抽取模块从检索到的相关网页中抽取出相关的答案直接返回给用户。其中,答案抽取模块是问答系统中最为重要的处理模块,直接关系到问答系统返回答案的质量。在答案提取的实现技术方面，通常是计算用户问句和返回文档各个句子中相同的关键词数量来确定其相似度，并将相似度最大的句子作为答案返回给用户[1-3]。关于关键词的权重，常用的方法是用关键词的词频值作为关键词的权重[1,2,4]。但是，在一个句子中各个关键词的重要程度有时是不一样的，因此，需要根据其对实际答案抽取任务的贡献来加以区分；另外，存在于同一个句子和同一个文档中的各个词语之间存在着某种潜在的语义结构，同义词之间具有相同的语义结构，多义词之间具有多种不同的语义结构，忽略这种语义结构，会影响答案抽取的性能。针对上述问题，本文基于潜在语义分析（Latent Semantic Analysis，LSA）理论和关键词权重计算方法，给出一种基于加权潜在语义分析的答案提取方法，并通过实验验证了该方法具有较好的效果。

2　潜在语义分析理论

潜在语义分析是 Landauer，Dumais 等人提出的[5,6]，它是为了改善向量空间模型的效果而提出的。LSA 的基

① 基金项目:安徽省高校省级自然科学基金(KJ2010B223)
收稿时间:2011-04-26;收到修改稿时间:2011-06-05

础是文档和句子中的词与词之间存在一定的潜在的语义结构，同义词之间具有相同的语义结构，多义词之间具有多种不同的语义结构[4,7]。这些语义结构与其在文档和句子中的权重有关，因此通过构造文档的 m×n 词-句子矩阵来量化这些结构，从而消除同义词和多义词的影响。设词-句子矩阵为 S=（s_{ij}）m×n，s_{ij} 表示词语 i 在句子 j 中的权重。由于句子集合通过分词后得到的关键词是非常多的，所以，词-句子矩阵是个高维矩阵。又因为每个句子中只包含所有关键词中的一部分，导致这个高维矩阵是个稀疏矩阵。因此，要对这个高维稀疏矩阵利用奇异值分解技术进行降维处理，减少无用信息的干扰。

潜在语义分析是以奇异值分解技术为基础，将词-句子矩阵 S 分解为如下式的三个矩阵的乘积：

$$S=WAD^T \quad (1)$$

其中，W 是一个 m×m 的正交矩阵，它的每一列被称为左奇异向量；D 是一个 n×n 的正交矩阵，它的每一行被称为右奇值向量；A 是一个 m×n 的对角矩阵，且元素值按从大到小在对角线上排序。设矩阵 S 的秩为 r，则存在 k,且 k<r,通过把最小的 r-k 个奇异值置为零,S_k 就是原矩阵 S 的秩 k 的逼近公式:

$$S_k=W_kA_kD_k^T \quad (2)$$

式中，W_k 是 m×k 矩阵，是压缩到 K 维空间的词向量，m 是词的数量；A_k 是 k×k 的矩阵；D_k 是 k×n 的矩阵，是压缩到 K 维空间的句子向量，n 为句子数。由此就实现了对高维稀疏词-句子矩阵的降维处理。

这样得到的矩阵 S_k 是 S 的一个 k 秩最优近似矩阵，它保持了 S 中所反映的词和句子之间联系的内在结构（潜在语义）[6,7]，同时它又去除了词项使用上的因同义或多义而产生的噪音数据。将词的 k 维空间理解为概念空间，依据 S_k 矩阵，就可以将句子的词空间转化为语义概念空间。

潜在语义空间的转换实质上是降维的过程，即如何选择 K 值非常关键，K 值太大，则语义空间接近于标准的向量空间模型；K 值太小，则保留下来的重要语义结构太少，无法把握运算结果，且不能适应样本误差。因此，根据相关文献和具体实验选取 K 值，取 K 值在 100-200 之间。

3 关键词权重计算

潜在语义分析的基本思想是[4,7]：文档中的同义词之间具有相同的语义结构，多义词的使用具不同的语义结构，词汇间的这种语义结构与其在文档中的权重有关，通过计算文档中的词汇的权重来量化这种潜在语义结构，进而消除同义词和多义词的影响，提高了文档表示的准确性，因此，本文给出了文档中词汇权重的计算方法。同时为了提高答案抽取的准确率，本文也利用了文献[8]中关于用户问句中词汇权重的计算方法。

3.1 用户问句中关键词的权重计算

传统的关键词权重计算采用的是布尔模型，即用 0 或 1 表示相应关键词在句子中出现与否。为了进一步反映句子中各个关键词的重要程度，文献[8]利用信息熵的原理，分别基于训练问句库和未标注问句库给出相应的关键词权重计算方法：

1) 基于训练库的权重计算，将用户问句训练库 Q 中的问句通过预处理（分词、词性标注、去停用词、问句分类）分成 N 个问句类型集合 Q={Q_1，Q_2，…，Q_N}。设 q_i 为词语 t 的权重，f_{ti} 为词语 t 在集合 Q_i 中出现的频率，n_t 为词语 t 在训练库中出现的频率，则

$$q_i=\{1+\frac{1}{\log N}\sum_{i=1}^{N}[\frac{f_{ti}}{n_t}\log(\frac{f_{ti}}{n_t})]\} \quad (3)$$

2) 基于未标注问句库的权重计算，将问句集 Q* 通过预处理（分词、词性标注、去停用词等）不给出问句类型集合，而是利用 KNN[9]聚类技术将数据分成不同的簇集。同一簇集的问句是相似的，可以看成是同一类型，而不同簇集的问句是相异的。这样将问句库 Q*分成 K 个簇集 Q*={Q_1*，Q_2*，…，Q_K*}，其中 k 是簇集的数量。设 q_i*为词语 t 的权重，f_{ti} 为词语 t 在集合 Q_i*中出现的频率，n_t 为词语 t 在 Q*中出现的频率，则

$$q_i^*=\{1+\frac{1}{\log k}\sum_{i=1}^{N}[\frac{f_{ti}}{n_t}\log(\frac{f_{ti}}{n_t})]\} \quad (4)$$

综合公式（3.1）和（3.2）可以计算出 t 的最终权重,在此定义为 q_i 和 q_i*的线性组合：

$$W^i=\alpha\, q_i+\beta\, q_i^* \quad (5)$$

式中的 α 和 β 是平衡因子，且 α+β=1。在实际计算中可以适当调节二值的大小，以得到更加精确的结果。在文献[8]中，α 和 β 的值都取 0.5。

考虑到该方法的有效性，本文以下在计算问句中每个关键词的权重时也是采用这种方法。

3.2 文档中关键词的权重计算

众所周知，每一个关键词在句子中的重要程度是不一样的，同样，每一个关键词在文档中的重要程度也是不一样的，所以本文在计算文档中关键词的权重时既考虑了关键词在句子中的重要程度，同时也考虑了它在文档中的重要程度[10-12]。对于问答系统，用户问句通过预处理后得到关键词集合，将关键词集合提交给搜索引擎，返回的是与问句相关的文档，将每个文档进行预处理（分句、分词、词性标注等），就可将每个文档分成若干个句子，而每个句子又是由若干个关键词组成的，因此可以用下面的这个矩阵 S 来表示：

$$S=\begin{pmatrix} s_{11},s_{12},\cdots\cdots,s_{1n} \\ s_{21},s_{22},\cdots\cdots,s_{2n} \\ \cdots\cdots\cdots\cdots \\ s_{m1},s_{m2},\cdots\cdots,s_{mn} \end{pmatrix} \quad (6)$$

其中的 m 是每一个句子中的关键词的个数，n 是文档中所包含的句子数，s_{ij} 表示词语 i 的权重。这里的 s_{ij} 的既考虑了关键词在句子中的重要程度，也考虑了它在文档中的重要程度。参照文献[11，12]中对词语权重的改进方法，我们给出如下计算公式：

$$s_{ij}=AL(i,j)\times AG(i)\times AP(i,j)\times AW(j) \quad (7)$$

与文献[11,12]不同的是，在文献中各个量反映的是关键词在多文档或多段落中的权重，而本文反映的是关键词在句子和文档中的权重。突出了关键词在句子和单文档中的重要性。

下面分别讨论上式中的每一部分的意义。首先给出以下几个量的解释：

tf_{ij}：关键词 i 在句子 j 中出现的次数；

qf_i： 文档中出现关键词 i 的句子数；

tqf_i：关键词 i 在文档中出现的次数；

atqf：文档中所有关键词出现的次数之和；

其中，AL（i，j）表示关键词 i 在句子 j 中的重要程度，它和 tfij 有关，本文用如下式子计算其值：

$$AL(i,j)=\log_2(tf_{ij}+1) \quad (8)$$

AG（i）是关键词的全局权重，表示关键词 i 在整个文档中的重要程度，它和文档中出现关键词 i 的句子数 qf_i 有关，n 表示文档中的句子数。

$$AG(i)=\log_2\left(\frac{n}{qf_i}\right) \quad (9)$$

AP（i，j）表示句子 j 在文本中分辨词语 i 的能力，它和关键词 i 在句子中出现次数及在文档中出现次数有关：

$$AP(i,j)=\log_2\frac{tf_{ij}}{tqf_i} \quad (10)$$

AW（j）表示句子的全局权重，表示句子在文档中由于提供词语互信息量的差异而具有的不同重要程度。参照文献[11]词语全局权重中的熵定义方法，将 MI（word）- MI（word/qj）看成 q_j 确定出现后，文本中的词语消除了的不确定性，即 q_j 提供给 word 变量的信息量。其中 MI（word/q_j）是句子确定条件下的词语条件分布熵，描述某个句子对消除词语不确定程度的。它的定义如下：

$$AW(j)=1-\frac{MI(\frac{word}{q_i})}{MI(word)} \quad (11)$$

其中，

$$MI(word)=-\sum_{i=1}^{n} q(word_i)\times\log_2 q$$

$$(word_i)=-\sum_{i=1}^{n}\frac{tqf_i}{atqf}\times\log_2\frac{tqf_i}{atqf}$$

$$MI(word/q_j)=-\sum_{i} p(i/j)\times\log_2 p$$

p（i/j）是条件句子 j 出现的情况下，词语 i 出现的的概念，其计算方法如下：

$$p(i/j)=\frac{tf_{ij}}{tqf_i}$$

4 基于加权潜在语义分析的答案抽取

基于前文介绍的潜在语义分析（Latent Semantic Analysis，LSA）理论和关键词权重计算方法，本文给出一种基于加权潜在语义分析的答案提取方法。具体的思路是，对于用户提交给问答系统的的问句，通过问题分析阶段的处理，得到由关键词构成的查询集，将查询集提交给搜索引擎进行相关文档的检索。为了提高系统执行效率，本文只利用返回结果中的摘要部分作为答案抽取的资源，而不是利用摘要对应的网络源文件。选取检索返回结果中的前 100 个摘要，对选取的 100 个摘要信息分别进行分句处理，将摘要信息拆分成由若干个单个句子组成，然后将拆分后的所有句子分词，词性标注及去停用词，最后利用问句分类结果，将不含预期答案类型的句子排除掉，得到答案

的候选集合。在这个答案候选集合中，每个文摘被拆分成由若干个句子组成，而每个句子又是由若干个关键词构成，每个文摘就是一个词-句子矩阵，这样答案候选集就是由 100 个词-句子矩阵构成的集合，记为：$D=\{D_1, D_2, \ldots, D_{100}\}$。对这个集合进行相应的处理就可抽取出答案，并返回给用户。以下为该方法的形式化表述：

对集合 D 中的每一元素 Di（$D_i \in D$，$i \in [1 \ldots 100]$）进行如下操作：

1) 利用公式 3.5 计算矩阵 D_i 中每个关键词的权值，得到了关于词-句子关键词的权重的矩阵，记为 S, 矩阵 S 中的任一元素 s_{ij} 表示词语 i 在句子 j 中权重及其在文档中的权重。

2) 利用上文的潜在语义分析理论，将矩阵 S 转换成一个潜在的词-句子语义空间 S_k。

3) 将用户问句 P 通过预处理形成关键字集合，根据公式 3.3 计算该集合中各个关键字的权重后，并利用下式映射到第 2)步的语义空间。

$$P^*=P^T W_K A^{-1} K \tag{12}$$

其中 $A^{-1}{}_K$ 是 AK 的逆矩阵，P 是映射前的问句向量。

4) 利用下面公式计算问句 P 与候选答案句子 s_j 的相似度：

$Sim（P，s_j）= Sim（P^*，s_j^*）=$

$$\frac{\sum_{m=1}^{k} p^*_{im} \times s^*_{jm}}{\sqrt{\sum_{m=1}^{k} (p^*_{im})^2 \times \sum_{m=1}^{k} (s^*_{jm})^2}} \tag{13}$$

其中 P*为问句 P 的映射后的语义向量，s^j*为句子 s_j 映射后的语义向量，k 为语义空间的维数，p_{im}*和 s_{jm}* 分别是 P*和 s_j*向量中的第 m 维的权值。

5) 在每篇文摘中选取 Sim 值最大的五个句子作为答案候选句子，所有的答案候选句子构成答案候选集，再将答案候选集中的句子按其 Sim 值的从大到小排序。

6) 选取排在前面 5 个句子作为答案返回给用户。

5 实验结果和分析

5.1 实验数据

本实验的问句集的数据来自于以下两个部分：第一部分是选自于哈尔滨工业大学信息检索研究室和中科院自动化研究所模式识别国家重点实验室，并在此基础上加上本课组自己扩充的部分，共 4500 个问句；第二部分是根据实验的需要从网上下载的 4500 个问句。从这两个部分各选取 500 个问句，共 1000 个问句构成测试集。对第一部分剩下部分和测试集中的问句通过分类方法识别出每个问题对应的问题类型[8,13]。这样第一部分剩下的问句就构成了已标注类型的问句集，而第二部分剩下的问句构成了未标注类型的问句集。测试集通过分类得到人物类型 93 个，地点类型 176 个，时间类型 135 个，数量类型 156 个，实体类型 172 个，其他类型 268 个。本文没有对测试集中所有问句进行测试，只对测试集中五种类型共 732 个问句进行测试。另外，根据前文介绍的理论和方法，在由搜索返回的文档中生成词-句子矩阵设为 3000×2000，使用 MATLAB 软件软件工具包进行 SVD 分解，生成语义空间，在此取 K 值为 140。答案抽取评估的方法采用 MRR（Mean Reciprocal Rank）标准，$MRR=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{ri}$，其中 n 为所有测试问句的数量, ri 为第 i 个问题的第一个正确答案的位置。

5.2 实验结果和分析

为了评价本文提出方法的性能，实验中对三种方法做了比较。方法 1 是基于 VSM（向量空间模型）的答案抽取方法，方法 2 是基于潜在语义分析的答案抽取方法（利用文献[4]介绍方法进行实验），方法 3 是本文提出的加权潜在语义分析方法。表 1 给出了三种方法答案提取的实验结果。

表 1 三种方法答案提取的实验结果

类型	问句数	VSM 方法的 MRR 值	LSA 方法的 MRR 值	加权 LSA 方法的 MRR 值
人物	93	0.52	0.59	0.69
地点	176	0.25	0.32	0.34
时间	135	0.29	0.41	0.46
数量	156	0.49	0.55	0.62
实体	172	0.36	0.42	0.49

从表 1 可以看出，在五种类型问题的答案抽取实验中，采用 LSA 方法要比 VSM 方法的 MRR 的值要高，这是由于 LSA 方法对 VSM 方法进行了变换处理，经过变换得到的潜在语义空间，不仅对 VSM 空间进行了降维处理，去除了 VSM 空间中的一些不必要的数据；而且，LSA 方法能够更准确反映词与词之间的

相关度。另外从表中也可看出加权 LSA 方法的答案抽取比 LSA 方法的答案抽取的 MRR 值也要高，主要是因为：突出了句子中各个关键字的权重的不同：一方面，考虑了用户问句中各关键字的权重的不同，利用信息熵来量化问句特征对问句分类的重要程度，将对问句特征具有重要作用的关键字赋予较高的权值；另一方面，对于通过搜索引擎检索返回的文档中，既考虑了各关键字在各个句子中的重要程度的不同，同时也考虑了各关键字在整篇文档中的重要程度的不同，并通过权值反映这些不同。这样对句子特征起重要作用的关键词被赋予较高的权重，所以加权潜在语义分析的答案抽取方法的答案抽取的准确率也得到了一定的提高。

6 总结

本文提出了基于加权潜在语义分析的答案抽取技术，计算不同关键词在用户问句中和检索返回的文档中的重要程度，给出相应的权值计算方法；然后利用潜在语义分析方法来实现答案提取。实验证明，该方法提高了答案抽取的准确率。但是，本文提出的方法由于计算关键词的权值和 LSA 方法的运算导致计算量大，所需存储空间也很大；对于地点类型由于分词等因素造成准确率没有达到相应的效果；另外，在文献[12]中将问句类型分为七大类，除了上述类型外，还包括描述类型，因为描述类型问句的答案往往要一段文字才能回答准确，在本文中没有对这方面进行研究；下一步的工作将对这些方面进行更加深入的研究。

参考文献

1 郑实福,刘挺,秦兵,李生.自动问答综述.中文信息学报,2002,16(6):46–52.

2 崔桓,蔡东风,苗雪雷.基于网络的中文问答系统及信息抽取方法研究.中文信息学报,2004,18(3):24–31.

3 余正涛,樊孝忠,宋丽哲,等.汉语问答系统答案提取方法研究.计算机工程,2006,32(3):183–185.

4 余正涛,樊孝忠,郭剑毅,耿增民.基于潜在语义分析的汉语问答系统答案提取.计算机学报,2006,29(10):1889–1893.

5 Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J. Amer Soc. Info. Sci., 1990, 41: 391–407.

6 Dong A. The latent semantic approach to studying design team communication. Design Studies, 2005,26(5).

7 董杰,王怡,武港山.基于潜在语义分析的信息检索.计算机工程,2004,30(2):58–60.

8 黄鹏,卜佳俊,陈纯,康志明,陈伟,胡洪涛.利用加权特征模型改进问句分类.浙江大学学报,2009,43(6):994–998.

9 MacQueen J. Some Methods for Classification and Analysis of Multivariate Observations. Proc. of the 5th Berkeley Symposium on Mathematical Statist and Probability. California, USA: University of California Press, 1967,1:281–297.

10 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.东南大学学报(自然科学版),2004,34(5):609–612.

11 刘云峰,齐欢,等.潜在语义分析权重计算的改进.中文信息学报,2005,19(6),64–69.

12 何媛媛.基于潜在语义分析的多网页自动文摘研究.硕士论文.上海师范大学.2008.

13 文勖,张宇,刘挺,马金山.基于句法结构分析的中文文问题分类.中文信息学报,2006,20(2):33–39.