Wrapper Induction of News Information for Feeding to Social Networking Service on Smartphone

Zhong-Liang Xiang*, Xiang-Ru Yu*, Dae-Ki Kang**

*Computer Software Institute, Weifang University of Science & Technology, Shouguang 262-700, Shandong, China

** Division of Computer and Information Engineering, Dongseo University, Busan 617-716, South Korea

ugoood@163.com, yuxiangru1119@163.com, dkkang@dongseo.ac.kr

Corresponding Author: Dae-Ki Kang

Abstract—**In this paper, we propose NewsFeedAndroid, a novel system that interconnects a social networking service and online newspaper sites in order to extracts news articles from the online news sites and to perform feeding of news articles to social network service (SNS) users. In NewsFeedAndroid, news information agents extract news article information from the news and portal sites using Minimum Description Length (MDL) wrapper induction algorithm. The news document collecting module regularly gathers news list information from news list page in the news sites and portals. In the collected documents, the document preprocessing module removes tags that are unnecessary for news information extraction. Lexical analyzer converts the rest text information and tags to a sequence of tokens, and news information is obtained by matching token patterns to the sequence. Those extracted news information from the various sites are integrated in the system and supplied to the end users through the social networking service on a smartphone. NewsFeedAndroid demonstrates a novel usage of integrating social networking services and online newspaper sites.**

Keywords— **NewsFeedAndroid, Minimum description length, Smartphone, Cellphone, Social network service, Wrapper**

I. INTRODUCTION

In Web mining, automated generation of wrappers has been one of important topics [1-9]. Wrapper induction is important because wrappers can bridge between HTML based hypertext pages on the Web and business applications that need useful information from the HTML pages in a structured form.

With the advent of smartphones [10-12] and social networking services(SNS) [13-15], we have seen huge potentials in academic and industrial research stemmed from the fusion between smartphone and SNS.

For example, feeding appropriate news information to the end users over SNS can be an interesting topic, because SNS clients are light-weight and works as an independent mobile application on the smartphone and can be connected to text message service on a phone. It is worth noting that previous research on wrapper induction primarily concern the Web accesses on desktop computers, and the previous applications usually work on client program or Web browsers on desktop computers. For the appropriate information delivery, it is necessary to have effective wrappers that interconnect data flows from the Web to SNS end users over smartphone. There have been considerable amount of research on application of wrapper induction [16-19], however there are no applications so far on wrapper induction for smartphone and SNS.

With these backgrounds, we propose an efficient and accurate wrapper induction technique for news article extraction that elicits concise but accurate patterns that occur frequently in the HTML pages using minimum description length (MDL) principle [20] and a suffix-tree sequence storage mechanism.

To induce accurate and concise wrapper patterns from Web pages, our proposed algorithm uses MDL principle as a tradeoff criterion between the number of occurrence of important patterns and the length of the patterns. The estimation of the occurrence is efficiently calculated by and obtained from suffix tree storage mechanism.

The remainder of this paper is as follows: Section 2 describes related work; Section 3 explains the methods; and Section 4 summarizes with conclusion.

II. RELATED WORK

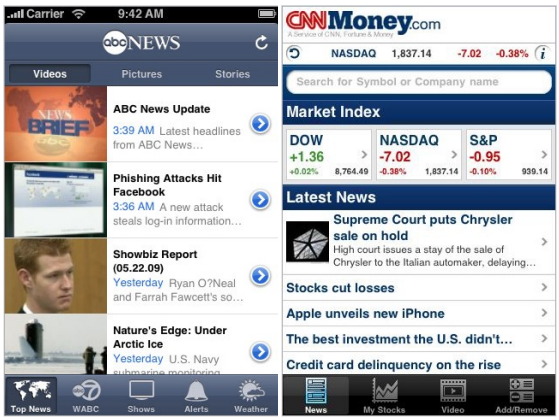

Figure 1. News feeding systems on Smartphone

As for news feeding on the smartphone, there have been several systems available (two examples shown in Figure 1). However, there have been no smartphone applications available that can aggregate news articles from multiple sources using wrapper induction techniques.

We explain related work on wrapper induction techniques. In [1], Kang and Choi developed MetaNews that uses noise removal and string matching algorithm for hyperlinks of

newspaper articles. Kushmerick [2] evaluates efficiency and expressiveness of six wrappers, and compares the results with PAC learning models. Senellart et al. [3] present a wrapper incuction algorithm that uses domain knowledge expressed as a set of concept names and concept instances. Pazienza et al. [4] propose CROSSMARC, an application of ontological knowledge in their wrapper induction, but their experimental results are not so sufficient and systematic to assess the effectiveness of their approach. Muslea et al. [5] propose STALKER, a hierarchical information extraction system that induces rules as finite automata from the data by repeating candidate generation and refinement. Doorenbos et al. [6] introduce Jango, one of the first successful wrapper induction systems that collect shopping information with prices from online stores. Cohen [7] designs WHIRL, a deductive database system for information retrieval with database-like query that can integrate information from multiple Web sites. Hammer et al. [8] propose TSIMMIS, a template based wrapper system with parser, matcher, and engine that transforms native query to application query and vice versa. Liu et al. [9] design and implement XWrap, an XML-based wrapper generator with structure analyzer that handles HTML tree. We have not aware of any applications that incorporate wrapper induction technology on smartphones.

III. METHOD

We explain NewsFeedAndroid. Firstly, we describe tokenizer with removal of needless tags. Secondly, we analyze Minimum Description Length (MDL) principle with respect to wrapper induction. Thirdly, we detail suffix tree data structure. Finally, we explain MDL-Wrapper algorithm.

A. Tokenizer with Noise Removal

Because HTML web pages are generated for visual presentation purpose, the pages mostly have HTML tags used for decoration of their contents. Although, the tags are for visual decoration, a few of them sometime imply the systematic structure of the contents. For example, <*table*>, <*th*>, and <*td*> tags are for tables, but those tags strongly imply that they are used to present relational data, i.e. multiple records.

With these considerations, we remove most of these HTML tags which are not suitable for extracting wrapper patterns [1]. This removal (or abstraction) stage also makes the wrapper robust to small changes of HTML documents, because most HTML tags for visual presentation are removed and disregarded in wrapper induction.

After removing the HTML tags for visual presentation, we generate one token sequence from one HTML document. In the sequence, each token denotes a HTML tag or text data.

After the generation of a token sequence, the problem of wrapper induction is reduced to the problem of finding a token sub-sequence that, when converted back to the original tags and text, covers relational data in the HTML page.

Note that, from the rules in table 1, we treat all general texts equally (as 'X'), but maintain the HTML tags for table structure because they provide significant clues for relational data.

TABLE 1. TOKEN CONVERSION

Token	Character
<TABLE>	T
</TABLE>	t
<TR>	R
</TR>	r
<TH>	H
</TH>	h
<TD>	D
</TD>	d
<A>	A
</A>	a
<P>	P
</P>	p
 	B
</BR>	b
hyperlink (HREF=…)	U
text	X

B. Suffix Tree

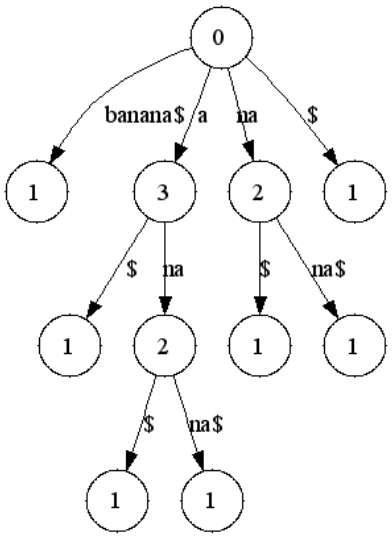

Figure 2. Suffix Tree

To store the generated sequences by the tokenizer, we use suffix trees. Figure 2 shows an example suffix tree for a string "banana$", where '$' denotes the end of the string. A number in each node represents a number of occurrences of patterns. For example, in the string "banana\$", 'a' occurs three times and 'na' occurs twice.

As for its complexity, Ukkonen[21] has devised a linear time algorithm for constructing the suffix tree. When the length of a string is n, then it takes a linear *O(n)* time to build a suffix tree for the string. Once a suffix tree is generated, then it takes *O(m)* time to find a pattern with length m. Also, with edge label compression, it only needs *O(n)* space for a suffix tree. In practice, to store multiple strings, a generalized suffix tree is used.

C. Minimum Description Length (MDL) for Wrapper Induction

With minimum description length (MDL) principle [20], our wrapper (MDLWrapper) controls the extent of generalization during the wrapper induction. MDL is a criterion that trades off the accuracy and the size of a theory. That is, MDL principle chooses the theory (or model) from a set of data that minimizes (1) the sum of the length of the theory and (2) the length of the data that are encoded according to the theory. In our setting, the theory is an inferred pattern of the wrapper, and the data is the tokenized sequence converted from an HTML document and stored in a generalized suffix tree.

Following those ideas, we briefly define the problem of wrapper induction from HTML documents as follows: Let D(i) be a document i that consists of words from finite alphabet Σ. Suppose a document is composed of a finite number of letters from Σ, then the document $D(i) \in \Sigma^*$. Similarly, let Σ_{TOK} be a set of token alphabet, then the tokenizer can be defined as a function named TOK. Then $TOK: \Sigma^* \rightarrow \Sigma^*_{TOK}$ is a tokenizing function that removes needless HTML tags and tokenized HTML tags and words into tokens $\in \Sigma_{TOK}$, and the goal is to find a token subsequence that maximizes the MDL formula for evaluating conciseness and accuracy.

Considering that we are interested in the patterns that occur frequently and we want the pattern to be meaningful (i.e. long enough), minimum description length for wrapper induction can be formulated as follows:

$$\mathrm{MDL(tok)} = \#(\mathrm{tok}) \times \sum_{i} w(\mathrm{tok}_i) + \alpha \times l(\mathrm{tok})$$

where *w(toki)* is a user-specified weight for a token which reflects the user's domain knowledge, *#(tok)* is the number of occurrence of a token sub-sequence, *l(tok)* is the length of a token sequence, and α is a user-supplied parameter.

D. MDL-Wrapper Algorithm

We explain our wrapper algorithm. The major steps for MDLWrapper are as folllows:

MDLWrapper(D)

1. For each document Di $\in$ D, remove needless HTML tags and normalize relative URL's.
2. Generate token sequence from the HTML documents and insert the token sequences into a generalized suffix tree.
3. Sort all the nodes in the suffix tree in descending order, and for each suffix node, obtain the pattern p corresponding to the node.
4. Choose the node and its pattern that maximizes the MDL score.

E. NewsFeedAndroid

The preliminary experimental results show that MDL-Wrapper is efficient and effective for wrapper induction tasks. For seventeen news sites, we have found that very simple patterns like 'AUXa', 'XAUa', and 'AUaX' occur frequently and are effective as wrapper patterns for news articles, although we have found more diverse patterns for online stores's price information.

Figure 3 shows the running screen of NewsFeedAndroid developed by the ideas we have explained as an application on Android OS.

Figure 3. NewsFeedAndroid

IV. CONCLUSIONS

In this paper, we have proposed NewsFeedAndroid that interconnects a social networking service and online newspaper sites in order to extracts news articles from the online news sites and to perform feeding of news articles to social network service (SNS) users. Preliminary experimental results have shown that MDL-Wrapper we incorporated is efficient and effective for wrapper induction tasks for news sites. We have found that very simple patterns like 'AUXa', 'XAUa', and 'AUaX' occurs frequently and are effective as wrapper patterns for news articles.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their useful comments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2013R1A1A2013401).

REFERENCES

[1] D. Kang, and J. Choi, "MetaNews: An Information Agent for Gathering News Articles on the Web," *In Proc. International Symposium on Methodologies for Intelligent Systems*, Maebashi City, Japan, Oct. 2003, pp. 179-186.

[2] N. Kushmerick, "Wrapper Induction: Efficiency and Expressiveness," *Artificial Intelligence*, Vol. 118, No. 2 (2000), pp. 15-68.

[3] P. Senellart, A. Mittal, D. Muschick, R. Gilleron, and M. Tommasi, "Automatic Wrapper Induction from Hidden-web Sources with Domain Knowledge," *In Proc. 10th ACM Workshop on Web information and Data Management*, Napa Valley, California, USA, Oct. 30, 2008, pp. 9-16.

[4] M. T. Pazienza, A. Stellato, and M. Vindigni, "Combining Ontological Knowledge and Wrapper Induction Techniques into an e-Retail System," *In Proc. Workshop on Adaptive Text Extraction and Mining* held with ECML/PKDD 2003.

[5] I. Muslea, S. Minton, C. A. Knoblock, "Hierarchical Wrapper Induction for Semistructured Information Sources," *Journal of Autonomous Agents and Multi-Agent Systems*, Vol. 4 (2001), pp. 93-114.
[6] R. B. Doorenbos, O. Etzioni, and D. S. Weld, "A Scalable Comparison-Shopping Agent for the World-Wide Web," *In Proc. 1st Internation Conference on Autonomous Agents*, 1997.
[7] W. Cohen, "A Web-based Information System that Reasons with Structured Collections of Text," *In Proc. 2nd International Conference on Autonomous Agents*, 1998, pp. 400-407.
[8] J. Hammer, H. Garcia-Molina, S. Nestorov, R. Yerneni, M. Breunig, and V. Vassalos, "Template-based Wrappers in the TSIMMIS System," *In Proc. ACM SIGMOD International Conference on Management of Data*, 1997, pp.532-535.
[9] L. Liu, W. Han, D. Buttler, C. Pu, and W. Tang, "An XML-based Wrapper Generator for Web Information Extraction," *In Proc. 1999 ACM SIGMOD International Conference on Management of Data*, 1999, pp. 540-543.
[10] R. Ballagas, J. Borchers, M. Rohs, and J. G. Sheridan, "The Smart Phone: A Ubiquitous Input Device," *IEEE Pervasive Computing*, Vol. 5, No. 1, Jan.-Mar. 2006, pp. 70-77.
[11] M. Abramsky, "Smartphones to Outnumber PC Sales," RBC Capital, 2009.
[12] J. Aguero, M. Rebollo, C. Carrascosa, and V. Julian, "Does Android Dream with Intelligent Agents?," *In Proc. International Symposium on Distributed Computing and Artificial Intelligence*, 2008, pp. 194-204.
[13] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of Topological Characteristics of Huge Online Social Networking Services," *In Proc. 16th International Conference on World Wide Web*, Banff, Alberta, Canada, May 2007, pp. 835-844.
[14] A. Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter: Understanding Microblogging Usage and Communities," *In Proc. 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, 2007, pp. 56-65.
[15] "A World of Connections: A Special Report on Social Networking," *The Economist*, Jan. 2010.
[16] J. Gibson, B. Wellner, and S. Lubar, "Adaptive Web-page Content Identification," *In Proc. 9th Annual ACM International Workshop on Web Information and Data Management*, 2007, pp. 105-112.
[17] J. Prasad, and A. Paepcke, "CoreEx: Content Extraction from Online News Articles," *In Proc. 17th ACM Conference on Information and Knowledge Management*, 2008, pp. 1391-1392.
[18] S. Kofler, "Improving of Web-based News Articles Retrieval with Social Software," Master Thesis, Technical University Graz, 2007.
[19] D. Kang, and K. Sohn, "Wrapper Induction Based on Minimum Description Length using a Suffix Tree," *In Proc. 22nd International Technical Conference on Circuits/Systems, Computers and Communications*, Busan, Korea, July 2007.
[20] M. N. Garofalakis, A. Gionis, R. Rastogi, S. Seshadri, and K. Shim, "XTRACT: Learning Document Type Descriptors from XML Document Collections," *Data Min. Knowl. Discov.*, Vol. 7, No. 1, pp. 2003, pp. 23-56.
[21] E. Ukkonen, "On-line Construction of Suffix-Trees," *Algorithmica*, Vol. 14, 1995, pp. 249-260

Zhong-Liang Xiang is a candidate Ph.D. student in computer science at Dongseo University in South Korea. He received a science master degree in computer science at Ocean University of China in 2010 and a Bachelor of Science (BS) degree in computer science at Mudanjiang Normal University in 2003. His research interests include data mining and machine learning.

Xiang-Ru Yu received a science master degree in computer science at Ocean University of China in 2010 and a Bachelor of Science (BS) degree in computer science at Mudanjiang Normal University in 2003. Currently, she is a Lecturer at Weifang University of Science and Technology. Her research interests include data mining and machine learning.

Dae-Ki Kang is a professor at Dongseo University in South Korea. He was a senior member of engineering staff at the attached Institute of Electronics and Telecommunications Research Institute in South Korea. He earned a Ph.D. in computer science from Iowa State University in 2006. His research interests include intrusion detection, security informatics, ontology learning, and relational learning. Prior to joining Iowa State, he worked at a Bay-area startup company and at the Electronics and Telecommunication Research Institute in South Korea. He received a science master degree in computer science at Sogang University in 1994 and a bachelor of engineering (BE) degree in computer science and engineering at Hanyang University in 1992.