CONTEXTUAL NEWS INFORMATION RETRIEVAL

Zain Ul Abedin

Faizan Ahmad

Annas Israr

Yaldram Shahzad

Supervised By:

Engr. Muhammad Umer Haroon

Submitted for the partial fulfillment of BS Software Engineering degree to the Faculty of Engineering & Computer Science

NATIONAL UNIVERSITY OF MODERN LANGUAGES ISLAMABAD

OCTOBER, 2019

ABSTRACT

Data generation and its growth rate is an abrupt process these days and will grow exponentially with each passing day. Users on the internet can enjoy abundant services and information in e-commerce websites, electronic newspapers, blog & social networks. Although this data is available for its consumption by users, quite an amount of time is spent retrieving this information and processing it. This has favored the research in several fields such as web scrapping. Web scraping, a process of extracting useful information from HTML pages, which is the main formatting tool of information on the internet today. Web scraping is a hot topic in today's perspective and it has multi faced applications. But two of the most important utilities of scraping are information retrieval for personal usage and for analytical purposes.

In this project, the aim is to do a survey of personalized information retrieval for statistical purposes, a specialized and crucial subsection of information retrieval and propose a system that will do this job on behalf of user. The proposed system will solve the above mentioned problem by searching the web pages for the relevant information and extracting the information that is relevant to the user's context. Methods that are choosed for information retrieval as Web Scraping, a technique that is extremely popular and is proven to have multi-domain usage these days. The proposed system is currently limited to Pakistani news websites only.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

Technology has entered every aspect of life these days, with a profound impact on people's way of living. In the recent years, almost every business is trying to provide their services online. Almost everyone is trying to discover approaches to do their work in a more convenient way. Everyone is looking forward to do their work efficiently and save their time. Although most of the services are now available on a single tap, there is a need of a system that promotes or allows you to extract data relevant to what you want, for your knowledge, because in this era people don't have time to go on each and every website and then search for the required resources and then compile it.

As the information become more and more online, billion articles of data are available for users to search from it, so it become more time consuming and frustrating. This system will help the news researchers to find their required data for analysis purposes just by entering the keywords related to respective domains. This system is strictly specific to finding and extracting data related to newspapers. It helps journalists as well as students related to specific field to find the accurate and previous data available online by entering specific keywords, which makes it much helpful for them to analyses, to make their reports, presentations and so on.

2. Existing systems

Systems related to the proposed system are listed below.

2.1 Google News

Google News is a news aggregator app developed by Google. It presents a continuous, customizable flow of articles organized from thousands of publishers and magazines. [1]

2.1.1 Features:

- Daily updates.
- It provides different categories for the user to look up to information.
- It provides a search facility to the user.
- It categorizes the articles when the user searches something.

2.1.2 Limitations:

- It does not have any E-newspapers only contains articles and blogs.

- It provides a wide list of articles that may be impossible for the user to read.
- It provides links of news, does not have any mechanism to document the information or download it, user has to visit links individually and extract relative information manually.

2.2 ABC News

ABC News is the news division of the American Broadcasting Company. [2]

2.2.1 Features

- It has tabs for videos and live news.
- It allows the user to read top stories.
- It provides a search facility which allows the user to search the required news.

2.2.2 Limitations

- It shows links of sources that contain information.
- No mechanism to download information.
- It only provides the links that take you blog sites, does not analyze newspapers or pictorial information.

3. Proposed System

This proposed system will work on the limitations found out in the similar existing systems. The proposed system will be a web based application that would be of major assistance to news researchers, news reporters and to the general public by providing them with their relevant information, by searching the internet, finding newspapers and articles and extracting relevant data from them on behalf of users according to their query. This will save a lot of time.

3.1 Features of the Proposed System

- The proposed system will provide a new visitor with the top trending news.
- The proposed system will enable the user to register on this application and will save user's preferences according to choices made at the time of registration.
- The proposed system will provide a functionality to the user to look up news on their interests.

- The proposed system will display any past events of user's interests on that particular date.
- The proposed system will search and provide relevant content according to user's query.
- The proposed system will provide news from events of approximately past 15 years.
- The proposed system will provide an option to import the compiled information on users system in the form of a document.
- The proposed system is not only limited to textual data on websites but it can also analyze articles and pictorial information.
- The proposed system will generate a document or report of the collected data along with references.

3.2 Limitations of proposed system

- The system is for audience who can read and understand English language.
- The system is currently limited to 6-8 Pakistani newspaper websites only.

3.3 Flow chart

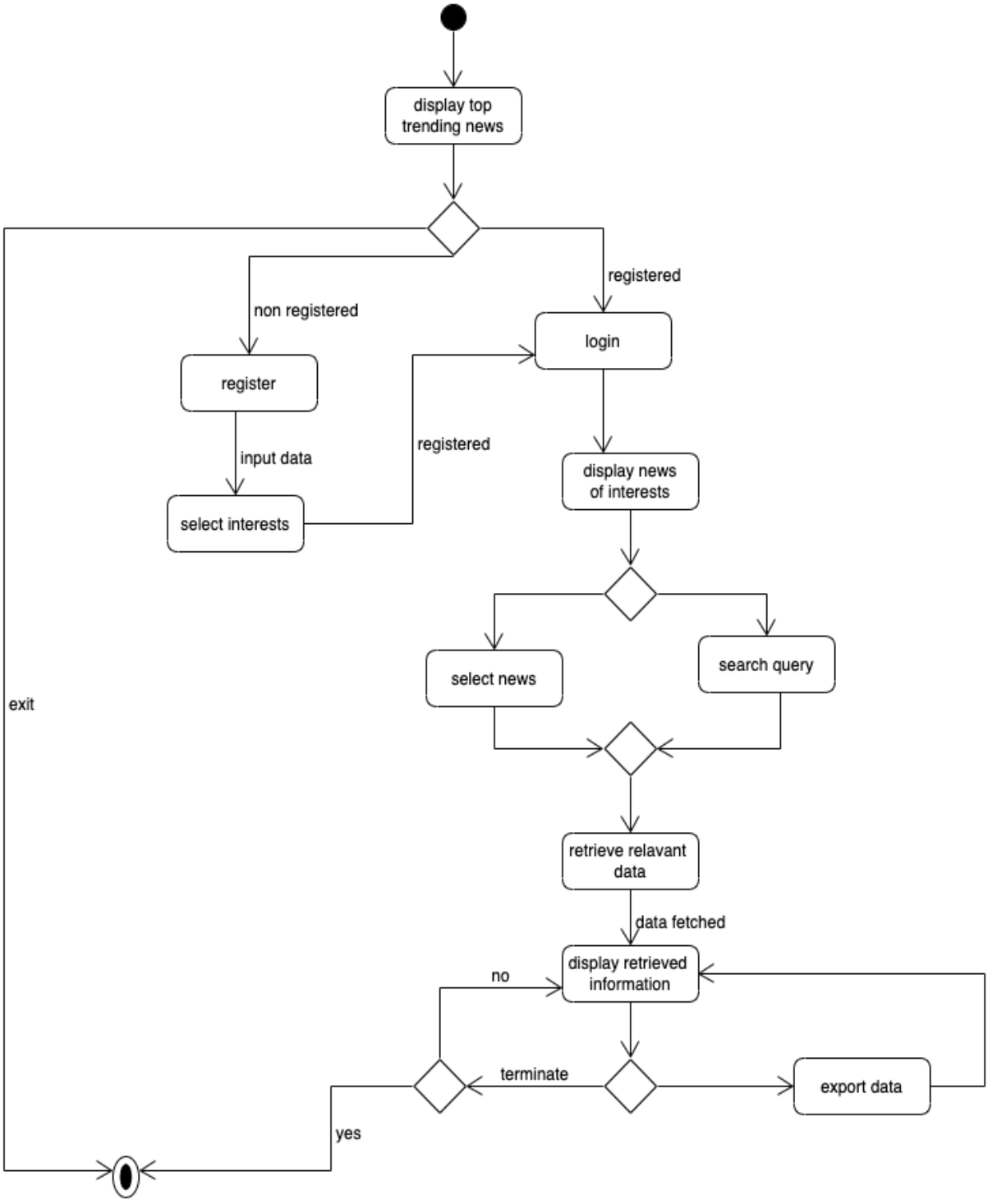

Figure 1: Flow chart of proposed system.

4. Proposed plan for implementation

The work time required for this project is split into following phases:

4.1 Idea exploration

Idea exploration phase was started in October.

4.2 Requirement Specification

Requirements gathering will be completed in 20 days after idea's acceptance.

4.3 Modeling Phase

System modeling includes data flow paths and system interface. It will be completed in 1 months after completion of requirements specification.

4.4 System Development

After successful completion of modeling phase, development will start from December 2019 to May 2020.

4.5 System Testing

Testing will be performed during each phase to maintain quality of the product. After all modules are developed, they will be integrated together and integration testing of the system will be performed.

4.6 Documentation

The project's progress reports will be submitted time by time to give a complete overview about the progress of the project. Documentation will be started during requirements specification and completed after application development.

5. Scope

This project will be a web-based application, that will work on systematic browsing of news on internet based on related input scripts and will retrieve the information in a most relative and accurate way. This project will help journalists, students and public to extract the news from available online data to minimize their time and effort and provide the required desired information with the media having consider as fourth pillar of any society, this will optimize media components. The project will also generate a document or report of the collected data

along with references. The main thing about the project is to save the data permanently in the user's local device.

6. Aims and Objectives

The objective of this web-based application is to help the researchers to find the relevant and accurate data with minimum effort. This system helps to find the available online data by entering specific keywords related to that field (currently limited to news). The aim of this system is to facilitate journalists, students and any other person related to it to solve their case studies in short span of time. We mainly aim to provide best, qualitative and relevant information from bulk amount of data available online.

- This web application aims to help researchers to find relevant and accurate data about news with minimum effort.
- This web application aims to provide the data demanded by user without having the need to physically going through all the related links.
- This web application aims to facilitate journalists, reporters, students and general public looking forward to get information about a specific issue, people or event.
- This web application aims to provide best, qualitive, specific and relevant information from bulk amount of data available online.
- This web application aims to provide a base for a non-professional to know each and everything related to a specific information that is published by different journalists or news websites.

7. Comparison of Existing Systems and Proposed System

The Comparison of proposed system with existing system is given below.

Table 1: Comparison of existing and proposed system

Features	Features availability		
	Google News	**ABC News**	**Proposed system**
Web Application	Yes	Yes	Yes
Textual extraction	No	No	Yes
Text extraction from pictures	No	No	Yes
Multimedia supported extraction	No	No	Yes
Import data to local disk	No	Yes	Yes

8. Resources

The main resources required to complete the proposed system are given below.

8.1 Software Requirements

The main software requirements for proposed system are given below:

Table 2: Software Requirements

	Tool	**Version**	**Rationale**
Tools and Technologies	Adobe Dreamweaver	2017	An IDE for Web Designing.
	Sublime Text	3	Clean, functional, and fast code editor.
	Adobe Photoshop	2017	A simple feature enriched tool for illustrations related to application.
	Flask	1.1x	A micro web framework written

			in Python. It is classified as a microframework.
	Pytorch	1.3	An open source machine learning library based on the Torch library.
	MySQL	8.0	Open-source relational database management system.
	Technologies	**Version**	**Rationale**
	HTML	5	A mark-up language to define page structure.
	CSS	3	Used for the styling of web pages.
	Python	3.7	Commonly used language for client and server side programming.

8.2 Hardware Requirements

The minimum hardware requirements of the system are below.

Table 3: Hardware Requirements

Processor	1.0 GHz dual core
Memory	4GB
Internal storage	30GB
Web Browser	Any web browser that supports JS

8.3 Functional Requirements

8.3.1 User Registration

System should be able to register a new user to the database and ask for users interests to provide a better user experience.

8.3.2 User Dashboard

On successful login, system should show latest news about users interest area, this should be updated every 4 hours and new topics should be added, if any. User can click the news and can read further information about it.

8.3.3 Analyze User Script

System should analyze the keywords in the users query using NLP (Natural Language Processing), to get the semantic annotation, input of the user will be based on some predefined parameters.

8.3.4 Scrapping of pages

System should scrap the pages that have the relevant information, this can be done by using Selenium Library.

8.3.5 Past News

System should show past news or events related to current date on the landing page of the application, user can click and read in detail.

8.3.6 Extraction of Text from HTML Pages

System should extract the text from html pages, using Scrappy [3].

8.3.7 Extraction of content from Pictures

System should analyze the text relevant to the query from pictures using OCR.

8.3.8 Extraction of text from tables

System must be able to extract and analyze text from tables using Tabula-Py, which allows you to read data from tables and other pages and export their data to external files.

8.3.9 Saving Data on User System

System should be able to show the extracted data with references(sources) and should allow the user to download the compiled information on his system, this can be done using Pandas.

8.4 Non-Functional Requirements

8.4.1 Security

The system's back-end servers shall only be accessible to authenticated administrators.

8.4.2 Reliability

The system provides storage of all databases on redundant computers with automatic switchover. The reliability of the overall program depends on the reliability of the separate components. The main pillar of reliability of the system is the backup of the database which is continuously maintained and updated to reflect the most recent changes.

8.4.3 Availability

The system may be available to access by user by using an application, only restricted by the down time of the server on which the system runs. In case of a hardware failure or database corruption, an alert dialog will be shown. Also in case of a hardware failure or database corruption, backups of the database shall be retrieved from the server and saved by the administrator. Then the service will be restarted. It means 24x7 availability.

8.4.4 Maintainability

MySQL database is used for maintaining the database and the application server takes care of the system. In case of a failure, a re-initialization of the program/server will be done. Also the software design is being done with modularity in mind so that maintainability can be done efficiently.

8.4.5 Portability

The application is web based, so any device using any world wide web browsers should be able to use the features of the system, including any hardware platform that is available or will be available in the future.

9. Gantt chart

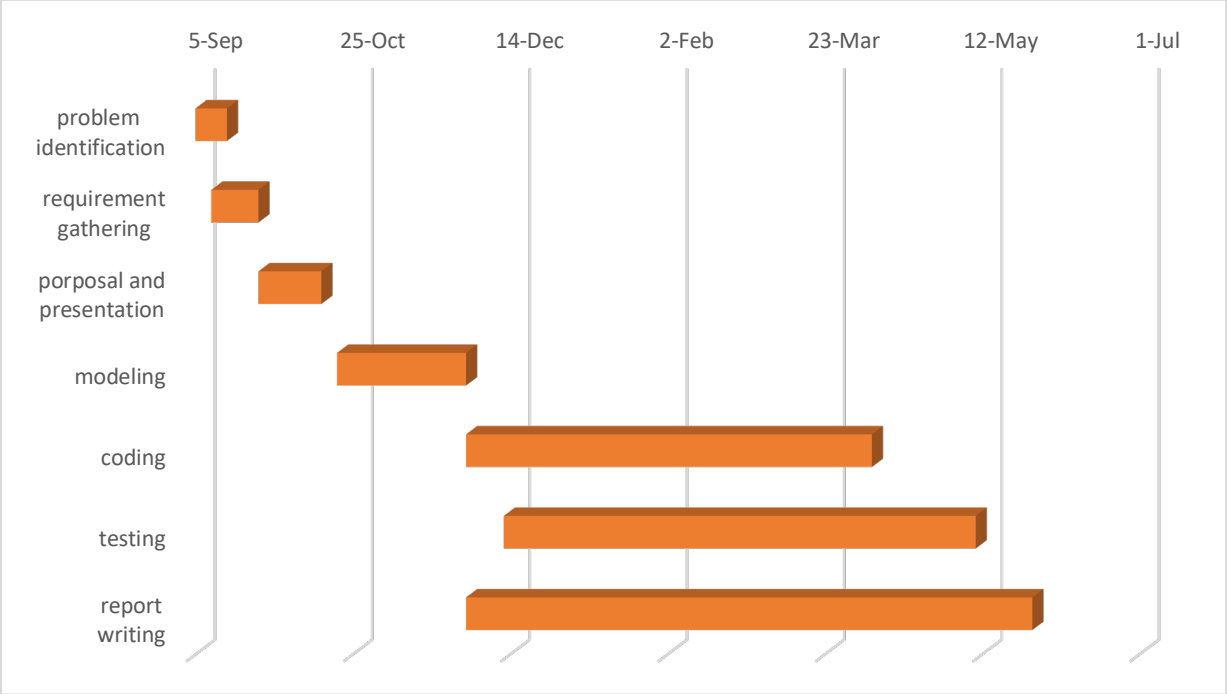

Figure 2: Gantt Chart

10. Conclusion

Now-a-days where media is considered as a building block for each society. Just because of its importance people are investing a lot, making it a great business hub and competition among various channels, journalists need a lot of information and they are in need of information at each minute to make the competition alive, for this purpose they research a lot to find relevant information about any case, which consume a lot of their time, so in order to provide relevant information from previous past newspapers, journals and articles in minimum span of time just by entering some keywords related to specific domain and the desired information will be in front of his/her web page in minimum amount of time.

11. Deliverable Outcomes

The outcome of the system will be in the form of a web based application for the proposed system “Personalized News Information Retrieval”. Final Project report will also be compiled after the complete development, execution and working of the proposed system.

References

[1] "Google News," Google, [Online]. Available: news.google.com. [Accessed 25 10 2019].

[2] "ABC News," American Broadcast Company, [Online]. Available: www.abcnews.go.com. [Accessed 25 10 2019].

[3] "Scrapy," [Online]. Available: https://scrapy.org/. [Accessed 20 10 2019].

[4] "DIFFBOT," [Online]. Available: https://www.diffbot.com/dev/docs/article/. [Accessed 20 10 2019].