RSS FEEDS AND NEWS CONTENT CATEGORIZATION

[1]SV. Shri Bharathi, [2]Angelina Geetha

B. S. Abdur Rahman University, Department of Computer Scinence and Engineering, Department of Computer Science and Engineering, Chennai India.
[1]shribharathi01@gmail.com, [2]anggeetha@yahoo.com

***Abstract*- Really Simple Syndication (RSS) or Rich Site Summary is a Web feed format used for publishing frequently updated content on the Internet, such as blog, news, audio and video in a standardized format. Content categorization is the process in which the contents are grouped into categories, usually for some specific purpose. In the existing system, users' do not have control over the information and up-to-date content information is also not possible. In order to overcome this, the proposed paper puts forward a web information extraction method which is based on the RSS feed reader which helps to categorize the News articles (informative content) in an effective manner. RSS is a spam free, quick and efficient way to read the news and weblogs. The experimental study was carried out for content categorization using the RSS feed Data sets. RSS news feeds with 2658 web pages with articles of different category are used as training data set and 300 web page news contents are considered as testing dataset. The extensive experiment carried out proves the effectiveness of categorization of this method.**

***Keywords* -RSS, Content Categorization, Syndication, RSS Reader, RSS feeder**

I. INTRODUCTION

The World Wide Web has grown from a few thousand pages in 1993 to more than billions of pages at present. Moreover, all these pages that exist on the WWW are neither static nor stable, but they form a continuously changing system. Today there are several billions of documents, pictures and other multimedia files available via Internet and the number is still rising. But considering the impressive variety of the Web, retrieving interesting content has become a challenging task.

Really Simple Syndication (RSS) is a format for delivering regularly changing Web content. Many news-related sites, Weblogs and other online publishers syndicate their content as an RSS Feed to whoever wants it. RSS takes the latest headlines from different Web sites, and pushes those headlines down to your computer for quick scanning. RSS mostly, uses XML to deliver updated content on the Web. The biggest advantage of monitoring the RSS content is that users do not have to provide personal information such as email address there by reducing the possibility of virus infection. RSS is also called web feeds and content delivery vehicle. It uses some format to syndicate the news and the Web contents from blogs.

This paper proposes content categorization of News articles with the help of RSS feed. From the web pages, RSS feed reader reads the required content e.g. title, description, date, author, link etc in the format of XML document and then, the specific contents are extracted and categorized as per the available category.

This paper is organized as follows: Section II surveys the related work. Section III formalizes the working principle of Content categorization techniques based on RSS feed. Section IV presents the experimental results. Finally, Section V gives the conclusion of this research area.

II. RELATED WORK

The use of RSS, which is the XML-based format for syndication and subscription of information, is diffused by Gill et al [1]. In Wikipedia [2], Really Simple Syndication, (RSS) uses some format to syndicate the news and the web contents from News articles.

In order to avoid the noisy data that exist in the news and to acquire the theme information Zheng Rui-Juan et al [3] introduces Web news information collection technology to obtain news theme information from RSS website. Filtering the noise and locating the information more accurately done by index of RSS are explored by Qingcheng et al [23].

To improve the missing postings in RSS Aggregation policy Geun Han et al [4] explored new aggregation policy to minimize the number of missing postings within an aggregation. Due to inefficiency of general search engine and slow updates, Jian Zhu et al [5] introduces RSS search engines to overcome the shortcomings to achieve high efficiency, high-speed searching of the page and also applies the concept in e-

commerce applications. Different RSS aggregator types were analyzed and various applications of RSS among the users are discussed by Phoey Lee et al [6].

Independent of news page layout is analyzed by Han et al [7] and introduces the extraction of news article contents in effective and efficient way. However, most of the RSS readers only display items in chronological order, which doesn't work well when users are inundated with too many items in the feeds, so Cansheng Ji et al. [8], explored the recommendation system to help people to find items in an RSS reader. Specifically, profiled based features, and also updates frequency as well as Post Rank values for RSS recommendation system.

Due to the heavy growth of RSS, feed overload problem occurs. In order to avoid this problem, Burnham et al. [9] evaluated keyword based searches that filter posts based on keywords and also introduces a Meta feed for RSS feed. Today, there is a broad range of search and discovery of online news sites, e.g., Google News [10], Allin One News [16] available.

To evaluate a possible improvement to search methods Lindholm et al [15], extracts the news articles and describes the implementation techniques which is generic and easily adopted to new data sources.

In general, certain issues arises like search engine percentage coverage over the Internet and to find the less popular contents from the Web through search engines. So, Ying Zhou et al [12] introduces self-organizing search engine for RSS syndicated Web data. It is built on structured peer-to-peer technology and also enables indexing and searching of frequently updated Web information described by RSS feed.

The similarity of the twin-pages collected from the same topic section of a site and published on the same/near date is introduced and applied in the algorithms is explored by Qiujun et al [11] which is much less complicated , and its accuracy and efficiency are fairly high, its complexity about the pages size is just linear.

Instead of the hunt and peck of Web surfing, users can download or buy a small program that turns their computer into a voracious media hub, letting them snag headlines and news updates was introduced by Lassica et al [13]. Baker et al. [14] introduced the integration of advertising in RSS feed aggregators and blog search listings and also offers paid sponsored links in its search results which are powered by Yahoo.

Langford et al [17] explored that the Weblogs are written as personal diaries that sometimes combine subjective thoughts with journalistic reporting and also noticed that entries usually focus on specific topics and may contain hyperlinks to other relevant resources. Various heuristic technique for extracting the main article from news site Web pages formalized by, Jyotika et al [18].

The common background of most of the existing approaches is their capability of performing automated aggregation of information from a single domain (i.e., the Web). So far, however, only a few attempts have been made to investigate the possibility of merging information from many application domains. As an example, Simon et al. [19] claimed that RSS feeds not only can set up to deliver headlines but also to display a few lines of stories. Luminita et al [20] discusses some RSS directions and ideas for teachers and instructors in order to get benefits of using RSS in education.

George Adam et al [21] explored that every major and minor portal use RSS feed, to keep the crawler up to date and retain a high freshness of the "offline content" and also it observes the temporal behaviour of each RSS feed.

RSS news feeds deliver the categorized news items depends on the request generated from the client application was introduced by Subrata Saha et al [22]. An unsupervised framework was presented by Messina et al [24] for content-based Web newspaper articles and broadcast news stories aggregation and retrieval.

In general, users receive a huge amount of news that may contain a large number of irrelevant bits of information. To address this problem Pinheiro et al [25] provides an autonomic way to reduce it through the use of data killing operators.

Yi et al [26] describe how to remove irrelevant information in Web pages in order to increase the quality of subsequent data mining. 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.

III. PROPOSED WORK

RSS is an abbreviation for Rich Site Summary, Really Simple Syndication, RDF Site Summary, or a variation on any of these names. It is an XML document that facilitates content syndication. RSS allows an information publisher to easily syndicate (feed) content headlines or blurbs; other Web sites can publish this information at no cost to either party. RSS is a spam free, quick and efficient way to read the news and weblogs.

In the proposed paper, RSS is considered as the reliable way to extract the Web content to Internet and the RSS data is small and fast-loading, it can be used with services like cell phones or PDA's, voice mails, and email ticklers. Unlike email an RSS feed is zero maintenance, the messages will never get blacklisted or filtered. With RSS, users can (finally) separate wanted information from unwanted information (spam). RSS Allows users to generate up-to-date news and postings, as information and content in the RSS readers or aggregators.

RSS Web structure contains the mixed links content of XML and RSS. First, get the RSS source. RSS source consists mainly of summary of News and the original URL of the News site information. So extract the summary and the site's URL, and then by resolving the URL to obtain the News content information. Finally, through the content extraction and categorization techniques the contents are retrieved and categorized from the Web.

a] Basic Structure of RSS 2.0

RSS documents use a self-describing and simple syntax. Figure 1 shows the basic structure of RSS 2.0. The first element in the document is the <rss> element. This includes a mandatory version attribute. The subordinate to the <rss> element is the <channel> element that contains metadata about RSS feeds. The <title> element is the title, either of the entire site (if it's at the top) or of the current item (if it's within an <item>). Next to <title> is the <link> element indicates the URL of the Web page that corresponds to the RSS feed, or if it's within an <item>, the URL to that item. The <description> element describes the RSS feed or the item. <item> element is the meat of the feed. These are all the headlines (<title>), URLs (<link>) and descriptions that will be in users' feed. <item> is an element that has information on postings. A channel may contain any number of <item>s.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<rss version="2.0">
<channel>
  <title>Google Home Page</title>
  <link>http://www.Google.com</link>
  <description>Web tutorials</description>
  <item>
    <title>RSS Tutorial</title>
    <link>http://www.Google.com/rss</link>
    <description>New RSS tutorial

    </description>
  </item>
  <item>
    <title>XML Tutorial</title>
    <link>http://www.Google.com/xml</link>
    <description>New XML tutorial
```

Fig.1. Basic structure of RSS 2.0

RSS feeds filtering out large numbers of noise data that present in the website, reducing the appearance of spam and providing high-level network information view services. RSS has overcome the low efficiency, slow updates, and then achieved the high efficiency, high-speed updates and retrieval of information.

In this proposed approach, a channel can contain channel information and content <item>. < Title > represents the name of the channel, < link > said news of the original URL, contains a full web content, < description > is a brief description of site content. Element < title >, < link > and < description > are usually included in < item >

element, <item> element contains the actual news item information used to reflect the specific news summaries. Item in all sub-elements are optional, however, must at least there is a title or description. The proposed paper extracts the content through the RSS reader - link, title, description and other information.

b] Proposed Architecture

The proposed architecture consists of multiple subsystems which are assigned with specific roles in order to achieve the effective news content and to perform the categorizing task efficiently. The following Figure 2 shows the overall architecture of the proposed work.

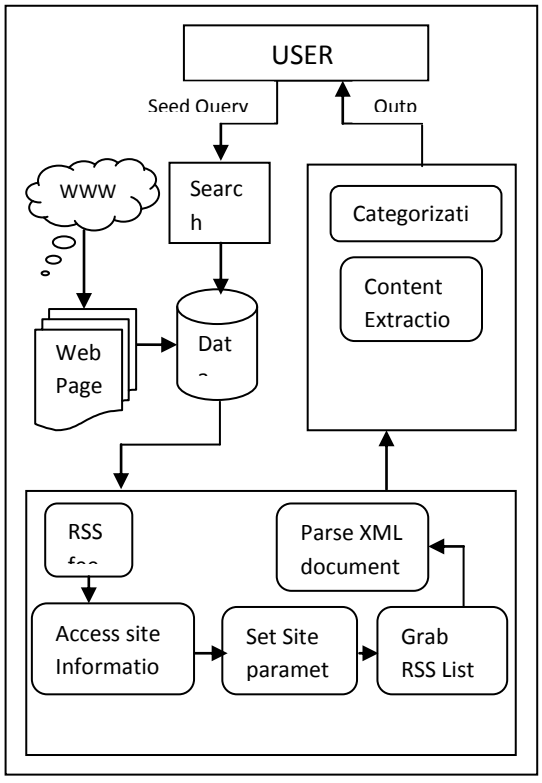

Fig. 2. Proposed System Architecture

The basic parts of the system are (a) User based Search engine (b) the centralized database, (c) RSS Feed Reader (d) User Interface that execute the extraction and categorizing techniques.

- Search Engine

A Search Engine is a user based program that visits Web sites and reads their pages and other information in

order to create entries for a search engine index. The major search engines on the Web all have such a program, which is also known as a "spider" or a "bot." These Search Engines are typically programmed to visit sites that have been submitted by the users' as new or updated. Entire sites or specific pages can be selectively visited and indexed.

- Data Base

The database is used for storing permanent information; the database is used in order to store the starting URL, which is links to XML files, and the results of the parsing procedure. The XML files are RSS feeds which means communication channels provided by news portals. Additionally, the database stores information concerning the articles that are fetched from the Search Engine.

- RSS Feed Reader

An RSS Reader, also called a feed reader, is a browser add-on program designed to gather and display RSS feeds according to user-definable parameters. An RSS reader can reduce the time and effort needed to check online publications for updates. It creates, in effect, a personalized news subscription for the Internet user. It accesses the Site information like Title, Description, Author, Link, URL address etc. It grabs the final information of title and description after setting up the site parameter. Finally feed reader parses the XML document from RSS feed list.

In general, a typical RSS reader gathers content in the background as the users surf the Web. The program checks for RSS feeds at regular intervals and generates a pop-up when a new article or message of potential interest is available. Instead of a generalized document arriving once a day, an RSS reader provides specific content as soon as it makes the news. When a pop-up appears, users' can ignore it, read a summary or view the entire file.

- User Interface

The User Interface Display consists of two types of techniques. First one is content extraction and another one is Categorization techniques. From the RSS Feed Reader the Contents are extracted by applying the preprocessing techniques like stop word removal, Stemming.

Content categorization includes analysis of URL links embedded in the content. Such analysis can provide more accurate categorization of certain types of content. In this method the contents are divided into different categories. Similarity measures are used to compare the contents and based on the frequency range the contents are grouped as different categories. Finally the output is displayed to the user.

c] Data Indexing

Data Indexing is a data structure that is added to a file to provide faster access to the data. It contains Search key and pointer. Each <item> element of a feed is treated as a document. The <item> element corresponds to either a news article or a weblog entry. Instead of automatically generating globally unique document ID, <link> value of an <item> as document id in constructing term-document inverted index. The <link> value is the unique permanent link of web content.

Indexing either the whole content page or only the <title> and <description> information. This will not make much difference if the <description> contains the full body of the content. This paper takes the approach of indexing <title> and <description> section for the following reasons. This approach is both simple and low cost since no extensive crawling is required.

experiments and results

A working knowledge base of news articles from some major news portals from the U.K. and the U.S are gathered. These pages were crawled and indexed from the web without any filtering or manual discarding of anyone. In this way, this paper built a collection that is very representative of the reality of the on-line news area.

Six distinct news categories: business, entertainment, health, politics, science and sports, are defined for organizing the captured texts. Afterwards, using the categorization mechanism, 50% of the keywords are extracted and categorized for each keyword with the text's category using the absolute frequency as a relativity measure

The following Figure 3 shows the sample CNN News articles which consists of News Title, Description, Author, Link etc.

Fig. 3. Sample of Web News page

TABLE I A List of RSS News Sites

Country/ Region	News Site	URL
United States	CNN ABC News New York Times USA Today Fox News	http://www.cnn.com/ http://www.abcnews.go.com/ http://www.nytimes.com/ http://www.usatoday.com/ http://www.foxnews.com/
United Kingdom	BBC Guardian Unlimited	http://www.bbc.co.uk/ http://www.guardian.co.uk/

The above Table I shows List of RSS News sites for the collection of News Articles dataset.

The following Figure 4 shows the number of daily News contents in all RSS feeds. The average of daily articles is 1083 (1,160 on weekdays, 892 on weekends). The number of postings generated daily ranges from 600 to 1400. The distribution of daily articles rate for total RSS feeds is shown in Figure 11. Approximately 94% of total RSS feeds generate less than 10 postings per day.

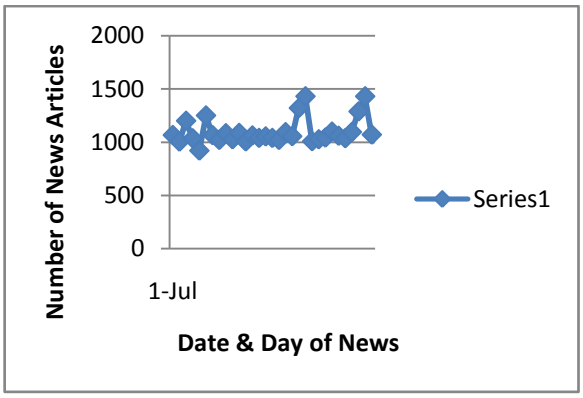

Fig 4. Number of Daily News articles in all RSS Feeds

The Proposed system consists of two phase. These phases are training and testing phases.

In training phase, before the selection of news contents, removing punctuations and tokenization steps are applied. Then the news contents are categorized according to the defined category like business, entertainment, health, politics, science and sports.

In testing phase, the selected news articles are compared with already trained set of categories. The Categories are subdivided and users can get their favorite categorization such as Health, politics, and so on. Users can also give keywords for filtering certain topics. Table II shows the Data set for training and testing phase of the News articles.

The following Figure 5 shows the Title and Description of the extracted News articles.

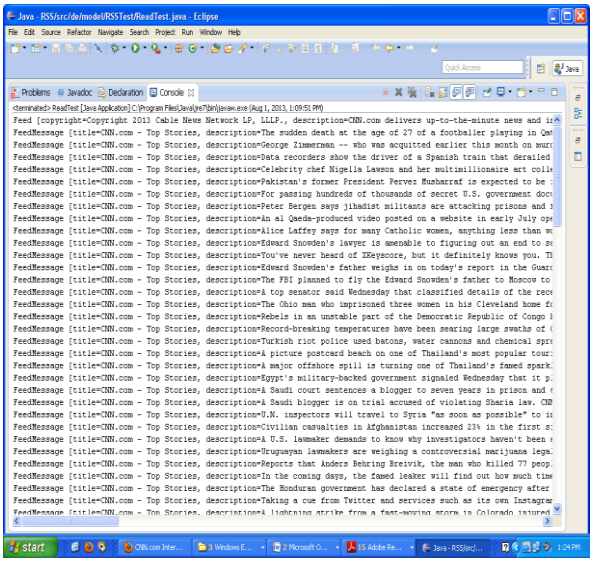

Fig.5. News Items Extracted from the CNN News

TABLE II DATA SET TABLE FOR TRAINING AND TESTING

Data Set	Training	Testing
Business	200	100 (CNN,ABC News)
Entertainment	240	120 (New York Times)
Health	170	70 (ABC News, Fox News)
Politics	253	100 (BBC)
Science	166	80 (USA Today, BBC)
Sports	200	140 (BBC)
Business	300	206 (New York Times, CNN)

TABLE III Final Classification of Accuracy Measures.

Category Name	Total	Actual Positive	Precision
Business	300	288	0.96
Entertainment	340	325	0.96
Health	200	197	0.99
Politics	250	240	0.96
Science	230	226	0.98
Sports	300	291	0.97

The above Table III shows the accuracy measure of the content categorization. The accuracy of the measure is defined by precision and recall.

Precision p is defined as the number of correctly classified positive examples divided by total number of examples that are classified as positive.

$$\text{Precision} = \frac{tp}{tp + fp}$$

IV. CONCLUSION

This paper presented an effective approach to realize the automatic news article contents extraction and categorization using the news RSS feeds. This paper proposed a novel algorithm applicable to the general news pages, which can extract the news paragraphs automatically, accurately and consistently. Hence RSS feed plays a vital role in extracting the contents of the News articles effectively and categorizing the contents efficiently as per the available category.

V. REFERENCES

[1] Gill, K.E. "Blogging, RSS and the Information Landscape: A Look At Online News", WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (Chiba Japan, May 2005).

[2] RSS (protocol)," from Wikipedia. ND.

http://en.wikipedia.org/wiki/RSS_(protocol)

[3] Zheng Rui-Juan, Zhang Yang-sen,"Design and implementation of news collecting and filtering system based on RSS",9 th International conference on Fuzzy Systems and Knowledge Discovery(FSKD),2012.

[4] Young Geun Han, Sang Ho Lee, Jae Hwi Kim, Yanggon Kim, "A New Aggregation Policy for RSS Services", International workshop on Context enabled source and service selection, integration and adaptation (CSSSIA '08) 2008.

[5] Jian Zhu, Hanshi Wang,"Application of E-Commerce personality searching based on RSS", 2nd IEEE International Conference on Information Management and Engineering (ICIME), 2010.

[6] Teh Phoey Lee , Abdul Azim Abdul Ghani , Abdul Azim Abdul Ghani ,"Survey on application tools of Really Simple Syndication (RSS): A Case Study at Klang Valley",International Symposium on information technology, 2008.

[7] Hao Han, Tomoya Noro and Takehiro Tokuda ,"An Automatic Web News Article Contents Extraction System Based on RSS Feeds", Journal of Web Engineering, Vol. 8, No. 3 (2009).

[8] Cansheng Ji ,Jingyu Zhou, "A Study on Recommendation Features for an RSS Reader", International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery 2010.

[9] Burnham, Bill. "Saving RSS: Why Meta-feeds will triumph over Tags", Burnham's Beat. 25 January 2005.

http://billburnham.blogs.com/burnhamsbeat/2005/01/saving_rss_why_.html

[10] Google News. http://news.google.com.

[11] Qiujun LAN,"Extraction of News Content for Text Mining Based on Edit Distance" , Journal of Computational Information Systems, (2010).

[12] Ying Zhou, Xin Chen, Chen Wang,"A Self-Organizing Search Engine for RSS Syndicated Web Contents" , Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), 2006.

[13] Lassica, JD. "News That Comes to You," Online Journalism Review 2009. http://www.ojr.org/ojr/lasica/1043362624.php

[14] Baker, Loren. "News, RSS, and Blog Search - Feedster, Topix.net, and Yahoo," Search Engine Journal. 4 March 2005. http://www.searchenginejournal.com/index.php?p=1382

[15] Sigrid Lindholm, "Extracting Content from Online News Sites", Master's Thesis in Computing Science, UMEA University, Sweden 2011.

[16] AllInOneNews. http://www.allinonenews.com.

[17] Langford.L.K..,"Surf's up: social network services and analyses", IEEE Transactions on Engineering management review,vol.38,2010.

[18] J. Prasad and A. Paepcke. CoreEx: Content extraction from online news articles. In The Proceedings of the 17th ACM conference on Information and Knowledge Mining, pages 1391–1392, 2008.

[19] Simon S., (2007), "RSS Feeds", Available at:

http://www2.warwick.ac.uk/services/library/main/research/researchers/literaturereview/keepuptodate/rss/.

[20] Luminita Giurgiu, Ghita Barsan , Dan Mosteanu, "Web Syndication in Educational Environment," 50th International Symposium, ELMAR-2008,Sep 2008.

[21] George Adam, Christos Bouras, Vassilis Poulopoulos ,"Utilizing RSS feeds for crawling the Web", Fourth International Conference on

Internet and Web Applications and Services (ICWI'09), 2009.

[22] Subrata Saha, Atul Sajjanhar, Shang Gao, Robert Subrata Saha, Atul Sajjanhar, Shang Gao, Robert Dew, Ying Zhao, "Delivering Categorized News Items Using RSS Feeds and Web Services", 10th IEEE International Conference on Computer and Information Technology (CIT), 2010.

[23] Li Qingcheng, Li Youmeng, "Extracting Content from Web Pages Based on RSS", International Conference on Computer Science and Software Engineering, 2008.

[24] Alberto Messina, Maurizio Montagnuolo, "Content-Based RSS and Broadcast News Streams Aggregation and Retrieval",Third International Conference on Digital Information Management, (ICDIM 08') 2008.

[25] Wallace A. Pinheiro, Thiago de S. Rodrigues, Marcelo A. R. da Silva, Marcio A. N. da Silva, Marcelino Campos Oliveira Silva, Geraldo Xexéo, Jano M. de Souza, "Autonomic RSS: Discarding Irrelevant News",Fifth International Conference on Autonomic and Autonomous Systems, 2009.

[26] L. Yi, B. Liu, and X. Li, "Eliminating noisy information in web pages for data mining", In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003.

◈◈◈