Software Proposal Document for project News Aggregator

Alaa Mohamed, Marwan Ibrahim, Mayar Yasser,Mohamed Ayman

October 9, 2019

Abstract

News Aggregator is simply an online software which collects new stories and events around the world from various sources all in one place. News aggregator plays a very important role in reducing time consumption, as all of the news that would be explored through more than one website and social media platforms will be placed only in a single location.Also summarizing of all of the aggregated content absolutely will save reader's time, instead of reading from more than one source,a summarized page will be shown which contains the required article with the relevant articles all in one page.The main goal of this project is to develop a news aggregator with machine learning approach able to aggregate relevant articles of a certain input article and summarize all this information in one page.

1 Introduction

1.1 Background

In the last few years,The world had an incredible and huge growth of the rate of news,so readers need to keep track of what events are happening.There are so many examples of what was said about the news growth in the internet like the number of people using social media and how they give their opinions through it which will also give information, but let's be more concentrated in the scope of this proposal and talk about the news industry which is lately becoming all way online,for example if you want to search about a certain event,you'll see lots of electronic publishers who have talked about that event or that topic and that's proof that electronic versions of news are getting more important than old fashioned or traditional paper versions.So there were a lot of services were built to satisfy the presence of the electronic versions of the news as portals which aggregate news from plenty of sources.

Despite of the pros of the presence of lots of information to the people through the internet,it will get us another problem which is information overload,there will be too much information that is in front of the user and might be not his interests,which we can say that our news aggregator system will solve it by two main functionalities, the first one is that the system will have a major requirement which is summarizing relevant articles from the various sources talking about the event and writing this summarization in one page,which will have a lot of pros like reducing the time of reading to the user and getting only the useful and real news. The second requirement is knowing the user's interests and according to this interests the user will see a particular type of news.There are three types of news which could be shown to the user:Personal news,Stereotype news and Related news.

1.2 Motivation

1.3 Market Motivation

Online social networks a useful tool for collecting, aggregating and consuming the specific or general contents for various purposes in a certain period of time.Newspapers are in competition with modern online media. Among online media sources, news aggregators appear to be the most significant. An Outsell report (2009), 57 percent of news media clients go to computerized sources, and they are too more likely to turn to an aggregator 31 percent than to a newspaper site8 percent or other news sites 18 percent.[4]

1.4 Academic Motivation

Hamborg et al. [1] accumulate news articles in HTML organize from news websites employing a Python framework specifically scrapy. The most objective is to extract distinctive components of the web-page, such as title, the content of news, lead passage, distribution information, the news creator and the related image. Felix Hamborg, Norman Meuschke, and Bela Gipp [2] developed a web-based system to integrate different perspective news article

1.5 Problem Definitions

1. Not all of the news aggregators have different perspectives and also not all of them give the user variety of options through exploring, for example some news aggregator have just the title of the event with some details, but it would be better for the reader if he has the option of watching a video of this story or read it in an informal way as a blog.

2. User interface and its language are also a big problem which faces a lot of news aggregators, most of those systems are having an English language as a major language and some of them do not have any other language, and that will reduce a number of users who would use the system but can't do it because of that user interface language, so from the challenges that this system would do it is the ability to convert from English to other languages like Arabic for example.

3. Taking into consideration that also accuracy enhancement is a target in detecting and classifying articles to its categories

4. Most of the news aggregators don't have sources Ranking System that can avoid contradiction

2 Project Description

2.1 Objective

Our main objective is to reduce the time estimated to search for a topic at a website as the users spend a lot of time finding articles on different websites so they can quickly access only one article summarized and filtered. In addition, There will be a ranking system for Websites agencies and the Relevant Articles and Media

2.2 Scope

1. Aggregate news, articles and tweets all in one place.

2. The system will crawl the relevant articles when user input some article to read about it.

3. Summarize the aggregated information from various sources in one page in order to reduce reading time.

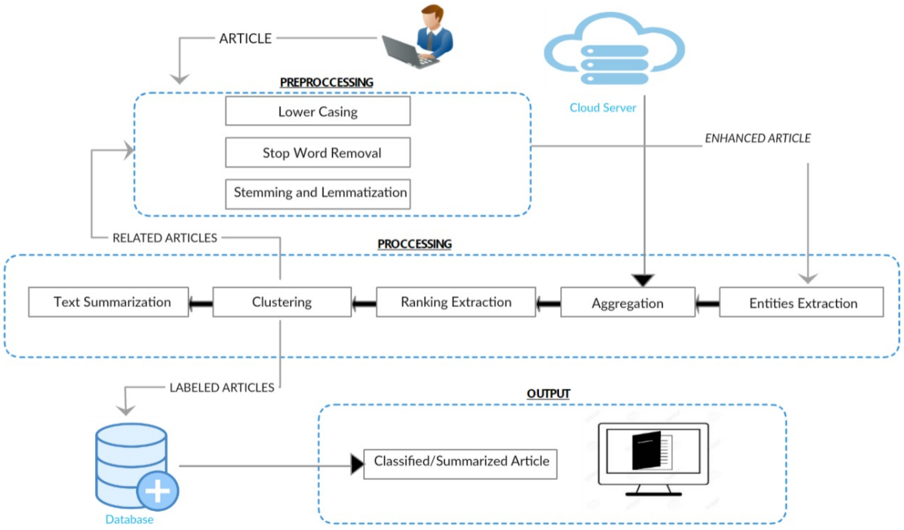
4. Adding Ranking System for the websites agencies/sources in order to avoid contradiction of news

5. Another Ranking system for viewing the most relevant article

6. Adding the Most relevant Media(Images-Videos) through the article according to the website agencies

2.3 Project Overview

Fast-Text was used at first which is a library for text classification and word embedding which was created by Facebook.It's used in the system for classification a supervised data set and classifying its text.



- The user enter an article news or search for a certain event

1. PreProcessing:

    - Lowercase: used to reduce the size of the vocabulary in our data that cause multiple copies of the same word meaning
    - Stop-word Removal: Done to remove small information of a text in order to focus on important words
    - Lemmatization and Stemming: Remove inflection and map the word into original/root form

2. Processing:

    - Entities Extraction: classifying key elements from a text into predefined categories
    - Aggregation: Collection of ranked news articles from cloud service.
    - Clustering: Finding out which articles are similar to each other then label and group them, afterwards it send the related articles back to prepossessing and send the labeled articles to the database
    - Summarization: Getting the main and important information.

3. Output: Summarized and well classified news article for the user to read

3 Similar System Information

- Matrix-based News Aggregation: Exploring Different News Perspectives [2]:

1. Reducing media bias
2. The main problem is that writing and presenting the news could affect the reader perspective
3. By using the MNA approach they explore both common and different information in related articles
4. MNA groups news articles into the cells of a matrix spanned over two dimensions, which are selected to maximize the expected diversity in the resulting cells
5. MNA leads to higher accuracy for highlighting different perspectives on topics

- The Improvement of Indonesian News Curator Classification in Twitter [7]:

  1. Twitter user as news curator would be a critical an valuable client for making a difference news making handle particularly as news source
  2. There was 2 kind of news guardians on twitter To begin with, news story curators, news curator as human client that made tweet by himself. The second one is, news aggregator, news curator as bot (programmed) client that made tweet automatically from tools
  3. By using API streaming and collecting 12 features ((location, follower, profile description, verified, website, URLs, mention, hashtag, retweet, general tweet, number of RT and number of like) and saved it in database
  4. URLs feature the most effective and verified feature the lowest effectiveness in our framework. For the best feature combination.
  5. Help us to select the best combination of features that gives the highest accuracy which was 95.55 percent.

- NEWSONE- AN AGGREGATION SYSTEM FOR NEWS USING WEB SCRAPING METHOD [8]:

  1. In an attempt to extract the news from multiple sites, newspapers, magazines, and television and merge them all in a single platform. It is also categorized into many categorize and classify them according to the pieces of information. The aim of this project is to save readers' time, in addition, to make them know all the information about a specific topic with the latest news about it.
  2. The web pages become full of a large amount of information which leads to inconsistency in knowing information about a topic and not all the websites will have the same data. Sometimes reader wastes a lot of time without reaching the data that he wants to know.

3. The web pages become full of a large amount of information which leads to inconsistency in knowing information about a topic and not all the websites will have the same data. Sometimes reader wastes a lot of time without reaching the data that he wants to know.
4. The web pages become full of a large amount of information which leads to inconsistency in knowing information about a topic and not all the websites will have the same data. Sometimes reader wastes a lot of time without reaching the data that he wants to know.
5. The web pages become full of a large amount of information which leads to inconsistency in knowing information about a topic and not all the websites will have the same data. Sometimes reader wastes a lot of time without reaching the data that he wants to know.

- Atlas: News Aggregation Service [1]:

  1. The researchers aim to improve the news aggregator platform by merging the RSS news with social media like twitter and news web-services that provides API. This platform collects data from more than one website and provides people to read articles in any language. Also, they want to apply a new idea that is to create an account that will help each reader will have his own account and he can favorite the articles he wonders.
  2. News aggregator is to drag together content from numerous channels such as news features, articles from blogs or social systems posts but without getting what is written in twitter and Reddit.
  3. They solve it by using news web-services that supply Twitter Application Protocol Interface so this requires the usage of HTTP requests. Also, they apply reading articles in more than one language. They use REST APIs to help them in creating accounts.
  4. The application creates accounts for users also presents for them a lot of options that help them to reach the article quickly. The interface uses simple and easy components for Designing. The reader can use the search bar that lies at the top of the page in order to filter articles by words.
  5. This project uses Twitter API and Reddit API so this will help us in concatenate the news information with the tweets that are written on twitter. In addition, The idea of creating an account is very good. [5]:

- THE RISE OF THE NEWS AGGREGATOR:Legal Implications and Best Practices. [3]:

  1. According to the study and statistics, the researchers reach that the internet has become a very important root to know the news. So they decided to make one single platform that collects the pieces of

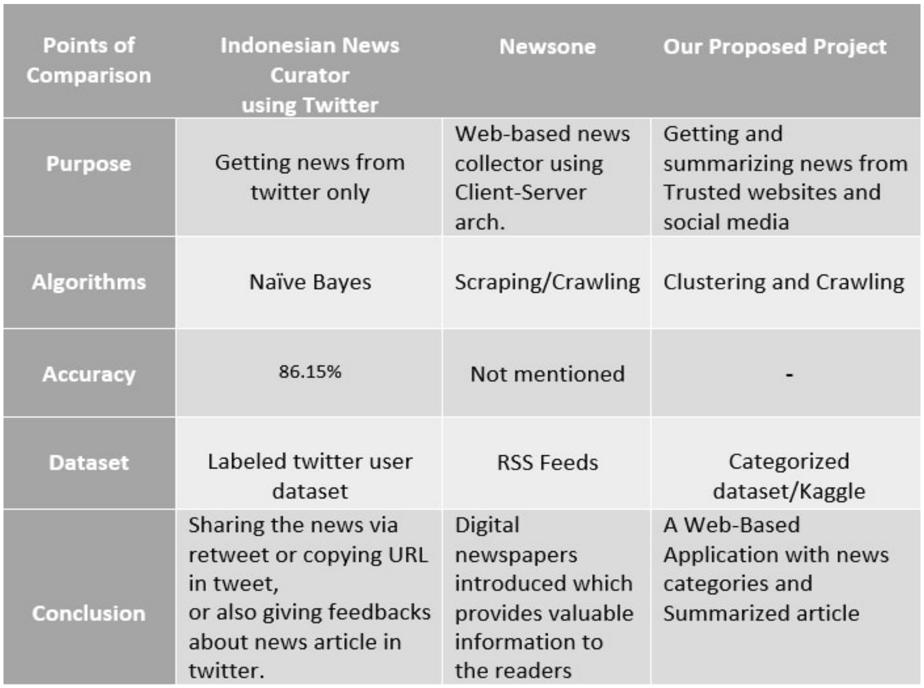
information that are written on different websites.to be easy instead of opening each website and read the news from it.

2. When the user wants to read an article, he will access a lot of sites to know everything about it and if he used Google News or Yahoo News, these sites wouldn't contain all the news.
3. They invented this project that is divided into various categories to help users to select which category they want to read in a single site that will save users' time.
4. Website with all information about news in summarization and the information is analysis between different categories of news aggregators.
5. The website is divided into categories and this makes the site more simple and easy.

3.1 Similar System Description

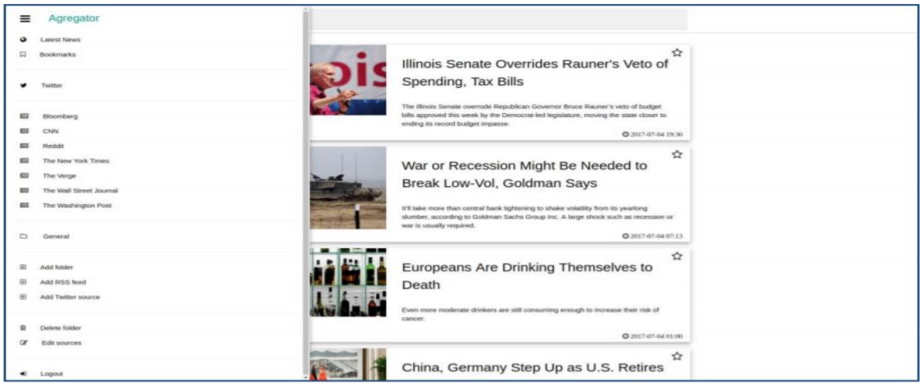
NewsOne is a news aggregator website, It provides a adequate way to keep up with updates on news sites, without re-visiting news sites manually. It provides news and content in categories (Just-In, Technology, Business and Economics, Sports, Science, Entertainment, Health).

3.2 Comparison with Proposed Project



Points of Comparison	Indonesian News Curator using Twitter	Newsone	Our Proposed Project
Purpose	Getting news from twitter only	Web-based news collector using Client-Server arch.	Getting and summarizing news from Trusted websites and social media
Algorithms	Naïve Bayes	Scraping/Crawling	Clustering and Crawling
Accuracy	86.15%	Not mentioned	-
Dataset	Labeled twitter user dataset	RSS Feeds	Categorized dataset/Kaggle
Conclusion	Sharing the news via retweet or copying URL in tweet, or also giving feedbacks about news article in twitter.	Digital newspapers introduced which provides valuable information to the readers	A Web-Based Application with news categories and Summarized article

3.3 Screen Shots from previous systems



Illinois Senate Overrides Rauner's Veto of Spending, Tax Bills

The Illinois Senate overrode Republican Governor Bruce Rauner's veto of budget bills approved this week by the Democrat-led legislature, moving the state closer to ending its record budget impasse.

2017-07-04 19:30

War or Recession Might Be Needed to Break Low-Vol, Goldman Says

It'll take more than central bank tightening to shake volatility from its yearlong slumber, according to Goldman Sachs Group Inc. A large shock such as recession or war is usually required.

2017-07-04 07:13

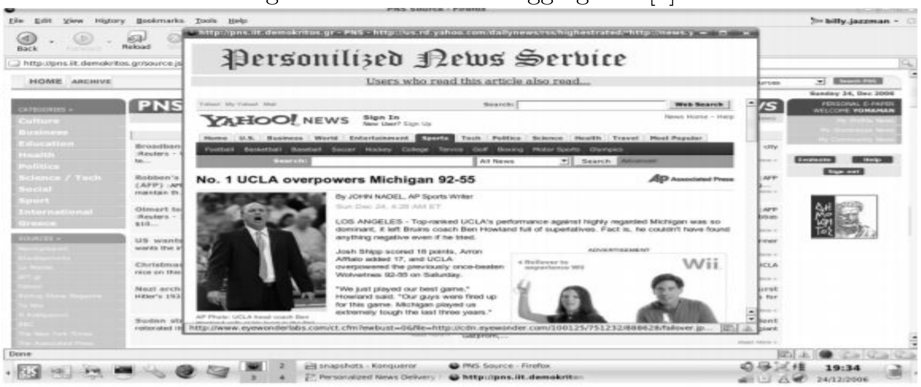
Europeans Are Drinking Themselves to Death

Even more moderate drinkers are still consuming enough to increase their risk of cancer.

2017-07-04 01:00

China, Germany Step Up as U.S. Retires

Figure 1: Atlas: News Aggregation[1]


File Edit View History Bookmarks Tools Help
billy.jazzman
Back
Reload
http://pns.iit.demokritos.gr/source.js
HOME ARCHIVE
PNS
CATEGORIES
Culture
Business
Education
Health
Politics
Science / Tech
Social
Sport
International
Greece
SOURCES
Sunday 24, Dec 2006
PERSONAL E-PAPER
WELCOME YOMAMAN
http://pns.iit.demokritos.gr - PNS - http://us.rd.yahoo.com/dailynews/rss/highestrated/*http://news.y
Personilized News Service
Users who read this article also read...
YAHOO! NEWS
Sign In
New User? Sign Up
Search:
Web Search
News Home - Help
Home U.S. Business World Entertainment Sports Tech Politics Science Health Travel Most Popular
Football Basketball Baseball Soccer Hockey College Tennis Golf Boxing Motor Sports Olympics
All News
Search
Advanced
AP Associated Press
ADVERTISEMENT
Wii
http://www.eyewonderlabs.com/ct.cfm?ewbust=0&file=http://cdn.eyewonder.com/100125/751232/888628/failover.jp...
Done
snapshots - Konqueror
PNS Source - Firefox
Personalized News Delivery
http://pns.iit.demokritos
19:34
24/12/2006

No. 1 UCLA overpowers Michigan 92-55

By JOHN NADEL, AP Sports Writer

Sun Dec 24, 4:28 AM ET

LOS ANGELES - Top-ranked UCLA's performance against highly regarded Michigan was so dominant, it left Bruins coach Ben Howland full of superlatives. Fact is, he couldn't have found anything negative even if he tried.

Josh Shipp scored 18 points, Arron Afflalo added 17, and UCLA overpowered the previously once-beaten Wolverines 92-55 on Saturday.

"We just played our best game," Howland said. "Our guys were fired up for this game. Michigan played us extremely tough the last three years."

AP Photo: UCLA head coach Ben

Figure 2: PNS: A personalized news aggregator on the Web[6]

4 Project Management and Deliverables

4.1 Tasks and Time Plan

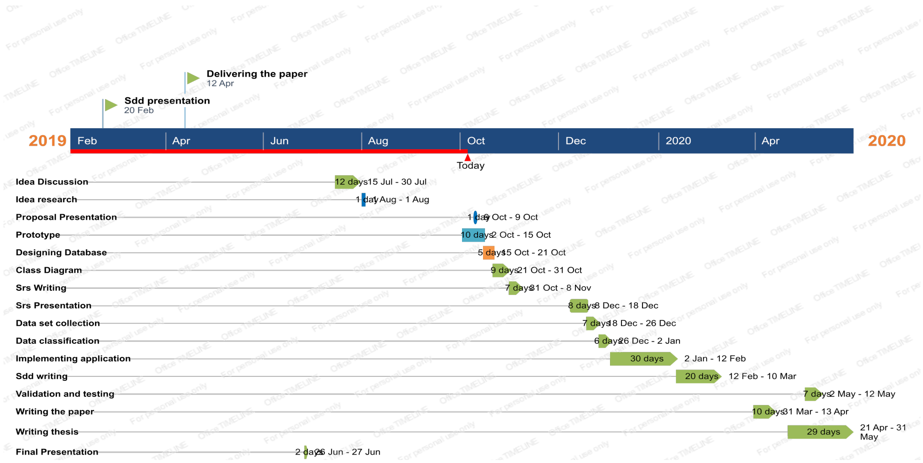
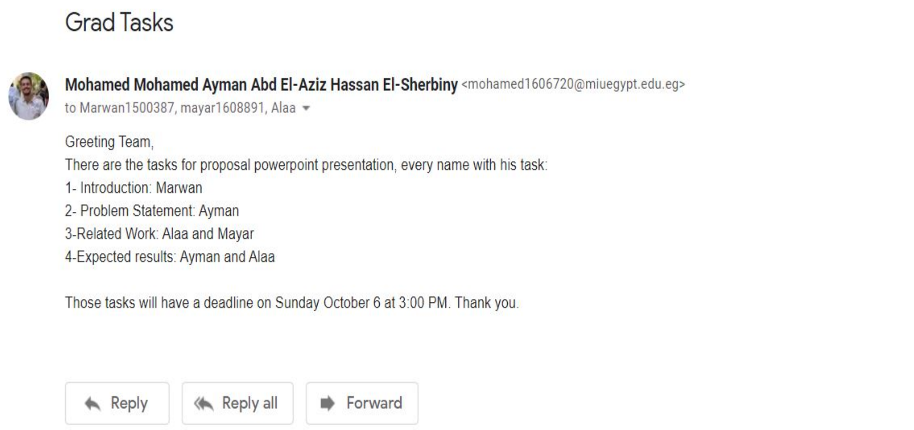


Figure 3: Gantt Chart for time plan



Grad Tasks

Mohamed Mohamed Ayman Abd El-Aziz Hassan El-Sherbiny <mohamed1606720@miuegypt.edu.eg>
to Marwan1500387, mayar1608891, Alaa

Greeting Team,
There are the tasks for proposal powerpoint presentation, every name with his task:
1- Introduction: Marwan
2- Problem Statement: Ayman
3-Related Work: Alaa and Mayar
4-Expected results: Ayman and Alaa

Those tasks will have a deadline on Sunday October 6 at 3:00 PM. Thank you.

4.2 Budget and Resource Costs

Our graduation project is a web application that will be accessed directly from mobile phones,PCs or laptops and no extra budget will be needed for development.

References

[1] Cosmin Grozea et al. “Atlas: News aggregation service”. In: *2017 16th RoEduNet Conference: Networking in Education and Research (RoEduNet)*. IEEE. 2017, pp. 1–6.

[2] Felix Hamborg, Norman Meuschke, and Bela Gipp. “Matrix-based news aggregation: exploring different news perspectives”. In: *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*. IEEE Press. 2017, pp. 69–78.

[3] Kimberley A Isbell. “The rise of the news aggregator: Legal implications and best practices”. In: *Berkman Center Research Publication* 2010-10 (2010).

[4] Doh-Shin Jeon and Nikrooz Nasr Esfahani. “News Aggregators and Competition Among Newspapers in the Internet (Preliminary and Incomplete)”. In: (2012).

[5] Aibek Musaev et al. “Fast text classification using randomized explicit semantic analysis”. In: *2015 IEEE International Conference on Information Reuse and Integration*. IEEE. 2015, pp. 364–371.

[6] Georgios Paliouras et al. “PNS: A Personalized News Aggregator on the Web”. In: vol. 104. Jan. 1970, pp. 175–197. DOI: `10.1007/978-3-540-77471-6_10`.

[7] Jaka E Sembodo, Erwin B Setiawan, and ZK Abdurahman Baizal. “The improvement of Indonesian news curator classification in Twitter”. In: *2017 5th International Conference on Information and Communication Technology (ICoIC7)*. IEEE. 2017, pp. 1–7.

[8] K Sundaramoorthy, R Durga, and S Nagadarshini. “NewsOne—An Aggregation System for News Using Web Scraping Method”. In: *2017 International Conference on Technical Advancements in Computers and Communications (ICTACC)*. IEEE. 2017, pp. 136–140.