An Information Concierge for the Web

Feifei Li Zehua Liu Yangfeng Huang Wee-Keong Ng
Centre for Advanced Information Systems, School of Computer Engineering
Nanyang Technological University, Singapore 639798, SINGAPORE
awkng@ntu.edu.sg

Abstract

*WWW Information Collection, Collaging and Programming (*Wiccap*) system is a software system for generation of logical views of web resources and extraction of the desired information to a structured document. It is designed to enable people to obtain their interested information in a simple and effective manner as well as to make information from the WWW accessible to applications, in order to offer automation, inter-operation and Web-awareness among services. A key factor in making this system useful in practice is that it provides tools to automate and facilitate the process of constructing the logical representation of Web Sites, defining the interested information and subsequently retrieving them. In this work, we present the design of the* Wiccap *system and its two main components, namely Mapping Wizard and Network Extraction Agent.*

1 Introduction

In the past couple of years, the World Wide Web has completely reshaped the Internet. The Web has been so successful in information delivering and sharing that people are getting used to searching for information that they are interested from WWW. The information available in the Internet keeps expanding in a surprising speed. Users don't face the problem of information unavailability in the net. Instead, they are challenged by the problem of information overloading.

In the search for answers to the issues mentioned above and to investigate the effects of agent technology on network content extraction and filtering areas, the Wiccap project was started in July 2000 at the School of Computer Engineering, Nanyang Technological University, with the objectives of designing and implementing a software infrastructure for a large scale, distributed, open agent-based Intelligent Information Retrieval (IIR) applications. The approach taken by this project is to introduce two agent modules to automate the process of content extraction from Internet. These two modules are Mapping Wizard, and Network Extraction Agent. The combination of the two will provide users with a robust, reliable, effective Information searching, filtering and extracting software application. The information retrieved will be also re-processed to a structured XML document, which make those data accessible not only to users but also applications.

2 Related Work

This section briefly reviews the related systems from other commercial and non-commercial organizations to identify useful features that could be incorporated into the Wiccap system.

2.1 WysiWyg Web Wrapper Factory (W4F)

W4F (World-Wide Web Wrapper Factory) [8] [7] is a toolkit to generate Web wrappers. These wrappers consist of three independent layers. The retrieval layer is in charge of fetching the HTML content from a Web data source. The extraction layer extracts the information from the document. The mapping layer's role is to specify how to export the data. These layers are working together as a network extraction toolkit. At the end it will produce a java object based on the content extracted from web data source.

The toolkit consists of an HTML parser that generates parse trees out of HTML pages (using various heuristic to handle ill-formed pages), a compiler to produce Java code for each layer and various visual wizards to assist user in writing the specifications.

2.2 Network Query Language (NQL)

Network Query Language [6] is a tool for rapid and simple development of intelligent agents, bots, spiders, middleware and scalable business to business content aggregation applications. Network Query Language does for network programming what SQL did for database programming, creating a unified layer across conflicting network standards and software.

NQL gains certain popularity as it is capable of extracting data from web pages in a scripting manner. But it also has its limit as users have to get familiar with the scripting language itself. And it's difficult and troublesome for users to write scripts for every web page that they are interested in order to perform information retrieval.

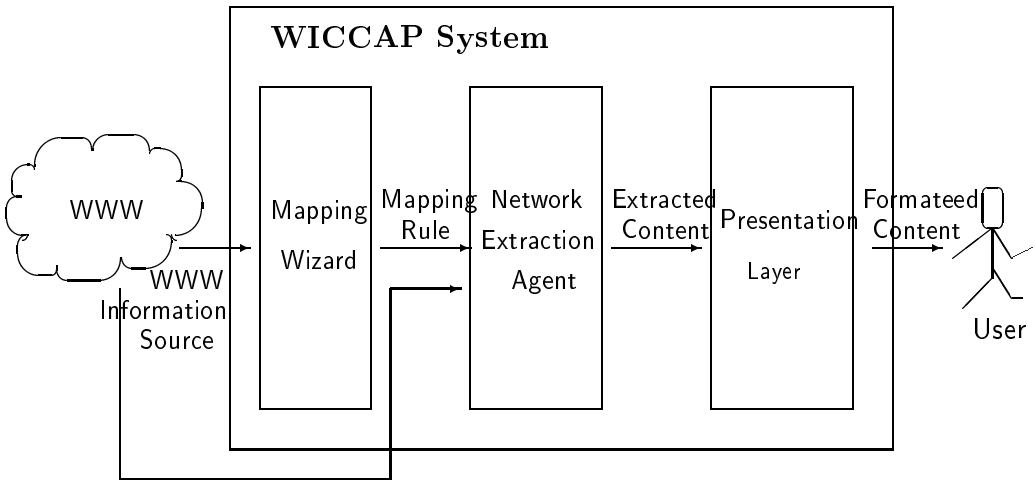

Figure 1: Architecture of the WICCAP System.

2.3 ARANEUS

The model-based web site development in ARANEUS means that all aspects of the web site design process, namely managing data, hypertext and presentation, are based on the adoption of suitable models.

ADM data model gives a compact, intentional description of a site structure at an abstract level, and provides a basis for reasoning about the effectiveness of the chosen hypertext organization. ARANEUS projects has enlighten to us that a data model for mapping a web site logically is essential to a successful network content extraction system, such as WICCAP system.

3 WICCAP System Overview

The aim of the WWW Information Collection, Collaging and Programming (WICCAP) project is to design and implement a software system to streamline and facilitate the process of content extraction and presentation from the World Wide Web. In this section, we present an overview of the WICCAP system architecture and propose a data model for mapping a given web resource from its physical view to a logical one.

3.1 WICCAP Architecture

Figure 1 shows the three main components of the architecture of the WICCAP system: *Mapping Wizard*, *Network Extraction Agent*, and *Presentation Toolkit*. These three tools together form a channel, through which the information in the WICCAP system flows. The intermediate outputs resulting from each layer of the system are stored in XML format, which makes it possible for other applications or agents to make full use of these information. The final output, Formatted Contents, to the user could be of any format, ranging from static HTML Web Page to animated Flash clip.

We will present the these three layers in the remaining sections. But before that, we will first define the data model for the intermediate outputs between the layers.

3.2 WICCAP Logical Representation

To extract data from a website, we have to first derive a logical representation from the physical organization of the website. The use of a specific data model is central in our approach.

3.2.1 WDM and Mapping Rule

The WICCAP *Data Model* (WDM) defines the basic data elements and how these elements are structured to form a logical view of a website. It concerns not only about a single web page but a set of web pages, which we call it *WWW Information Source*.

The WDM provides a set of basic elements that define the basic data structure and at the same time allows extension to these basic elements to achieve a more specific and accurate modelling. The WICCAP system categorizes web sites into different types, such as Online Newspaper, Digital Library, Product Category, and Stock Information. For each type of website, a specific WDM is defined.

A *Mapping Rule* of a WWW Information Source refers to a specific XML file that describes that web resource's logical structure using the WDM of the type of that web source. Mapping Rules can be viewed as instances of a specific WDM, analogous to the Object-Class relationship in Object-Oriented concept.

The final logical view of a web site is usually a tree structure, as shown in Figure 2.

In the WICCAP system, WDM is defined using XML Schema and Mapping Rule as normal XML document. The

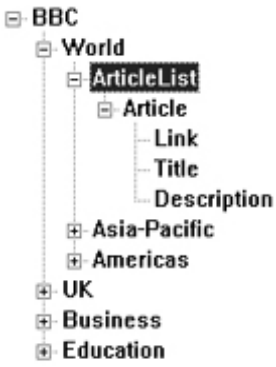

Figure 2: Logical View of BBC Online News

WDM defines how Mapping Rule should look like and hence the structure of the logical representation.

3.2.2 WDM Elements and Construction

WDM defines several general basic elements, including *Locator*, *Link*, *Form*, and *Mapping*.

A *Locator* is the fundamental element of WDM, which helps to locate a portion within a webpage. A *Link* can be *static*, a simple fixed URL, *dynamic*, which is obtained by using a Locator, or *form*. *Form* is defined to cater for a special category of HTML tags: FORM and other related HTML elements.

Locator allows us to navigate anywhere within a single page, while Link enables us to jump from one page to another. *Mapping* combines these two basic elements to enable navigation throughout a whole web resource. Mapping allows a WDM element to jump through multiple levels of links and essentially makes it possible to represent the logical structure instead of the physical structure defined by the hierarchy of level of links.

Every main element, which is shown in the logical tree view as a node, should have a Mapping as its attribute to indicate how to get to this node in the physical website.

Based on all the basic elements, we can define specific WDM for different types of web sites. Some new elements that are specific to the type of website that we are interested in have to be defined first. We then organize all the defined elements into certain hierarchical structure. This is part of what we mean by *Collaging*, as the second *C* in the word WICCAP. Besides constructing the hierarchy, we also associate attributes with elements.

4 Mapping Wizard

The aim of the Mapping Wizard is to facilitate and automate the process of producing a Mapping Rule for a given WWW Information Source, typically a website.

The architecture of the Mapping Wizard consists of four components, or stages: Basic Logical Structure Construction, Content Extraction Definition, Mapping Rule Generation, and Testing and Finalization.

Basic Logical Structure Construction For everything to start with, we have to supply a starting addressing of the WWW Information Source, which is usually the home URL of a website, such as "http://news.bbc.co.uk" or "http://www.acm.org/dl". We then decide on the type of the website and select an appropriate WDM for that website. The Wizard will automatically identify the structure that the website would follow and provide online assistance to us. We can then figure out the basic logical organization either using our knowledge about the website or by navigating through the website.

The Mapping Wizard provides a set of tools for manipulation of the logical tree to further simplify the construction of the tree. With the help of these tools, a tree that represents the basic organization of the website can be quickly built up.

Content Extraction Definition In this stage, we have to supply the Mapping Wizard the necessary mapping information for it to link the logical structure to the physical website. Typically, this includes specifying the Mapping attributes of various main tree nodes.

The Mapping Wizard is equipped with the ability to figuring out the mapping information with certain human guidance, based on a set of predefined heuristics. By providing certain feedback information, we are able to extract the mapping information to associate the logical tree with the actual website quickly and accurately.

As we have seen in the sample Mapping Rule earlier, an Item tells a piece of information of our interest. The Mapping Wizard comes with some assistant tools to help to identify these pieces of information, in particular various means of specifying how to locate the Items. It also provides Graphical User Interface (GUI) for modifying the properties that have been specified. We can fine tune or optimize the mapping information. Quick manipulation of properties among similar type of nodes is provided by applying the "copy and paste" semantics.

Mapping Rule Generation Now that we have all the information ready, and what we need to do is to simply click on the "Generate" button. The Mapping Wizard will generate the Mapping Rule according to the tree structure and all the tree nodes' properties. This process is fully automated. The Mapping Wizard will validate all the information and produce the Mapping Rule according to the syntax defined in the WDM.

Testing and Finalization Once the Mapping Rule is generated, the Mapping Wizard can perform the actual extraction using the generated Mapping Rule and show the result to the user for verifying the correctness of the Mapping Rule. Debugging information will be provided, if any error occurs.

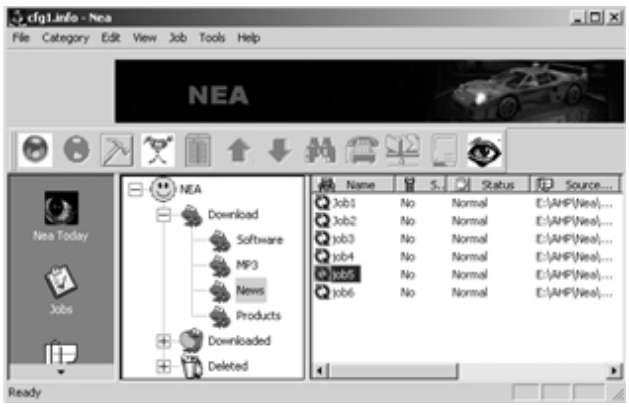

Figure 3: Network Content Extraction Agent

If we are sure that the Mapping Rule that we defined represents the logical structure of the web site correctly, we can finalize the Mapping Rule. We can either put the finalized Mapping Rule into a Mapping Rule repository or deliver it to the user of the Network Extraction Agent.

5 Network Extraction Agent

The Network Extraction Agent(NEA) is the agent component in the WICCAP system that is responsible for managing extraction jobs and retrieving information from web sites based on the Mapping Rules defined by Mapping Wizard System. It also provide the post-processing features. NEA stores the retrieved information into XML documents complied with WICCAP Information Storage Data Model.Below is the screen shot of the NEA system.

5.1 Data Model and Mapping Rule

We have introduced the fundamental concepts in WICCAP system,namely, Data Model and Mapping Rule. NEA system will retrieve information from Mapping Rule using XML Parser. Based on the information from Mapping Rule, NEA system is able to perform the extraction smoothly.

For example, NEA system could get the base link to the web site from Mapping Rule by querying the node listed below:

```
<Mapping>
  <Link Type="Static">http://news.bbc.co.uk</Link>
</Mapping>
```

The dynamic link for a web site could be located in the web site on the fly by querying its mapping pattern:

```
<Mapping>
  <Link Type="Dynamic">
    <Locator Type="ByPattern">
      <LocatorPattern>
        <BeginPattern>
          <![CDATA[<TD CLASS="rootsection"
            BGCOLOR="FFFFFF"><A HREF="]]>
        </BeginPattern>
        <EndPattern>
          <![CDATA[" CLASS="index">]]>
        </EndPattern>
      </LocatorPattern>
    </Locator>
  </Link>
</Mapping>
```

After querying the begin pattern will be

```
<TD CLASS="rootsection" BGCOLOR="FFFFFF"><A HREF="
```

and the end pattern will be `CLASS="index"`. With these information the actual dynamic link could be located in the corresponding HTML page easily.

Similarly NEA could process the Mapping Rule in a recursive manner to retrieve information dynamicly from website as specified by the Mapping Rule.

5.2 Logical View of the Web site

It's impossible to perform good extraction directly on the physical view of web site. In NEA system a logical view of the web site being concerned is constructed from the Mapping Rule. A logical view of the web site here represents the user defined structures for the selected portion from the web site. In this logical view user will see parts from web site that are only to his interest. The parts are organized into a hierarchy tree structures. The whole tree will be traversed using Depth-First Algorithm.The algorithm allows recursive traverse which increase the power of the logical view dramatically.

5.3 Supporting Form Process

The Data Model in Mapping Wizard has defined an element for form to provide necessary information about the form from the web site.With these information NEA will construct a dialog box in run time with respective windows controls. For example, the radio buttons in web site's form will be represented by radio group control; the drop down list will be represented by combobox control;etc. The form element in Mapping Rule will provide the action of the form. NEA will get user's input from the dynamic dialog box and then construct the complete query strings. With the action specified NEA then could simulate a HTTP post request to the right server program. The response from the server program will be processed as normally.

5.4 Information Storage

The information retrieved by NEA is stored into structured XML document according to Information Storage Data Model (ISDM). ISDM defines several general basic elements, including *Item Record*, *Section*.

Figure 4: WIPAP System

A sample XML document which stores the information extracted from BBC web site has been shown. This document was generated by Network Extraction Agent based on the Mapping Rules. The document strictly follows the ISDM, hence those information stored in the document are easily accessible by any other applications.

```
<?xml version="1.0" encoding="UTF-8"?>
  <Wiccap Name="BBC" Group="News">
    <Section Name="World" NumOfRecord="3">
      <Record NumOfItem="3">
        <Item Type="Link">
            /hi/english/world/middle_east
                /newsid_1106000/1106019.stm
        </Item>
        <Item Type="Title">
          Mid-East deal 'unlikely'</Item>
        <Item Type="Description">
          Washington says the prospects of a
          Middle East peace deal before President
          Clinton leaves office are slim, but he
          has vowed to keep trying.
        </Item>
      </Record>
      ...
      <Section Name="Asia-Pacific" NumOfRecord="3">
      ...
    </Section>
    <Section Name="Business">
    ...
</Wiccap>
```

6 Presentation Toolkit

The Web Information Programmer And Player (WIPAP) is the software component in the WICCAP system to present the extracted information to the end users. One objective of WICCAP system is to enable fast and effective information access. Once the desired information have been extracted into XML documents in ISDM format, user could view these information through WIPAP. WIPAP has the features so that program the viewing process becomes possible. Flash has been used to present the information in animated and appealing ways. User could use WIPAP system to create his own program to view the information. There are several flash templates available in the WIPAP system for user to choose from. More templates are possible to be added in at runtime. Below is a snapshot of the WIPAP system.

7 Conclusions

WICCAP project looks at the Information Overloading problem and focuses on Intelligent Information Retrieval area. The aim of WICCAP project is to provide a simple yet powerful software tools for Network Content Extraction.

In the following phase, we will keep refining the data models in WICCAP system and adding more functionality into main modules. In particular, enhancement on intelligence of the Mapping Wizard and post-process capabilities of the Network Extraction Agent will be the major concern of the project in the coming days.

References

[1] Naveen Ashish and Craig A. Knoblock. Semi-automatic wrapper generation for internet information sources. In *Proccedings of the Second IFCIS International Conference on Cooperative Information Systems*, Charleston, SC, 1997.

[2] Alper Caglayan and Colin Harrison. *Developing Intelligent Agents for Distributed System.* Computing McGraw-Hill, 1998.

[3] Paolo Atzeni Giansalvatore Mecca and Paolo Merialdo. To Weave the Web. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, pages 206–215, Athens, Greece, 1997.

[4] Giansalvatore Mecca and Paolo Atzeni. Cut and Paste. In *Proceedings of 16th ACM SIGMOD Symposion on Principles of Database Systems*, pages 144–153, Tucson, Arizona, 1997.

[5] BBC Online News. http://news.bbc.co.uk/.

[6] NQL: Home Page. http://www.nqli.com/.

[7] Arnaud Sahuguet and Fabien Azavant. Looking at the Web through XML glasses. In *Proceedings of the Fourth IFCIS International Conference on Cooperative Information Systems*, Edinburgh, Scotland, September 1999.

[8] Arnaud Sahuguet and Fabien Azavant. WysiWyg Web Wrapper Factory (W4F). *Proceedings of WWW Conference*, 1999.