WebKER: A Wrapper Learning-based Tool for Extracting Knowledge*

Xi Bai and Jigui Sun

College of Computer Science and Technology, Jilin University,
Changchun 130012, China
xibai@email.jlu.edu.cn, jgsun@jlu.edu.cn

Abstract. This paper proposes WebKER, a system for extracting knowledge from Web documents based on domain ontologies and suffix-array-based wrapper learning. It also gives the performance evaluation of this system and the comparison between querying information in the Knowledge Base (KB) and querying information in the traditional database. Moreover, the *outstanding*-wrapper selection and a novel method for linking and merging the knowledge are also presented.

1 Introduction

Wrappers are widely used for converting Web pages to structured data [1]. However, the layouts of Web pages from different websites are usually different. In addition, the layout of a specific Web page may be updated from time to time. To the best of our knowledge, among published works little is dealing with proposing complete frameworks or systems for extracting the knowledge from diverse Web pages in Chinese. Due to the lack of semantics, the traditional information depositories (XML files, tables in database etc.) are not convenient for users to do queries and the search results are usually not sufficient enough. Domain ontologies can provide machine-readable representation of meaning of information to aid the extraction [2]. Motivated by the above analysis, we propose WebKER, a domain-ontology-aided system for extracting knowledge from Chinese Web pages by employing suffix-array-based wrapper learning. Moreover, our system can be modified to deal with the Web documents in other languages if appropriate domain ontologies and the Name Dictionary (ND) are provided.

2 Our Approaches and Experiments

After normalizing the Web documents using Tidy[1] package, we first introduce a method of finding out the Information Content Blocks of Interests (ICBIs). We use regular expressions [4] to match and omit nonsignificant blocks. For instance, we can find out *Script* blocks using regular expression "<SCRIPT[^>]$*$(> |\\$n*$)" and fulfill the omission task. Then, we compare each two lines of two documents generated from the same template. The same *lines groups* will be neglected and the ICBIs will be retained. We abandon the attributes within the tags, replace the literal content with a new tag </txt> and map each tag to an identical character according to a predefined Mapping Table (MT). Regarding documents as

* This work is supported by the NSFC (grant no. 60496321) and European Commission under the project EASTWEB (grant no. 111084).

[1] http://www.w3.org/People/Raggett/tidy

strings, we create their suffix arrays [3] which help us find out the maximal hidden repeats. Since most of the extracted patterns are incomplete or redundant, we develop the criteria for selecting the *outstanding* pattern according to *tag loop*, *pattern length*, *the first tag* and *coverage percentage*. Then the *outstanding* pattern is automatically transformed into original tags according to MT. Meanwhile, </txt> is replaced with a regular expression "> [^<^>]+ <". We can use the *outstanding* pattern to extract data. However, the data newly extracted is raw since one resource (class, property or individual) may have several synonymous names on different websites. Based on Hownet[1], we establish an ND to map the original names of resources to the names predefined in domain ontologies using *binary balance tree* structure. By querying the domain and the range of a specific property, we can identify its subject and object and then generate the knowledge described with OWL[2]. Finally, we propose a similarity-based approach for linking and merging the knowledge before adding them to the KB.

The domain ontologies employed in our system are created by four people within ten months, including 200 classes, 265 properties and 1332 individuals. We collect Web documents randomly from five large portals and four professional websites covering finance information in Chinese. For each website, we retrieve 20 documents to extract information about corporations. We ask ten domain experts to extract information from these documents manually and finally get 894 individuals. Focusing on five websites, we ask 40 users to generate 120 queries randomly. According to the experimental results, we find that the performance of extracting information using WebKER is superior and the $F_{measure}$ of querying the KB are all higher than those of querying the traditional database.

3 Conclusions and Future Work

We propose a wrapper-driven system WebKER. Based on domain ontologies and suffix-array-based wrapper learning, WebKER learns wrappers from Web documents and generates knowledge for users to do queries. Experiments on documents from large Chinese websites yields promising results. Our current-research goal focuses on improving our pattern processor by adding heuristics to the pattern discovery and selection. Our long term goal is to improve our system and make it semiautomatically extend domain ontologies within the knowledge-extraction process.

References

1. N. Kushmerick. Wrapper induction: Efficiency and expressiveness. Artificial Intelligence, 118(1-2), pages 15-68, 2000.
2. L. K. McDowell and M. Cafarella. Ontology-driven information extraction with OntoSyphon. In *Proceedings of the ISWC 2006*, LNCS 4273, pages 428-444, 2006.
3. V. Makinen. Compact suffix array. In *Proceedings of the CPM 2000*, LNCS 1848, pages 305-319, 2000.
4. L. Ilie, B. Shan and S. Yu. Fast algorithms for extended regular expression matching and searching. In *Proceedings of the STACS 2003*, LNCS 2607, pages 179-190, 2003.

[1] http://www.keenage.com

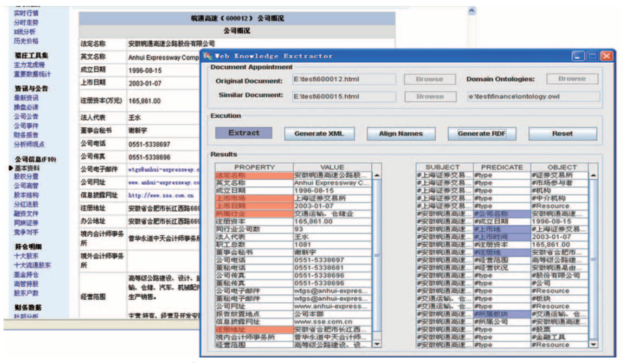
[2] http://www.w3.org/TR/owl-features/

Explanations for WebKER

In this appendix we describe the implementation and the demo of our approach covered in the main part of this poster description.

Implementation and Example

WebKER is implemented as a part of CRAB[1]. CRAB contains a set of tools such as the Web-document crawler, the page filter, the knowledge extractor, the reasoner and the report generater. After the Web documents are denoised, WebKER can automatically learn wrappers and generate knowledge. The framework for WebKER contains following blocks. *Retriever* block retrieves Web documents from the Web. Then these documents are transformed from the ill-formed expression into the well-formed expression by *Transformer* block. Two documents of the same layout are compared by *Information Block Extractor* block and the generated HTML snippets will be sent to *Translator* block. According to the MP, tags are replaced with specific characters. Then *Suffix Array Constructor* block discovers the repeated patterns hidden in the Web documents. According to the predefined criteria, *Selector* block finds out the *outstanding* pattern. Then resources can be extracted and stored in XML files. Based on the ND, these names are replaced with the predefined ones in domain ontologies by *Name Mapper* block. Finally, the knowledge is generated and merged by *Knowledge Generator* block and *Knowledge Merger* block respectively.



实时行情
分时走势
K线分析
历史价格
猎庄工具集
主力龙虎榜
重要数据统计
资讯与公告
最新资讯
操盘必读
公司公告
公司事件
财务报告
分析师观点
公司信息(F10)
▶ 基本资料
股权分置
公司高管
股本结构
分红送股
融资文件
同族证券
竞争对手
持仓明细
十大股东
十大流通股东
基金持仓
高管持股
股东户数
财务数据

皖通高速（600012）公司概况

公司概况

法定名称	安徽皖通高速公路股份有限公司
英文名称	Anhui Expressway Comp
成立日期	1996-08-15
上市日期	2003-01-07
注册资本(万元)	165,861.00
法人代表	王水
董事会秘书	谢新宇
公司电话	0551-5338697
公司传真	0551-5338696
公司电子邮件	wtgs@anhui-expressway.c
公司网址	www.anhui-expressway.co
信息披露网址	http://www.sse.com.cn
注册地址	安徽省合肥市长江西路669
办公地址	安徽省合肥市长江西路669
境内会计师事务所	普华永道中天会计师事务所
境外会计师事务所	
经营范围	高等级公路建设、设计、… 输、仓储、汽车、机械配件… 生产销售。 主营:持有、经营及开发安…

Web Knowledge Exctractor

Document Appointment

Original Document: E:\test\600012.html
Similar Document: E:\test\600015.html
Domain Ontologies: e:\test\finance\ontology.owl

PROPERTY	VALUE
法定名称	安徽皖通高速公路股...
英文名称	Anhui Expressway C...
成立日期	1996-08-15
上市市场	上海证券交易所
上市日期	2003-01-07
所属行业	交通运输、仓储业
注册资本	165,861.00
同行业公司数	93
法人代表	王水
职工总数	1081
董事会秘书	谢新宇
公司电话	0551-5338697
董秘电话	0551-5338681
公司传真	0551-5338696
董秘传真	0551-5338696
公司电子邮件	wtgs@anhui-expres...
董秘电子邮件	wtgs@anhui-expres...
公司网址	www.anhui-express...
报告放置地点	公司本部
信息披露网址	www.sse.com.cn
注册地址	安徽省合肥市长江西...
境内会计师事务所	普华永道中天会计师...
经营范围	高等级公路建设、设...

SUBJECT	PREDICATE	OBJECT
#上海证券交易...	#type	#证券交易所
#上海证券交易...	#type	#市场参与者
#上海证券交易...	#type	#机构
#上海证券交易...	#type	#中介机构
#上海证券交易...	#type	#Resource
#安徽皖通高速...	#公司名称	安徽皖通高速...
#安徽皖通高速...	#成立日期	1996-08-15
#安徽皖通高速...	#上市地	#上海证券交易...
#安徽皖通高速...	#上市时间	2003-01-07
#安徽皖通高速...	#注册资本	165,861.00
#安徽皖通高速...	#注册地	安徽省合肥市...
#安徽皖通高速...	#经营范围	高等级公路建...
#安徽皖通高速...	#经营状况	安徽皖通是由...
#安徽皖通高速...	#type	#股份有限公司
#安徽皖通高速...	#type	#公司
#安徽皖通高速...	#type	#Resource
#交通运输、仓...	#type	#板块
#交通运输、仓...	#type	#Resource
#安徽皖通高速...	#所属板块	#交通运输、仓...
#安徽皖通高速...	#所属公司	#安徽皖通高速...
#安徽皖通高速...	#type	#股票
#安徽皖通高速...	#type	#金融工具
#安徽皖通高速...	#type	#Resource

The left figure shows an excerpt of a Web page associated with the basic information for "安徽皖通高速公路股份有限公司" (Anhui Expressway Company Limited) retrieved from Yahoo China. Besides, a screen snapshot for our interface after the knowledge extraction from this Web page is also shown in this figure. Given the similar documents and domain ontologies, we can get raw information by pressing the *Extract* button. When *Align Names* button is pressed, WebKER maps the resources' original names to the ones predefined in domain ontologies and creates a new version of the XML file to describe this corporation. In the above figure, after mapping, the names with the orange background in the bottom left table are replaced with the ones with the blue background in the bottom right table. Then the new XML file is used for extracting the knowledge based on domain ontologies. After being linked and merged, these knowledge are added to the KB finally.

[1] CRAB is a project aimed at providing an integrated multifunctional system that can crawl Chinese Web documents, extract knowledge, and finally provide users with human-readable reports.