●章成志 侯汉清

面向概念挖掘的文本层次模型研究*

摘 要 针对当前Web文本挖掘工具的不足之处,提出了一种基于层次结构、面向概念挖掘的模型,即文本层次模型。该模型具有数据源适应性强、结构灵活、可操作性强、用途广泛优点,具有很强的实用性和一定的可扩展能力。图2。参考文献13。

关键词 文本层次模型 Web文本挖掘 概念挖掘 关键词 自动标引
分类号 G254

ABSTRACT The authors propose Text Mining Model, a model based on layer structure and oriented to concept mining to solve the present problems in Web text mining. It has the characteristics of data source adaptability, flexible structure, strong operability, wide application, practicality and extendibility. 2 figs. 13 refs.

KEY WORDS Text layer model. Web text mining. Concept mining. Keyword. Automatic indexing.
CLASS NUMBER G254

当前国内外对网络信息资源进行研究和系统开发的机构层出不穷,相关的研究课题有:网络信息整序、自动标引、自动分类、网络信息检索、Web概念挖掘等等。其中Web概念挖掘主要有两种策略:直接挖掘文档的内容,或在其他工具搜索的基础上改进。当前国内外开发了一些Web挖掘系统,但是存有以下几个不足之处:(1)目前大部分搜索引擎在信息提取时,不是基于全文,而是截取文档的前几百字和几十行,因此挖掘和检索的效果难以令人满意。(2)智能化的搜索工具所用的知识是由用户自己提供的,即使采用机器学习方法,也需要用户事先提供大量具有代表性的学习样本,方能通过自学习而获得有用的信息识别模式知识。这就使得这种知识获取方法的实用性大打折扣。(3)一些搜索引擎通过手工操作对Web上的文档进行处理和分类,如Yahoo,自动化程度不高,需要大量的人力物力资源,并大大影响了索引库的规模。(4)一些对HTML页面内容进行挖掘的系统,对数据源的格式要求高,即缺乏数据变化的灵活性。(5)现有的Web内容挖掘系统基本上是基于关键词索引原理而设计的,一般只考虑到关键词词频、位置等因素,很少考虑到关键词之间的同义关系和其他语义关系,离概念提取还有一定距离,因此很难实现概念检索。

针对以上不足,我们提出一个适用于WWW上各类文档自动标引和自动分类的概念挖掘模型,即文本层次模型[1]。通过此模型,对WWW的资源进行面向主题的、层次化的信息提取,包括文档格式识别、文本段落识别、文本关键句识别、文本关键词提取、文本主题概念提取、文本自动分类,以此来提高挖掘的效率,实现个性化Web概念挖掘。

1 文本层次研究综述

文本是指文件的某种本子(多就文字、措辞而言),也指某种文件。中外文论史上,都曾有人把文学产品的构成,看成是一个由表及里的多层次审美结构。中国古代的《周易》在探讨哲学思想的表达问题时,曾提出"言、象、意"3个要素。后来,三国时期的著名经学家王弼,详明地理清了三者之间的关系,认为"言、象、意"是一个由表及里的审美层次结构。这种构成观,西方早在古希腊时期也有萌芽。不过把它当做一种理论提出来的是德国著名哲学家黑格尔。黑格尔认为:一件艺术作品,我们首先见到的是它直接呈现给我们的东西,然后再追究它的意蕴和内容。黑格尔把"直接呈现给我们的东西"称为"外在形状",它的作用是"能指引到一种意蕴",而"意蕴"是一种内在的东西。黑格尔虽然已朦胧地意识到"形状"与"意蕴"的关系,可惜他对"象"并没有王弼那样清晰地认识。不过他提出的"意蕴"说,却为文本层次的探讨提供了一个重要概念。

19世纪30年代,印度著名图书馆学家阮冈纳赞创制了《冒号分类法》,提出了"分面组配"的理论和方法。他明确提出了3个结构层面——概念层面、词语层面和标记层面的假设和理论,并用它指导分面分类法的编制。其中,所谓"概念层面"是指据概念本身所考虑的层面,不能取决于代表它们的词汇和表示它们的号码。词语层面的特定功能是以同音异义词、同义词、单词、多字词来转换文献主题。标记层面指表示概念的数字或其他符号的层面,直接涉及分类号的构成,专业分类表中表示概念的标记符号的理论与方法[2]。

* 本文系国家社科基金项目"基于知识库的中文信息自动分类和自动标引"(02BTQ012)的研究成果之一。

对文本层次进行深入研究的还有波兰现象学派理论家R. Ingarden，他把文学作品的文本由表及里地分成声音层面、意义单元的组合层面等5个层面。

国内文学理论界对文本层次也有较深入研究，典型代表人物如童庆炳。他综合古今中外对文本层次的探讨，从总体上将文本分为3个大的层次：文学话语层、文学意象层和文学意蕴层[3]。

国外计算语言学界对文本层次的研究源于20世纪六七十年代，其中美国学者Hind Join于70年代就指出，“不同的文章类型有不同的组织形式，大多数文章的自然段有很好的组织层次”[4]。

国内计算语言学界对文本层次的研究起步较晚。为了适应因特网迅速的发展，信息处理研究者90年代以来不断探索信息提取、半结构化数据模式识别、信息过滤、数据挖掘等等。这些研究基本上都是基于文本而进行的。国内学者撰文指出，对文本的信息检索、提取、识别、过滤、挖掘等等，都应该以分析文本的层次结构为前提[5~11]。

尽管国内外对文本层次的研究已有多年，但研究的深度不够，处理的文本比较单一，且离真正的实用化还有一段距离。本项研究尝试提出一种基于层次结构、面向概念挖掘的模型，即文本层次模型。

2　文本层次模型的提出与实现

在计算机通信模型中ISO/OSI七层协议，互联网通信中TCP/IP协议，都是基于层的概念。ISO提出OSI（Open System Interconnection）模型是一个定义连接异构计算机的标准主体结构，以方便计算机遵循这种标准（协议）进行通信。该模型的基本构造技术是分层，各层中，每层的目的都是为上边的层提供某种服务，把这些层与提供服务的细节分开就形成了结构化模型[12]。

本项研究结合自然语言处理中的文本层次概念和计算机通信模型，提出文本层次模型（Text Layer Model，TLM）。

2.1　文本层次模型基本概念

文本（Text）：在这里主要是指WWW上出现的文档格式，包括结构化数据，如XML文档；半结构化数据，如HTML、PDF、PS、DOC等文档；非结构化文档，如TXT文档等。

层次（Layer）：上述不同格式的文本具有各自的物理结构或逻辑结构，在进行信息处理时不便于统一化处理，就文本在表达概念上的不同作用，将文本分成若干层面，即文本的层次。

（N）层：即某一特定层；（N+1）层：即相邻的高层；（N-1）层：即相邻的低层。

属性（Attribute）：即文本中的每个层次所具有的相关特性。

值（Value）：根据文本中的每个层次所具有的相关特性，通过一定的计算方法，得到属性的值。

2.2　文本层次模型各层的基本内容

根据结构中的不同层面在概念表达上的不同作用，将文本分为如下几个层次。

第1层：数据层，即最低层，是文本的逻辑载体。任何文档在计算机上都可以看成是数据流的形式。没有数据层，就无文本可言，更谈不上文本的主题概念了。该层的数据单位是字节，包含有相应的较原始的文本属性，如文本长度、文本的文件名、文本的格式等。

第2层：文字层，它是文本表达的基本形式。该层由文本中的文字、图形、图像、声音、视频等基本表达单元构成。这些单元也有各自的属性，如文本中文字具有的属性有：文字个数、文字编码、文字大小、文字颜色、文字样式等；图像具有的属性有：图像大小、颜色、位置等等；图形、声音、视频等其他多媒体信息也有相应的属性。

第3层：文本结构层（或称结构层）。文本结构层是由第2层各部分由语义联系所构成的主次关系，如标题、小标题、文摘、段落等，其中段落又分为首段、尾段、第2段、第3段等等。文本结构层的这些部分可以初步表达文章的主题概念，并且各部分表达能力大小各异，它们具有一定的主次结构。因此，这一层可以根据主题表达能力细分为标题、首段、尾段等等。段落的属性有：段落的数目、段落序号、段落句子数、段落长度等等。

第4层：句子层，来源于第3层各部分中的句子。本项研究为了研究方便，将标题、小标题等也作为句子来处理。该层的句子由于表达能力的不同，可以分为关键句（即论题句）、次要句、无关句等。在HTML文档中，广告栏、导航栏、联系信息栏等部分的句子一般为次要句和无关句。过滤这些句子，提取关键句对随后的关键词提取、概念提取有很大帮助。提取后的关键句可用于文摘的生成。句子包含的属性有：句子的长度、句子的分句数（该句由几个分句组成）、句子重要性（是属于关键句还是属于次要句或无关句）等。

第5层：短语层，来源于由句子层中的关键句，该层由短语构成。所谓短语，是有别于关键词的词或词组。对短语层进行分析处理可以得到一部分未登录词，发现同义词和准同义词、相关词，因此，此层的识别与处理对信息抽取、概念提取、Web挖掘、信息过滤很有意义。短语包含的属性有：短语出现的频率、短语出现的位置、短语的长度、构成短语的词个数、短语的同义短语、同义词、相关词、主题词、分类号等。

第6层：关键词层，即通常所说的来源于句子或短语的关键词所构成的层面。此层对文本的主题表达能力一般高于短语。关键词的属性有：关键词词频、关键词出现的位置、关键词的词长、关键词的同义词、相关词、概念

词、分类号等。目前的搜索引擎大多是基于关键词而进行标引的，对关键词层处理的好坏，直接影响到网络资源的查全率和查准率。

第 7 层：概念层，为最高层，它是表达文本主题最为重要的一个层面。一般说来，每个文本都应该有 1 个或者 1 个以上主题概念。用户关心的也就是文本的概念层，通过概念层可以实现概念标引和检索，可以提高网络资源的检全率和检准率。概念层是由概念款目构成，属性包括：概念词名称、概念词的专指度、概念词类别（即分类号）、概念词的相关词等等。

2.3　文本层次模型结构

上述的七层模型具体结构层次关系，如图 1 所示。

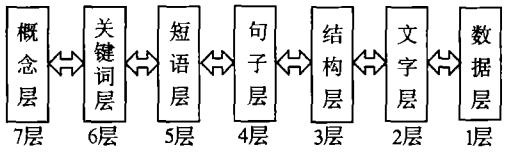

图 1　文本层次模型

2.4　文本层次模型的实现

实现 TLM 模型的每个层面的功能可以看成是对文本不同层面进行信息抽取或数据挖掘。我们称（N）层向（N+1）层提供的主题表达能力为（N）服务。在 TLM 模型中存在 6 种（N）服务，这 6 种类型的服务都属于信息抽取的范畴。服务的类型及实现方法描述如下。

（1）文字识别和抽取：即数据层向文字层提供的服务，文字层通过该项服务可将文本中的文字（图像）信息抽取出来。在文字层要对相应的属性（字符编码、文字大小、文字颜色等）赋值，必须有一个合理的文字识别算法。本项研究的对象主要为中文 HTML、DHTML，ASP、PHP、TXT、PDF 等文档都会涉及到字符编码的问题。其中根据 HTML 等半结构化文档的特点，不仅可以识别出文字的字体大小、字体颜色等信息，还可以识别出图像的有关属性及值。

（2）结构识别和抽取：即文字层向结构层提供的服务，结构层通过该项服务可以将文本的结构层次识别出来，从而抽取出相应的段落属性值。本项研究是通过一组启发式规则来对结构层的标题、小标题、首段、尾段等进行识别和相关属性值提取的。

（3）句子识别和抽取：即结构层向句子层提供的服务，本项研究中识别和抽取的主要是关键句，关键句的抽取也是根据一定的启发式规则结合关键词的词频来进行的。相关的调查表明：一般每段首句、尾句的表达能力较强，分句少的句子较分句多的句子表达能力强，包含较多关键词句子的表达能力比包含较少关键词句子的表达能力强。

（4）短语识别和抽取：即句子层向短语层提供的服务，本项研究通过该项服务识别出文本中的短语。通过共现频率、互信息、基于字的匹配、基于词的匹配等算法实现未登录词的识别[13]。另外借助于短语，也可以对 WWW 资源进行描述，从而扩大用户检索 WWW 资源的检索入口。通过改进的字面相似度算法可以将短语转换到关键词或概念款目。

（5）关键词抽取：即短语层向关键词层提供的服务。本项研究是依据关键词词典，采用改进的分词算法对文本结构或句子、短语进行关键词的抽取并实现同义词的识别。抽取过程中对关键词的词频、词长、出现位置、词的分布性等几个属性进行了相应的赋值操作。

（6）概念识别和挖掘：即关键词层向概念层提供的服务，本项研究是依据关键词与主题词（概念）对应转换的知识库来实现概念的识别，并通过概念及其标识（即分类号）完成文本的自动分类。对于不出现在关键词—主题词对应转换知识库的关键词，即未登录词，是通过局部统计与规则约束来进行挖掘的。

图 2 简单表示了 TLM 模型各层的操作。

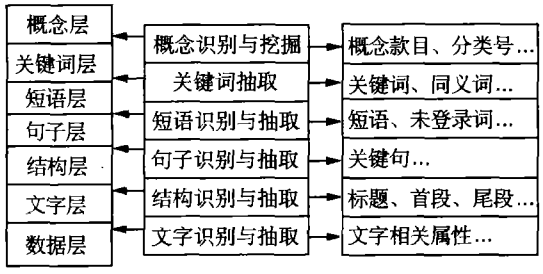

图 2　文本层次模型中的各层操作

2.5　文本层次模型噪声分析及处理方案

TLM 模型中的每个层或多或少地存在着和主题表达不相关的非属性物质，这里称之为噪声。

数据层的噪声主要来自原始文本本身的某些错误，使得格式难以读取。

文字层的噪声主要是指文本中的文字由于乱码不能完全被识别；图像的属性有时很难识别，也将它纳入到噪声的行列。

结构层的噪声比较明显，如 HTML 中的 Frame，通常对主题表达的作用不是很大而将其作为噪声考虑，TXT、HTML 中的一些联系信息，HTML 中的导航信息、广告信息对主题表达能力很弱，也应视为噪声。

句子层的噪声来自无关句。无关句是相对于关键句而言的，所以它具有一种相对性。

短语层的噪声来自无关短语，在此层中将停用词也视为噪声。

关键词层的噪声也是相对于关键词而言，具有一定的相对性，可以考虑将权重低于一定阈值的关键词作为噪声。

概念层的噪声主要来自概念转换时出现的一些错误。

针对以上各层所出现的噪声，我们可以考虑利用适当的过滤算法对噪声进行消除。例如在结构层，利用 Frame 过滤算法，可以过滤掉框架，从而方便后续的信息提取；在短语层，利用停用词过滤方法，可以提高短语识别的准确性，提高后续处理的效率等等。

3　文本层次模型的应用

我们利用 TLM 模型进行了基于概念语义网络的自动标引和自动分类的研究，即依据此模型构建了基于文本层次模型的 Web 概念挖掘系统，对因特网上的 Web 文本进行自动标引和自动分类。在每个层次上的信息提取都是面向主题的，即将每层中和文本主题相关的信息都提取出来进行分析，如文字层中，文字的大小、颜色等信息对提取主题概念有重要启发作用。在文本结构层中，我们将文本具体分为：网页题名、文章标题、文章第一段首句、尾句等部位，并根据对它们的主题表达能力的调查统计，设计了各部位的权重，以便进行有针对性的自动加权主题提取。进行文本的关键词提取时，引入同义词进行扩展，以更好地发现文本的主题。

实验结果表明，依据基于文本层次模型的自动标引和自动分类方法，从效率、兼容性及系统挖掘质量上都达到预期的效果，此模型具有很强的实用性和一定的可扩展能力。

4　结束语

本项研究提出的 TLM 模型是针对目前 Web 概念挖掘存在的诸多不足而提出的。在 Web 概念挖掘中，采用 TLM 的优点为：(1) 数据源适应性强。TLM 可以从不同层面对文本进行信息提取而不受文档格式、文本编排方式等因素的影响，能适应 WWW 上多变的、不同结构的数据源。(2) 结构灵活，可操作性强。TLM 模型是针对一般文本而言的，根据 WWW 中某一类文档的具体特点，我们可以将 TLM 模型进行变形或简化。(3) 用途广泛。

TLM 模型中的 6 种 (N) 服务，都属于信息提取和数据挖掘的范畴。这些服务都是当前自然语言处理的热点。若能很好地识别和抽取概念、关键词、短语、句子、结构、文字，将对信息检索有很大帮助，例如对识别和抽取出各层属性值的两个文本，可采用相似度算法实现基于层匹配比较，实现基于层的文本自动归类。所以实现这些服务，即信息提取，将具有很大的挑战性。

TLM 模型需要改进的地方主要是：在结构层引入语义网络的概念，在短语层、关键层进行词性标注，这样才能更好地揭示文章的主题概念。

Web 概念挖掘是一个多学科交叉领域，涉及到数据库技术、人工智能、机器学习、神经网络、统计学、模式识别、知识库系统、知识获取、信息提取、高性能计算和数据可视化等学科领域。本项研究只是提供一种可供参考的模型，虽然引入了结构层概念，但离真正的文本语义网络还有一定距离，另外层次的设计是否合理还有待进一步深入研究。

参考文献

1,13　章成志. 基于文本层次模型的 Web 概念挖掘研究——基于概念语义网络的自动标引和自动分类研究. 侯汉清指导. 南京农业大学硕士毕业论文, 2002
2　宋克强, 许墀基译著. 冒号分类法解说及类表. 北京: 书目文献出版社, 1986
3　童庆炳. 文学理论教程. 北京: 高等教育出版社, 1992
4　Hind Join. Organizational patterns in discourse, Syntax and Semantics: Discouse and Syntax. New York: Academic Press, 1979
5　张晓龙, 姚天顺. 基于文本句法的文本生成模型. 中文信息学报, 1995(1)
6　迟呈英, 麻志毅. 文本理解与汉语文本结构分析. 中文信息, 1997(1)
7　朱靖波, 姚天顺. 中文信息自动抽取. 东北大学学报, 1998(1)
8　林鸿飞, 战学刚等. 文本层次分析与文本浏览. 中文信息学报, 1999(4)
9　林鸿飞, 战学刚等. 基于概念的文本结构分析方法. 计算机研究与发展, 2000(3)
10　林鸿飞, 战学刚等. 文本结构分析与基于示例的文本过滤. 小型微型计算机系统, 2000, 21(4)
11　薛翠芳, 郭炳炎. 语文本结构的自动分析. 情报学报, 2000(4)
12　胡道元. 网络设计师教程. 北京: 清华大学出版社, 2001

章成志　南京农业大学信息管理系硕士生。通信地址：南京。邮编 210095。

侯汉清　南京农业大学信息管理系教授，博士生导师。通信地址同上。

(来稿时间：2004－04－05)