《智能信息处理》课程作业

基于形式概念分析的问答系统答案抽取

田延凯

作业	分数[20]
得分	

2021 年 11 月 11 日

基于形式概念分析的问答系统答案抽取

田延凯

(大连海事大学计算机科学与技术辽宁省大连市中国 116026)

摘　要　为了在问答系统中尝试回答更加复杂的问题,有必要存在一种原理性的方法来动态产生不同的问答策略。提出在问答系统中使用形式概念分析(FCA)来抽取答案。在抽取过程中,首先在常问问题(FAQs）中寻找已经存在的问题,如果在 FAQs 中的答案不满足用户的需求,再通过搜索引擎获取相关的文档。接着利用这些文档中前N个文档构建概念格,进而采用概念匹配在格中抽取答案。对于不同的问题,使用不同的策略进行匹配和抽取答案。

关键词　形式概念分析；概念格；问答系统；答案抽取

Answer extraction of question answering system based on formal concept analysis

Tian Yankai

(Dalian Maritime University, Computer Science and Technology, Dalian, Liaoning Province, China 116026)

Abstract　In order to try to answer more complicated questions in a question answering system, it is necessary to have a principled method to dynamically generate different question answering strategies. It is proposed to use Formal Concept Analysis (FCA) to extract answers in the question answering system. In the extraction process, first look for existing questions in frequently asked questions (FAQs). If the answers in FAQs do not meet the needs of users, then obtain relevant documents through search engines. Then use the first N documents in these documents to construct a concept lattice, and then use concept matching to extract answers from the lattice. For different questions, different strategies are used to match and extract answers.

Key words　formal concept analysis; concept lattice; question answering system answer extraction

1 引言

自从文本检索会议（TREC)在 1999 年的TREC8会议上引入了对问答系统的评测后,越来越多的基于自然语言的问答系统产生。随着自然语言处理、信息获取技术,信息抽取等技术的发展,问答系统逐渐向问题多样化、问题复杂化和评估精确化方向发展。近年来,问答系统渗入学习机制,向多种语言、多领域发展。近期的研究集中在为问题获取答案,而不是获取与查询词相关的文档或者最佳的匹配章节[1]。现有的问答系统所采用的方法主要包括自然语言处理方法,冗余技术[2]、基于频率统计[3]和多策略方法[4]。我们的问答系统是一个基于语料和 Web 的多策略方法,试图回答更复杂

的问题。

2 基本概念

现实世界由各种各样的对象组成,每个对象有自己的一组属性或者特征,概念格结构是反映对象与属性之间联系以及概念泛化与例化关系的一种完备的概念层次结构[5]。下面给出概念格的形式化描述。

定义 1 一个形式背景是一个三元组 T=(O，D，R),其中 o 是对象集合,D 是特征集合，R 是O和 D 之间的二元关系，即 REU×D。g R m 表示 g∈ o 与 m∈ D 之间存在关系 R。形式背景可以用一个数据表来表示,它描述了对象及其特征之间的自然分组和关系的有序集[5]。

定义 2 在形式背景中,对于对象集 A∈ P(D)和特征集 B∈ P(O)可以定义下面的两个函数 f 和 g[5]：

$B = f(A)= \{y \in D \mid x \in A, x\ R\ y\}$

$A = g(B)= \{y \in O \mid y \in B, x\ R\ y\}$

简记为 A'= B,B'=A。通常称函数 f 和 g 为 D 的幂集 P(D)和 o 的幂集 P(O)之间的 Galois 连接。定义从形式背景中得到的每一个满足 B=f(A)和 A = g(B)二元组(A，B)为一个形式概念。其中 A 称为概念的外延，B 称为概念的内涵。

显然,概念的内涵是概念外延中所有对象的共同属性的集合,而概念的外延是概念内涵可以确定的最大的对象集合，一个概念是一个完备的二元组。

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 $C_1 = (A_1，B_1)$和 $C_2=(A_2，B_2)$,则 $C_1>C_2$, $B_1 \in B_2$,领先次序意味着 C_1 是 C_2 的父节点或称泛化。则称 C_1 是 C_2 的直接超概念，C_2 是 C_1, 的直接子概念,记为$(A_1，B_1) >(A_2，B_2)$[5]。

根据偏序关系可生成概念格的 Hasse 图。如果有 $C_1>C_2$,在 H as se 图中将存在一条边从 C_1 到 C_2，C_1 是 C_2 的直接超概念，C_2 是 C_1 的直接子概念,形式背景 T =(O,D,R)中,满足直接子概念-超概念关系的所有概念节点的集合是一个完全格,称为 Galois 概念格,简记为概念格。

3 问答系统体系结构

传统的问答系统采用了典型的管道方法,主要包括问题理解、信息获取、答案抽取三个部分。我们的问答系统是一个基于语料和 Web 的多策略方法，试图回答更复杂的问题。首先,我们对问题进行分析,将问题基本分为三类:仿真陈述型、列表型和定义型。对于不同的问题类型,利用形式概念分析来处理不同的答案抽取策略（如图 1)。

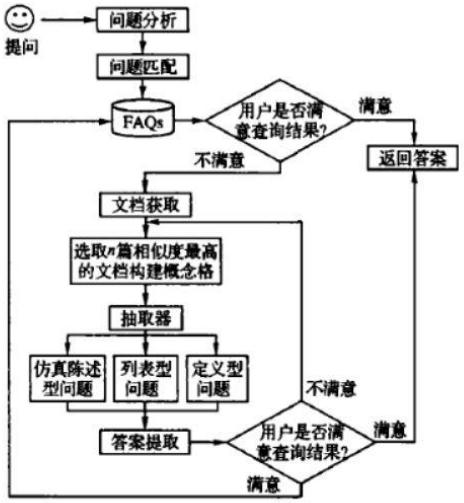

图 1 系统结构

首先,用户用自然语言输入一个问题,问题分析模块将首先处理这个问题,去掉停用词,将查询项归类。这个模块的输出为问题表达式,它将被用到 FAQs 中的问题匹配和从 W eb 中获取信息。我们使用问题的表达式和向量空间模型余弦值通过获取子模块得到相关的文档。下一步,我们使用获取的前 n 个最相关的文档形成形式背景,从而构建概念格。在执行子模块,我们分类问题,对不同的问题类型使用不同的识别方法来得到答案实体集,最后,我们在答案抽取模块中使用概念匹配来抽取答案。

4 答案抽取

一旦获取了初始的文档集,我们的问题回答系统将试图从这些文档中抽取答案。首先,我们选择相似度最高的文档构建概念格,这里我们使用格蒸馏代理和最小浏览域两个评估方法来评估我们的短语选择策略，自动化地选择名词短语作为文档的描述来建立基于 FCA 的信息获取框架。接下来，使用执行器识别问题的实体,将问题分为不同的问题类型,选择适合的答案抽取方案。最后，我们对于不同类型的问题,使用不同的抽取器来抽取答案。

4.1 构建概念格

1)信息组织模型

通过一个搜索引擎（如: Google)获取一个有序文档列表,组织这些搜索结果来组建一个概念格。格的产生式基于一个形式背景 K =(G,M,I),这里 G=fdoe1,doe2,…" , docn}表示获取的文档集，M-{ descl,dese2,…",desck}作为文档的描述子集，1 表示 G 和 M 的关系.

2）短语选择

对于名词短语的处理包括两步:短语抽取和短语选择。短语抽取短语抽取是为了发现所有可能的与文档相关且能描述文档的候选短语，它们与文档的域相关。这些短语将依次与术语表中的短语匹配。选择策略选择策略是为了选择能最好描述获取文档的短语。我们使用查询特殊平衡策略来选择至少包含一个查询项的短语作为短语组成，它们都有最高的文档频率值作为文档的描述〔6]。

4.2 问题概念节点

本文主要围绕下面几种类型的问题。问题被划分为不同的类型,从而使用不同的处理策略和答案样式。

1)仿真陈述型问题:

H ow far is it from Earth to M ? Is Bin Laden still a live

2)列表问题:

List names of cell phone manufact arsurers

3)定义型问题:

W ho is V lad the Imn palef?

接下来,我们找到问题的主题和问题的焦点，从而构建问题概念节点。问题的主题是指关于问题的对象或者事件。问题的焦点是指由问题找到的相关性质和实体。

例如: In wha t s ta tep is the G rand Canyon?

What is the population, of Bulg ria-?

What color, is a pomegranate?

这里 state,是指 state 是问题的焦点，Grand Canyon,是指 Gand Canyon 是问题的主题。另外,我们还要确定问题的相应的答案类型。

例如:PER SON: who… ?

LOCATION: where… ?

DATE: when…?

NUMBER: how m any…?

因此,问题格通过问题的主题、问题的焦点,问题的类型综合形成。我们用问题的主题作为对象集合，用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:

问题: How many people in China?

相关的问题概念节点为:{ China} → { people,number}。

4.3 使用概念匹配来抽取答案

与信息获取(Information Retrieval)系统相似,我们的抽取策略需要对每个问题答案对获取情况值计算相似度，来表示对于一个问题,答案的相似性和适当性。我们使用两个概念的部分匹配,将两个概念共同的对象集合作为新概念的对象集（外延),将两个概念共同拥有的属性作为新概念的属性集(内涵)。下面的例子说明了一个问题和文档之间的节点匹配情况。

5 问答系统相关研究

随着网络和信息技术的快速发展，同时人们想更快的获取信息的愿望促进了自动问答系统技术的发展。最近越来越多的公司和科研院所参与了自动问答技术的研究。比如:微软和 BM 等著名的跨国公司。麻省理工就开发出一个问答系统 STAR[9],从 1993 年开始发布在 Intenet 上,可以回答一些有关地理、历史、文化.科技.娱乐等方面的简单问题。在 2000 年,MT 开发的 START 是世界上最早基于 Web 的 QA 系统,返回段落或者句子。AnswerBus[10]是一个多语种的自动问答系统,它不仅可以回答英语的问题,还可以回答法语、德语、意大利语和葡萄牙语的问题,返回的是段落或者句子。AskJeeves 的返回结果与普通的搜索引擎很相似,都是网页。特点是允许用户用自然语言句子提问,检索系统会自动分析用户的提问,然后通过反问,即人机交互方式,准确地辨识用户的意图,这样用户就能够充分表达他的检

索需求,虽然比关键词检索方式有了明显的进步。但它并不是真正意义上的问答系统。ASKMSR,由微软研究院研制开发,为了快速查找相关文档的能力,建立在 Google 搜索引擎之上,返回简短词语或短语。ASKMSR 是基于答案顿率统计的问答系统。把答案用 Answer-Tiling 的方式组合,没有使用词性信息(但是,对中文来说,词性对答案很有用),简单使用出现频次和 N-GRAM 模型来匹配答案。

6 结束语

本文我们使用了 FCA 来处理问答系统的答案抽取。在抽取处理中,首先在 FAQs 中寻找问题,如果该问题相应的答案不能满足用户的需要，再通过搜索引擎从网上获取相关的文档,从而使用返回的最相关的前 n 个文档建立概念格。最后,利用概念匹配在格中抽取答案。对于不同的问题,使用了不同的抽取策略。通过实验,这种混合的策略和多种资源,使我们的答案抽取系统取得了初步的成效。

参考文献

[1]夏平,向学军,吉培荣,等. 基于自适应提升方案的激光探测系统消噪[J 微计算机信息，2016. 22(25): 274—275.

[2]MALLATS,Zeo-c rossing of a wave le t transfonn[EEETrans-ac tions on Information Theory,2017,37(4): 1019—1033

[3]UNSERM,ALDROUBIAB-Splnesignal processing part I-theory.

[4]JIEEETransactions on Signal Processing 1993 41(2): 821—833MALLAT S, HWANG WL. [5]魏海,沈兰荪．反对称双正交小波应用于多尺度边缘提取的研究[J 电子学报, 2012.30(3):313—316.

[6]刘曙光，刘明远，何钺．基于 Canny 准则的基数 B 样条小波边缘检 2011,17(5):418—423

[7]钟平. 小波边缘检测算子的构造与应用 Dl 长沙:国防科学技术大学，2013

[8]ECHHABIA,MARCUDAnoisy-channel approach to question swering[A]. ACL[Cl]2003

[9]BOUMAG,MURINOORDGV. LCAIW ok shop, Ed inbu rghScotland,20055]WLLER Restructuring lattice theory: an approach based on hier-archies of concepts[Al RALI edordered Sets[C].445—470.

[10]ZHENG ZP Answer Bus Question Answering System [A] Human Language Technology Confrence[CL] 2012 24—27.