文章编号:1673-9469(2008)02-0101-03

一种基于概念匹配度模型的中文问答系统

李 静,宋振明
(西南交通大学 数学系,成都 610031)

摘要:答案抽取是问答系统的核心部分,本文根据汉语本身的特点,采用了本体论方法,借鉴了概念格的思想,提出了一种基于概念匹配度模型的中文问答系统。该系统通过计算与问题概念节点的匹配度,就可以抽取出查询问题的答案。该模型采用了定量计算,缩短了问题—答案匹配时间,并通过实例证明了该模型是有效的。

关键词:概念匹配度;概念格;问答系统;相似度
中图分类号: O153 **文献标识码**:A

Chinese question answer system based on the model of the conceptual matching degree

LI Jing, SONG Zhen-ming
(Department of Mathematics, Southwest Jiaotong University, Chengdu 610031, China)

Abstract: Answer extracting is central to question answer system. Chinese question answer system based on the model of the conceptual matching degree was put forward by using the ontological approach and using for reference from the concept lattice. This model can extract satisfied answer by calculating matching degree. The model adopts quantitative calculation and reduces the problem - - answer matching time.

Key words: conceptual matching degree; concept lattice; question answer system; similarity

近年来,问答系统深入学习机制向多种语言多领域发展。国外研究英语方面的QA系统已经有了很大进展,而我国由于汉语自身的语言特点,在实践汉语QA系统方面有自身的困难和特点。首先,中文词语没有空格,在操作之前需要进行切分;其次,汉语不像英语那样可以通过词性和时态的变换来表达意思,而是以词作为实体表现各种关系,是通过语义建立起相应的关联,由于语序的变化等造成中文句子结构的千变万化,给中文的机器理解带来更多的困难。因此,中文问答系统可根据汉语本身的特点,利用句子的组成词语和语义信息,计算用户问题与答案之间的相似度。

中文问答系统中,对于用户输入的问题,首先,我们在文献[1]常问问答集(Frequently Asked Questions,FAQs)中查找,如果能够找到相应的问题就可以直接将相应的答案返回给用户;如果没有或者答案不满足用户的需求,再通过搜索引擎从Web中搜索相关的文档,本文基于本体论的概念,将问题和返回的前n个文档通过相应的映射投影到本体论中,并构建概念格,最后利用概念匹配度模型抽取出用户所需求的答案。

1 概念匹配度模型

1.1 概念格

概念格是由德国的Wille.R在世纪年代初期提出的一种形式化概念分析方法,由外延和内涵组成。外延表示属于这个概念的所有事物的集合,而概念的内涵表示所有这些事物所共同具有的属性集合。概念格是一种二元关系,它所对应的图形象地揭示了概念之间的泛化和特化关系,实现了对数据的可视化。

收稿日期:2007-09-17
作者简介:李静(1984-),女,江西宜春人,硕士研究生,从概念格、优化与决策的研究。

定义1 一个形式背景是一个三元组(G,M,R)其中G是对象的集合,M是属性的集合,R是G和M的二元关系,$\forall d\in G, t\in M$,若d具有属性t,则记为:dRt或$(d,t)\in R$

定义2 形式背景(G,M,R)的一个形式概念(简称概念)是一个二元组(A,B),它满足$A'=B$,$B'=A$,其中$A\in G, B\in M$ $A\to A'=\{t\in M\mid \forall d\in G:(d,t)\in R\}$, $B\to B'=\{d\in G\mid \forall t\in M:(d,t)\in R\}$。$A$是概念$(A,B)$的外延,$B$是概念$(A,B)$的内涵。所有满足这样条件的全体概念的集合称为概念格。

定义3 在概念节点之间能够建立起一种偏序关系,对于给定的(A_1,B_1)和(A_2,B_2),假设$(A_1,B_1)\leq(A_2,B_2)\Leftrightarrow A_1\subseteq A_2(\Leftrightarrow B_2\subseteq B_1)$成立,则称$(A_1,B_1)$是$(A_2,B_2)$的子概念,$(A_2,B_2)$是$(A_1,B_1)$的超概念。

根据偏序关系可生成概念格的Hasse图。如果有概念$C_1>C_2$,并且不存在另一个元素C_3使得$C_1>C_3>C_2$,则从C_1到C_2就存在一条边,即C_1是C_2的直接超概念,C_2是C_1的直接子概念。

1.2 基于概念匹配度模型的答案抽取

1)构建概念格:以问题的主题作为对象集合,问题的焦点和问题的类型作为属性集合,这样可以确保答案的精确率。例如:中国什么时间举行奥运会?中国是问题的对象,奥运会是问题的焦点,时间是答案的类型,根据概念格的规定,相关的问题概念节点为{中国,奥运会\时间}。用户输入一个问题,使用返回的前n个文档作为形式背景$K=(G,M,R)$,这里$G=\{d_1,d_2,\cdots,d_n\}$表示获取的文档集作为对象集,$M=\{t_1,t_2,\cdots,t_m\}$表示描述文档的的关键词语集作为属性集,文档与词语之间的关系作为值(布尔值)记为a_{ij}。$a_{ij}=1$表示文档i具有关键词j;$a_{ij}=0$表示文档i不具有关键词j。

问答系统的答案抽取关键在于怎样把最满意的答案从相关文档中抽取出来,即怎样建立问题与答案之间的概念匹配。这里我们根据文献[2]中提到的根据相关的映射将概念分别映射到本体上,然后将两个概念共同的对象集合作为新概念的对象集(外延);再将两个概念共同拥有的属性作为新概念的属性集(内涵),从而对象集就变成用户所提的问题外加返回的前n个文档所对应的象。由上述形式背景我们可以画出概念格相应的Hasse图,本文按属性数相同分层来构图。

1.3 概念匹配度模型

定义4 以连接两个词语t_i,t_j节点间的边数为距离,用D表示,概念格的层数为H,则两个词语间最长距离为$2(H-1)$,两个词语间的语义相似度记为$S(t_i,t_j)=\dfrac{2(H-1)-D}{2(H-1)}\times 100\%$,式中距离$D$的计算方法是根据$t_i,t_j$在概念格中的首次出现的位置信息先求出其与最近共同超概念的距离,然后两者相加。遍历词语t_i所有出现位置,分别计算出其与词语t_j的距离,取其中最小值为词语t_i,t_j间的距离。一般地,两个词语相同,其相似度为1;如果两个词语距离为无穷大,其相似度为0;两个词语的相似度随其之间的距离增大而减小。

一个问题通常由多个词语组成,首先对问题进行切分,但是每个词语的分量(重要性)是不一样的,可用权重这一特性。在信息论中,信息熵是系统无序程度度量,其定义为

$$H(t_j)=-\sum_{i=1}^{n}y_{ij}\ln y_{ij} \tag{1}$$

式中$y_{ij}=\dfrac{a_{ij}}{\sum_{i=1}^{n}a_{ij}}$, $j=1,\cdots,m$。

一般某关键词的变异程度越大$H(t_j)$就越小,该关键词提供的信息量就越大,权系数也相应越大,反之亦然。因此可以根据各个关键词的变异程度,利用信息熵可计算出熵权,各关键词差异度$E_j=1-\dfrac{H(y_j)}{\ln n}$,最后我们易得第$j$个关键词的熵权

$$W_j=\frac{G_j}{\sum_{i=1}^{m}G_i} \tag{2}$$

定义5 概念匹配度(Conceptual Matching Degree,CMD):计算概念C_i中的每个关键词t_k与C_j中所有词语相似度最大值乘以相应的权重W_k所得积的和。概念C_i与C_j的匹配度为$CMD(C_i,C_j)$,记为CMD_{ij}公式如下

$$CMD_{ij}=\sum_{k=s}W_k\times\max(t_k,t_s) \tag{3}$$

其中$s=j,j+1,\cdots$

我们已知通过某一映射将问题与返回的文档都映射到本体后,我们再由它们共同构成的形式背景可以构造出概念格,在概念格上就可以找到用户提出的问题的相应的概念节点$C_i(d_i\cdots,t_i,t_{i+1},\cdots)$,用概念匹配度就可以计算出与$C_i$匹配度最高的另一概念节点$C_j(d_j\cdots,t_j\cdots)$。最后按照

与概念节点 C_i 的匹配相似度排序,从而可以找到最满足用户问题的答案。

抽取算法如下:(1)从问题处理部分得到的问题焦点、主题和答案类型。(2)在 FAQs 中查找,有则直接返回答案,否则转到下一步。(3)通过搜索引擎返回 n 个文档。(4)将用户输入的问题和前一步得到的文档构成的概念分别映射到本体后,构建概念格。(5)计算在概念格中出现的关键词的熵权。(6)计算其它关键词与问题中出现的关键词的词语相似度。(7)求出问题概念节点中的每个词语 t_k 与其他概念节点中所有词语相似度最大值相乘后再乘以相应的熵权 W_k 之和。(8)取上述和中最大值即可抽取所要的答案。

2 实例

用户输入一个问题,假设在 FAQs 中没有满足用户需求的答案,为方便起见由算法我们将映射到本体后的文档集合$\{d_1, d_2, d_3\}$记为$\{123\}$,而描述文档的关键词 $t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8$ 分别记为 $t_1t_2t_3t_4t_5t_6t_7t_8$,所得的形式背景如表 1。

表 1 形式背景

Tab.1 The formal context

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
d_1	1	1	0	0	0	0	1	0
d_2	1	1	1	0	0	0	1	1
d_3	1	1	0	1	0	1	0	0
d_4	1	1	1	1	0	1	0	0
d_5	1	0	1	1	1	0	0	0
d_6	1	0	1	1	0	1	0	0

我们令 C_9 为用户输入的问题通过相应的映射得到的象,下面求 $C_i(i = 1, \cdots, 8, 10, \cdots, 13)$与 C_9 概念匹配度,通过排序就容易得到与 C_9 最匹配的概念节点,即满足用户最需要的答案。

根据(2)式我们可以求得

$W_1 = 0$, $W_2 = W_3 = W_4 = 0.061\ 511\ 71$, $W_5 = W_8 = 0.271\ 821\ 623$, $W_6 = 0.105\ 154\ 96$, $W_7 = 0.166\ 666\ 67$。

根据(3)式我们可以求得

$CMD_{19} = 0$, $CMD_{29} = 0.061\ 511\ 71$, $CMD_{39} = 0.055\ 360\ 539$, $CMD_{49} = 0.061\ 511\ 71$, $CMD_{59} = 0.211\ 511\ 713$, $CMD_{69} = 0.116\ 872\ 249$, $CMD_{79} = 0.166\ 666\ 67$, $CMD_{89} = 0.116\ 872\ 249$, $CMD_{10,9} = 0.222\ 027\ 209$, $CMD_{11,9} = 0.334\ 329\ 547$, $CMD_{12,9} = 0.484\ 329\ 55$, $CMD_{13,9} = 0.283\ 538\ 919$。

由上述分析我们可知 $CMD_{12,9}$最大,即概念 C_{12}与 C_9 最匹配,从而我们可以得知文档 d_2 是最满足用户问题的答案。

3 结束语

借鉴概念格的思想,设计出满足用户要求的查询系统,该系统通过计算与问题概念节点的匹配度,就可以抽取出查询问题的答案,通过实例证明该模型是有效的。

参考文献:

[1] 唐 娟,杜亚军. 基于 Web 的问答系统答案抽取的研究[D]:成都:西华大学,2007.

[2] 夏 虹,李增智,陈彦萍. 基于概念格的语义 Web 服务匹配研究[J].北京邮电大学学报.2006, 29(9):185 - 188.

[3] 张 珂,沈夏炯,董 鑫. 基于概念格的语义相关度计算[J].郑州轻工业学院学报.2007, 22(6):178 - 181.

[4] 刘云峰,齐 欢. 潜在语义分析权重计算的改进[J].中文信息学报.2004, 19(6):64 - 69.

(责任编辑　刘存英)

一种基于概念匹配度模型的中文问答系统

作者： 李静， 宋振明， LI Jing， SONG Zhen-ming
作者单位： 西南交通大学,数学系,成都,610031
刊名： 河北工程大学学报（自然科学版） ISTIC
英文刊名： JOURNAL OF HEBEI UNIVERSITY OF ENGINEERING(NATURAL SCIENCE EDITION)
年，卷(期)： 2008,25(2)

参考文献(4条)

1. 唐娟;杜亚军 基于Web的问答系统答案抽取的研究 2007

2. 夏虹;李增智;陈彦萍 基于概念格的语义Web服务匹配研究 2006(09)

3. 张珂;沈夏炯;董鑫 基于概念格的语义相关度计算[期刊论文]-郑州轻工业学院学报 2007(06)

4. 刘云峰;齐欢 潜在语义分析权重计算的改进[期刊论文]-中文信息学报 2004(06)

本文读者也读过(10条)