基于多种规则的课程元数据自动抽取*)

杨 宇[1] 张 铭[1] 周宝曜[2]

(北京大学信息科学技术学院计算机科学与技术系 北京 100871)[1] (惠普中国实验室 北京 100871)[2]

摘 要 在线课程组织和管理系统就是为了使学习更加便利而提供的一个教育资源的集成平台。作为系统中重要环节的元数据抽取模块,需要对半结构化网页能够达到较好的抽取精确性,并具有处理结构松散文档的能力。

本文设计并实现了一种按照指定规则自动抽取的元数据方法。该方法能够按照多优先级规则匹配网页元数据,并按照两步抽取的方法进行精细化处理。针对不同的问题域使用不同规则抽取,不需对程序进行特定修改。实验证明,这种方法能够很好地处理半结构化网页,F测度达到85%以上,具有较好的实用价值。

关键词 元数据抽取,正则表达式,信息精化

A Rule-based Metadata Extractor for Learning Materials

YANG Yu[1] ZHANG Ming[1] ZHOU Bao-Yao[2]

(School of Electronics Engineering and Computer Science, Peking University, Beijing 100871)[1] (HP Labs China, Beijing 100871)[2]

Abstract Integrating all kinds of learning material is becoming more and more significant for the teachers and students to take advantage of the online E-learning courses. As the key part of the whole Online Course Organization System, Metadata Extraction function needs to be accurate enough when dealing with semi-structured documents, even those incompact ones. We design and implement a Metadata Extractor to compare between several rules ordered by priority, and there is another step of information refinement to help improving the final accuracy. When domain changes, users just need to input specific rules, without considering the program. The experiment shows that our new method can perform very well with those semi-structured documents, with F measure higher than 85%, which indicates that this method is quite feasible in reality.

Keywords Metadata extraction, Regular expression, Information refinement

1 网页元数据抽取简介

元数据是描述数据的数据(data that describes data),用于描述数据的特征和属性,也是描述和组织 Internet 信息资源、发现 Internet 信息资源的工具。

元数据抽取在整个信息组织与检索中处于数据准备的基础地位。抽取过程的数据来源首先经过必要的预处理,剔除在格式、内容、语言等方面存在问题或严重缺失的文档,产生格式相对规整的文本文档;其次经过元数据抽取模块的处理,生成符合规范定义的文档元数据,并将结果存储在与具体系统相关的数据库中。文档元数据也可依据具体系统的差异,采用不同的规范格式进行组织,以方便数据的共享与信息交流。而随着网络的进一步发展,网页元数据已经成为存储有用信息的最主要方法之一,因此在网页元数据抽取方面的科研工作也越来越广泛,越来越受到重视。

2 研究背景及应用范围

在科学日益发达的今天,大学教育成为了时代的焦点,如何将大学教育合理化,如何将大学教育普及化,成为了整个社会关注的话题。本项目"在线课程组织和管理项目"(http://fusion.grids.cn:8080/ocos/, http://162.105.146.245:8080/ocos/index_en.jsp)的研究目标在于将尽可能多的大学教学资源整合在一起,特别是将那些可以公开的教学课件和视频资料整合起来,通过免费发布这些资源来促进教育发展。教师可以借鉴使用这些资源,服务于自己的教学;自学者可以获得个人知识发展。

到目前为止已经收录了来自十多个国内外知名高校的教育元数据,最主要的几个来源包括:中国教育科研网格"大学课程在线",其中收录了很多国内知名大学的教学课件;麻省理工学院的 OpenCourseWare(http://www.core.org.cn/OcwWeb),其中涵盖了 MIT 所有的本科生和研究生阶段的课程内容;美国伊利诺斯大学(UIUC)的 3000 余门课程资源;以及威斯康星大学麦迪逊分校(WISC)的千余门课程资源。

为检验程序的抽取效果,本文选取了 MIT 的 OPEN COURSE WARE(以下简称 OCW)项目的开放课程资料,伊利诺斯大学课程资源,以及威斯康星大学麦迪逊分校课程资源进行评测。

3 国内外相关研究

随着 Internet 的快速发展,对 Web 元数据抽取的内在需求的增长,在元数据抽取领域,科研工作者已经进行了大量的理论研究,也提出了很多元数据抽取的技术和方法,并根据这些理论开发出了不少可用的工具。

文[1]中根据每种工具采用的技术将 Web 元数据抽取的方法大致分为六类。其中基于语言的元数据抽取方法和基于 NLP 的元数据抽取方法是现阶段研究得比较多的两种方法。基于语言的元数据抽取采用特定的开发语言帮助用户制定 Wrapper。一般多采用通用 Perl 和 Java 语言。采用这类方法

*)本文研究得到国家自然科学基金"网络计算环境综合试验平台"(编号 90412010),惠普大学合作基金"在线课程的组织与管理"项目,国家自然科学基金(编号 60573166)以及广东省网络重点实验室基金的支持。杨 宇 硕士研究生,主要研究领域为典型应用领域的元数据抽取与管理;张 铭 教授,主要研究领域为网格计算、语义网和数字图书馆。

的典型工具包括：Minerva[2]，TSIMMIS[3]，Web-OQL[4]，Jedi[5]等。基于 NLP 的元数据抽取方法是将 NLP(Natural Language Processing)技术应用于从自然语言描述的文档中学习抽取规则。这类方法主要采用短语句法及语义分析技术，包括句法成分的识别与标记，关键词抽取，检索特征的抽取、索引等。基于 NLP 的元数据抽取方法比较适用于从结构松散甚至是纯文本中抽取有用信息。采用这类方法的典型工具包括：RAPIER[6]和 SRV[7]等。

4 基于规则的全自动元数据提取模块

在在线课程组织和管理系统中，主要针对教育类网页进行分析，即一些大学或教育机构的公开网站，这类网站大多为半结构化的网页，并且广告类信息较少，可以找到较为统一的网页规则，因此选择基于规则制定的提取方法。同时，为避免程序的局限性，本文中的提取模块可以根据任意制定的规则全自动地进行分析，当规则改变时，程序不需重新编写仍可正确提取出所需资源。

整个提取模块主要分为两个阶段：粗糙网页预处理阶段和信息提取阶段。

4.1 粗糙网页预处理

通常情况下，抽取算法比较适应于处理 xml 文档和格式良好的 html 文档。这里所说的“格式良好”，指的就是要在网页编写时，严格要求网页标签前后一一对应，有开始的标记就要有终结的标记。由于 html 网页并没有严格要求代码中的网页标签是前后对应的，因此，流行于网络中的各式网页的代码格式可能相差甚远。爬虫程序抓取到的都是一些毫无规范性可言的粗糙网页。

在提取工作正式开始之前，需要对粗糙网页进行一定的预处理。将粗糙网页中缺少的标签补全，将所有网页归一为比较规范的格式，方便后期处理，并且在这个阶段，网页中的所有图片将会被自动过滤。这里使用了开源工具包 NekoHTML 进行处理，这个工具是 J. Andrew Clark 用 Java 写的系列工具(Java APIs)中的一个。它是一个简单的 HTML 扫描器和标签补偿器(tag balancer)，使得程序能解析 HTML 文档并用标准的 XML 接口来访问其中的信息。经过以上处理的网页就能够达到基本的要求，转化为符合基本 XML 文档规范的网页了。

4.2 信息提取阶段

在第一阶段的网页预处理之后，就进入了信息提取阶段。为了提升提取准确度，将提取阶段分为三个步骤：信息预定位，基于多种规则制定的信息提取，以及信息精化。

(1)信息预定位

信息预定位是指对整个网页文档中最有价值的部分，也就是网页的核心内容进行初步定位。在待处理网页中往往既包含核心内容区域，也包含结构化内容区域。核心内容通常包括所要提取的信息，例如课程名称、课程编号、课程特点等；而结构化信息区域包括导航条、搜索区导航等。在处理过程中，这些结构化内容是没有意义的，但是却是所有页面都会包含的内容。因此，在第一步提取过程中，要将有价值的内容区域先整体提取出来，但并不对核心区域内的标签作具体区分和剔除。在这里，使用的是 Apache 的另一项开源工具：xalan-java，这是一套 XSLT 处理器。其中，XPath 是 XSLT 用于对 XML 文档各部分进行定位的语言，它可描述如何识别、选择、匹配 XML 文档中的各个构成元件，包括元素、属性、文本内容等。

预定位提取出的内容将交给基于多种规则制定的信息提取步骤来处理。对于每次处理，只要指定相应的处理规则，程序就能够根据要求进行提取。这里的规则描述形式，采用 java. regex 提供的正则表达式工具包进行描述及处理，能够方便地根据给定正则表达式匹配目标字符串中的指定内容。

(2)基于多种规则制定的信息提取

在实际提取中，不可能只使用一套 XPath 或正则表达式就能够处理所有的网页提取，为了使程序能够更加灵活，本文提供了一个选项，可以针对同一类网页，设定不同预定位 XPath 路径，以及一系列不同规则，并设定每种 XPath 路径和规则的优先级，进而更加精确地进行提取。程序会按优先级辨别出该网页符合那种预定位路径，以便提取出核心内容区域，而后再根据不同优先级匹配各个规则，直到提取出所要内容。

所有适合于特定网页类型的一系列不同优先级 XPath 路径和规则，都需要在 xml 文件格式的配置文件中写明，经过文件解析形成 DOM Tree 的格式进行处理。配置文件的书写格式为：

```
〈domain name="www. FirstWebSite. com"〉
    〈urlPattern regex="~/title/tt[0-9]{7}$" rank="1"〉
      〈entity name="title"〉
        〈location xpath="//BODY[1]/DIV[2]" rank="1"/〉
      〈/entity〉
    〈/urlpattern〉
    〈urlpattern regex="" type="ONE_MOVIE" rank="1"〉
      〈entity name="title"〉
        〈location xpath="//BODY[1]/DIV[2]" rank="1"〉
    〈/entity〉
    〈/urlPattern〉
    〈urlPattern regex="~/ti/t[0-9]{7}" rank="1"〉
      〈entity name="title"〉
        〈location xpath="//BODY[1]/DIV[1]" rank="1"/〉
      〈/entity〉
    〈/urlPattern〉
    〈urlPattern regex="~/nowplaying/$" rank="2"/〉
    〈urlPattern regex=". *" rank="3"/〉
  〈/domain〉
〈domain name="www. SecondWebSite. com"〉
```

(3)信息精化

经过预定位，提取结果中含有很多不相关的内容。众所周知，对于较为通用的元数据提取程序，是不可能做到百分之百的准确提取、准确分类的。为保证精确性，需要尽可能地将其中不相关的词语都剔除掉。在这一步骤中使用到基于机器学习的方法。经过观察发现，一般情况下那些多余的信息在内容和形式上都是经常重复的，因此，本模块中的训练数据将着重用来辨别哪些是多余信息。

首先确定哪些数据域需要信息精化步骤，因为有些数据域能够根据规则精确提取，例如：课程 id、课程名称、课程简介等，这些数据域的现实规则比较固定，因此提取后不会出现不相关词语混杂的情况，直接将第二步的结果作为最终提取结果输出即可。但是对于另外一些数据类，如：参考读物作者、参考读物题名、参考读物出处等，如果不进行信息精化，就会出现大量的冗余或杂乱信息，而机器学习的结果自然会将这些数据域指定为需要精化的类型。

在学习过程中，采用二元组〈word，feature〉的形式表示遇到的多余词语。其中，word 即多余词语本身的记录，而 feature 是依据规定好的分类方法，对词语的类型判断和标记。在本模块中，对遇到的词语划分为如下 4 种类型：(1) 所有字母均为小写，lowcase 型；(2) 包含数字和字母，或者只包含数字，in_number 型；(3) 该数据域的第一个单词，或一句话的第一个单词，并且该单词的第一个字母为大写，first_word 型；(4) 单词的第一个字母为大写，并且其前面一个单词或后面一个单词也为大写字母开头的单词，con_capital 型。

在学习过程中，如果某个单词或某种单词类型在一个特定位置作为杂质出现的概率大于一个特定域值，则将单词或

单词类型以及出现位置记录下来。在提取过程中，如果在固定匹配位置出现了杂质单词，即匹配〈word，feature〉中的第一项相同，则直接将其过滤掉。若出现了与杂质单词不相同，但与杂质类型相同的单词，则也将其判断为杂质并过滤掉。

因为本文针对的都是半结构化的教育类网站和资源，网页内容比较规范，很多时候，冗余信息是相似的或者相同的，因此，将针对 word 本身的判断作为首要判断标准，如果发现多余信息与已有记录相同或相似，则可以简单地做出判断，而不需要额外的工作。

5 性能评测

衡量一个元数据提取系统好坏的指标主要有以下三个：

查全率(Recall) $R=\frac{\text{正确提取结果数}}{\text{所有可能提取结果数}}$

查准率(Precision) $P=\frac{\text{正确提取结果数}}{\text{所有提取结果数}}$

F 度量(F-measure) $F=\frac{(\beta^2+1)PR}{\beta^2P+R}$

F 度量让用户在查全率和查准率上求得平衡。而我们认为元数据提取的查准率比查全率要重要一些，因为准确的数据是服务质量的保证，而数据的完备应该建立在数据准确的基础之上。因此，选择查准率和查全率的重要程度值 $\beta=0.5$，代表 P 的重要程度是 R 的 2 倍。

表 1

	OCW	UIUC	WISC
课程 id′	0.995	0.998	0.996
课程名称	0.995	0.998	0.990
课程描述	0.890	0.936	0.997
参考读物信息	0.821	—	0.897
教师信息	0.857	0.890	—
学时信息	0.902	—	0.953
教学大纲	0.798	—	—

测试数据包括 MIT 的 OCW 网站中 1900 门课程信息；美国伊利诺斯大学(UIUC)的 3000 余门课程资源；以及威斯康星大学麦迪逊分校(WISC)的 1000 门课程资源。表 1 中列出了不同元数据信息域在三所大学课程网页中提取效果的 F 度量值。

由于不是所有本项目中需要的元数据域都能在各个大学课程网页中涉及，因此上表中有些数值无法统计，例如 UIUC 的课程网页中没有包含学时信息和教学大纲等信息，而 WISC 网页中没有包含教师信息和教学大纲。

基于以上图表可以看出，对于一些格式比较固定的信息域，如课程 id，课程名称，课程描述等，本项目的提取效果能够保证在 90%以上，为实际应用提供了非常好的元数据资源。对于一些变化较大的信息域，如参考读物信息，教师信息，教学大纲等，在描述格式较为固定的课程网站中(如 UIUC 和 WISC)，本项目仍然能够达到 90%左右的准确效果，而在格式不是非常固定的情况下，提取效果也能够达到 80%～85%的准确程度，同样能够满足实际应用的需要，完成绝大多数网站元数据的整合任务。

总结与展望 本文在单一规则匹配提取方法的基础上，设计实现了一种按照优先级处理多种规则的元数据提取方法，并且对抽取的结果进行信息精化。本项目的创新之处在于，能够全自动地按照不同规则对不同网页进行元数据提取，而不需要用户了解并修改程序内部。实际应用的结果表明，该方法能够很好地处理半结构化网页。

现有系统仍然有需要进一步完善的地方，例如如何实现自动发现规则，以适应海量环境多种异构数据源的快速集成处理。

参 考 文 献

1 刘世杰，唐世渭，杨冬青，王腾蛟，李立宇. 基于 XML 技术的 Web 信息提取和集成. 见：第二十届全国数据库学术会议，2003
2 Crescenzi V, Mecca G. Grammars have Exceptions. Information Systems 1998,23 (8): 539～565
3 Garcia-Molina H, Papakonstantinou Y, Quass D, et al. The TSIMMIS Approach to Mediation: Data Models and Languages (extended abstract). In NGITS, 1995
4 Arocena G, Mendelzon A. WebOQL: Restructuring Documents, Databases, and Webs. In: Proc. ICDE '98, Feb. 1998
5 Huck G, Fankhauser P, Aberer K, Neuhold E J. Jedi: Exchanging and Synthesizing Information from the Web. Coopis,1998
6 Califf M E, Mooney R J. Relational Learning of Pattern-Match Rules for Information Extraction. In: Proceedings of the Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence, Orlando, Florida, 1999. 328～334
7 Freitag D. Machine Learning for Information Extraction in Informal Domains. Machine Learning, 2000,39(2-3):169～202

(上接第 86 页)

表 2 KF 代表基于八段骨骼特征和关键帧提取的检索算法，SF&DT 是基于时空特征和检索树学习的检索算法

运动	Recall		Precision	
序列	KF	SF&DT	KF	SF&DT
Walk	0.85	0.89	0.89	0.91
Run	0.70	0.85	0.83	0.93
Punch	0.40	0.75	0.59	0.81
kick	0.50	0.75	0.66	0.84

总结和展望 为了有效地对通过运动捕获设备得到的大规模的三维运动数据库进行分析和处理，我们针对运动捕获数据中的人体运动的特征进行了研究，提出了一种在三维空间考虑运动特征的方法，并在关节点的时空特征相对独立的特性前提下生成有关各关节点重要性的决策树，然后在决策树的指导下完成检索过程。实验证明，本文提出的空间特征高效而且有效地反映了运动的特性，并且通过决策树学习的方法也使检索过程简单和快速。

参 考 文 献

1 Zhuang Yueting, Rui Yong, Huang T S. Adaptive Key Frame Extraction Using Unsupervised Clustering [A]. In: IEEE ICIP'98, Chicago, USA, Oct. 1998
2 Wolf W. Key frame selection by motion analysis [J]. Acoustics, Speech, and Signal Processing, 1996,2: 1228～1231
3 Lim I S, Thalmann D. Key-Posture Extraction Out of Human Motion Data by Curve Simplification [C]. In: Proc. EMBC2001, 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2001. 1167～1169
4 Lee Jehee, Chai Jinxiang, Reitsma P S A, et al. Interactive Control of Avatars Animated with Human Motion Data [A]. In: Proceedings: SIGGRAPH 2002 [C], San Antonio, Texas, 2002. 491～500
5 Chui Y, Chao S P, Wu M Y, et al. Content-based Retrieval for Human Motion Data [J]. Journal of Visual Communication and Image Representation, 2004,16(3): 446～466
6 Liu F, Zhuang Y, Wu F, et al. 3d Motion Retrieval with Motion Index Tree [J]. Computer Vision and Image Understanding, 2003, 92(2-3): 265～284
7 Hunt E B, Marin J, Stone P T. Experiments in induction. New York: Academic Press,1966
8 Quilan J R. Generating Production rules from decision tree. In: Proceeding of IJCAI-87,1987
9 Quilan J R. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann,1993