第36卷 第14期 计 算 机 工 程 2010年7月
Vol.36 No.14 Computer Engineering July 2010
·软件技术与数据库· 文章编号：1000—3428(2010)14—0029—03 文献标识码：A 中图分类号：N945

基于概念图的中文FAQ问答系统

卜文娟，张 蕾
(西北大学信息科学与技术学院，西安 710127)

摘 要：提出一种利用概念图计算问句相似度的方法，并在此基础上实现基于概念图的中文FAQ问答系统，在该系统中采用概念图的形式表示用户问句及在FAQ库中找到的候选问句集中的问句，通过改进的概念图语义相似度计算问句相似度，在候选问句集中找到相似的问句并将答案返回给用户。该系统能够自动更新和维护FAQ库。实验结果表明，与基于关键词的句子相似度相比，基于语义的句子相似度提高了问题匹配的准确率。

关键词：概念图；相似度；常问问题集；问答系统

Chinese FAQ Question-answering System Based on Concept Graph

BU Wen-juan, ZHANG Lei
(College of Information Science & Technology, Northwest University, Xi'an 710127)

【Abstract】A method of computation sentence similarity based on concept graph is proposed and the Chinese Frequently Asked Question(FAQ) set system based on concept graph is implemented. The concept graph is used to express users query and candidate questions. The semantic similarities of sentences are computed between users query and candidate questions by semantic similarity based on concept graph. The answer in correspondence with the most similar query is returned to the user. This system can also automatically update and maintain FAQ. Experimental results show the new computing method gets better performance than the keywords-based approach.

【Key words】concept graph; similarity; Frequently Asked Question(FAQ) set; question-answering system

1 概述

FAQ(Frequently Asked Question)问答系统是一种在已有的“问题-答案”对集合中找到与用户提问相匹配的问句，并将其对应的答案返回给用户的问答式检索系统[1]。在该系统中，如果用户的提问与以往的记录相符，可直接将对应的答案提交给用户，免去了重新组织答案的过程，可以提高系统的效率。其中关键的问题是要计算用户问句与“问题-答案”对集合中问句的相似度，并把最佳结果返回给用户。

目前，国内外关于FAQ问答系统中问句相似度的计算方法的研究，主要有基于统计的方法和基于语义概念的方法。在国外，Robin D B等人利用Word-Net计算问句的语义相似度，并与基于TF-IDF的句子相似度的方法相融合，在FAQ集合中寻找相似度最大的5个问句；在国内，秦兵等人利用知网(How-Net)采用计算句子的语义相似度的方法来找出匹配的问句。

本文在以上研究的基础上提出一种基于概念图的计算问句相似度的方法，并在此基础上实现基于概念图的中文FAQ问答系统。

2 概念图

概念图[2]是一种描述复杂对象结构的知识表示工具，其思想来源于Pierce C S的存在图和菲尔墨的语义网络，是一种具有一阶谓词逻辑的完整表达能力的图形化表示，通常由一组分别称作概念和概念关系的节点之间以有向弧相连而构成。这种图的形式有利于处理汉语中复杂的语法现象，通过意义上的理解建立严格的概念图结构避免歧义的产生，从而减少信息处理中的消极工作量。

3 基于概念图的FAQ问答系统

FAQ自动问答系统的核心问题是如何快速地将用户所提出的问题与FAQ数据库中的问题比较，进而确定与其最相似的问题，如果有，则将对应的答案作为结果返回给用户。从数学的角度看，可以用2个映射表示[3]：$f_1:Q_1\rightarrow Q_2$，$f_2:Q_2\rightarrow A_2$。其中，Q_1表示用户提问的问题；Q_2表示FAQ库中的问句；A_2表示FAQ库中的问句答案。本文所提出的基于概念图的FAQ问答系统的流程如图1所示。

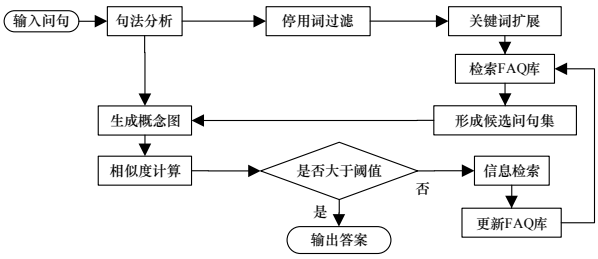

图1 基于概念图的FAQ问答系统流程

在本系统中，用户问句在进入基于概念图的FAQ问答系统之前，首先要进行句法分析，根据句法分析的结果一方面将问句表示成概念图的形式，另一方面根据停用词表对分析结果进行停用词过滤，并抽取问句中的关键词，利用知网对

基金项目：陕西省教育厅专项科研基金资助项目(HD0130)
作者简介：卜文娟(1984－)，女，硕士研究生，主研方向：人工智能，自然语言理解；张 蕾，教授、博士
收稿日期：2009-12-03 **E-mail：**buwenjuan_521@163.com

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

抽取的关键词进行同义词的扩展，根据扩展过的关键词利用信息检索系统对 FAQ 库进行检索，得到候选问句集，这主要是使后续的句子相似度的计算能在一个较小的范围内进行，然后将得到的候选问句分别表示成概念图的形式放入概念图库，并与用概念图表示的用户问句进行相似度的计算，如果计算出的相似度大于系统设置的阈值，则表示在 FAQ 库中有与用户问句相似的问句，并将问句的答案返回给用户，在本系统中设定将相似度大于阈值的前 5 个问句的答案都返回给用户，让用户挑选最合适的答案。而如果计算的相似度小于系统设定的阈值则表示在 FAQ 库中没有与用户问句相似的问句，则这时要通过网络进行信息检索找出答案，并将用户问句和相应的答案放入 FAQ 库中，对 FAQ 库进行更新。在整个过程中最重要的问题是概念图的形成和相似度的计算，这也是本系统中的难点所在。

3.1 概念图的生成

概念图的生成过程中主要的问题是概念结点和关系结点的形成以及它们之间弧的连接。本文结合本体知识库对概念图的生成做了研究，具体的算法思想为：

(1)利用哈工大 IR-Lab 的语法分析系统对用户问句进行句法分析，包括分词、词性标注等，再根据停用词表过滤句中的停用词(停用词指句中的客套词(请问、请问一下等)、助词(的、吗、啊等)等对句子意义关系不大但出现频率很高的词)，然后抽取出句中的名词、动词、形容词、代词等作为关键词并根据本体知识库对关键词进行同义词和相关词的扩展，然后根据概念-词汇表将扩展后的关键词映射为本体中的概念。

(2)同一个关键词扩展的同义词映射为本体中的同一个概念，将这个概念作为概念图中的一个概念结点，而同一个关键词扩展的相关词则映射为本体中不同的概念，但这些概念之间及与关键词对应的概念之间具有一定的关系，而这个关系在本体中已经标注，则将这些关系抽取出来作为关系结点，体现在概念图中。

(3)句中不同的关键词之间的关系，则依靠抽取句中的虚词、功能词等作为关系结点在概念图中的体现。这里的虚词指副词、连词、介词，这些词可以直接表示成概念图中的结点。比如介词，介词用来做动词、形容词的附加成分，表示时间、处所、方式、条件、对象等的虚词，如表示处所的介词可以表示成关系结点 LOC，而表示对象的介词则可表示成关系结点 OBJ。

功能词语是一种特殊意义的词语，一般不能充当句法成分，并且大都以表示语法意义为主，所以，可以直接映射为概念图中的关系结点，如功能词“也”，从概念层次上讲对应于概念项中的动态角色，在概念图中可以表示成关系结点 AGNT。概念图中的基本关系如表 1 所示。

表 1 概念图的基本关系

CG 关系	CG 名称	CG 关系	CG 名称
AGNT	动作	MANR	方式
OBJ	对象	INST	工具
MATR	材料	CHRC	特征
POSS	具有	HAS	作用
LOC	地点	DEST	目的
PART	状态	ISA	类事

(4)所有的概念结点按照本体中的 4 种基本关系，即部分与整体的关系(part-of)、继承关系(kind-of)、实例与概念间关系(instance-of)和属性关系(attribute-of)中的继承关系建立概念类型层次，概念结点之间插入相应的关系结点，并用弧线连接，最后利用十字链表的方式存储概念图。

3.2 改进的相似度的计算方法

本系统在对问句进行相似度的计算时，首先将用户问句和候选问句集中的问句分别表示成概念图的形式，然后对 2 个概念图进行不完全匹配[2](投影匹配和最大连接匹配)的运算，如果不完全匹配失败，则利用文献[4]提出的概念图相似度的计算公式对概念图进行相似度的计算，但是，此公式进行计算时，对于 2 个概念结点的关系不论相同与否，此算法都是通过分别计算 2 个结点的子图的相似度，取其中的最大者，然而对于 2 个概念结点的关系不同的情况，按照此算法就会降低运算的效率，本文针对以上问题对文献[4]中提出的概念图的相似度的算法做出了改进。

用户问句概念图和候选问题集中的问句概念图进行相似度计算时，对概念图入口结点两边的关系先判断再计算，如果两边的关系结点相同则对关系结点所连接的概念结点分别进行相似度的计算并取其最大值返回，如果两边的关系结点不完全相同，则只计算相同的关系结点下的概念结点的相似度(因为关系结点如果不同，则关系结点的相似度为零，关系结点下的概念结点的相似度不需要再计算)，通过减少计算次数提高系统的效率，具体的算法表述如下：

```
double SimilarGraph(Graph user,Graph candidate)
{ Sim=Sim + Sim(user.node, candidate.node); // 计算用户问句概
//念图和候选问题集中的问句概念图的入口概念相似度
 If(user.LeftRelation= = user.RightRelation) // 判断用户概念图
//中的概念结点两边的关系是否相同
  { if (candidate.LeftRelation= = candidate.RightRelation) //判断
//候选问题集中的问句概念图的概念结点两边的关系是否相同
   { if(user.Relation= = candidate.Relation) // 判断 2 个概念图
//的两边的关系是否相同
      Sim=Sim+max{SimilarGraph(user.child, candidate.child)};
//计算概念图子图的相似度，并取相似度的最大值
      Else
       Return Sim; // 只计算概念的入口结点的相似度
    }
   }
 Else //用户问句概念图的概念结点两边的关系不相同
{ if (user.LeftRelation= = candidate.LeftRelation)
   If(user.RightRelation= = candidate.RightRelation)
    {Sim=Sim+SimilarGraph(user.leftchild,
candidate.leftchild)+
      SimilarGraph(user.leftchild, candidate.leftchild) // 用
//户问句概念图左边的结点和右边的结点分别相同，则左子图与右子
//图分别匹配
    Else // 只进行左子图匹配
    }
 Else // 用户问句概念图左边的子图与候选问题集中的问句概
//念图右边子图匹配，用户问句概念图右边子图与候选问题集中的问
//句概念图左边子图匹配
}
}
```

3.3 FAQ 库的更新

在本系统中由于初始的 FAQ 库中的问句-答案对的数量是有限的，用户输入的问题在 FAQ 库中有时会找不到答案，这时就要采取其他的方法来寻求答案如网络信息检索、答案抽取等，并将所获得的答案及相应的问句添加到 FAQ 库中，但随着问题的不断加入，FAQ 库越来越大，从而导致问题检索和推理的效率降低，因此还需要把 FAQ 库中的旧问题从库中删除，只有这样能使 FAQ 库的内容更加丰富和完善，同时

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

保持最有利于用户的状态。本系统在对 FAQ 库设计时，对每个问句都设置了记录访问次数的标记，每次对问句及答案访问成功(指对于用户的问题在 FAQ 库中能找到相似的问句并能返回给用户答案)后，访问次数加 1，这样通过对访问次数的查询就能得到 FAQ 库中哪些是用户经常询问的问题，而哪些问题用户不常提问，而对于那些访问次数低于本系统设置的最少访问次数的问题，本系统自动将其删除，这样就能保证FAQ库中总是用户最为关注的问句-答案对，并能保证FAQ库中问句-答案对的数量不会太大，从而提高系统的效率。

4 系统评测

中文 FAQ 问答系统的设计与传统的信息检索系统有很大的区别，因此，系统性能的评价也相对复杂，传统的用于评测信息检索的召回率和准确率并不能准确地评价中文FAQ问答系统，因此，对于本系统的评测，本文采用文献[5]中提出的评测方法，在文献[5]中，作者修改了召回率的计算方法，具体的计算公式为

$$recall = \frac{c}{n} \tag{1}$$

其中，*recall* 表示系统的召回率；*n* 表示 FAQ 库中用户问句的所有正确答案的个数；*c* 表示返回给用户的正确答案的个数，本系统中要求返回给用户的答案必须是其对应的问句与用户输入的问句的相似度大于系统所设置的阈值，并且是相似度最大的前 5 个问句所对应的答案。

而对于正确率，在文献[5]中没有使用传统的准确率，而是提出了系统的不匹配率来进行代替，文中对系统的不匹配率的定义如下：

$$rejection = \frac{c}{n} \tag{2}$$

其中，*rejection* 表示系统的不匹配率；*n* 表示用户问句的个数(其正确的答案都不在 FAQ 库中)；*c* 表示系统判定 FAQ 库中没有正确答案的问句的个数。文中提出的不匹配率的方法能很好地评测系统特别是系统中 FAQ 库的性能，因此，在本系统中也采用这种方法来进行评测。

5 实验结果与分析

本系统所用的测试集是在参考 TREC 的问句和哈工大信息实验室问句的基础上，再加上在本实验室中收集的一些问句加工生成。本系统问句集的数量为 800 条，本文在实验中，随机地从这 800 条问句集中抽取了 300 条作为实验测试问句。其中 200 条给出了相应的答案，剩下的 100 条没有作答，这 100 个问句用来测试系统的不匹配率，另外还构造了 100 条与 FAQ 库中问句语义相似的问句，用来测试系统的召回率。本文根据以上系统评测的方法选取系统设置的相似度阈值为 0.65(经多次实验得出阈值为 0.65 时系统性能最高)进行实验，实验结果如表 2 所示。

表 2 实验结果 (%)

问句相似度计算方法	*recall*	*rejection*
基于 TF-IDF 方法	70	72
基于分解向量空间与语义的方法	73	78
基于概念图的方法	75	80

从以上实验结果中可以看出，本文提出的基于概念图的句子相似度的计算方法能使 FAQ 问答系统的召回率与不匹配率有所提高。

6 结束语

本文提出并实现一个基于概念图的中文 FAQ 问答系统，实验结果表明，该系统具有较好的召回率和准确率，但是由于系统在检索时计算量比较大造成系统检索速度较慢，在以后的研究中还需要对系统的检索速度做进一步提高。

参考文献

[1] Jijkoun V, Pijke M. Retrieving Answers from Frequently Asked Questions Pages on the Web[C]//Proc. of the 14th ACM Int'l Conf. on Information and Knowledge Management. [S. l.]: ACM Press, 2005.

[2] 张 蕾, 李学良. 概念结构及其应用[D]. 西安: 西北工业大学, 2001.

[3] 王继成, 潘金贵. Web 文本挖掘技术研究[J]. 计算机研究与发展, 2000, 37(5): 514-516.

[4] 朱海平, 俞 勇. 基于概念图匹配的语义搜索[D]. 上海: 上海交通大学, 2006.

[5] Burke R D, Hammond K J, Kulyukin V, et al. Question Answering from Frequently Asked Question Files: Experiences with the FAQ Finder System[D]. Chicago, IL, USA: University of Chicago, 1997.

编辑 陈 文

(上接第 25 页)

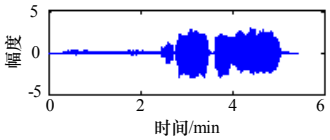

图 4 *N*1 分离后波形

6 结束语

本文采用 onset/offset 作为 CASA 的声音分离线索进行混合语音分离的研究，通过实验可以得出，onset/offset 线索无论对于音乐、语音还是噪声都可以进行分离处理；并且 onset/offset 对于清音和浊音同样适用，避免了用不同的声音线索处理清音和浊音，使得算法更为简单。由于对声音的起始和结束时刻检测可能会不准确，容易导致形成的片段出现同一声音元素丢失或者将不同的声音元素混合在一起的现象，这将是以后将要研究解决的问题。总体来说，onset/offset 算法对混合语音可以得到很好的分离效果，不但分离后的 *SNR* 平均提高了 5 dB，而且系统运行时间也得到了明显的改善。

参考文献

[1] Bregman A S. Auditory Scene Analysis[M]. Cambridge, MA, USA: MIT Press, 1990.

[2] Wang Deliang, Recazone G H. Cocktail Party Processing[C]//Proc. of IEEE World Congress on Computational Intelligence. Hong Kong, China: [s. n.], 2008.

[3] Patterson R D, Moore B C J. Auditory Fiters and Excitation Patterns as Representations of Frequency Resolution[M]. London, UK: Academic Press, 1986.

[4] Hu Guoning, Wang Deliang. Auditory Segmentation Based on Onset and Offset Analysis[J]. IEEE Transactions on Audio Speech and Language Processing, 2007, 15(2): 396-406.

[5] Hu Guoning, Wang Deliang. Monaural Speech Segregation Based on Pitch Tracking and Amplitude Modulation[J]. IEEE Transactions on Neural Networks, 2004, 15(5): 1135-1150.

编辑 索书志

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net