●周宁　余肖生　刘玮　张芳芳

基于 XML 平台的知识元表示与抽取研究*

摘　要　互联网上检索信息,查准率、查全率不高的主要原因是信息组织的深度仅停留在文献层次。解决的根本方法是将信息组织深入到知识元层次。为此就要解决知识元表示与抽取正确与否的问题。图 4。表 2。参考文献 4。

关键词　XML 平台　知识元　知识元表示　知识元抽取
分类号　G254

ABSTRACT　The unsatisfactory hit rates and coverage of Internet searches are resulted from the low level of information organization at the level of documents. To solve the problem, we should consider the level of knowledge units and their correct representation and extraction. 4 figs. 2 tabs. 4 refs.

KEY WORDS　XML platform. Knowledge unit. Knowledge unit representation. Knowledge unit extraction.
CLASS NUMBER　G254

在互联网上检索信息采用关键词匹配的检索方法,将文本当做一个无序的字符集,导致了两个问题:一是一词多义现象,降低了查准率;二是多词一义,导致了很低的查全率。导致查准率、查全率都不能令人满意的主要原因是当前信息组织的深度仅仅停留在文献层次。有专家认为:解决的根本方法是要将信息组织深入到知识的最小的独立单位——知识元这个层次上[1]。而以知识元为组织单位的信息检索系统建立的难点在于知识元表示与抽取的正确与否。本文在对知识元概念剖析的基础上,讨论了基于 XML 平台的知识元表示和知识元抽取模型。

1　知识元的框架

1.1　知识元的概念

关于知识元的概念,目前还没有统一的定义。有学者认为,知识元是人的知识结构中的基本元素,知识元的构成为:知识元 = 信息元 + 经验 + 智慧 + 问题的解决[2]。也有人认为知识元是构造知识结构的基元[3]。笔者认为知识元是一个有确定意义的词组集合,是不可再分的知识单位,它们的结构可以抽象为 4 种数学结构:孤立点,线性表,树,图。如图 1 所示。我们一般可以采用 *ku*(*Name*, *Value*)来标记一个知识元,每个知识元(*ku*)有一个名(*Name*)和一个值(*Value*)。它的名表示它的内容,是这个数据的意义,它的值是被抽取的信息。例如:从一则房地产广告中,我们抽取的知识元是:*suburb*, *price*, *size* 和 *type*,且可标记为 *ku*(*suburb*, *ringwood*), *ku*(*price*, 105), *ku*(*size*, 1)和 *ku*(*type*, *flat*)。

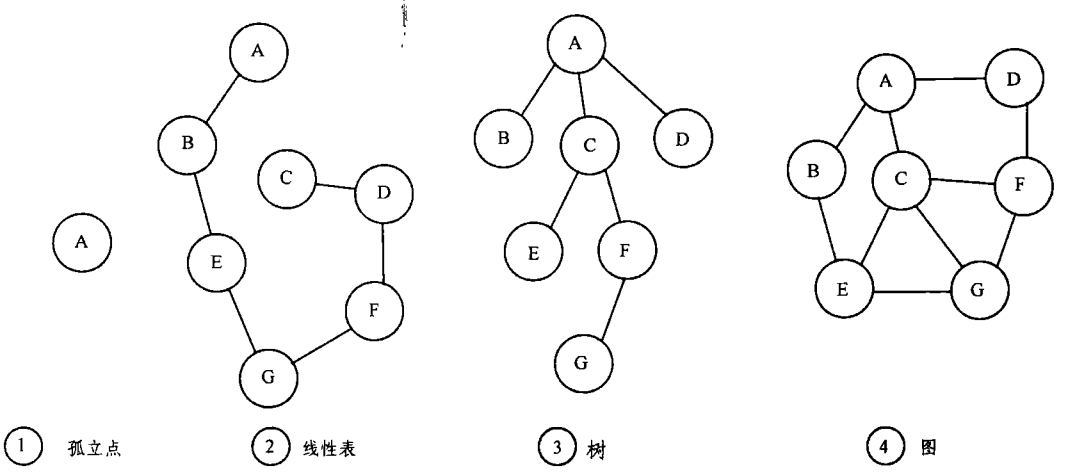

图 1　知识元的数学结构

* 本文系国家自然科学基金项目"我国数字化信息资源管理的可视化模型研究"(批准号 70473068)的研究成果。

1.2 XML平台上的知识元表示(框架表示法)

XML以一种开放的自我描述方式定义了数据结构,在描述数据内容的同时能突出对结构的描述,从而体现出数据之间的关系。加之与平台无关性等特点,使其较好地实现了异构数据源数据的集成。因而为不同的网络环境下,知识元的统一表示和抽取提供了可能。我们在此是以XML平台为基础讨论知识元的一些基本问题。

框架表示法是以框架理论为基础发展起来的一种结构化的知识表示方法。这里使用框架来表示知识元,有利于抽取领域知识的知识元。知识元框架的巴科斯范式(BNF, Backus Normal Form)格式如下:

<知识元框架> ::= <框架头> <槽部分>

<框架头> ::= 框架名 <框架名的值>

<槽部分> ::= <槽>[,<槽>]

<框架名的值> ::= <符号名> | <符号名>(<参数>,[<参数>])

<槽> ::= <槽名> <槽值> | <侧面部分>

<槽名> ::= <系统预定义槽名> | <用户自定义槽名>

<槽值> ::= <静态描述> | <过程> | <谓词> | <框架名的值> | <空>

<侧面部分> ::= <侧面>,[<侧面>]

<侧面> ::= <侧面名> <侧面值>

<侧面名> ::= <系统预定义侧面名> | <用户自定义侧面名>

<侧面值> ::= <静态描述> | <过程> | <谓词> | <框架名的值> | <空>

<静态描述> ::= <数值> | <字符串> | <布尔值> | <其他值>

<过程> ::= <动作> | <动作>,[<动作>]

<参数> ::= <符号名>

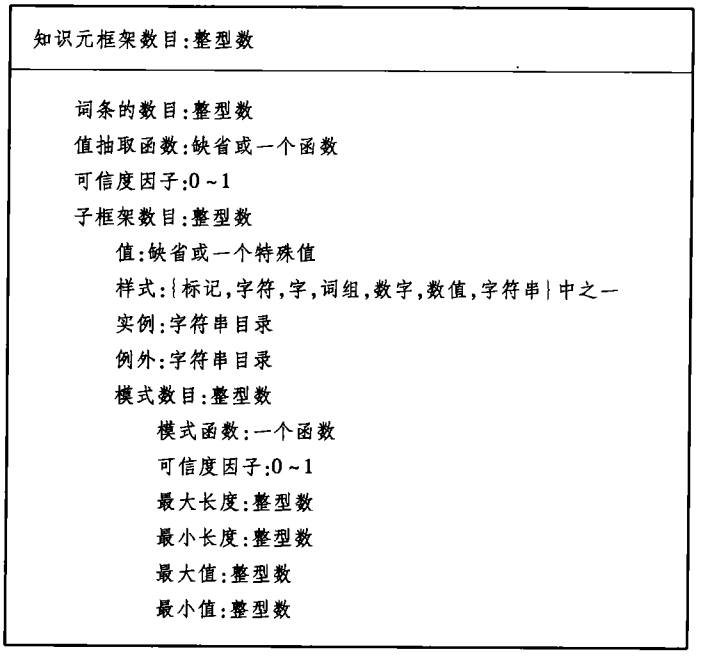

图2 知识元框架的语法

如图2所示,一些重要特征对知识元的认知和选择抽取的值组成槽是十分有用的。每个框架由多个槽和1个子框架集组成。一个框架的词条数定义了子框架的数目(每个词条表示成1个子框架)。值抽取函数定义了如何抽取知识元的值。可信度因子(*certainty factor*)提供了1个在多个框架之间进行选择的标准。每个子框架可以有多个槽。值槽指明了通过这个子框架返回的值。当槽被忽略时,缺省值是这个抽取的字符串。如果值被指明,那么这个子框架将返回这个特定的值。例如,假定一个子框架能抽取flats, flat, SC FLAT, Apartment, apart。如果我们希望词条的值是flat,那么我们指明这个值的槽为flat;如果这个值槽被忽略,这些抽取的字符串如flats, flat, SC FLAT, Apartment, apart中的任何一个都能作

为值返回。样式槽指明了词条的数据类型,包括标记、字符、字、词组、数字、数值、字符串等。在这里,标记是指XML标记,它能用来定义词条。实例槽包括正确的关键词目录,表明知识元是现成的。例外槽包括错误的关键词目录,表明知识元不是现成的。词条抽取模式槽能有许多模式,每个模式有两个子槽:模式函数确认模式值,例如,函数any_number(X)为抽取"123"和"5000",且可信度因子表明模式的优先权。最大长度和最小长度槽表明了字符串长度约束。最大值和最小值槽表明数值词条的范围约束。下面以价格知识元的表示作为一个实例,如图3所示。这个框架充分表达了从"$ 120PW"、"800PW"、"$ 1200/pw"、"$ 300p/w"、"420 PWeek"中抽取的价格知识元。

<table>
<tr><td colspan="2">框架:价格</td></tr>
<tr><td colspan="2">1号知识元框架</td></tr>
<tr><td colspan="2">词条数目:4
值抽取函数:single (v(2))
可信度因子:1. 0</td></tr>
<tr><td>1号子框架
样式:字符
实例:["$"," "]</td><td>2号子框架
样式:数字
模式数目:1
模式函数:any_number(X)
可信度因子:1. 0
最大值:9999
最小值:0</td></tr>
<tr><td>3号子框架
样式:字符
实例:[" ","/"]</td><td>4号子框架
样式:词组
实例:["pw","per week","pwe","p/we","p/w","perweek","pweek","p/week"]</td></tr>
</table>

图3　知识元框架的实例

2　基于XML平台的知识元抽取模型

为了改变人们检索时,获得的信息很丰富,而知识却很缺乏的状况,我们有必要将信息组织的单位延伸到知识元。先将文档分解为许多段落,对每一段中根据需要解析出相应的基本知识元,对每个基本知识元用3个约束(结构约束、长度约束、内容约束)来表示。基本知识元的抽取主要包含3个步骤:结构解析,长度解析,内容解析。先搜索到所需求的文档且根据结构约束来抽取片断,然后检查片断的长度和它的内容。分解算法如图4表示。

(1)如果知识元有结构约束,结构解析器开始工作。结构解析器查看文本且根据结构约束来抽取片断。

如果使用一个字符模板来指明结构约束,那么文本字符串一个字符接一个字符解析成仅含标记的结构字符串、字符标签("c"表示大写字母,"l"表示小写字母,"n"表示数字)、标点和特殊字符。

在结构字符串或文本字符串检查结构约束来设定开始和结束点。

使用开始和结束点从文本字符串抽取片断。结构字符串仅用于设置断点。输出文本通常从文本字符串中获得。

(2)如果知识元有长度约束,长度解析器来检查最大和最小长度。

(3)如果知识元有内容约束,这个文本被解析成一个仅包含单词(没有标记和标点)的单词线性表。这时内容解析器被激发,使用DCG(Definite Clauses Grammar)规则自顶向下内容解析器来解析生成知识元。

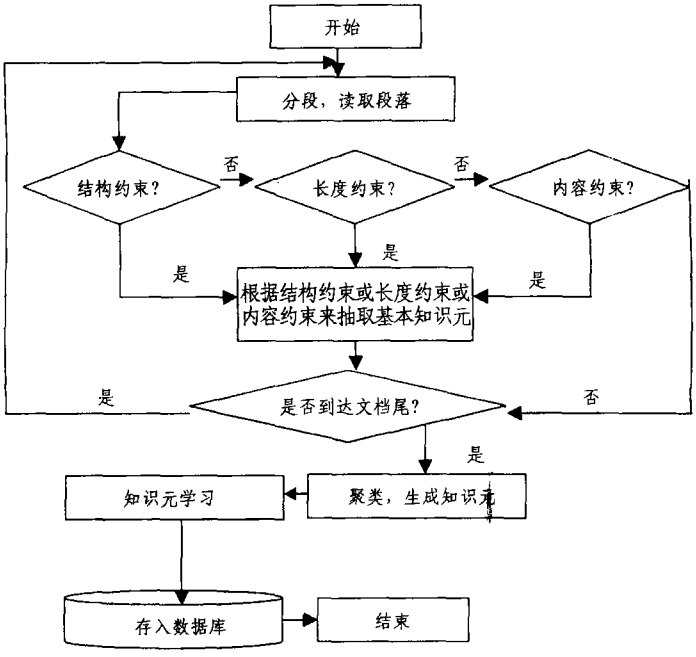

图4 基于XML平台的知识元抽取模型

(4)经过上述三步,看文档是否到达文档尾。如果没有则重新执行上述三步,如果已经到达文档尾,则进行第(5)步。

(5)基本知识元被抽取出来后,经过聚类和知识元学习再存入到相应的数据库中。

例如:在一则房地产广告中,先根据知识元框架将城市、价格、尺寸和类型等基本知识元抽取出来,然后基本知识元聚类,组成房地产广告这个知识元。通过相关的知识元学习,然后将它们存入相应的数据库中。

3 实例系统

为了提高检全率和检准率,只有用知识元来替代目前普遍采用的关键词作为检索单位,已经成为这一领域专家的普遍共识。关于这一方面的研究也越来越受到人们的重视,其中较为有代表性的研究成果包括:Nodose, MANIC, CASA 等。这里重点介绍 CASA。

CASA(online Classified Advertisement Search Agent)是由澳大利亚 Melbourne 大学计算机科学系人工智能实验室开发出来的。它采用传统信息检索评价标准来评价文本解析的性能。在一个网页的静态集合中,通过 CASA 与手工解析结果的比较来计算查全率和查准率。通过对这个文本解析器从两个网站上下载的6个网页中进行测试,这6个网页包含来自维多利亚"Houses to let"广告。前3个广告来自 Newsclassifieds,另3个来自 Fairfax Market。4个主要知识元是 suburb, size, price, type。CASA 文本解析结果如表1所示。由表1可知整体查准率为96%,整体查全率为78%。

表1 文本解析结果 (%)

知识元	查准率	查全率
Suburb	90	63
Price	99	88
Size	97	80
Type	96	78
Overall	96	78

注　查准率 = N_C/N_R,查全率 = N_C/N_K,其中 N_R 是返回的知识元的数目,N_K 是这些文档中知识元的总数,N_C 是返回的正确的知识元的数目。表1和表2的数据来源:文献4。

为了评价CASA的性能,我们对CASA与*Newsclassifieds*搜索引擎两者的搜索结果在返回知识元的总数、返回正确知识元的总数、返回错误知识元的总数三方面进行比较,比较结果如表2所示。

表2　*Newsclassifieds*搜索引擎与CASA之间的比较

	Newsclassifieds	CASA
返回知识元的总数	186	44
返回正确知识元的总数	51	40
返回错误知识元的总数	135	4

结果显示:CASA比*Newsclassifieds*搜索引擎有更高的查准率,CASA的高查准率表明在某些特定领域中以知识元匹配为基础的搜索策略是非常成功的。

参考文献

1　温有奎,徐端颐,潘龙法. 基于XML平台的知识元本体推理. 情报学报,2004,23(6)
2　孙成江,吴正荆. 知识服务战略:创建增值联盟. 情报科学,2002,20(10)
3　温有奎,徐国华. 知识元链接理论. 情报学报,2003,22(6)
4　Xiaoying (Sharon) Gao. A Knowledge-Based Approach for Searching Semi-Structured Documents. http://www.cnts.ua.ac.be/conll98/pdf/a21024ga.pdf (Accessed 2005-04-04)

周　宁　武汉大学信息管理学院教授。通信地址:武汉。邮编430072。
余肖生,刘玮,张芳芳　武汉大学信息管理学院博士生。通信地址同上。
(来稿时间:2005-10-09)