MetaNews: An Information Agent for Gathering News Articles On the Web

Dae-Ki Kang[1] and Joongmin Choi[2]

[1] Department of Computer Science
Iowa State University
Ames, IA 50011, USA
dkkang@cs.iastate.edu
[2] Department of Computer Science and Engineering
Hanyang University
Ansan, Kyunggi-Do 426-791, Korea
jmchoi@cse.hanyang.ac.kr

Abstract. This paper presents *MetaNews*, an information gathering agent for news articles on the Web. MetaNews reads HTML documents from online news sites and extracts article information from them. In order to recognize and extract articles from an HTML document, MetaNews removes redundant HTML tags, builds *a reference string* which is a sequence of semantic components of a noise-removed document, and performs pattern matching between the reference string and each of predefined *information patterns* for articles. With a few training inputs from the operator with intermediate-level skills, MetaNews is capable of adding new sites easily and extracting articles in real time. By reducing the complexity of designing and creating wrapper interfaces, the proposed techniques in MetaNews are useful for many information-mediator applications such as meta-search engines, information-push solutions, and comparison-shopping agents.

1 Introduction

The World Wide Web has been growing rapidly, and as a result, it is becoming more difficult to find the right information that users really need. The main issue that is raised commonly from many applications such as online shopping and meta-search systems is how to integrate semi-structured and heterogeneous Web information sources and provide a uniform way of accessing them. *Wrapper* has been suggested for this kind of integration[7, 8].

Most wrapper interfaces have been implemented in an ad-hoc way that the knowledge about information sources is obtained manually and hard-coded into the program. In this approach, however, whenever the information structure of a site is changed, the corresponding wrapper should be rewritten accordingly. The modification of a handwritten wrapper is not trivial, since it may include the rewriting of program codes that requires at least a moderate level of programming skills. Furthermore, handwritten wrappers are not scalable. That is, since

a handwritten wrapper corresponds only to a single Web information source, adding a new Web site requires building a new wrapper by creating a program code for the site. For these reasons, automated methods for wrapper generation have been suggested[3, 5, 7]. However, most previous studies related to the automatic wrapper generation have some drawbacks in actual applications. First, a small change in the corresponding Web site such as changing text colors or font sizes might affect the wrapper significantly. Second, pattern-matching process based on regular expression is complex. Only a few experts who are knowledgeable on both domain features and regular grammars can build patterns in a regular-expression form for new application domains.

In this paper, we present an efficient method for relaxing the overheads in generating wrapper interfaces and promoting the scalability of adding and changing semi-structured Web information sources. *Effective noise removal* and *fast pattern matching of strings* are the two key features in this method. With these features, information-mediator applications can retrieve the needy information more efficiently in real time, and are immune to small format changes in retrieved documents. To show the effectiveness of our approach, we developed *MetaNews*, an information agent for gathering online news articles that can manipulate over 100 international news sites. By categorizing various news sites in a systematic way, MetaNews provides the knowledge about each site and up-to-date articles. In addition, MetaNews has an interesting feature for ubiquitous Web browsing that transmits collected articles to a PDA so that users can read them even when they are away from their desktops.

The characteristics of MetaNews can be described in four ways. First, the back-end analyzer of MetaNews can focus only on meaningful information through preprocessing that removes noisy HTML tags and annotations. Second, even with a small amount of knowledge, users can find the right information they need. Third, the wrapper interface is not affected by trivial changes in the corresponding Web site. Fourth, it is quite easy to add a new site to the system with only a small amount of additional information. The method is effective not only for the MetaNews agent but also for other information-mediator applications such as meta-search engines, push solutions, and comparison-shopping agents[3].

This paper is organized as follows. In Section 2, we describe the system architecture and the features of MetaNews, focusing on noise removal and pattern matching. In Section 3, we present some empirical results to show the practical merits of MetaNews. Finally, in Section 4, we conclude with a summary and future direction.

2 MetaNews

MetaNews is an information agent for gathering news articles on the Web. The goal of MetaNews is to extract news articles from periodically updated online newspapers and magazines. More specifically, MetaNews collects HTML documents from online newspaper sites, extracts articles by using the techniques of

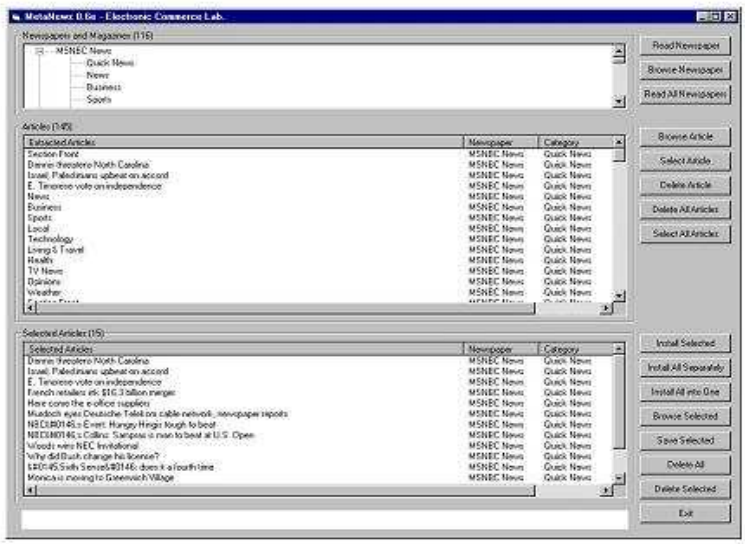

Fig. 1. Main interface of MetaNews

noise removal and pattern matching, and provides the user with the titles of extracted articles and the hyperlinks to their contents.

Figure 1 shows the main interface of MetaNews which has three display windows. The upper window provides a hierarchical view in a way that each top-level node corresponds to a news site and its subnodes denote topic categories. In the figure, the "MSNBC News" site is displayed with its categories including "QuickNews", "News", "Business", and "Sports". The center window lists the extracted news articles. The user can see the titles not the contents, and select interesting articles from several news sites. The selected articles are displayed at the bottom window, where the user can save them to files or install them in a PDA. At any window, the user can see the content of an article by double-clicking its title that invokes an ActiveX-controlled Web browser.

Figure 2 shows the architectural diagram of MetaNews. The control flow of MetaNews can be explained in three different stages. At the *initial configuration stage*, MetaNews is given with the URLs of a site's homepage and its subcategory pages and also with the information patterns for article extraction. This information is used for the addition of new sites. At the *article retrieval stage*, the user selects specific news in the control panel, and then MetaNews gets HTML documents from the selected Web sites in real time. After that, noise removal and substring pattern matching are performed on the documents, and the informative records of articles in the matched documents are extracted. MetaNews has an embedded Web browser with ActiveX control for convenient article display. At the *article saving stage*, the user can select interesting articles and save

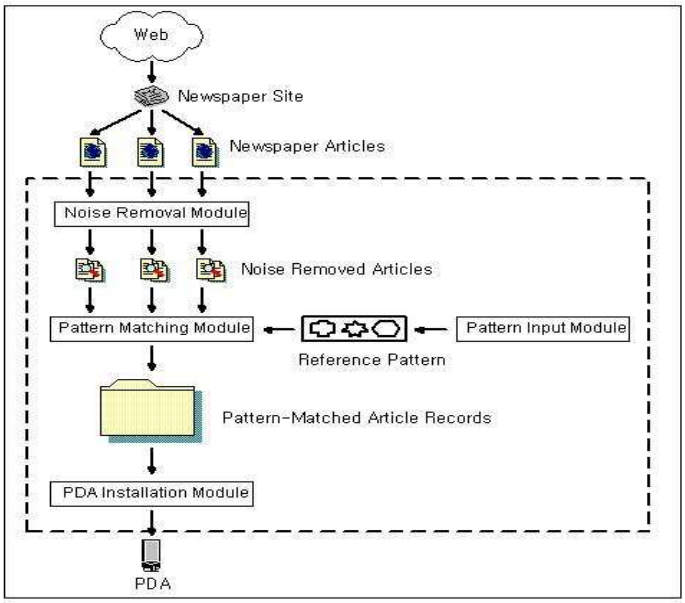

Fig. 2. Architecture of the MetaNews agent

them to files or install them in a PDA. With a minimal editing work, a new site can be easily added to MetaNews.

We now discuss in more detail about the key techniques in our method: noise removal and pattern matching.

2.1 Noise Removal

MetaNews extracts the articles' titles and relevant URLs from the category pages of a news site. Since HTML is mainly used for displays and hyperlinks, most HTML tags are irrelevant to the information content. Although some of these tags may be useful for emphasizing or giving weights to terms, we are certain from the empirical results that removing these tags is not harmful in analyzing the document for the extraction. Furthermore, for the applications like MetaNews in which the priorities of information in the document is not so important and most pieces of information are equally useful and needed, removing tags is always advantageous. This process is called *noise removal*.

A category page of a news site contains the list of news article records, and each article record contains a hyperlink surrounded by some anchor tags. These anchor tags are treated as useful HTML tags, and the URL in each anchor tag of a hyperlink is converted into a simplified canonical form by the anchor tag conversion module. Also, some of table-related tags including `<table>`, `<tr>`, and `<td>` are also considered as useful tags. Other redundant tags are simply removed by the tag removal module.

The noise remover in MetaNews is implemented in Fast LEX (FLEX), a clone of LEX which has more expressive power and generates an efficient binary executable. Each regular expression in the FLEX code can be regarded as a rule/action pair for noise removal purpose. Fig. 3 shows an example of noise-removed part of a document.

```
<TABLE>
<TR>
<TD>
</TD>
<TD>
<A>
HREF=303767.asp
With a heavy heart, crew leaves Mir
</A>
</TD>
</TR>
<TR>
<TD>
</TD>
<TD>
Amid sadness and ceremony, a Russian-French crew left the
13-year-old Mir space station unoccupied and returned
to earth.
</TD>
</TR>
</TABLE>
```

Fig. 3. Noise removed document

Note that the noise removal module makes it possible to use the same wrapper interface for different news sites. This advantage also applies to the situation when the external layout of a news site is changed.

2.2 Pattern Matching

In MetaNews, *pattern matching* is performed to recognize and extract articles from the noise-removed document. Pattern matching is composed of three subprocesses. First, the noise-removed document is converted into a *reference string* which is a sequence of semantic components. Next, the string matching is performed between the reference string and each of the *information pattern* strings. An information pattern (also called a reference pattern) is also a sequence of semantic components that comprises an article. Information patterns for articles are predefined by the operator. Finally, the substrings in the reference string that are matched with any information pattern are extracted and displayed.

To help understand the operation of this module, we revisit the example in Fig. 3 which represents a single record. This record can be converted into a reference string "**TRDdDAUXadrRDdDXdrt**", by applying a simple conversion rule described in Fig. 4 to each line .

In contrast with the reference string that is generated automatically by the conversion rules, the information pattern for an article is built manually. For

`<TABLE>`	→	**T**
`</TABLE>`	→	**t**
`<TR>`	→	**R**
`</TR>`	→	**r**
`<TD>`	→	**D**
`</TD>`	→	**d**
`<A>`	→	**A**
`</A>`	→	**a**
hyperlink (`HREF=...`)	→	**U**
other (*general text*)	→	**X**

Fig. 4. Rules for reference string generation

example, the title of a news article is normally surrounded by hyperlink tags, so we might induce that the information pattern for an article can be "**AUXa**". In this pattern, '**A**' indicates the beginning of an anchor tag (`<a>`), '**U**' is the URL of an article (`href=http://...`), '**X**' is the title text of an article, and '**a**' is the ending anchor tag (`</a>`). With some intuitive heuristics, this simple pattern can be used effectively for locating the title of a news article with the corresponding hyperlink. Since this information pattern is a substring of the reference string for the record in Fig. 3, MetaNews can recognize it as an article. Figure 1 shown in the previous section is the result of applying pattern matching algorithm to the QuickNews section of the MSNBC News.

Note that, in the online shopping domain, each record corresponds to a product description with the product name and the price. In this case, the information pattern for a record can be "**DXd**", which consists of table tags and the text for product description. The information pattern denotes how the desired information is encoded in a noise-removed HTML page, so it may be changed in different domains.

Our pattern matching scheme has advantages in terms of simplicity and scalability. Especially, at the stage of pattern matching, extracting information can be done by just one string matching which is supported by most programming languages. Also, since these information patterns are encoded in a data file instead of being hard-coded into the program, a trained operator can add new sites easily by just inserting less than fifty lines of data including the URLs and the information patterns.

3 Evaluation

The MetaNews agent can extract articles and their hyperlinks from 116 news sites. MetaNews is scalable in a sense that a new site can be added by inserting a few lines of pattern data, even for a large news site. For our actual experiments of reading news, only three information patterns were needed for all cases: "**AUXa**", "**XAUa**", and "**AUaX**".

Since MetaNews does not depend on a morphological analyzer or a stemming program, it does not have the language limit. Therefore, without changing any program code, MetaNews can be used for sites with various languages such as English, Korean, or Japanese.

In order to evaluate the performance of MetaNews, *precision* and *recall* values are estimated by using positive false and negative false data.

Positive false refers to the data that are not news articles but extracted by MetaNews since they contain fragments matched with the information patterns. Examples of positive false are quick links for other sections, cool links, advertisement links, etc. We have tested MetaNews for 116 sites, and Fig. 5 shows the comparison between the number of total data and the number of positive false data gathered by MetaNews for each site. Note, in the graph, that the first 90 sites are the news sites which have the 3-level information structure consisting of the main page, the category pages, and the actual article pages, and the last 26 are the magazine sites which have the 2-level information structure with the main and article pages, without category pages. MetaNews extracts more articles from 3-level sites since the category page plays an important role in pattern matching. Positive false affects the precision value of the system, and MetaNews shows the average precision value of 88% for the news sites, and about 82% for the magazine sites.

On the other hand, *negative false* occurs when MetaNews fails to recognize and extract news articles, and this phenomenon affects the recall value of the system. With just three information patterns, however, MetaNews does not produce any negative false data for all 116 sites, so the recall can be said as 100% in our experiments.

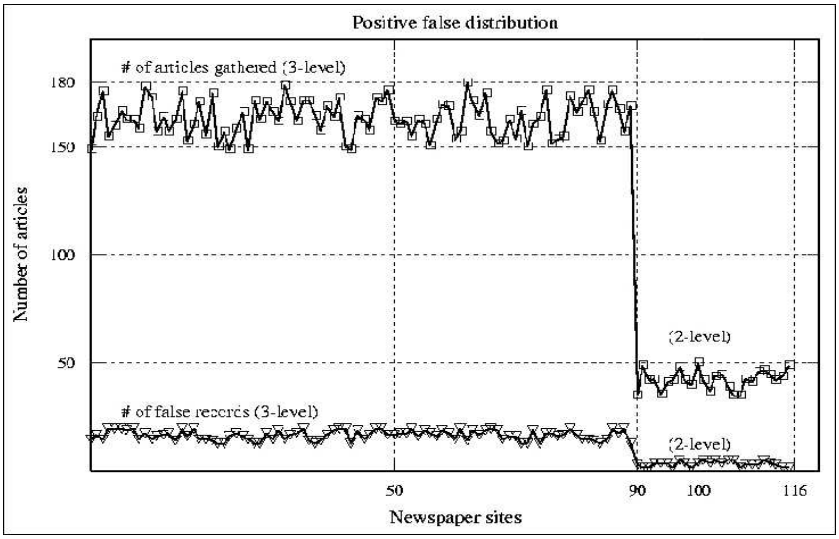

Fig. 5. Positive false distribution on test sites

4 Conclusions

We have presented MetaNews, an information gathering agent for news articles on the Web. With a few training inputs from the operator with intermediate-level skills, MetaNews is able to add a new site easily and extract articles in real time. Our method is simple but powerful enough to reduce the complexity in designing and creating wrapper interfaces, and is useful for many information-mediator applications such as meta-search engines, information-push solutions, and comparison-shopping agents.

Several problems exist for MetaNews that will be handled in the future. First, we need to work out for the removal of false record with post-processing after pattern matching. If the information pattern is not too long or too general, false links can be regarded as news articles. The three information patterns used in this paper can be generalized but it is difficult to set more constraints to our pattern matching approach. A keyword filter can be helpful for reducing the ratio of false records. For example, if the user sets a keyword filter as *Clinton*, only the articles with the title containing *Clinton* will be gathered. Also, a deny list could be maintained in post-processing. Second, we need to search for the *back issues* to extract past articles. MetaNews should be able to deal with the CGI scripts of news sites to achieve this. Finally, porting MetaNews to the conduit mechanism for mobile communication will be an interesting topic.

References

1. D. Angluin, "Interface of reversible languages", Journal of ACM, vol.29, no.3, pp.741–765, 1982.
2. N. Ashish and C. Knoblock, "Wrapper generation for semi-structured Internet sources", Proc. Workshop on Management of Semistructured Data, Tucson, Arizona, 1997.
3. J. Choi, "A customized comparison-shopping agent", IEICE Trans. Comm., vol.E84-B, no.6, pp.1694–1696, 2001.
4. W. Cohen, "A web-based information system that reasons with structured collections of text", Proc. 2nd International Conf. on Autonomous Agent, Minneapolis/St. Paul, Minnesota, pp.400–407, May 1998.
5. R. Doorenbos, O. Etzioni, and D. Weld, "A scalable comparison-shopping agent for the World Wide Web", Proc. 1st International Conf. on Autonomous Agent, Marina del Rey, California, pp.39–48, Feb. 1997.
6. J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, and A. Crespo, "Extracting semistructured information from the Web", Proc. Workshop on Management of Semistructured Data. Tucson, Arizona, 1997.
7. N. Kushmerick, "Wrapper induction: efficiency and expressiveness", Artificial Intelligence, vol.118, pp.15–68, 2000.
8. I. Muslea, S. Minton, and C. Knoblock, "A hierarchical approach to wrapper induction", Proc. 3rd International Conf. on Autonomous Agents, pp.190–197, 1999.