ISSN NO. 2320-5407

Journal homepage:http://www.journalijar.com *INTERNATIONAL JOURNAL*
Journal DOI:10.21474/IJAR01 *OF ADVANCED RESEARCH*

RESEARCH ARTICLE

NEWS CONTEXTUALIZATION AND AGGREGATION, IMPLEMENTED USING A WEB APPLICATION SYSTEM.

Prathmesh Achyut Kestikar, Atul Gutal, Vrushali Karne, Akshata Kasliwal, Prof. Shraddha Toney.

Manuscript Info

Manuscript History:

Received: 18 March 2016
Final Accepted: 19 April 2016
Published Online: May 2016

Key words:

Corresponding Author

Prathmesh Achyut Kestikar.

Abstract

News contextualization provides the users with general news as well as news about a particular requested topic, on a single page. This will help the users to know about any news topic on a single page and save the time required to search on different websites. Also the news provided will be provided from well-known and reliable news websites and hence will provide different perspectives about a single topic. Also, the feeds from social websites will help the users know about public response to a particular topic. News will be provided from specified number of news websites like The Times of India, NDTV, BBC News and The Economic Times. Also Videos from The Times of India and Tweets from twitter will be provided. This will help the users in gaining insight about public response for a particular news topic. The Facebook pages of news websites will be displayed. To display the news in a user friendly way, the page is divided in sections. Different sections are provided for different websites. The system involves use of PHP, AJAX, JavaScript, and JSON. Web scraping tools will be used for retrieving data from different news websites. APIs will be used for extracting data from social websites.

Introduction:-

A system can be implemented where the news can be provided to the user in an integrated manner and in a single place. We have developed such a system (a web application), by using various web development languages such as php, javascript, jquery, HTML, CSS. This paper provides an insight as to how we implemented this system as a web application mainly using javascript, AJAX and php for the functionality of the system and HTML, CSS, jquery for the user interface of the system.

This web system is considered to be adopting a client-server software architecture, where the server processes the data and provides it to the client (the user). This software architecture is adopted to reduce the processing overload on the client machine which may make this system faster in processing the web application and getting the final results.

Scope:-

The system will help not only general users get an overview of news but also help a user research a particular topic. As the system provides data from different sources and in different formats such as text, images, videos, posts and tweets, the website can be used for many other purposes. If used for educational purposes, the system can provide voluminous information of a topic. The system can also help business entrepreneurs in learning about public response to a particular strategy or product.

But the system has some limits due to lack of hardware and software resources. If a requested topic is not available on the specified websites, data will not be displayed. Also the data is scraped from websites and displayed without any changed in the content. Hence, the system is not responsible for any problems related to the content. The speed

of data retrieval depends on the availability of hardware and internet resources. Also if there are changes in the website format, these changes will have to be reflected in the system as well.

Implementation overview:-

The proposed system provides a website that displays news content from a specific number of news websites and social websites. The content displayed will be relevant to the user request and fetched from specified websites only. The content fetched will be displayed on a single page to save the user the hassle of searching different web pages. The news articles, tweets, videos and posts will be displayed in a proper fixed format with different sections for different websites. The sections will contain a brief matter of each article fetched. The website will display the entire article if further requested by the user.

The news websites included are The Times of India, NDTV, Economic Times, BBC News, whereas the social websites used are Twitter and Facebook. The system will take input from user and search for data on these websites using keyword provided. The system will then return relevant and latest news articles and tweets from the respective websites. The proposed system will use web scraping methodologies for fetching data from news websites. APIs will be used for fetching data from social websites. The Facebook content displayed will be Facebook pages of specified news websites. The website will display videos from 'The Times of India'.

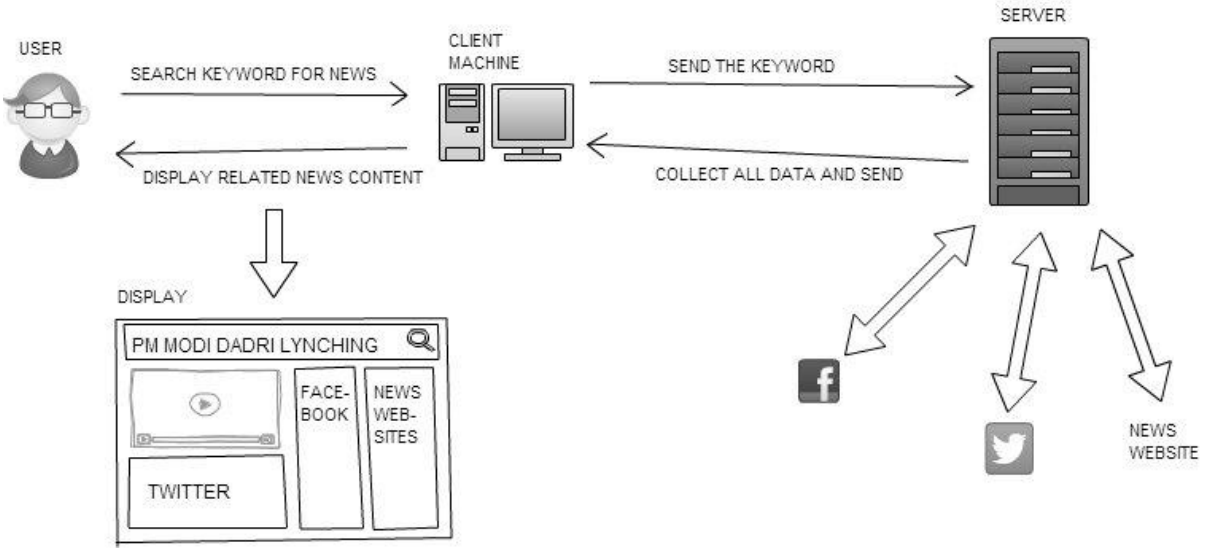

Fig.1 Architecture overview of the system.

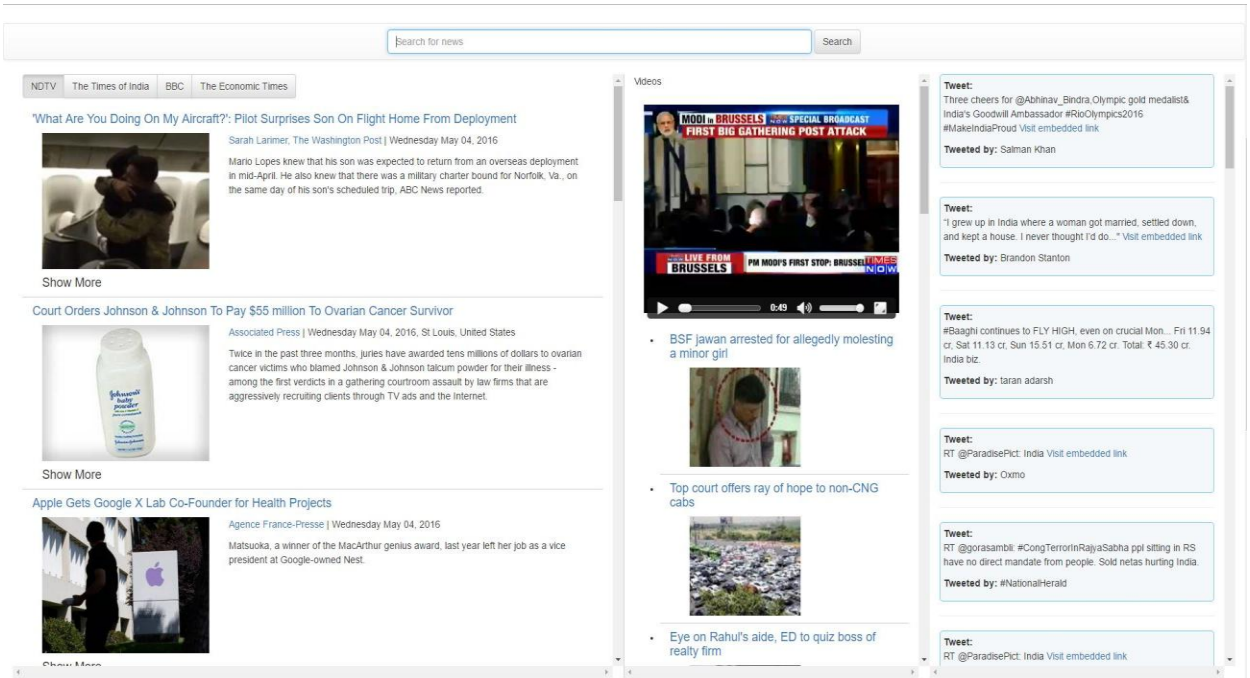

Fig 2 User interface overview of the system

Methodology:-

Initially, the website will display recent news articles from specified news websites in respective sections. Also tweets and Facebook pages of news websites fetched dynamically will be displayed. For this, different functions are created for different websites. There is function that executes the web scraping php code for the respective news website for each different website. The structure of every news website is different. This makes the web scraping aspect of the system a little complicated. Web scraping is mainly dependent on the website structure. If the website to be scraped is developed in a simple and traditional manner, then the web scraping job gets easier.An example of one ofthese functions is 'function get_latest_news_ndtv()'.

Initially, when the page loads, there is no user input provided from the user. Hence, the recent news need to be displayed to the user instead of keeping the display sections blank. This is done by scraping the page of the news websites where the recent and latest news articles are provided. This job is done by the functions like one mentioned above.These functions fetch the recent news articles available on respective websites.

The main page of our system is developed mainly in javascript and HTML. The functions which execute the php web scraping code use AJAX to get the results from the news websites dynamically.

The use AJAX (Asynchronous JavaScript and XML) helps to update a particular content of the web pageinstead of sending the whole page to the server. Hence the user can read recent news faster. For displaying the content of a particular news topic, the user is expected to enter the search keyword.On receiving a keyword the system stores the keyword in a variable called 'query'. Then a new variable is created that needs to be passed as a URL (that has the search query) on the internet. This variable contains the user query along will the website name. This variable is created based on the way a search query is passed on a particular news website. Hence,we get the news articles relevant to the user search keyword for a specific website. Example of a variable created for NDTV website is:

```
query                //contains the search keyword entered by the user
$search = " ";       // used for creating the url for searching on website
$replace = "-";      // used for creating the url for searching on website
$ndtv_search = "http://www.ndtv.com/topic/";
```

$new_string = str_replace($search,$replace,$query); //This function replaces the spaces in the search query with dashes, as it needs in an URL

$ndtv_final_search_query = $ndtv_search.$new_string.'/news';

Suppose a user searches for the keyword “NIT Srinagar”

- query= “NIT Srinagar”;
- search= “ ”
- replace= “-”
- ndtv_search= http://www.ndtv.com/topic/
- new_string= NIT-Srinagar
- ndtv_final_search_query= http://www.ndtv.com/topic/NIT-Srinagar’/news’ - This is the final URL that will provide us with the news search results from the NDTV news website.

In this way, the search query provided by the user is used to pass it on to the different news websites from out system. Hence, “$ndtv_final_search_query” is the variable used for web scraping. This variable (containing the final URL) is then passed to a function called ‘curl()’. This function is used for web scraping from the website. The contents fetched are stored in a variable ‘$ndtv_get_content’. Hence now the system has the data fetched by the curl function. The data fetched is the actual HTML structure of the webpage.

But the data fetched has to be specific and relevant to the data that needs to be displayed. Hence formatting is an important task before displaying the data. The data fetched is in HTML format and contains everything available on the website. Hence selection of precise news article is required. A function called ‘scrape_between()’ is used for this purpose. This function extracts the required part of the webpage by taking two arguments, the starting point from where the data needs to be extracted and the ending point till where the data is needed. It basically scrapes the page from one point to another.

$ndtv_data = scrape_between($ndtv_get_contents, ' ', ' ');

$ndtv_data = scrape_between($ndtv_data, '
', '
');

On the NDTV website the data we require is inside the ‘ ’ tag. Hence, to get the data from just this tag, we pass it as the starting point and also provide an appropriate ending point for the scrape_between() function.

This gives a set of recent news articles from a particular website. The data to be displayed has to be in a precise format and arranged. Most of the further work in the system involves formatting of obtained data. Here each news article is separated, traversed and divided into various sections of an array ($article).The articles are divided into various sections like $newstitle, $newsimg, $newssource, $excerpt and $newslink. Information about each article contains an image, the excerpt (small introduction of the article), the link of that article page, the title of the article. This information is stored separately and sent back to the main page so that they can be easily accessed separately to make the user display easier to design. By doing this, we have the liberty to display the image, title, etc. according to our design requirements.These sections are created and sent to the main page. The data sent to the main function is in the form ofJSON (JavaScript Object Notation). JSON helps in data-interchange and provides ease of formatting. The sections in array help in formatting the data retrieved from news website in the main page. Finally all the articles collected are displayed in a particular format.Further if a user wishes to view the entire article and clicks on “Show More”,the main page calls the function ‘showArticle()’. In this case, the entire article is displayed with a scrolling bar using the ‘getArticle()’ function.

Content from other different news websites included is extracted using the same way, except the web scraping start and end points differ. Each news website had to be analyzed (its HTML structure) for us to be able to extract data from them using this web scraping method.

In case of Twitter, API is used for retrieving data. For accessing the interface an account is created where the account holder is provided with authentication keys to access the public tweets. Four keys provided by Twitter are

$consumer, $consumersecret, $accesstoken and $accesstokensecret. The 'get()' method returns a collection of relevant Tweets matching the query provided in the variable "search".

$tweets=$twitter->get('https://api.twitter.com/1.1/search/tweets.json?q='.$search.'&result_type=mixed&count=35');

Resource URL:-https://api.twitter.com/1.1/search/tweets.json.

Resouce Information:- Response format specified is JSON.

Parameter:- "q" specifies search query of 500 characters maximum which is UTF-8and URL-encoded.

Result type specifies what type of search results are preferred. "mixed", "recent" and "popular" are the 3 result types available. The system includes Tweets which are both popular and real time that is "mixed".

"count" specifies the number of Tweets to return per page. The system provides 35 as count.

The results are further encoded in JSON and returned to the main page.

For videos, the system follows a similar procedure for fetching the data search results. The 'curl()' function returns the block of data including the video results and their information. 'scrape_between()' function scrapes the data to be displayed from the entire block. Links are obtained and data is separated for formatting. 'startvideo()' function is used to start a particular video.

We have used a simple HTML5 video player on the page for playing the acquired videos. The HTML5 player needs a link from where it streams the video to be played. In our case, we need to get the 'data link' of the video for successfully playing it.

The news videos in this system are retrieved from 'The Times of India' news webpage. On The Times of India page, each and every video that plays on their website has been placed into an '<iframe>' html tag. The <iframe> tag specifies an inline frame.An inline frame is used to embed another document within the current HTML document. So this means that the video is entirely a different document that needs to be scraped separately. The link of this iframe document (found in the 'src' attribute of the iframe), is passed to the curl function for scraping. This document contains the data link of the video we need to stream on our page, which is written is JavaScript.

The link is stored in to an array named URL. By using a regular expression and parsing the document for this name we found the link of the video for its data.
Finally, we assigned this link to the 'source' attribute of our HTML5 player and the video becomes playable on our page directly.

Design:-

The webpage application is designed using HTML, CSS (Cascading Style Sheets), jQuery, JavaScript and php for designing the functionality.

The use of 'Bootstrap' has also been included to make the webpage responsive to any device it may be displayed on. Bootstrap allows the webpage to dynamically scale and position the webpage content elements depending on the device the webpage is being used on. For example, smaller devices get an easy to navigate design whereas larger displays get the normal original design.

The main page of the system is divided into different sections. Each section displays a respective news website search results so that it's simpler for the user to navigate through them.

The sections of the system i.e. the structure is designed using HTML. The way this structure looks is developed in CSS, whereas, the behavior of any UI element is developed in JavaScript and jQuery.

The functional design includes extensive use of php. The system is designed in such a way that each result element in the system has its own data request. This is done using AJAX, so that the entire system does not get reloaded when the user inputs another search query.

Conclusion:-

Hence the system provides a web application that displays news from different news websites and social websites. Displayed content is in text, image, video formats. Its saves the trouble of looking for a single news topic on different news websites. It also provides public view of the news topic and how the public respond to it. It provides overall view of the topic and hence provides the context of the news and not just the updates. The system can further be improved by embedding it with a framework that integrates all the news and provides the summary and history of the news topic.

References:-

1. Arno Scharl, Alexander Hubmann-Haidvogel, Albert Weichselbraun, Gerhard Wohlgenannt, Heinz-Peter Lang, Marta Sabou, "Extraction and interactive exploration of knowledge from aggregated news and social media content"(2012), Proceedings of the 4th ACM SIGCHI symposium on Engineering interactive computing systems.
2. Zechao Li, Meng Wang, Jing Liu, ChangshengXu and Hanqing Lu, "News contextualization with geographic and visual information" (2011), MM '11 Proceedings of the 19th ACM international conference on Multimedia.
3. B. Batrinca , P. C. Treleaven , "Social media analytics: a survey of techniques, tools and platforms", Received: 25 February 2014/Accepted: 4 July 2014/Published online: 26 July 2014 at Springerlink.com.
4. PrashantKhare, Pablo Torres, Bahareh R. Heravi, "What just happened? A Framework for Social Event Detection and Contextualisation",1530-1605/15 $31.00 © 2015 IEEE DOI 10.1109/HICSS.2015.190 .
5. IvánCantador, Pablo Castells, "Semantic Contextualisation in a News Recommender System", CENIT-2007-1012.
6. Schrenk, M. Webbots, spiders, and screen scrapers: a guide to developing Internet agents with PHP/CURL. No Starch Press, 2007
7. Rahul Dhawani, Mrudav Shukla, PriyankaPuvar, BhagirathPrajapati, "A Novel Approach to Web Scraping Technology", Volume 5, Issue 5, MAY 2015, International Journal of Advanced Research in Computer Science and Software Engineering