INTERACTIVE NEWS FEED EXTRACTION SYSTEM

Prerna¹, Sanjay Singh², Rajesh Singh³, Monika Jena⁴

¹ Student M.Tech. (CSE), B. S. Anangpuria Institute of Technology and Management, Faridabad, India
² Student M.Tech. (CSE), Amity University, Noida, India
³ Assistant Professor, B. S. Anangpuria Institute of Technology and Management, Faridabad, India
⁴ Assistant Professor, Amity School of Computer Sciences, Noida, India

ABSTRACT

Our Interactive News Feed Extraction system approach is designed to provide feeds automatically for a given topic on demand of user. It is a dynamic as well as interactive approach that requires no offline data and feeds are generated online only. Thus, it is able to adapt efficiently to the dynamic information space. Interactive News Feed Extraction system is based on peer knowledge that is given by the user online to the system. This system integrates feed from different news sources and users get a relevant set of new feeds on their demand.

Keywords – Extraction, Architecture, Algorithms, Aggregates

I. INTRODUCTION

Our system is based on automatically finding of essential news articles from heterogeneous sources. Consider an example, given a news website comprising different kinds of web pages. Besides news pages, there are no news pages also. These news sites are crawled to find a relevant page which is a difficult task to recognize and acquire all news pages quickly from a large number of news websites. Also different news sites have different news page layout.

RSS feed aggregators allow a user to subscribe read and access feed content from different news sources. But feed becomes difficult to manage due to addition of different sources containing relevant information.
In this paper, we propose an approach to construct an Interactive News Feed Extraction system based on RSS feeds. RSS news feeds are basically text content rich heterogeneous and dynamic documents.

While reading a news article, topics of interest would be title, guid, subject, summary, link etc. It is useful if a user is able to specify what’s interesting to him on a web page with an easy way to extract them. Example, news sites consists of guid, title, subject and link which needs to be extracted from the page and parsing algorithm is applied to extract them.

In the following sections we will discuss parsing algorithm using the library of basic python parsing functions. Then we will discuss Interactive News Feed Extraction system for news extraction from RSS feeds.

The rest of this paper is organized as follows. Section 2 briefly introduces the related approach of news extraction using RSS feeds. In section 3, we introduce our novel method of Interactive News Feed Extraction system. Section 4 summarizes the paper and outlines some interesting directions for future research.

II. RELATED WORK

An approach was designed by Yi et al. to describe [16] how to remove irrelevant information in web pages in order to increase the quality of extraction. Their goal is to remove advertisements, navigation fields, copyright information, etc. This is achieved by detecting common elements in different pages belonging to the same site. Bar-Yossef and Rajagopalan in [5] present methods to extract informative information from web page tables. Ramaswamy et al. in [3] also presented the same method. An approach to detect content structure on web pages based on visual representation was presented by Cai et al. [10]. Embley et al. [15] present heuristics for extracting records from web pages which is a domain specific approach.

Well-known search engines like Google and Yahoo also extract information from web pages and categorize them according to topic.

The novel method to extract information from web pages is to develop wrappers. The wrapper takes as input a web page containing information, and creates a mapping from the page to another format. Laender et al. [17] developed this wrapper based system. Shinnou et al. gave an extraction wrapper learning method and expected to learn the extraction rules which could be applied to news pages from other various news sites [1]. An Automatic Web News AZheng et al. presented a news page as a visual block tree and derived a composite visual feature set by extracting a series of visual features, then generated the wrapper for a news site by machine learning [8]. Dong et al. gave a generic Web news article contents extraction approach based on a set of pre defined tags [9].

III. PROPOSED WORK

A. Parsing

Interactive News Feed Extraction system collects news articles form news sources. User specifies his topic of interest, from which relevant news articles are passed using parsing algorithm. Elements of parsing includes:-
1) Parsing Library: It is a library of parsing function that provides extraction rules to extract guid, title, subject and summary and provides a list of news stories. These rules specify what is interesting to a user and extract portions they are interested in.

2) News Story Object Model: For each news article, a set of guid, title, subject, and summary are formulated as shown in Fig 1 and this encapsulation of news articles of interest and corresponding feed extraction forms a news story object model.

![Fig 1 News Story Object Model Attribute](image)

Guid = getGuid (Self)
Title = getTitle (Self)
Subject = getSubject (Self)
Summary = getSummary (Self)

**Fig 1 News Story Object Model Attribute**

**B. News Feed Extraction Architecture**

A news story object model consists of a set of attributes shown in Fig 1 and corresponding parsing function which extract them from news sites. This news story object model is fed as input to the News engine extractor as shown in Fig 2. The entry point of extracted feeds is based on triggers. These triggers are passed on to the news articles, which identify the relevant articles. These triggers proceed to recursively identify relevant articles.

![Fig 2 News Feed Extraction Architecture](image)

Extraction rules that are followed by News feed extractor are:-
1) Single parsing function: It identifies the exact phrase of interest.
2) Multiple parsing function: After identifying an item of interest, parsing function will continue to search through the entire document for similar items of interest.
News story object model extracts guid, title, subject, summary and link of each news article. News Feed Extraction Architecture process web pages based on News story object model using following triggers:-

1) Word Trigger: Entry point to a news article would identify text without including the unimportant words, punctuations that are removed. After identifying text, title trigger, subject trigger and summary triggers are used.

Title trigger checks for the title of news articles by comparing with triggers. Subject trigger checks for the title of news articles by comparing with triggers. Summary trigger checks for the title of news articles by comparing with triggers.

2) AND Trigger: This function searches for the occurrence of all triggers in the text. Function searches in all news articles. If either of the trigger is not present in a news article, then that article is not selected.

3) OR Trigger: This function searches in the news article if either of the trigger exists then that is selected.

4) NOT Trigger: This function searches in the news article if either of the trigger does not exist then that news article is not selected.

5) Phrase Trigger: This function searches in the news article for exact phrase rather than words.

![Fig 3 Triggers used by News Engine Extractor](image)

### IV. EXPERIMENT AND EVALUATION

Consider an example in which News object model was derived by referring to news articles obtained from news.google.com and news.yahoo.com. The news article is described by a set of four variables guid, title, subject and summary using library parsing functions based on user input. Many news articles are given as input to the extraction engine; the results of Interactive News Feed Extraction system are measured in terms of recall and precision.

Recall is a measure of how well the proposed system finds all relevant news feeds based on a user topic for search, even to the extent that it includes some irrelevant news feeds.

Precision is a measure of how well such system finds only relevant news feeds based on a user topic for search, event to the extent that it skips irrelevant news feeds.

Example. If the Interactive News Feed Extraction system retrieves A relevant news feeds, B irrelevant news feeds and misses C relevant news feeds. The Interactive News Feed Extraction system’s performance for Yahoo and Google news are shown in fig 4 and 5. Fig 4 shows the output of Interactive News Feed Extraction system that displays news feeds from Google and Yahoo top news based on user’s input. Fig 5 shows the performance of given proposed system in terms of recall and precision.
V. CONCLUSION

This paper presents an interactive and dynamic approach to extract news from RSS feeds. It can be considered as a simplified version of wrapper. It serves as an easy to use system for the user to quickly extract the needed information. Multiple parsing functions allow the recursive search of relevant news feeds through triggers. As future work, we will modify the system to improve the accuracy rate.

REFERENCES


AUTHORS PROFILE

Sanjay Singh received his B.E degree (2009) from the MRCE; Faridabad affiliated to MD University and M.Tech scholar (2010-2013) from Amity University. He joined as the Faculty of the Department of CSE/IT at the ACEM, Faridabad in 2009, where he is now working as Sr. Lecturer. He has total 3.5 years of teaching experience.

Prerna received his B.Tech (2011) from the BSAITM; Faridabad affiliated to MD University and M.Tech scholar (2011-2013) from BSAITM; Faridabad.

Monika Jena is working as Assistant Professor in Amity School of Computer Sciences. She has 12 years of teaching experience. Her current research interests include QoS routing, multimedia communication and network computing.

Rajesh Singh is working as Assistant Professor in BSAITM Faridabad. He has 12 years of teaching experience.