Research on browse tendency of web users based on Web data

Cao Zhen Jie, He Yuan, Li Yi Hang

Beijing University of Posts and Telecommunications, Beijing102209

huner2011@foxmail.com

Keywords: XML, Information extraction, XSLT

Abstract. There is one problem that to be solved based on rapid development of Web, that is how to obtain the information required through it, therefore, it is very necessary to accomplish information extraction. Extraction from web pages adopts Wrapper to finish the best architecture of Wrapper in accuracy, robust degree as well as universal property so that to prevent influences of different website architectures and page structures on it, at the same time reduce human involvement to the maximum extent. This is a problem that must be solved in research on information extraction. This thesis puts forward a Web information extraction platform based on XML technology and conducts search and recognition of users browse tendency through application of inductive learning algorithm.

Introduction

Massive and dynamic change and isomerism are characteristics of the Internet, therefore, Web information extraction is more difficult, which is different from traditional information extraction. Specifically, firstly, how to deal with the enormous amount of information automatically and efficiently aiming at geometric growth of Web information in information space; secondly, how to realize correct recognition of information point required aiming at Web page in isomerism; thirdly, how to keep adaptability of information extraction aiming at real-time dynamic updates of Website.

Based on continuous increase of demand, extraction technology has also been enriched, methods of information extraction are various home and abroad, based on different extraction principals and methods mainly including induction method based on wrapper, structure method based on HTML, treatment method based on natural language as well as method based on ontology. Intuitively, emphases of these methods are different in solving information extraction, results obtained are good generally but not so improved, that is, there are more or less deficiencies or disadvantages.

To better deal with many problems confronted with Web information extraction, it is very necessary to conduct further intensive research on Web information extraction.

Web information extraction based on HTML structure

Positioning of the information based on Web page structure is the character of this information extraction technology. First, accomplishing analysis of syntactic tree on Web documents through the resolver, producing extraction planing with automatic or semi-automatic methods and realizing information extraction through operating the syntactic tree. Lixto(commercialized), XWRAP(non commercialized),RoadRunner and SG-WRAM as well as W4F are all using this technology.

Lixto: the user could mark the information in sample page through visual and interactive method aiming at this system, the system produces rules of extraction through “system default” or “user customization” methods based on records of marked information by the user to realize information extraction of similar pages(for instance definition 1 has explained “similar page” here). Figure 1 is the system diagram of Lixto, the user adds semantic information of the system into sample learning stage, firstly the user could define the mode through visual user interface, at the same time express complex structure of semantic pattern, the storage form of data extraction is XML document. This system could simplify the steps of information extraction to some extent and strengthen the practicability of information extraction technology. However, this system is difficult to realize

realization and optimization(Elog language describes its extraction rules based on Datalog), at the same time the extraction rules have not fully described the information extracted. While two methods of production of extraction rules have their respective deficiencies: “system default” method has high level of automation, however, the robustness is not so good enough; “user customization” method has low level of automation and requires the user with certain standard, improper operation will also influence the robustness of rules.

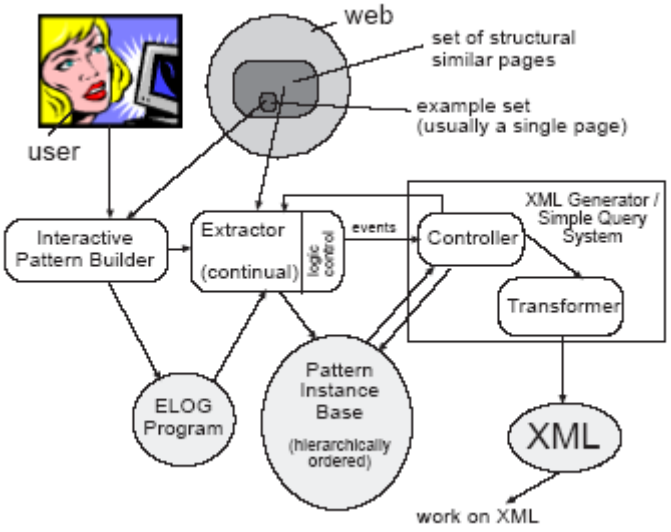

Fig. 1 Diagram of Lixto implementation

Knowledge base and database in platform

Constructing domain knowledge base. Functions of domain knowledge base include mainly as follows: 1) Providing the user with query and navigation services to make the user not feel confused at beginning. The method is to add some URLs of important websites into corresponding fields.
2) Providing supports for management of rules in logic and method, the method is to store the extraction rules by classification according to sub-fields.

Domain in this thesis refers to specialized websites which publish similar information, domain knowledge base is the basic concept, property, entity as well as rules included in the domain where the information contained. For instance, English books website publishes information on books. Its domain knowledge base includes various basic concepts and properties. This thesis regulates that various domains in domain knowledge base form a hierarchical tree according to the affiliation, the root is virtual which could also be referred to as “root domain”. For instance, different sub domains which are subject to the same books domain “China-Pub” and “wikipedia” are located in one layer under the root domain.

Extraction rules library. Extraction rules library restores extraction rules that have been learned and the extraction rules are the knowledge of recognition mode to be extracted. Rules adopted are different for different domains and websites, with the operation of extraction system, it would produce many rules and the system needs one library to restore these rules naturally. When the system needs information extraction, first it could search in the rules library to see if there are rules in repeated use, if yes, it could extract the corresponding rules directly from the rules library and it is not necessary to reproduce new rules for similar websites or pages.

Extraction result database and Web page database. 1) As previously mentioned, the result finally extracted in this thesis is the XML document that includes the information point that interests the user. Pages restored in the extraction result database are these XML pages. Native XML DataBase, also called XML source database is specially designed to restore XML documents database, it restores XML documents in the form of its own form, the difference from other databases is that its internal model is basically the XML format. 2) Web page database restores source documents taken from Web pages. This part has been processed by the system including page documents that have been cleaned, repaired as well as marked.

Positioning of information bars. After obtaining the route assemblage of information block to be extracted, the information extraction has changed into extraction of assemblage of internal information block.

The similar patterns in information block in this thesis is called “format”, each piece of information in the information block can have this format. Furthermore, in similar information block, there is always certain character which could guide positioning of information block to be extracted. Therefore, preorder traversal of subtree of DOM tree contained in the information block, it could obtain Xpath expression of information block to be extracted.

Production of XSLT. Combining with the positioning information in information block and information bars in the block, next it only requires combination of the above-mentioned positioning information, forming XSLT document according to Xpath on each node. This XSLT document is the extraction rule.

Due to information structures in respective page in respective website are not different, the order of its information points may not conform to requirements of the users, in addition, certain information in the information block is not the information that interests the users, therefore, before the extraction rule has been formed, it requires a mapping file containing format and content that conforms to requirements predefined by the user. We could write extraction rule XSLT documents referring to this mapping file. In this example, the information to be extracted in the information block are TITLE, AUTHOR, PUBLISHER as well as PRICE, which is shown in the result document according to this order.

Description of information extraction procedure

When the extraction rule XSLT document has been obtained, it shall construct one wrapper conducting information extraction to execute this XSLT only. This thesis uses Xalan-J as the execution engine of XSLT.

The extraction result is shown in Fig. 2:

```
- <books searchkeys="java">
  - <book>
      <TITLE>ADVANCED JAVA HOW TO PROGRAM 1/E</TITLE>
    - <AUTHORS>
        <AUTHOR>DEITEL</AUTHOR>
      </AUTHORS>
      <PUBLISHER>PEARSON EDUCATION</PUBLISHER>
      <PRICE>$109.00</PRICE>
    </book>
  - <book>
      <TITLE>ENTERPRISE JAVABEANS 3E</TITLE>
    - <AUTHORS>
        <AUTHOR>MONSON-HAEFE</AUTHOR>
      </AUTHORS>
      <PUBLISHER>O'REILLY ASSOCIATES</PUBLISHER>
      <PRICE>$120.00</PRICE>
    </book>
  - <book>
      <TITLE>JAVA COOKBOOK</TITLE>
    - <AUTHORS>
        <AUTHOR>DARWIN</AUTHOR>
      </AUTHORS>
      <PUBLISHER>O'REILLY ASSOCIATES</PUBLISHER>
      <PRICE>$120.00</PRICE>
    </book>
  - <book>
      <TITLE>JAVA CRYPTOGRAPHY</TITLE>
    - <AUTHORS>
        <AUTHOR>KNUDSEN</AUTHOR>
      </AUTHORS>
      <PUBLISHER>O'REILLY ASSOCIATES</PUBLISHER>
      <PRICE>$31.95</PRICE>
    </book>
  + <book>
</books>
```

Fig. 2 Result extracted by Web

Summary

Development of Internet makes the amount of information online has increased rapidly, Web has become the major channel for humans to acquire information, however, difficulty in searching for useful information has also increased daily for human beings, useful information is always submerged in enormous amount of irrelevant information. To acquire information required rapidly and accurately, information extraction technology on Web has played its important role gradually.
Appearance and development of XML technology are making up deficiencies in HTML language. The largest character of XML is its marks are semantic and can be defined by users and reflect implication of data. In addition, to some extent, XML is a kind of semi-structure data model and could show a large amount of semi-structure data on Web. Furthermore, XML is more operable than HTML, which could support more accurate search request. Therefore, XML provides important supports for data extraction and other applications on Web.
Reference:

References

[1] Randolph E.Bucklin and Catarina Sismeiro. A Model of Web Site Browsing Behavior Estimated on Clickstream Data. The Anderson School at UCLA

[2] P.Baldi, P.Frasconi and P.Smyth. Modeling the Internet and the Web：Probabilistic Methods and Algorithms. Published by John Wiley & Sons,Ltd. 2003,Chapter1、Chapter2、Capter7

[3] Wendy W.Moe and Peter S.Fader. Dynamic Conversion Behavior at e-Commerce Sites. 2003.3

[4] Alan L.Montgomery. Using Clickstream Data to Predict WWW Usage. 1999.8

[5] Mandel Naomi and Eric J. Johnson (1999), “Constructing Preferences Online: Can Web Pages Change What You Want?” , Working Paper, The Wharton School, University of Pennsylvania

[6] Moorthy S.B.T. Ratchford and D.Talukdar (1997) “Consumer Information Search Revisited: Theory and Empirical Analysis,” Journal of Consumer Research, 23(4), 263－277.

[7] Catarina Sismeiro and Randolph E.Bucklin. Modeling Purchase Behavior at an E-Commerce Website：A Conditional Probability Approach. 2002.3