基于常见问题集的 OTC 问答系统的设计与实现

Design and Implementation of OTC Question Answering System Based on Freqnently Asked Questions

樊康新　(南通大学 计算机科学与技术学院 江苏 南通 226019)

摘　要： 为帮助人们合理选择和安全使用 OTC 药品,设计并实现了一个基于 FAQ 的 OTC 问答系统。描述了该系统的工作流程和系统结构。详细阐述了系统实现的关键技术,包括问句特征向量的提取、基于倒排索引的查找算法、根据用户问题建立候选问题集和基于知网的语义相似度计算等。运行结果表明,对于有关 OTC 的常问问题和普遍性问题,该系统具有很高的准确率。

关键词： 问答系统 常见问题集 语义相似度 OTC

随着我国医疗体制的改革和卫生知识的普及,人们的医疗保健观念和用药意识日益增强,"大病进医院、小病找药房"已逐步成为人们的共识。但是,由于大多数患者医药知识不足,加之未能得到医师或药师的用药指导,因此在实施自我药疗时,往往会步入误区。随着非处方药(Over The Counter,简称 OTC)的进一步推出,自我药疗中存在的一些似是而非、一知半解的认识导致了 OTC 药品使用的不安全性。为了避免人们盲目诊断、主观选药用药所带来的危害性,开发一个 OTC 问答系统,对于帮助人们合理地选择和安全地使用 OTC 药品,增强自我保健和自我药疗意识具有重要的现实意义。

问答系统是目前自然语言处理领域的一个研究热点,它允许用户用自然语言提问,又能为用户返回一个简洁、准确的答案,而不是一些相关的网页,因此这种方式更接近于人们的思维和习惯。目前问答系统的研究一般可分为三类[1]:限定域问答系统、开放域问答系统和基于常见问题集(Frequently Asked Question,简称 FAQ)的问答系统。基于 FAQ 的问答系统把人们经常咨询的问题和相关答案保存起来,对于用户输入的问题,首先在常问问题库中查找,如果能够找到相同或很相似的问题,就可以直接将该问题所对应的答案返回给用户,而不需要经过问题理解、信息检索、答案抽取等许多复杂的处理过程,从而不仅提高了系统的效率,同时还提高了答案的准确性。

本文作者与某医药机构合作,收集、整理了人们在选购 OTC 药品过程中遇到的大量问题作为 FAQ 集,在此基础上,利用自然语言处理技术开发了一个基于 FAQ 的 OTC 问答系统,为人们准确地选择和安全地使用 OTC 药品提供了一个方便的咨询平台。运行结果表明,对于有关 OTC 的常问问题和普遍性问题,该系统具有很高的准确率。

1　系统设计概述

系统接收用户用自然语言描述的问题后,首先对用户问句进行预处理,包括自动分词、过滤掉停用词、提取关键词等,以关键词组成用户问句的特征向量;根据特征向量从 OTC - FAQ 库中把相关度较高的问句选出来作为候选问题集;然后利用词语间的语义相似度和词语的权重计算出用户问句和候选问题集中各个问句的语义相似度,对于计算出来的语义相似度大于设定阈值的问句,就认为它们所对应的问句和用户问句是相同问题或最相似问题,可以直接将这个 OTC - FAQ 库中的问句所对应的答案输出给用户。如果计算出来的语义相似度均不大于设定的阈值,就认为 OTC - FAQ 库中不存在与用户问题相同或相似的问题,系统将规范用户问题,使用信息检索、答案抽取等方法找出答案,并形成索引,及时补充到 OTC - FAQ 库中。系统的工作流程如图 1 所示:

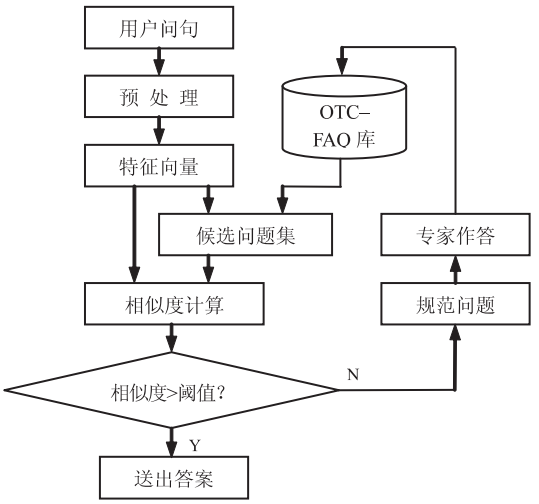

图 1　系统工作流程图

2　系统实现的关键技术

2.1　问句特征向量的提取

问句的特征向量是指在分词的基础上，去掉介词、叹词、象声词等虚词以及区分意义不大的高频词和低频词后形成的关键词序列。

自动分词是将由自然语言提出的问题根据词库分解成若干词语的集合。由于本问答系统关心的不是出现在问句中的每个词是否都能被准确切分而是那些对检索有意义的相关词语能否被快速准确切分，因此为提高系统效率，本系统不使用现有庞大的中文词库，而是根据系统特点设计了专用词库，包括专业词库和常用词库两类。专业词库主要涉及 OTC 等医药学内容，包含相应学科的概念、术语、符号等，它对分词具有决定性的作用。常用词库主要包含常见的名词、动词和少量的形容词等。由于专业关键词对问题检索的贡献往往大于常用词语，为提高系统的搜索精确度，对不同的关键词赋予不同的权值。通常，专业关键词的权值设定为 1，常用词的权值设定为 k（0 < k < 1）。专业词库和常用词库也将随着系统的使用不断地被补充、修正和完善。

常用的中文分词方法有：基于字符串匹配（或称为词典）的分词方法、基于统计的分词方法和基于理解的分词方法。针对 OTC 问答系统的特点，本文采用基于字符串匹配和基于统计相结合的分词方法，即首先按正向最大匹配法和逆向最大匹配法对用户问句进行分词，当切分结果出现歧义时，再考察相邻字的共现概率，选择共现概率大的相邻字组成词或词组。

自动分词时首先对用户问句进行预处理，去除无关的标点符号，将中英文字符分离，分割成若干个字串。再分别与专业词库和常用词库匹配，将匹配成功的词语组成问句的特征向量，剩余字符串则舍弃[2]。

例如，对于问题“我有一同事是前列腺肥大患者，近来患上感冒，请问专家应该选用何药？谢谢！”，经过上述分词等预处理后，提取的特征向量为（前列腺肥大，患者，感冒，应该，选用，何药），其对应的权重向量为（1，0.5，1，0.6，0.8，0.9）。

2.2　FAQ 库倒排索引结构的建立

查找相似问句时，一个最简单的方法是遍历 FAQ 中的每一个问句，计算其与用户问句的相似度，从中选出相似度最大的一个。这种方法虽然简单，但它使得与用户问句相似度为 0 和相似度很低的 FAQ 问句都要参与计算，因而检索效率十分低下。当 FAQ 库规模很大时无法在实际中应用。为解决这一问题，本文建立了基于单词倒排索引的 OTC – FAQ 结构（如图 2 所示），可以实现最相似句子的快速查找。

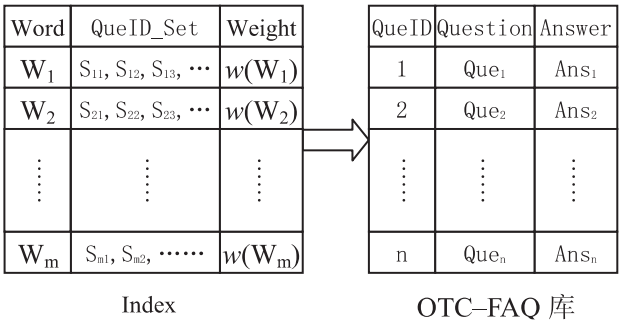

图 2　OTC – FAQ 库的倒排索引结构

图 2 中的 OTC – FAQ 库存放了本系统所有的问题及相应的答案。Question 和 Answer 分别为问题及其对应的答案，QueID 是问句 – 答案对的索引号，Index 表中的 Word 字段内容为 OTC – FAQ 库中的所有问句所包含的词语（有序排列），QueID_Set 字段记录了 OTC – FAQ 库中所有含有词语 Wi（1 < i < m）的问句的 QueID，Weight 字段内容为对应词语 Wi 的权重。

设 OTC - FAQ 库中有 n 个问句，则建立 OTC - FAQ 库的倒排索引结构的算法如下：

```
For i = 1 to n do
  {
   Locate to QueID = i,
    对问句 Quei 进行分词等预处理，提取特征向量
     Q = (w1, w2, w3, ……, wm),
    对应的权重向量 Q′ = (e1, e2, e3, ……, em)
    For j = 1 to m do
      {
        在 Index 中查找 Word,
        若找到某个 Wk，使得 Wk = wj，则
          locate to Word = Wk,
          let QueID_Set [k] = QueID_Set [k] ∪ { QueID }
        若找不到某个 Wk，使得 Wk = wj，则
          在 Index 中按序（设在 t 位置上）插入一条记录，
            let Word[t] = wj, QueID_Set[t] = { QueID }, Weight[t] = ej。
      }
  }
```

2.3 候选问题集的构造

为提高系统运行效率，从 OTC - FAQ 库中找出若干个相关问题组成候选问题集，以使后续的相似度计算等较复杂的处理过程都在候选问题集这个相对较小的范围内进行[3]。

设 OTC - FAQ 库中有 n 个问题，为建立候选问题集，需设置一个 n 行 2 列的数组 QuesArray[1..n,1..2]，数组的第 1 列存放 OTC - FAQ 库中的问句索引号 QueID，第 2 列记录相应问句中含有用户问句中词语的个数。初始时，QuesArray[k,1] = k (k = 1,2,3,……,n)，QuesArray[k,2] = 0。基于上述倒排索引结构，候选问题集 CandQueSet 的构造算法如下：

（1）对用户问句进行分词等预处理，提取特征向量 $U = (w_1, w_2, w_3, \cdots\cdots, w_m)$；

（2）对 S 作同义词扩展，得扩展向量 $U' = (w_{1.1}, w_{1.2}, \cdots, w_{i.1}, w_{i.2}, \cdots, w_{m.1}, \cdots)$

```
（3）For each w in U do
      {
        Locate to Word = w
        For each k in QueID_Set do
          QuesArray[k,2] = QuesArray[k,2] + 1
      }
```

（4）对扩展向量 U′重复步骤（3）；

（5）将数组 QuesArray 按 QuesArray[k,2] 作降序排序；

（6）由数组元素 QuesArray[1,1] ~ QuesArray[p,1]（例如本系统取 p = 30）中的问句号对应的问句组成的集合就是候选问题集 CandQueSet。

候选问题集的建立，使得在 OTC - FAQ 库中检索相似问句时，与问句相似度为 0 的句子不会参与计算，与问句相似度很低的句子只计算相同单词的个数，从而极大地提高了效率。

2.4 基于语义的问句相似度计算

计算问句间的语义相似度，需要一定的语义知识资源作为基础. 本文采用董振东和董强先生创建的知网（HowNet）作为系统的语义知识资源。

2.4.1 知网简介[4]

知网是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。在知网中，比较重要的两个概念是："概念"（也称义项）和"义原"。

"概念"顾名思义，就是对词语给一个定义，简单来说就是对词汇语义的一种描述。在实际的自然语言中，每个词语可以有多个语义，也就是说可以表达为几个概念。在知网中，最小的描述单位叫做"义原"，由"义原"按照一定的规则组织在一起的语言称为"知识表示语言"。"概念"就是用这种"知识表示语言"来描述的。

知网用一系列的"义原"对每个"概念"进行描述，不是将所有的"概念"归结到一个树状的概念层次体系中，因此知网同一般的语义词典如《同义词词林》或者 Wordnet 有着本质的不同。在知网中，是将所有的"义原"都归结到一个树状的层次体系中。

义原作为描述概念的最基本单位，相互之间存在着复杂的关系。知网作为一个知识系统，是一个名副其实的网而不是树。在知网中，有 8 种义原之间的关系，分别是上下位关系、同义关系、反义关系、对义关系、属性－宿主关系、部件－整体关系、材料－成品关系、事件－角色关系。义原之间通过这些关系组成一个复杂的网状结构。在这些关系中，义原的上下位关系是最基本的关系，也是最为重要的关系。根据上下位关系，可以将所有的义原组成一个树状结构。在知网中，有 6 棵义原树可以体现上下位关系，它们是：Entity、Event、Attribute、At 是一个 rttibuteValue、Qunatity 和 QunatityValue。根据这些义原树状结构，就可以进行语义相似度计算。

2.4.2　基于知网的语义相似度计算

使用知网进行问句语义相似度计算的主要步骤为：首先 T 使用知网的义原树计算两个词语间的语义距离；其次，根据词语间的语义距离，计算两个词语间的语义相似度；最后，在对问句进行分析的基础上，计算用户问句与候选问题集 CandQueSet 中问句的语义相似度。

（1）词语间的语义距离

我们将词语间的语义距离定义为两个词语对应的义原在义原树中的最短距离。设有两个词语 w1 和 w2，记其语义距离为 $Dis(w_1,w_2)$，则

$$Dis(w_1,w_2) = |T_1 \cup T_2| - |T_1 \cap T_2|$$

式中 T_1、T_2 分别为 w_1 和 w_2 两个词语所在义原树从树根到该节点语义元素集合，$T_1 \cup T_2$ 表示义原树中从树根到 w_1、w_2 各自语义节点包括的所有义原的集合，$|T_1 \cup T_2|$ 是该集合元素的个数。$T_1 \cap T_2$ 表示 w_1、w_2 对应的义原树中相同语义节点的集合，$|T_1 \cap T_2|$ 表示公共节点的个数。

由上式可知，$Dis(w_1,w_2) \in [0,\infty)$，即两个相同词语的语义距离为 0。如果两个词语中有一个词语的义原无法在 6 棵义原树中找到，或者两个词语的义原分别处于两棵不同的义原树，则认为这两个词语间的语义距离为 ∞。

（2）词语间的语义相似度

词语间的语义相似度与词语间的语义距离有着密切的关系：两个词语间的语义距离越大，则其语义相似度越低；反之，两个词语间的语义距离越小，则其语义相似度越大。

在很多情况下，直接计算词语间的语义相似度比较困难，通常可以先计算词语间的语义距离，然后再转换成词语间的语义相似度。

设有两个词语 w_1 和 w_2，记其语义相似度为 $S(w_1,w_2)$，本系统采用如下转换关系计算词语间的语义相似度：

$$S(w_1,w_2) = \frac{k}{Dis(w_1,w_2) + k}$$

式中 k 是一个可调节的参数。由于 $Dis(w_1,w_2) \in [0,\infty)$，由上式可知：$S(w_1,w_2) \in [0,1]$。即两个词语间的语义距离为 0 时，其相似度为 1；两个词语间的语义距离为无穷大时，其相似度为 0；两个词语间的语义距离越大，其相似度越小（单调下降）。

（3）问句间的语义相似度计算

有了词语间的语义相似度，就可以用它来计算用户问句与 CandQueSet 中的问句之间的语义相似度。设有用户问句 A 和 CandQueSet 中的问句 B，A 包含的词语为 A_1、A_2、……、A_m，B 包含的词语为 B_1、B_2、……、B_n，词语 A_i（$1 \leqslant i \leqslant m$）和 B_j（$1 \leqslant j \leqslant n$）之间的语义相似度为 $S(A_i,B_j)$，则两个问句中任意两个词语间的语义相似度矩阵 S_{AB}：

$$S_{AB} = \begin{bmatrix} S(A_1,B_1) & S(A_1,B_2) & \cdots & \cdots & S(A_1,B_n) \\ S(A_2,B_1) & S(A_2,B_2) & \cdots & \cdots & S(A_2,B_n) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ S(A_m,B_1) & S(A_m,B_2) & \cdots & \cdots & S(A_m,B_n) \end{bmatrix}$$

由此可以计算出用户问句 A 和 CandQueSet 中的问句 B 之间的语义相似度 Sim(A,B)：

$$Sim(A,B) = \frac{1}{2}\left(\frac{1}{m}\sum_{i=1}^{m} a_i w(A_i) + \frac{1}{n}\sum_{j=1}^{n} b_j w(B_j)\right)$$

式中：$a_i = \max(S(A_i,B_1), S(A_i,B_2), \cdots\cdots, S(A_i,B_n))$；$b_i = \max(S(B_i,A_1), S(B_i,A_2), \cdots\cdots, S(B_i,A_m))$；$w(A_i)$、$w(B_i)$ 分别为词语 A_i 和 B_i 在系统中的权重。

3 系统运行情况与分析

本系统收集、整理了众多患者购药过程中常见问题和普遍性问题共 2160 个，由医药学专家解答后存入 OTC – FAQ 库中。系统试运行半年来，接受各类用户咨询累计 5236 次，我们对此作了统计分析如表 1 所示。其中 λ 为系统设定的相似度阈值，N_i ($1 \leqslant i \leqslant 5$) 为出现在第 i 位置上最相似问句的问句数，$P_i$ 表示截止到第 i 个位置出现最相似问句的概率。

表 1　各位置出现最相似问句的数量及位置—概率信息统计

λ	N_1	N_2	N_3	N_4	N_5	P_1%	P_2%	P_3%	P_4%	P_5%
0.90	4198	188	26	0	0	80.2	83.8	84.3	0	0
0.80	4430	222	64	11	0	84.6	88.8	90.1	90.3	0
0.70	4597	251	73	26	10	87.8	92.6	94.0	94.5	94.7

由表 1 可以看出：①对于 λ 的不同取值，在第 1 个位置上出现相同问句或相似问句的数量远大于第 2 及以后位置上出现的数量；②最相似问句出现在第 1 个位置的概率很高，达到了 80%，说明本文采用基于知网的语义相似度计算方法有较高的准确性；③无论 λ 取值如何，位置 3 以后的概率变化很小，说明最相似问句出现在第 3 以后位置的可能性很小；④当 $\lambda = 0.8$，$i = 3$ 的时候，最相似问句出现的概率和达到 90% 以上。由此可见，系统已较好地满足了实际应用的要求。

试运行结果表明，系统对于目的明确、表达清晰的问题有非常好的效果，如“我最近患有感冒，发热恶寒、鼻塞、打喷嚏、无汗、轻微头痛，请问应选购何种 OTC 药品？”，经过相似度计算 ($\lambda = 0.92$)，系统给出了相关的 OTC 药品供用户选择；而对于那些含糊其辞、无针对性的问题，如“请问专家感冒吃什么药最好？”，系统 OTC – FAQ 库没有这样笼统的问题（因为感冒有多种类型、多种症状），因此计算出的相似度很低 ($\lambda = 0.17$)，系统无法提供确切的答案；同样对于一题多问、表述繁琐的复杂问题，如“请问专家，高血压病人患感冒时能否服用新康泰克？如不能，请推荐其它一些药物。您所推荐的药物能否与降压药一起服用？有什么副作用和不良反应？使用该药时应注意哪些问题？谢谢！”，系统计算出的相似度也很低。对于这种问题，因为涉及的语义理解成分较多，需要作进一步的语义处理。

4 结束语

本文充分利用了汉语本身的特点、句子的组成词语和语义信息，设计并实现了一个基于 FAQ 的 OTC 问答系统，经实际试用，系统除了可以回答正确答案外，还可以提供一些与用户问题相关的答案，而且查找速度快，准确率在 90% 以上。

本文提出的基于词语的倒排索引查找算法不仅高效，而且平均时间受问题库规模的影响很小；候选问题集的使用使得问句相似度的计算只涉及很小的范围；基于知网的语义相似度计算提高了问句查找的准确性。这些技术和方法对于中文信息处理的其他领域，如信息检索、基于实例的机器翻译、远程教育智能答疑等领域均具有一定的借鉴作用。

参考文献

1 吴友政，赵军，段湘煜，等. 问答式检索技术及评测研究综述. 中文信息学报，2004，19(3)：1 – 13.

2 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. 计算机应用，2005，25(2)：449 – 452.

3 秦兵，刘挺，王洋，等. 基于常问问题集的中文问答系统研究. 哈尔滨工业大学学报，2003，35(10)：1179 – 1182

4 董振东，董强. 知网. http://www.keenage.com.

5 郭鹏举，关德生. 中国非处方药完全手册. 陕西科学技术出版社，2005.