计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
47(5): 858-865, 2010

# D-EEM:一种基于 DOM 树的 Deep Web 实体抽取机制

寇 月[1] 李 冬[2] 申德荣[1] 于 戈[1] 聂铁铮[1]
[1](东北大学信息科学与工程学院 沈阳 110004)
[2](东软集团商用软件事业部 沈阳 110179)
(kouyue@ise.neu.edu.cn)

# D-EEM: A DOM-Tree Based Entity Extraction Mechanism for Deep Web

Kou Yue[1], Li Dong[2], Shen Derong[1], Yu Ge[1], and Nie Tiezheng[1]
[1](*College of Information Science and Engineering, Northeastern University, Shenyang* 110004)
[2](*Business Software Division, Neusoft Group Ltd., Shenyang* 110179)

**Abstract** With the increase of Web databases, accessing Deep Web is becoming the main method to acquire information. Because of the large-scale unstructured content, heterogeneous result and dynamic data in Deep Web, there are some new challenges for entity extraction. Thus it is important to solve the problem of extracting the entities from Deep Web result pages effectively. By analyzing the characteristics of result pages, a DOM-tree based entity extraction mechanism for Deep Web (called D-EEM) is presented to solve the problem of entity extraction for Deep Web. D-EEM is modeled as three levels: expression level, extraction level, collection level. Therein the components of region location and semantic annotation are the core parts to be researched in this paper. A DOM-tree based automatic entity extraction strategy is performed in D-EEM to determine the data regions and entity regions respectively, which can improve the accuracy of extraction by considering both the textual content and the hierarchical structure in DOM-trees. Also based on the Web context and co-occurrence, a semantic annotation method is proposed to benefit the process of data integration effectively. An experimental study is proposed to determine the feasibility and effectiveness of the key techniques of D-EEM. Compared with various entity extraction strategies, D-EEM is superior in the accuracy and efficiency of extraction.

**Key words** entity extraction; DOM-tree; Deep Web; data region location; entity region location

**摘 要** 随着 Web 数据库的不断增长,通过对 Deep Web 的访问逐渐成为获取信息的主要手段.如何有效地抽取 Deep Web 中结果页面所包含的实体信息成为一个值得研究的问题.通过分析 Deep Web 结果页面的特点,提出了一种基于 DOM 树的 Deep Web 实体抽取机制(DOM-tree based entity extraction mechanism for Deep web, D-EEM),能够有效解决 Deep Web 环境中的实体抽取问题.D-EEM 采用基于 DOM 树的自动实体抽取策略,利用 DOM 树中的文本内容和层次结构来确定数据区域和实体区域,提高了实体抽取的准确性;另外,提出了一种基于上下文距离和共现次数的语义标注方法,有效地将来自不同数据源的抽取结果进行合成.通过实验验证了 D-EEM 中所采用的关键技术的可行性和有效性,同其他实体抽取策略相比,D-EEM 在抽取效率及抽取准确性等方面具有一定的优势.

**关键词** 实体抽取;DOM 树;Deep Web;数据区域定位;实体区域定位

**中图法分类号** TP311.13

**收稿日期**:2008-10-29;**修回日期**:2009-05-08

**基金项目**:国家自然科学基金项目(60673139,60973021);国家"八六三"高技术研究发展计划基金项目(2008AA01Z146);中央高校基本科研业务费专项基金项目(NO90304005)

## 0 引　　言

随着 Web 数据库的不断增长,通过对 Deep Web 的访问逐渐成为获取信息的主要手段[1]. Deep Web 中的信息主要是通过向各个数据源提交查询而获得的,结果页面往往包含了现实世界中的某个或某些实体,如书籍、文章、商品等,它们由若干属性(如文章标题、作者姓名、出版日期等)描述. 在实际应用中,用户对页面中所包含的实体信息更感兴趣. 因此,能够从结果页面中自动地获取蕴含在 Deep Web 中有价值的实体数据将大大减少用户进行筛选、比较的负担. 然而,Deep Web 返回的查询结果主要是通过 HTML 页面来展现的,其内容多样、形式各异,这就造成了结果数据的异构性和缺乏结构性,使得自动从中获取有价值的信息变成一件具有挑战性的任务. 因此,如何有效抽取 Deep Web 中的数据资源成为一个值得研究的问题,其目标是将查询获取的结果页面中所包含的实体信息正确有效地抽取出来,并进行结构化表示.

目前,按照抽取原理和方式的不同,可以将实体抽取技术分为基于自然语言处理方式的实体抽取、基于归纳学习的实体抽取、基于视觉特征的实体抽取和基于 DOM 树的实体抽取. 例如,文献[2-3]将 Web 文档视为文本进行处理,通过对用户标记信息的语法成分进行语法分析生成抽取规则. 该方法比较适用于含有大量文本且句子完整的 Web 页面,但没有考虑 Web 文档的结构特点而缺乏健壮性. 文献[4]要求用户事先在样本实例上标记出感兴趣的语义项,然后采用机器学习的方式分析待抽取数据在网页中的结构特征,归纳生成基于定界符的抽取规则. 该方法减少了对句法分析等自然语言处理技术的依赖,但需要较多的人工干预. 文献[5-6]提出了基于视觉特征的实体抽取技术,利用 Web 页面的位置、样式、风格等视觉特征将页面分割成若干数据块,基于页面块之间的相似度来进行数据抽取. 该方法实现起来比较复杂,一旦有新的启发式规则加入将会对已经分析完毕的页面产生影响,因此不具备较高的适用性. 文献[7-11]提出了基于 DOM 树的实体抽取技术,其思想是将 Web 文档解析成语法树,根据语法树的内部结构产生抽取规则. 该方法充分考虑了 Web 页面的内部结构,通过 DOM 核心定义的基本接口可以方便地访问 DOM 树中的节点信息,但采用的大多数抽取策略具有一定的局限性,主要体现在缺乏对目标区域逐渐精化的处理,将 DOM 树中的结构标签和文本内容混淆在一起统一对待,因此抽取结果的准确性将受到影响;另外,由于这些技术缺乏语义自动标注过程,使抽取到的数据缺乏语义信息,因而不具有较强的实用性.

为此,本文提出了一种新型的基于 DOM 树的 Deep Web 实体抽取机制 D-EEM(DOM-tree based entity extraction mechanism for Deep Web),能够有效解决实体抽取中的区域定位及语义标注等问题;提出了基于 DOM 树的自动实体抽取策略,有效地利用 DOM 树中的文本内容和层次结构分别指导数据区域定位和实体区域定位,提高了实体抽取的准确性;提出了一种基于上下文距离和共现次数的语义标注方法,有效地将来自不同数据源的抽取结果进行合成;通过实验验证了 D-EEM 中所采用的关键技术的可行性和有效性.

## 1 D-EEM 的层次模型

从功能上可以将 D-EEM 划分为 3 个层次,自底向上依次为信息采集层、实体抽取层和外部表示层(如图 1 所示). 信息采集层用来从 Internet 中收集各种网页资源,作为数据抽取的对象,并将收集来的 HTML 网页转换为一种有效的数据结构,以方便后续组件对其进行分析处理;实体抽取层负责分析转换后的资源数据,确定数据区域及实体区域,从中抽取实体记录并存储;外部表示层用来为用户提供良好的外部交互界面,为数据分析、实体查询等其他应用组件提供功能调用接口.

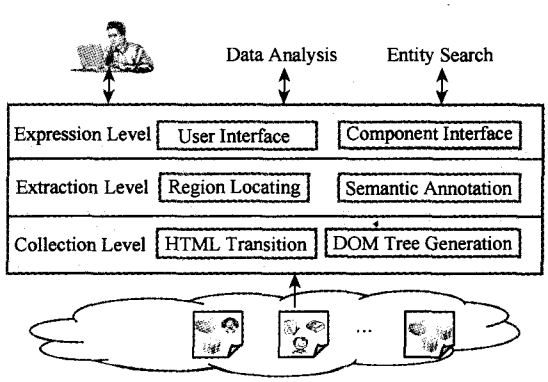


Fig. 1　The model of D-EEM.

图 1　D-EEM 的层次模型

在图 1 中,HTML 转换模块负责对 HTML 页面进行转换,即把 HTML 文档转化为符合 XML 语法的 XHTML 文档,这样可以利用各种强大的

XML 标准技术来操纵 XHTML 文档;DOM 树生成模块根据 XHTML 文档的结构将 Web 页面以树型结构进行表示,并提供一系列标准接口实现对树的查找、删除等操作;区域自动定位模块是 D-EEM 的核心模块,由数据区域定位和实体区域定位组成,主要负责对 Web 页面进行分析,确定各个实体在页面中的位置;语义标注模块负责将区域自动定位模块所获取的数据区域信息及实体区域信息进行形式化表示,对抽取结果的各个属性自动地添加语义标注,基于全局模式重现网页中所包含实体的模式结构信息;用户交互接口主要负责为用户提供外部的功能界面,友好地帮助用户完成相关参数的设置等任务;组件调用接口为外部的数据分析、实体查询等应用组件提供统一的函数调用接口,使外部组件能够透明地使用抽取到的实体数据. 本文将围绕核心模块——区域自动定位模块和语义标注模块的设计及实现策略进行详细的介绍.

## 2 基于 DOM 树的自动实体抽取策略

Deep Web 网页中的有用信息往往位于具有特定排列方式和次序的结构中,针对该特性本节提出了一种基于 DOM 树的自动实体抽取策略,主要包括数据区域定位和实体区域定位两个阶段.

### 2.1 数据区域定位

通常待抽取的实体记录聚集在网页中的某个区域范围内,将该区域定义为数据区域. 如果在结果页面中能够有效地将广告等无关信息加以过滤,那么实体的抽取操作将被限定在一个较精确的范围内,不但可以减少数据抽取的时间消耗,而且能够提高抽取的准确性. 因此,在进行实体定位之前有必要针对网页中的数据区域进行定位.

来自同一数据源的不同网页的 DOM 树(简称"同源 DOM 树")具有如下特点:同源 DOM 树的元素节点的层次嵌套关系基本一致;数据区域在 DOM 树中以子树的形式存在,同源 DOM 树之间的相异文本节点聚集于该区域范围内,将同源 DOM 树间对应文本内容相异的文本节点所组成的集合称为相异节点集合 $S_{diff}$. $S_{diff}$ 中的节点应集中地分布在数据区域内,仅有极个别的相异节点零散地分布于 DOM 树中数据区域以外的区域,因此,应优先选取相异节点所占比例较高的子树作为数据区域. 然而,如果将该比例设置得过高将会使数据区域的范围过小;相反,如果将该比例设置得过低将会使数据区域的范围过于广泛. 因此,在数据区域定位过程中需要权衡考虑两方面因素——相异度和覆盖度.

**定义 1.** 相异度 $Difference(T)$. DOM 子树 $T$ 中所有相异节点的数目占 $T$ 中节点总数的百分比(如式(1)所示):

$$Difference(T)=\frac{|\ S_{diff}\cap T\text{中所有节点}\ |}{T\text{中所有节点}}. \tag{1}$$

**定义 2.** 覆盖度 $Coverage(T)$. DOM 子树 $T$ 中所有相异节点的数目占整个 DOM 树中相异节点总数的百分比(如式(2)所示):

$$Coverage(T)=\frac{|\ S_{diff}\cap T\ |}{|\ S_{diff}\ |}. \tag{2}$$

从相异度和覆盖度的定义可知,相异度越大子树就越符合数据区域的特性;覆盖度越大子树所包含的实体信息就越全面. 然而,相异度与覆盖度呈互逆关系,相异度越高覆盖度就越低,反之亦然. 因此,本文定义了一个新的评测标准——聚集度来综合评价相异度与覆盖度,以保证数据区域定位的准确性. 聚集度用来衡量子树 $T$ 内相异节点的聚集程度,其定义如式(3)所示:

$$Aggregation(T)=\frac{2\times Difference(T)\times Coverage(T)}{Difference(T)+Coverage(T)}. \tag{3}$$

基于上述定义,数据区域定位的过程由以下步骤组成:首先,比较两棵同源 DOM 树,确定相异节点集合 $S_{diff}$;然后,将同源 DOM 树中的相异节点进行组合,针对每种组合确定候选子树——包含该组合的最小子树,针对每棵候选子树,依次计算其相异度、覆盖度以及聚集度;最终选取同源 DOM 树中聚集度较大的子树所在的区域作为数据区域.

### 2.2 实体区域定位

为了降低实体区域定位的复杂性,首先要对数据区域对应的 DOM 树进行最小化处理,提取 DOM 树中的有效结构信息,主要包括以下两方面:将任何与结构无关的文本节点从 DOM 树中去除,以减少需要遍历的节点数;若树中某父亲节点只有一个孩子节点,那么需要将父亲节点与孩子节点合并,以减少需要遍历的层数(合并后节点以父亲节点名+"/"+孩子节点名来命名). 例如,图 2 对比了简化前后的 DOM 树片段,经过化简处理后文本信息已经被过滤掉,存在冗余的节点被合并.

接下来要逐层对化简后的 DOM 树进行匹配,根据匹配程度可检测各级别子树是否频繁出现,若频繁出现则基于这些子树确定实体区域,否则针对当前子树的下一级子树重新进行匹配和检测,直到获

取到实体区域为止.本文采用基于序列的子树匹配策略,首先对子树进行先序遍历,可以确定一个由节点名称组成(由"/"将不同节点间隔开)的序列.通过计算字符串编辑距离可衡量序列间的相似性,子树 $T_1$ 与 $T_2$ 的相似度 $Sim$ 计算如式(4)所示,其中 $sequence$ 表示子树的先序遍历序列.各实体记录在结果页面中往往具有相似的表现形式,体现在 DOM 树中就是具有重复的结构模式,相关定义如下:

$$Sim(T_1,T_2)=1-\frac{ed(T_1.sequence,T_2.sequence)}{\max\{|T_1.sequence|,|T_2.sequence|\}}. \tag{4}$$

**定义 3.** 同级子树集相似度 $MultiSim(T_1,\cdots,T_n)$.若 $\{T_1,\cdots,T_n\}$ 是同级子树集,则 $T_1,\cdots,T_n$ 间的同级子树集相似度为

$$MultiSim(T_1,\cdots,T_n)=\frac{\sum_{i<j}Sim(T_i,T_j)}{C_n^2}.$$

**定义 4.** 同级频繁子树集.给定一同级子树集 $\{T_1,\cdots,T_n\}$,设 $minSim$ 是用户预先指定的相似度阈值($0\leqslant minSim\leqslant 1$).当且仅当 $MultiSim(T_1,\cdots,T_n)\geqslant minSim$ 时,$\{T_1,\cdots,T_n\}$ 是同级频繁子树集.其中,每棵子树 $T_i$ 被称为同级频繁子树.

**定义 5.** 最佳频繁子树.若 $T_i$ 是同级频繁子树,并且 $T_i$ 的任何超树及子树所对应的同级子树集相似度均小于 $T_i$ 的同级子树集相似度,则将 $T_i$ 称为最佳频繁子树.

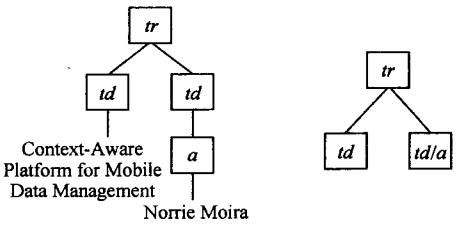


Fig. 2　Demonstration of DOM-tree simplification.

图 2　DOM 树化简示例

本文采用自顶向下的最佳频繁子树挖掘策略对最简 DOM 树中各个层次的同级子树进行匹配,检测该级别的各个子树是否为最佳频繁子树,以此来确定实体区域.图 3 描述了最佳频繁子树的挖掘过程,设 $minSim=0.6$.首先,以最简 DOM 树的根节点 *table* 作为入口点,获取以第 2 级节点为根节点的

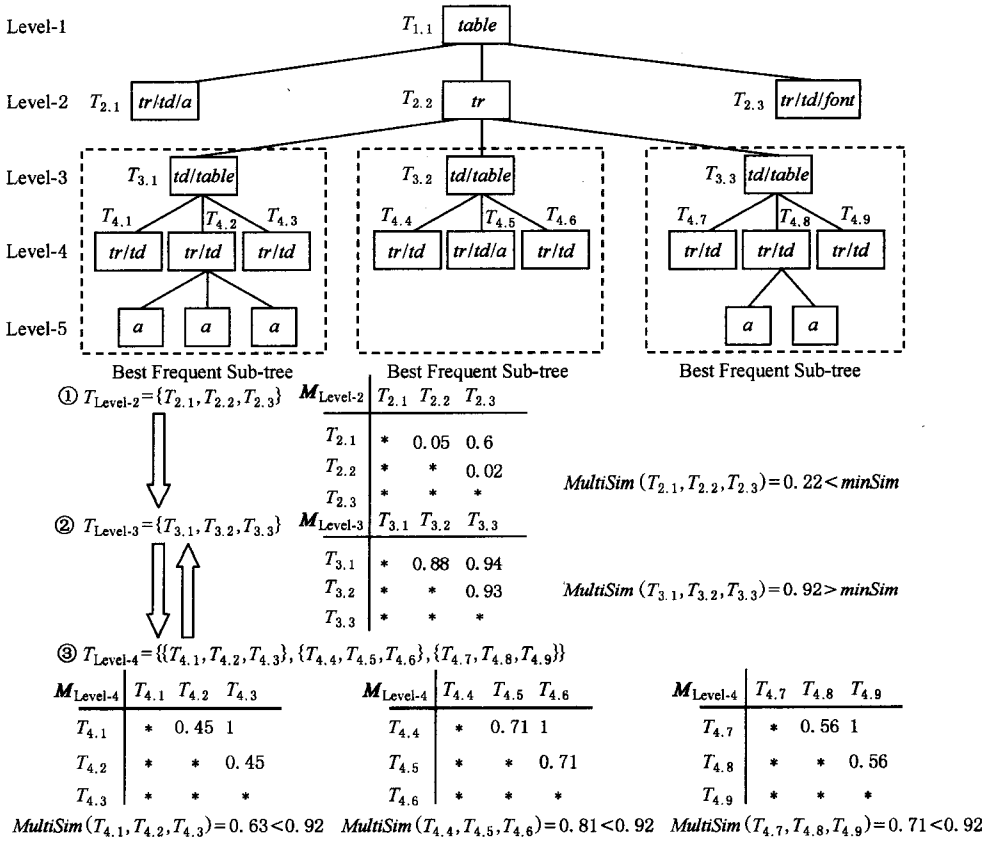


Fig. 3　Demonstration of mining best frequent sub-trees.

图 3　最佳频繁子树挖掘示例

子树所组成的同级子树集 $T_{\text{Level-2}}=\{T_{2.1},T_{2.2},T_{2.3}\}$. 将 $T_{\text{Level-2}}$ 中的子树两两比较,计算子树间的相似度,将结果存放于矩阵 $\boldsymbol{M}_{\text{Level-2}}$ 中,并计算同级子树集相似度 $MultiSim(T_{2.1},T_{2.2},T_{2.3})$. 由于 $MultiSim(T_{2.1},T_{2.2},T_{2.3})<minSim$,因此 $T_{\text{Level-2}}$ 不是同级频繁子树集. 接下来,考虑以第 3 级节点为根节点的子树所组成的同级子树集 $T_{\text{Level-3}}=\{T_{3.1},T_{3.2},T_{3.3}\}$,以同样的方式构建矩阵 $\boldsymbol{M}_{\text{Level-3}}$,由于 $MultiSim(T_{3.1},T_{3.2},T_{3.3})>minSim$,$T_{\text{Level-3}}$ 是同级频繁子树集. 因此需要进一步判断 $T_{\text{Level-3}}$ 中的子树是否为最佳频繁子树,再构建矩阵 $\boldsymbol{M}_{\text{Level-4}}$,由于 $MultiSim(T_{3.1},T_{3.2},T_{3.3})$ 大于 $MultiSim(T_{4.1},T_{4.2},T_{4.3})$,$MultiSim(T_{4.4},T_{4.5},T_{4.6})$,$MultiSim(T_{4.7},T_{4.8},T_{4.9})$,因此 $T_{\text{Level-3}}$ 中的各个子树均是最佳频繁子树.

## 3 实体语义标注

基于最佳频繁子树可以生成抽取规则,经过语义标注后以指导实体的最终抽取. 抽取规则在逻辑上包括两部分:语义项和实体项. 实体项用来描述各个最佳频繁子树的根节点,作为访问实体信息的入口点;语义项用来描述最佳频繁子树的叶子节点路径,对应于实体的各个属性. 以图 3 为例,最终获取的定位信息(以相对路径表示)如表 1 所示:

**Table 1 Demonstration of the Location Information of Entity Item and Semantic Items**

**表 1 实体项、语义项定位信息示例**

| Type | Path |
|---|---|
| Entity Item | *table/tr/td/table* |
| Semantic Item 1 | *table/tr/td/table/tr/td*[0] |
| Semantic Item 2 | *table/tr/td/table/tr/td*[1]/*a*[0] |
| Semantic Item 3 | *table/tr/td/table/tr/td*[1]/*a*[1] |
| Semantic Item 4 | *table/tr/td/table/tr/td*[1]/*a*[2] |
| Semantic Item 5 | *table/tr/td/table/tr/td*[2] |

利用抽取规则可以将具有相同路径的节点划分为同一语义类别中,也就是对应同一个语义项. 然而,数据抽取经常需要在多个数据源返回的结果页面中进行,这些数据源之间往往具有不同的结构特征. 因此,要想将多个数据源的抽取结果进行有效的集成,就必须为抽取到的数据分配一个有意义的标记来表示其语义,也就是要建立语义项与实际语义之间的映射关系,将这一过程称为实体的语义标注. 为了能够快速而有效地解决上述问题,本文提出了一种语义标注方法. 首先,根据领域特征定义全局模式,然后借助 Google API 将某类语义项的文本内容与全局模式中的各个属性逐一匹配,分别以语义项取值与各个属性名称作为查询关键字进行查询,根据 Google 返回的结果来衡量语义项与全局模式中某属性的语义关联程度. 直观来看,影响结果页面中语义项取值与全局模式中某属性名称的关联程度的因素主要有以下两方面:若二者的上下文距离越近关联就越紧密;若二者多次共同出现关联就越紧密.

因此,综合考虑以上两方面因素,本文定义了语义关联度的量化函数(如式(5)所示). 其中 $K\{k_1,\cdots,k_m\}$ 表示某个类别的语义项的取值集合,$Span(k_i,A_j)$ 表示 $k_i$ 与属性 $A_j$ 之间的上下文距离,$Itemspan(K,A_j)$ 表示 $K$ 中各元素与属性 $A_j$ 之间的上下文距离的平均值(如式(6)所示);$Times(k_i,A_j)$ 表示在 Google 返回的结果页面中 $k_i$ 与属性 $A_j$ 共同出现的次数,$Itemtimes(K,A_j)$ 表示在 Google 返回的结果页面中 $K$ 的各个元素与属性 $A_j$ 共同出现的平均次数(如式(7)所示);$Rel(K,A_j)$ 表示 $K$ 与属性 $A_j$ 的关联度. $\alpha$ 的取值在 0 和 1 之间,用来权衡上下文距离和共现次数对关联度的影响程度.

$$Rel(K,A_j)=\alpha\times\left(\frac{Itemspan(K,A_j)}{\max\{Itemspan(K,A_j)\}}\right)+(1-\alpha)\times\frac{Itemtimes(K,A_j)}{\sum_{j=1}^{n}Itemtimes(K,A_j)},\quad(5)$$

$$Itemtimes(K,A_j)=\frac{\sum_{j=1}^{m}Span(k_i,A_j)}{m},\quad(6)$$

$$Itemtimes(K,A_j)=\frac{\sum_{j=1}^{m}Times(k_i,A_j)}{m}.\quad(7)$$

依据这些量化函数,可以计算语义项与全局模式中各个属性的关联度,最终选取关联度最大的属性作为某语义项的语义标注. 若多个语义项被映射为相同属性时,还要将这些语义项的文本信息进行合并. 例如,经过语义标注,表 1 中各语义项与全局模式中属性的映射关系如表 2 所示:

**Table 2 Demonstration of Semantic Annotation**

**表 2 语义标注结果示例**

| Type | Annotation Result |
|---|---|
| Semantic Item 1 | Title |
| Semantic Item 2 | Author |
| Semantic Item 3 | Author |
| Semantic Item 4 | Author |
| Semantic Item 5 | Publisher |

## 4　实验测试

本文主要针对论文、图书及购物 3 个领域中的文章、图书及商品实体进行抽取,通过向一些大型的检索网站(如 ACM,Kluwer,Amazon 等)提交特定的查询请求来获取一定量的结果页面,作为测试数据集.为了有效地进行实体抽取,提交的查询请求应尽量满足以下两个条件:基于这些请求所得到的查询结果集具有较大的数据量;查询请求相互独立,以保证结果集之间不存在交集.实验环境设置如下:主机采用 Dell P4 2.4 GHz,内存容量为 512 MB,硬盘容量为 80 GB,操作系统为 Win2000.

首先,本实验采用均值测试法对 5 种大小级别的结果页面的区域定位时间代价进行评估,实验结果如图 4 所示.可以看出,实体区域定位花费的时间比数据区域定位花费的时间要长;另外,随着结果页面大小的增加,数据区域定位和实体区域定位所需的时间均随之增加.由于 D-EEM 在实体区域定位前进行了数据区域的定位,能够有效地过滤掉无关信息,只保留与实体相关的数据部分,因此大大减少了数据抽取的时间消耗.

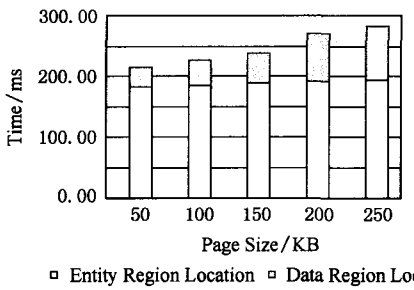


Fig. 4　Time cost of automatic region location.

图 4　区域自动定位的时间代价

另外,本文将 DERL 策略(在数据区域定位的基础上进行实体区域定位)与 ERL 策略(直接进行实体区域定位)针对查准率、查全率和 *F*-Measure 进行了比较,如图 5 所示.由于 DERL 策略对数据区域进行了定位,从而过滤掉一些无关信息,使实体抽取能够在较精确的范围内进行,因此,同 ERL 策略相比,D-EEM 在抽取性能上具有一定的优势.

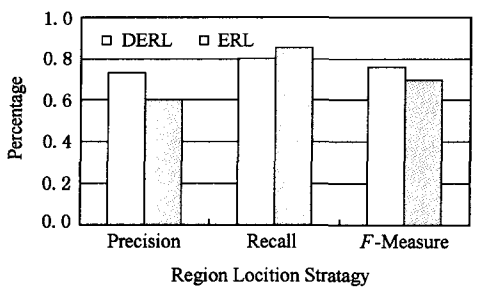


Fig. 5　Performance comparison between the DERL strategy and the ERL strategy.

图 5　DERL 策略与 ERL 策略的性能比较

图 6 表示 D-EEM 对 3 个领域的测试数据集进行实体抽取的性能比较结果.同论文和图书领域相比,针对购物领域的抽取性能相对较差.这是由于论文和图书领域中的大多数网站所提供实体记录的 DOM 树结构互相独立,而购物领域中的网站经常将多个商品放置于 HTML 页面中的同一行来显示,造成各个实体的 DOM 树结构互相参杂、难以区分.而本文提出的实体抽取策略是以实体记录的 DOM 树结构相互独立作为前提的,因而针对购物领域的抽取性能相对较低.

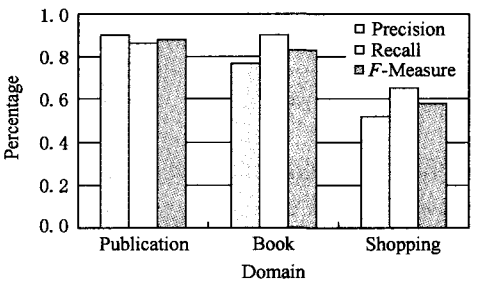


Fig. 6　Performance comparison for D-EEM among different domains.

图 6　D-EEM 针对不同领域的实体抽取的性能比较

## 5　结　　论

本文提出了一种基于 DOM 树的 Deep Web 实体抽取机制,能够有效解决 Deep Web 实体抽取中区域定位及语义标注等问题.通过模拟实验表明,由于 D-EEM 将实体抽取过程分为数据区域定位和实体区域定位两个阶段,使得实体区域的定位能够在一个较精确的范围内进行,提高了实体抽取的效率;其次,在实体抽取过程中充分考虑了 DOM 树内文本内容节点和元素节点的特征,提高了实体抽取的准确性;另外,由于添加了对抽取结果的语义标注过程,能够将来自不同数据源的抽取结果进行有效的集成.

今后拟在以下两个方面对现有工作进行改进:一方面需要针对一些特殊页面(如购物领域)的抽取作进一步的研究,以提高系统的通用性;另一方面将研究如何将 DOM 树结构同页面视觉特征相结合,以此来提高实体抽取的准确性.

## 参 考 文 献


[1] Chang KCC, He B, Li C, et al. Structured databases on the Web: Observations and implications [J]. SIGMOD Record, 2004, 33(3): 61-70

[2] Calife M, Mooney R. Relational learning of pattern match rules for information extraction [C] //Proc of the 16th National Conf on Artificial Intelligence and 11th Conf on Innovative Applications of Artificial Intelligence. Menlo Park, CA: AAAI, 1999: 328-334

[3] Soderlan S. Learning information extraction rules for semi-structured and free text [J]. International Journal of Machine Learning, 1999, 34(1-3): 233-272

[4] Muslea I, Minton S, Knoblock G. A hierarchical approach to wrapper induction [C] //Proc of the 3rd Conf on Autonomous Agents. New York: ACM, 1999: 190-197

[5] Liu Wei, Meng Xiaofeng, Meng Weiyi. Vision-based Web data records extraction [C] //Proc of the 9th SIGMOD Int Workshop on Web and Database. New York: ACM, 2006: 20-25

[6] Zhao Hongkun, Meng Weiyi. Fully automatic wrapper generation for search engines [C] //Proc of WWW'05. New York: ACM, 2005: 66-75

[7] Liu L, Pu C, Han W. XWRAP: An XML-enable wrapper construction system Web information sources [C] //Proc of the 16th IEEE Int Conf on Data Engineering. Washington: IEEE, 2000: 611-621

[8] Valter C, Giansalvatore M, Paolo M. RoadRunner: Towards automatic data extraction from large Web sites [C] //Proc of the 27th VLDB. San Francisco: Morgan Kaufmann, 2001: 109-118

[9] Li Xiaodong, Gu Yuqing. DOM-based information extraction for the Web sources [J]. Chinese Journal of Computers, 2002, 25(5): 526-533 (in Chinese)
(李效东, 顾毓清. 基于 DOM 的 Web 信息提取[J]. 计算机学报, 2002, 25(5): 526-533)

[10] Wang Ru, Song Hantao, Lu Yucang. Web pages data extraction based on tree automata [J]. Transactions of Beijing Institute of Technology, 2004, 24(9): 790-793 (in Chinese)
(王茹, 宋瀚涛, 陆玉昌. 基于树自动机的网页数据抽取[J]. 北京理工大学学报, 2004, 24(9): 790-793)

[11] Yang Shaohua, Lin Hailue, Han Yanbo. Automatic data extraction from template-generated Web pages [J]. Journal of Software, 2008, 19(2): 209-223 (in Chinese)
(杨少华, 林海略, 韩燕波. 针对模板生成网页的一种数据自动抽取方法[J]. 软件学报, 2008, 19(2): 209-223)



**Kou Yue**, born in 1980. She has been an assistant of the Northeastern University. She received her PhD degree in computer software and theory from the Northeastern University, China in 2009. She is a student member of CCF. Her main research interests include Web data management and information retrieval.

**寇 月**, 1980 年生, 博士, 助教, 中国计算机学会学生会员, 主要研究方向为 Web 数据管理和信息检索.



**Li Dong**, born in 1979. He is an assistant engineer of Neusoft Group Ltd. His main research interests include Web data extraction and information integration.

**李 冬**, 1979 年生, 助理工程师, 主要研究方向为 Web 数据抽取和信息集成.



**Shen Derong**, born in 1964. She is a professor of the Northeastern University. Senior member of CCF. Her main research interests include distributed and parallel systems, Web data management and data grid.

**申德荣**, 1964 年生, 教授, 中国计算机学会高级会员, 主要研究方向为分布式系统、Web 数据管理和数据网格等.



**Yu Ge**, born in 1962. He is a professor of the Northeastern University. Senior member of CCF. His main research interests includes data base theory and technology, distributed and parallel systems, embedded software, network information security.

**于 戈**, 1962 年生, 教授, 中国计算机学会高级会员, 主要研究方向为数据库、分布式系统、嵌入式软件和信息安全等.



**Nie Tiezheng**, born in 1980. He is an assistant of the Northeastern University. Student member of CCF. His main research interests include Web data management and information integration.

**聂铁铮**, 1980 年生, 助教, 中国计算机学会学生会员, 主要研究方向为 Web 数据管理和信息集成.


**Research Background**

With the increase of Web databases, accessing Deep Web is becoming the main method to acquire information. Because of

the large-scale unstructured content, heterogeneous result and dynamic data in Deep Web, there are some new challenges for entity extraction. Thus it is important to solve the problem of extracting the entities from Deep Web result pages effectively. In this paper, a DOM-tree based entity extraction mechanism for Deep Web (called D-EEM) is presented to solve the problem of entity extraction for Deep Web. A DOM-tree based automatic entity extraction strategy is performed in D-EEM to determine data regions and entity regions respectively, which can improve the accuracy of extraction by considering both the textual content and the hierarchical structure in DOM-trees. Also based on the Web context and co-occurrence, a semantic annotation method is proposed to benefit the process of data integration effectively. An experimental study is proposed to determine the feasibility and effectiveness of the key techniques of D-EEM. Compared with various entity extraction strategies, our approach is superior in the accuracy and efficiency of extraction. Our work is supported by the National Natural Science Foundation of China under grant No. 60673139 and the National High Technology Development 863 Program of China under grant No. 2008AA01Z146.

# 第 17 届全国网络与数据通信学术会议(NDCC2010)征文通知

## 2010 年 9 月 16—17 日　北戴河

由中国计算机学会网络与数据通信专业委员会主办、由东北大学秦皇岛分校和东北大学信息科学与工程学院联合承办的“第 17 届全国网络与数据通信学术会议”将于 2010 年 9 月 16 日到 17 日在美丽的海滨城市北戴河举行.

本次大会将围绕“网络与通信新技术及应用”这一主题展开,为来自国内外高等院校、科研院所、企事业单位的学者、教授、专家、工程师提供一个代表国内网络与数据通信产学研界高水平的高层信息交流平台,探讨本领域发展所面临的关键挑战问题和热点研究方向.

会议论文集将由《东北大学学报(自然科学版)》增刊出版(EI 检索源). 论文参照《东北大学学报》格式,字数一般不超过 6000 字,稿件通过投稿系统提交,具体参见会议网站 http://ndcc2010.neuq.edu.cn. 部分优秀论文将被推荐到《计算机学报》、《电子学报》、《计算机研究与发展》、《电子与信息学报》的正刊发表(均为 EI 检索源). 会议期间将评选会议优秀论文和优秀学生论文.

本次会议的主要征文范围包括以下领域(但不限于):

**新一代网络技术**:网络体系结构、路由/交换技术、协议工程、网络虚拟化、认知网络、IPv4/IPv6 过渡技术、NGN/NGI 平台应用; **新一代计算技术**:云计算/网格计算、并行/分布式计算、普适/效用计算、服务计算; **无线通信技术**:下一代移动通信技术、自适应信号处理、传感器网络、移动自组织网络、智能天线、卫星通信; **其他**:光通信技术、网络安全、网络管理、网络应用.

**投稿须知**

1) 投稿内容突出作者的创新与成果,具有较重要的学术价值与应用推广价值,未在国内外公开发行的刊物或会议上发表或宣读过.

2) 论文语言要求中文,字数一般不超过 6000 字,论文格式参照《东北大学学报》,投稿稿件用 Word 文件形式. 东北大学学报的网站地址如下: http://xuebao.neu.edu.cn/natural/index.asp.

3) 请在稿件最后附上第一作者姓名、性别、职务/职称、所属单位、通信地址、邮政编码、联系电话和 E-mail 地址,并注明论文所属领域.

4) 被录用的论文至少要有一位作者参加会议并发言才有资格参与优秀论文的评选.

**投稿方式**

论文投稿通过投稿系统进行提交,详见会议网站: http://ndcc2010.neuq.edu.cn 或者通过邮件地址 ndcc2010@mail.neuq.edu.cn 联系. 电子邮件请在邮件标题注明“NDCC2010 投稿”.

**重要日期**

论文提交截止日期:2010 年 5 月 15 日;论文录用通知日期:2010 年 7 月 1 日;会议注册截止日期:2010 年 8 月 15 日.

**联系方式**

联系电话:0335-8052155　　联系人:王翠荣　韩来权

邮件地址:ndcc2010@mail.neuq.edu.cn　　会议网站:http://ndcc2010.neuq.edu.cn