

补充反馈模块的深度标注框架研究

陶 皖¹, 廖述梅²

(1. 安徽工程科技学院 计算机科学与工程系, 安徽 芜湖 241000;

2. 江西财经大学 信息管理学院, 江西 南昌 330013)

摘要:由数据库生成的动态 Web 页是静态页面的数百倍, 直接针对 Web 页产生过程的深度标注可以提高动态 Web 页的标注效率。针对动态 Web 页以查询生成居多的特点, 提出标注与反馈相结合的深度标注框架, 即第一步通过标注模块进行初步标注, 并记录 Web 页面的查询要求; 第二步分析查询信息, 找出不同 Web 页的关系, 通过反馈模块进一步补充标注内容, 从而提高标注的质量。

关键词:动态 Web 页; 深度标注; 本体; 反馈模块

中图分类号:TP301.2; TP391

文献标识码:A

文章编号:1673-629X(2007)01-0018-03

Research on Deep Annotation Frame with Feedback Module

TAO Wan¹, LIAO Shu-mei²

(1. Dept. of Computer Sci. & Eng., Anhui University of Eng. Techn. and Sci., Wuhu 241000, China;

2. Info. Management Sch., Financial & Economical University of Jiangxi, Nanchang 330013, China)

Abstract: Dynamic Web pages generated from database are hundreds of static Web pages. The deep annotation dealing with generation process of Web pages can prompt the annotation efficiency of dynamic Web pages. In allusion to most dynamic Web pages are generated by query operation, a deep annotation frame with feedback module are put forward. Firstly prime annotation is done by annotation module with query requirement being recorded, secondly query information is analyzed, relations among different Web pages are found, and the additional content is appended to prime annotation by feedback module so as to increase the annotation quality.

Key words: dynamic Web pages; deep annotation; ontology; feedback module

0 引言

语义标注 (Semantic Annotation) 是在 Web 页中加入语义元数据信息从而使 Web 页的内容机器可识别。它是将现有 Web 提升为语义 Web 的有效方法之一。这一方法中有两个关键: 一是语义元数据 (即为通常所说的标注信息) 的产生; 二是实现标注的过程 (包括标注的对象、标注的方法等)。当前学术界提出了多种语义 Web 的标注方法, 如: MINDSWAP 研究组的 SMORE^[1]、AKT 项目下的 Melita^[2]、CREAM 框架下的 OntoMatAnnotator^[3] 等。这些方法主要是手动加入设计好的语义元数据的信息, 以标注静态 Web 内容为主。而有效地产生语义元数据信息, 及标注动态 Web 页 (如: 从数据库中动态生成的 Web 页的数量是静态 Web 页的 500 倍^[4]) 是完成语义标注的主要任务。因此, 研究数据库的深度标注^[5] (deep annotation, 即直接标注数据库的逻辑模式或间接标注从数据库

生成的 Web 显示内容) 显得很有必要。Steffen Staab 等在文献 [5~7] 中提出了一个基于深度标注的框架, 即产生映射规则后, 在服务器端进行标记, 在客户端进行语义标注。通过此框架数据库拥有者和/或标注人员均可以参与标注过程, 从其描述效果可以发现标注效率明显改善, 但此类标注仍局限于单独的 Web, 没有处理动态查询所产生的 Web 页之间可能存在的联系。文中提出带反馈模块的深度语义标注框架, 该设计通过信息分析找到不同 Web 页面信息的关联关系, 并将结果反馈到标注中, 从而提高标注的质量。

1 动态 Web 页产生及深度标注过程

动态 Web 页是根据查询要求从数据库中生成的页面, 其“动态”体现在它的产生过程, 而不是 Web 页面的自身。与静态 Web 页相比, 动态 Web 页数量大, 并且结构往往很相似, 如针对一个查询要求, 产生的页面往往结构类似, 仅仅是显示内容不同。因此直接针对动态 Web 页进行语义标注 (如图 1 的标注①), 一是工作量大, 二是产生的语义元数据信息经常因 Web 页内容的变化而变得不准确, 从而失去标注的意义。

收稿日期: 2006-03-24

基金项目: 安徽省高校省级自然科学基金项目 (2005KJ065); 安徽省高校青年教师科研基金资助项目 (2005jql069)

作者简介: 陶 皖 (1972-), 女, 安徽芜湖人, 讲师, 硕士, 研究方向为语义 Web、本体工程。

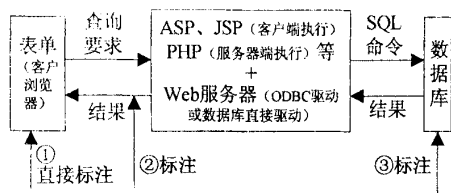


图 1 动态 Web 页标注生成方式

Carole Goble 在 WWW2002 年会议 Web 工作组上第一次提出了深度标注的概念,他指出深度标注是利用信息结构和信息上下文来推导出信息间映射的标注过程^[5],该过程将语义元数据的生成从 Web 页上迁移到 Web 内容产生过程(如图 1 的标注②)或数据库(如图 1 的标注③),进行的是数据库模式的标注中,从而提高了语义元数据的产生效率,并使语义元数据可以比较好地复用。

2 补充反馈模块的标注框架

针对图 1 的标注方式②,提出了一个补充反馈模块的深度标注框架(如图 2 所示):即第一步通过标注模块进行初步标注,同时记录 Web 页面的查询要求;第二步提取查询信息,分析信息后,通过反馈模块对标注做进一步的补充。

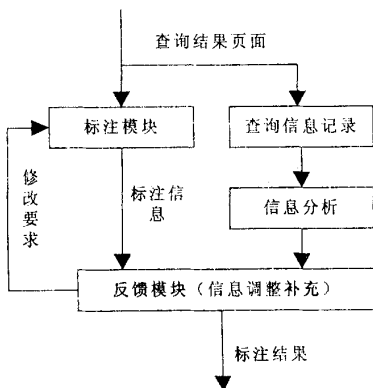


图 2 深度标注框架

2.1 标注模块 - 产生标注

标注离不开本体 (Ontology), 本体有类、属性和关系等词汇, 是共享词汇模型的明确的形式化规范说明。文中标注中的语义元数据描述即以 OWL 本体为例。下面简单说明 OWL 本体的形式化表示。

定义: 令 $Ont = \{C, O, D, R, H\}$ 代表一个本体。其中 C 是本体的概念集合; O 是概念的对象属性集合; D 是概念的数值属性集合; R 是本体的概念关系 (Relationship Between concepts) 和属性值 (Property Value) 的集合, 概念关系是三元组 $\langle C_1, O_p, C_2 \rangle$ (O_p 即是对象属性, 表示概念 C_1 实例的 O_p 取值是概念 C_2 的实例), 属性值是三元组 $\langle C, D_p, T \rangle$ (D_p 即是数值属性, 表示概念 C 实例的 D_p 取字面类型值 T); H 表示概念之间的层次结构关系, 用 $\langle C_1, H_R, C_2 \rangle$ 表示 C_1 和 C_2 的层次关系 H_R ^[8]。

图 3 表示以本体描述的标注, 其中 Ontology 部分是

Publication 本体片断, 矩形框表示词汇, 实形椭圆表示对象属性 (数据属性在表示中省略), 其形式化表示如下:

$$Ont_{Publication} = \{C_{pub}, O_{pub}, D_{pub}, R_{pub}, H_{pub}, X_{pub}\} \text{ where}$$

$$C_{pub} = \{Publication, Book, Person, Organization, Author, Publisher\}$$

$$D_{pub} = \{bookname, title, price, pubdate, kind, url, authname, \dots\}$$

$$O_{pub} = \{written_by, published_by\}$$

$$R_{pub} = \{< Book, written_by, Author >, < Book, published_by, Publisher > \} \vee \{< Book, bookname, string >, \dots, < Person, authname, string >, \dots\}$$

$$H_{pub} = \{< Author, rdfs:subClassOf Person >, < Book, rdfs:subClassOf Publication >, < Publisher, rdfs:subClassOf Organization >\}$$

图 3 的 Annotation 中分别表示了 Web 页面的 Book 信息和 Publisher 信息, 虚线箭头分别指示了标注关系, ①类指示了 owl:Class 的关系; ②类指示了关系属性 (如: A Book published - by some Publisher)。

实际的标注过程是: a. 通过动态 Web 页技术 (如: ASP) 将 HTML 和脚本 (反映动态页面的要求) 组合后生成动态 Web 页面后借助 Web 信息提取技术^[7-9], 抽取结构基本一致的各项记录; b. 根据本体生成映射规则^[8,9]; c. 通过标注算法^[6-10]自动生成标注。

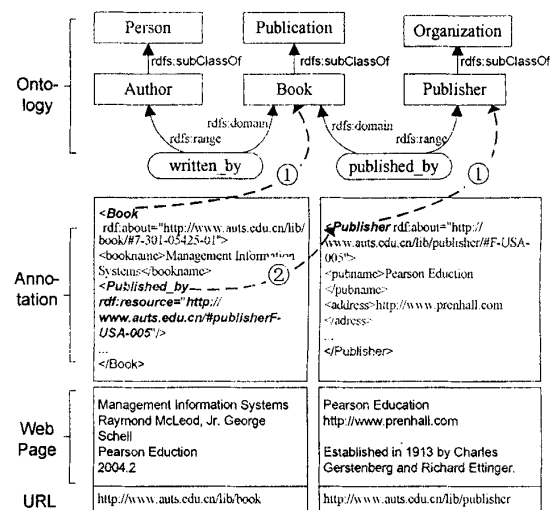


图 3 一般标注过程示例

2.2 信息分析、反馈 - 补充标注

通常在经过图 3 的标注过程后即产生页面的标注, 但分析以查询产生动态页面的情况后发现查询的关键字不同导致产生的页面不同, 但某些情况下其实质内容却类似, 通过记录下关键字, 分析界面不同但实际内容类似的实例并加以标注会有效地增加标注的信息量, 并可以补充本体的词汇。

如在 Amazon 书店 (<http://www.amazon.com>) 中输入查询关键字 Management Information Systems 后的查询结果如图 4 所示, 输入 MIS 后的查询结果如图 5 所示。虽

然页面信息不尽相同,但实际均为有关“管理信息系统”的信息,通过已有的本体(或补充新建本体)对这些页面建立映射。这其中的信息分析过程包括信息提取、信息比较、映射建立。

采用文献[7~9]中提出的信息提取技术(如将基于 DOM 树的结构化文本抽取方法)用于图 4 及图 5 的示例,并补充查询信息得到数据集如下(其中第一项记录查询关键字,之后是 bookname,author,price 和 URL):

$P_{11} = \langle \text{"Management Information Systems", "Management Information Systems Eight Edition", "Kenneth C. Laudon, Jane P. Laudon", \$140.80, "http://www.amazon.com/exec/obidos/search-handle-form/ref=sdp_tx_ti/102-7289903-9907322-01"} \rangle$

$P_{12} = \langle \text{"Management Information Systems", "Management Information Systems", "James A. O'Brien, George Marakas", \$134.38, "http://www.amazon.com/exec/obidos/search-handle-form/ref=sdp_tx_ti/102-7289903-9907322-02"} \rangle$

...

$P_{21} = \langle \text{"MIS", "Management Information Systems Eight Edition", "Kenneth C. Laudon, Jane P. Laudon", \$140.80, "http://www.amazon.com/exec/obidos/search-handle-url/102-7289903-9907322?url=index%3Dstripbooks%3Arelevance-above&field=keywords=MIS-01"} \rangle$

$P_{22} = \langle \text{"MIS", "Management Information Systems Ninth Edition", "Raymond McLeod, George Schell", \$90.20, "http://www.amazon.com/exec/obidos/search-handle-url/102-7289903-9907322?url=index%3Dstripbooks%3Arelevance-above&field=keywords=MIS-02"} \rangle$

...



图 4 “Management Information System”查询结果

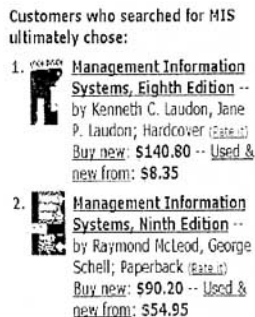


图 5 “MIS”查询结果

通过对页面信息进行比较,可以发现近 60% 以上的信息相同或相近(数据统计仍采用原始的方法,下一步的研究将结合 Web 页的数据挖掘方法进行页面信息分析),据此可以在页面的标注中补充相关联的 URL 页面信息。同时,效果可以是双向的,即可以通过建立的本体标注页面,也可以通过页面分析结果补充本体(如:使在本体词汇 Book 中计算机类的实例“Management Information System”和“MIS”产生相关的联系)。

(1) 反馈模块的工作过程。

a. 首先为 2.1 小节的 publication 本体补充对象属性 relation-with。

...

$O_{pub} = \{ \dots, \text{relation-with} \}$

$R_{pub} = \{ \text{Book relation-with Book} \} \vee \{ \dots \}$

...

b. 补充标注,结果如下:

$\langle \text{Book}$

rdf:about = "http://www.amazon.com/exec/obidos/search-handle-form/ref=sdp_tx_ti/102-7289903-9907322-01">

$\langle \text{bookname} \rangle \text{Management Information Systems Eight Edition} \langle / \text{bookname} \rangle$

...(内容与图 3 示例类似)

$\langle ! - - \text{反馈补充标注如下} - - \rangle$

$\langle \text{relation-with rdf:about = "http://www.amazon.com/exec/obidos/search-handle-url/102-7289903-9907322?url=index%3Dstripbooks%3Arelevance-above\&field=keywords=MIS-01"} \rangle$

...

$\langle / \text{Book} \rangle$

(2) 补充标注算法。

设页面分析结果集合为 AR,页面到本体的映射规则为 MR(文中未详细讨论页面信息提取及与本体的映射规则建立,相关内容参见文献[7,8]),本体库为 OntBase。

算法 AppendAnnotation(AR,MR,OntBase)

Begin

//将分析结果补充进本体,并建立映射关系

AppendOnt(AR,OntBase)

MapGeneration(AR,MR)

//自动生成补充标注

For ARI

AutomatedAnnotationGeneration(ARI,MR)

Next I

End For

End

限于篇幅,AutomatedAnnotationGeneration()过程参见文献[8]。

3 结束语

文献[5~7]中 Steffen Staab 等提出了一个基于深度标注的框架,文中的标注框架即受其启发,在补充信息分析及反馈模块后,通过添加相关的页面标注信息以改进标注的质量,从而提高计算机处理页面的效率。但现在研究的补充标注仍需手动进行,下一步的研究目标是进一步实现和完善补充标注算法,并结合数据挖掘技术中的信息分析处理,使分析结果的产生自动化或半自动化,此外,在标注产生后,对标注信息的利用(如:实现不同网站查询信息的关联)、中文页面的标注及反馈信息建立(如:建立类似“计算机”与“电脑”这样的信息联系)也将是研究的重点。

(下转第 23 页)

以突出重要属性的作用,消减不重要属性的影响,对 K-近邻算法进行了优化,提高了准确率。

3 实验结果及分析

同济大学信息安全实验室选取了安全领域和电子商务这两个特定主题作为实验的目标主题,在同样的实验环境下选取了不同 K 值和 N 值对文中所提出的网页归档算法进行了实验,训练集通过选取网络上 2000 多页网页先进行人工分类,再供系统进行自动学习,记录分类结果,并人工检验实验结果,判断其准确性,具体的实验结果见表 1。

表 1 网页自动分类系统实验结果

目标主题	网络安全	电子商务
K	20	20
N	10	15
训练集大小	1500	2000
测试集大小	1000	1000
准确率	86.54%	80.34%
目标主题	网络安全	电子商务
K	40	40
N	20	30
训练集大小	1500	2000
测试集大小	1000	1500
准确率	89.33%	85.32%

通过比较可以发现,应用该算法时,采用较高的 K 值和 N 值,将会获得更高的准确率,这是由于 K-近邻算法的准确率与选取的属性数量,和比较时选取的邻结点的数量成正比,与实验结果相符。通过实验结果,可以发现,该算法在确定主题的搜索中能够获得较高的准确率。具有

较高的实用价值。

4 结束语

文中描述了一种基于机器学习的文档自动分类系统,该系统能够在学习现有的已存档的网页的基础上,总结目标主题的特征,同时把它应用于新网页的自动分类过程中,该算法对现有的文档分类方法进行了一定的改进,并通过实验证明,该算法能够显著地提高针对于特定主题的网页搜索的准确率,为用户提供更好的搜索性能。同时,基于机器学习的方法还能通过不断学习来主动适应特定主题领域的发展,具有很强的适应性。

参考文献:

[1] 陈立孚,周 宁. 基于机器学习的自动文本分类模型研究[J]. 现代图书情报技术, 2005(10): 80-85.

[2] Dong Yan Shi, Han Ke Song. A Comparison of Several Ensemble Methods for Text Categorization[C]// Proceedings of the 2004 IEEE International Conference on Service Computing. [s.l.]: IEEE Computer Society, 2004.

[3] 何 清, 史忠植. 机器学习与概念语义空间生成, 中文信息处理若干重要问题[M]. 北京: 科学出版社, 2003: 266-277.

[4] Sebastiani F. Machine Learning in Automated Text Categorization[J]. ACM Computing Surveys, 2002, 34(1): 1-47.

[5] 赵国涛, 柯钦铭. 基于本体的异构文本分类系统[J]. 计算机工程, 2004(30): 21-25.

[6] Horho A, Madche A, Staab S. Text Clustering Based on Good Aggregations[J]. Künstliche Intelligenz, 2002(2): 4-9.

(上接第 20 页)

参考文献:

[1] Kalyanpur A, Hendler J, Parsial B. SMORE - Semantic Markup, Ontology and RDF Editor[EB/OL]. 2004-02-28. <http://mindswap.org/papers/SMORE.pdf>.

[2] Ciravegna F, Dingli A. User - System Cooperation in Document Annotation based on Information Extraction[C]// In Proc. of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02). Siguenza, Spain: [s.n.], 2002.

[3] Handschuh S, Staab S. Authoring and Annotation of Web Pages in CREAM[C]// In Proc. of WWW2002. Honolulu, Hawaii, USA: [s.n.], 2002.

[4] State of the art on Semantic Web languages. IST Project IST - 2001 - 34373 Esperonto Services D2. 1[EB/OL]. 2003. <http://www.esperonto.net/semanticportal/jsp/frames3.jsp/>.

[5] Handschuh S, Staab S, Volz R. On deep annotation[C]// In Proc. of WWW2003. Budapest, Hungary: [s.n.], 2003.

[6] Handschuh S, Staab S, Volz R, et al. Deep Annotation for In-

formation Integration[C]// In Proc. of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03). Acapulco, Mexico: [s.n.], 2003.

[7] Volz R, Handschuh S, Staab S, et al. Unveiling the hidden bride: deep annotation for mapping and migrating legacy data to the Semantic Web[J]. J. Web Sem, 2004(2): 187-206.

[8] 廖述梅, 徐升华, 陶 皖. 带模板的结构化 HTML 文档深度标注框架研究[C]// 中国系统工程学会. 信息系统协会中国分会第一界学术年会论文集(B 集). 北京: 清华大学出版社, 2005: 111-115.

[9] Mukherjee S, Yang G, Ramakrishnan I V. Automatic Annotation of Content - Rich HTML Document: Structural and Semantic Analysis[C]// In Second International Semantic Web Conference (ISWC2003). Sanibel Island, Florida, USA: [s.n.], 2003.

[10] Hyvonen E, Salminen M, Junnila M. Annotation of Heterogeneous Database Content for Semantic Web[EB/OL]. 2006-02-11. <http://www.seco.tkk.fi/publications/2004/hyvonen-salminen-et-al-annotation-of-heterogeneous-2004.pdf>.