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

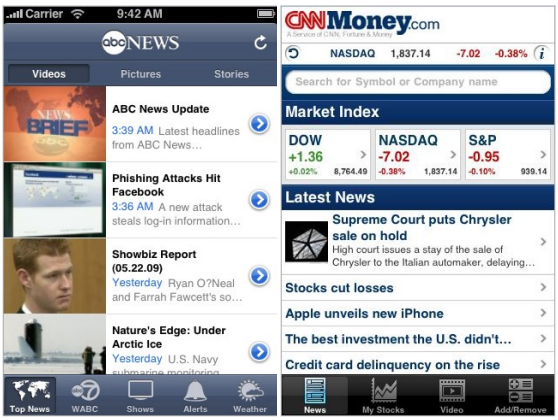

Figure 1. News feeding systems on Smartphone

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