News Aggregator and Efficient Summarization System

Alaa Mohamed[1], Marwan Ibrahim[2], Mayar Yasser[3], Mohamed Ayman[4],
Menna Gamil[5], Walaa Hassan[6]
Faculty of Computer Science
Misr International University
Cairo, Egypt

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 will be placed only in a single location. Also, summarizing this aggregated content absolutely will save reader's time. A proposed technique used called the TextRank algorithm that showed promising results for summarization. This paper presents the main goal of this project which is developing a news aggregator able to aggregate relevant articles of a certain input keyword or key-phrase. Summarizing the relevant articles after enhancing the text to give the reader understandable and efficient summary.

Keywords—*News aggregator; text summarization; text enhancement; textRank algorithm*

I. Introduction

In the last few years, the world had incredible and huge growth in the rate of news that is published [1]. People live in a time full of information, data, and news [2]. So, nowadays news has an important part and position within the community. As people read the news daily to keep up with the most recent data and inputs. These data may be about technology, sports, weather, food, and celebrities or many other fields [3].

With the development of the Internet, and lot of websites that provide the same data and information, getting this has become simpler getting to it has become simpler [4]. So, users frequently discover it troublesome to decide which of these websites can provide the specified data within the most valuable and effective way [2].

The conventional commerce model of daily papers has been threatened by the internet to lessening their advertising income and by presenting new online media, such as web-only news, blogs and news aggregators [5]. The online social system is a valuable instrument for collecting, aggregating and expending the specific or common contents for different aims in a certain period of time. Daily papers are in competition with present-day online media as shown in Fig. 1. Among online media sites, feeds aggregators show up to be more significant [5] [6].

An Outsell determination (2009) [5], 57% of feeds media clients go to computerized sites, and they are too more likely to turn to an aggregator 31 % than to a daily paper location 8% or other news sites 18% [5]

Feeds aggregator combines news data, and regularly briefs it in a good format and design for the reader, from various sites,
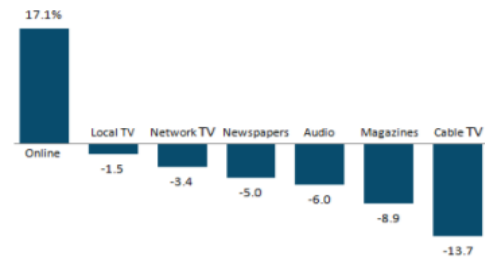

Fig. 1. Online feeds Develops Quickly contrast to Others

newspapers, and agencies [2]. News aggregators are frequently included in classifications such as "Websites Each Engineer Ought to Visit" [7].

Despite 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 not be his interests [8]. This problem can be solved throughout the proposed system. As News Aggregator looks like a gateway that integrates different feeds websites, it organizes feeds by subject [9]. It could be a site that takes data and news from numerous sources and displays it in a single site [10]. Which simplifies readers' search and reading time for news by gathering content based on viewing history [11]. Using news aggregation is one of the best ways to stay on top of the news and topics you want. They offer convenience and time-saving features [12].

News Aggregator system will have a major requirement which is Summarization. Summarizing articles from various sources talking about the same event then writing the content of this event on one summarized page with all perspectives [13]. Summarization is to create a shorter and smaller form of a text by protecting its meaning and the key substance of the initial content [14] [15]. So, summarization has a lot of pros like reducing the time of reading to the user and getting only useful and real news. Content summarization methods can be categorized into Extractive summaries and Abstractive summaries [16]. Extractive Summarization depends on extracting a few parts, such as phrases and sentences, from a piece of text and gather them together to form a summary. Therefore, identifying the right sentences for summarization is of the most extreme importance in an extractive method [17]. But Abstractive Summarization utilizes advanced NLP

methods to generate an completely new summary. A few parts of this summary may not indeed appear within the original text [17]. In this paper, we follow extracting summarization technique which gives better output and right sentence for Summarization.

The rest of the paper is structured as follows: Section II reviews related work on News aggregator based on summarization. The proposed system is presented in Section III. Section IV presents experimental results about the summarization and performance analysis of the system. Finally, Section V concludes the paper and discusses some future work.

II. RELATED WORK

In this section, we will discuss other news aggregator websites which based on summarization:

In [18], the authors were focusing on gathering news using matrix-based analysis (MNA) with 5 main steps as follows: the first and second steps are data gathering and extracting the article from the websites and save it in the database. The third step is grouping where they categorize the articles. The last two steps are summarization and visualization that view the important article to the user. Before the grouping step, they added the matrix-based analysis where the matrix has entity as row and the column is the states about the entities. When starting analysis, the user defines what he's looking for where MNA prepare the default values for this purpose. After that, the initialization of the matrix extends a matrix over the two required chosen dimension and look in each cell for the cell documents. The summarization phase is done according to the following steps: topic summary, cell summary and summarizing both by using TF-IDF for each cell in the matrix.

According to [11], the authors were aiming to accumulate the content from diverse websites such as articles fond moreover news headlines from blogs and websites. The belief that Rich Site Summary (RSS) gives us summarized and short data. Which is preferable for the news aggregator that they are still a successful solution for indexing articles. As reducing the time required for visiting some websites, subscribed users can quickly utilize Rich Site Summary feeds without wasting time going to numerous websites. Creating HTTP requests from the web-server is the primary step in the application and these requests are received from clients. At that point, they utilize Python to download Rich Site Summary feeds and extract articles from it according to the input. After periods of time, the web-server gives some requests to the subscribed users and in case there are any upgrades, it'll be stored and downloaded.

Author in [19] was aiming to use Rich Site Summary integrated with HTML by using wrappers (programs) and parser in order to extract the information from a specific source, then adjust them according to news categories and personalized web views via a web-based interface. They explain how they do the content scanner by using HTML and Rich Site Summary. The first step is wrapping (HTML/Rich Site Summary wrapper) which involves identifying the URL address of the new items from the source with category per the news, and the address is stored in the database as for each category pair and also combined with the corresponding wrapper. The second step of wrapping is getting information from the new items, that will be used for getting and indexing the article, for each article they obtain the first sentence and pass it to the corresponding HTML page.

According to [9], the authors were aiming to collect the news from multiple sites, newspapers, magazines, and television and merge them all in one summarized website. It progresses the goodness of results because the contents and data in it are brief and summarized. So, their work based on the Rich Site Summary fetcher for recovering Rich Site Summary reports from specific websites at a certain time. They also use web Crawling (Scraping) besides Rich Site Summary to get more accurate results. Web scraping may be a method utilized to collect huge amounts of information from websites.

From all the above mentioned researches on the news aggregator, the quality of the aggregator system is still an open area to be introduced.

III. PROPOSED SYSTEM

In this section, the basic structure of the proposed system is described, which will be able to aggregate online news from cloud service and summarize its content to reduce user's reading time.

Fig. 2 shows the main stages of the proposed system. It consists of four main stages as described below:
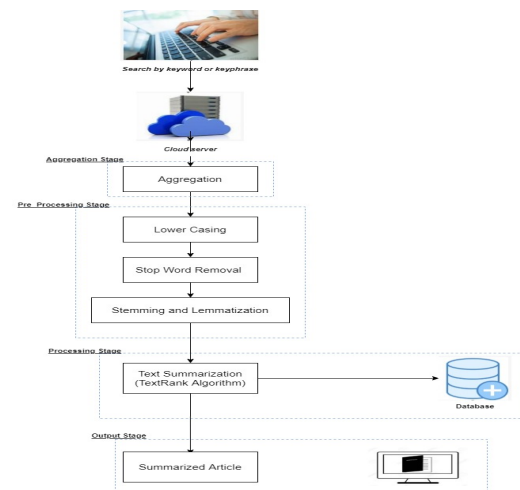

Fig. 2. Proposed System Overview

1) Aggregation stage: Aggregating related articles according to the keyword/phrase that the user enters. The articles are aggregated from different and credible sources like BBC, World Health Organization, etc.

2) Pre-processing stage: Consists of applying specific steps to the aggregated articles such as:

    a) Lowercase: used to reduce the size of the vocabulary in our data that cause multiple copies of the same word meaning.

b) Stop-word Removal: Done to remove small information of a text in order to focus on important words.

c) Lemmatization and Stemming: Remove inflection and map the word into the original/root form.

3) Processing stage: Consists of applying the summarization algorithm on the aggregated articles.
In this paper, we use the TextRank Algorithm [20] for summarizing articles which is a graph-based ranking algorithm that is used for summarization. Fig. 3 shows the steps of the textrank algorithm:
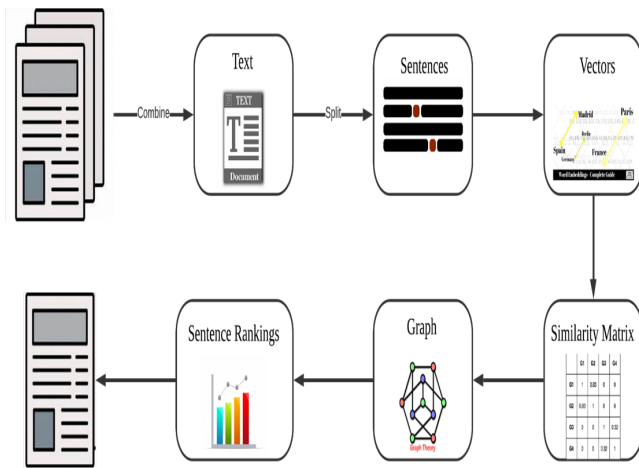

Fig. 3. Steps of TextRank algorithm

The steps can be illustrated below:

a) A set of articles are collected and combined as one one original article.

b) Tokenizing the original text as shown in Fig. 3 into sentences.

c) The Third step is vectorization. In this step, each word is represented by a vector based on the co-occurrence of a word with the others in a single sentence using Global vector algorithm (GloVe) [21]. Then we represent each sentence by a vector calculated from the mean of words vectors in a sentence.

d) In this step, we obtain a similarity matrix for all sentences using cosine similarity [22]. The similarity here refers to common content in sentences.

$$Cos\theta = \frac{\vec{a}.\vec{b}}{||\vec{a}|| * ||\vec{b}||} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i}\sqrt{\sum_{i=1}^{n} b_i}} \tag{1}$$

$$where, a.b = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \dots$$
$$+ a_n b_n$$

is the dot product of the two vectors.

e) After obtaining the similarity matrix in the previous step, we convert it into a Graph where the edges determined by a similarity relation between them. Those edges are used to obtain the vertices weight. The importance of a sentence is based on the number of edges that represented as a score for each vertex as shown in Fig. 4 using PageRank algorithm [23]. Let the directed graph,

$$G = (V, E) \tag{2}$$

where V represents set of vertices and E represents set of edges. The vertex score V_i is defined as follows:

$$S(V_i) = (1 - d) + d * \sum_{j \in (V_i)} \frac{1}{|(V_j)|} S(V_j) \tag{3}$$

where d is a factor that it's value is between 0 and 1 and usually the value is 0.85, which represents the probability of going to another random vertex from a given vertex in the graph.

```
{0: 0.0709214814920622,
 1: 0.06761143756846281,
 2: 0.06903441687181701,
 3: 0.06790578983952975,
 4: 0.06592006008715072,
 5: 0.07059702653895156,
 6: 0.07040802220461516,
 7: 0.0660010366566746,
 8: 0.04974690169623055,
 9: 0.06177281332647406,
 10: 0.07035545423941549,
 11: 0.0682389550735169,
 12: 0.06892467651993743,
 13: 0.0667355007117678,
 14: 0.0658264263890373}
```

Fig. 4. Sentence Scores

f) The final step is ranking the scores shown in Fig. 4 in descending order. The highest scores create the final summary shown in Fig. 5.

Cristiano Ronaldo scored yet another brace for Juventus Sunday to take his tally for the club to 50 goals. The Portuguese international only joined the Italian outfit in 2018 but has netted his half-century in just 70 games. Both of his latest goals came from the penalty spot as Juventus beat Fiorentina 3-0 to cement its place at the summit of Serie A. Dutch sensation Mattijs de ligt scored the game's only other goal with a late header. Ronaldo has now scored in nine consecutive Serie A games for Juventus, becoming the first man to do since David Trezeguet in 2005. He is also the second highest scorer.

Fig. 5. Illustrative example of TextRank algorithm

4) Output stage: Summary that contains key ideas to the topic is generated.

The proposed system overview will make the user able to explore news from more than one source. The system will also give the user the main feature which is a readable summary to all of the aggregated content of the same topic. In the next

section, an experiment will be discussed to compare between two summarization algorithms and to decide which one of them will fulfill the requirements of a good and accurate summary.

IV. Experimental Results

To validate the effectiveness of the proposed News Aggregator system, a set of experiments have been conducted with different keywords and key phrases. Also, the efficiency of the TextRank algorithm is tested against a common used algorithm which is Word Frequency [24]. Fig. 6 shows the original article that will be summarized, Fig. 7 and 8 are the output summary after applying summarization algorithms.

Fig. 6. Original Article

Fig. 7. Summary using Word Frequency

Fig. 8. Summary using TextRank

"Word Frequency" algorithm depends on more than one factor to summarize the input text as:

- The existence of the frequency table is the first step towards executing the algorithm, then every sentence will be tokenized.

- After tokenizing, we will have separated sentences, scoring every sentence will be the next step, which its formula is a division of every non-stop word in the sentence by the total number of words in the sentence.

- Getting average score of the sentences is the next step. A comparison will be done between every sentence and this average score, if the score sentence is larger, then this sentence will be considered as a summarized part of the input article.

In this paper, we used a summary evaluation tool named 'Rouge' for our summarization comparisons between TextRank and WordFrequency algorithms, it turns out that this algorithm has drawbacks. Drawbacks of Rouge tool is that its execution depends on the permanent existence of an expert one who knows the actual rules of summarization. Rouge executes by comparing the results of this expert to the system results and his existence might not be always available for us to have his/her consultation. This is the reason to find another evaluation criteria which was applying a survey with our social network, asking to rate each 2 summaries out of 5 which were implemented by 2 different algorithms. 1 refers to a very bad summary and 5 refers to an excellent one.
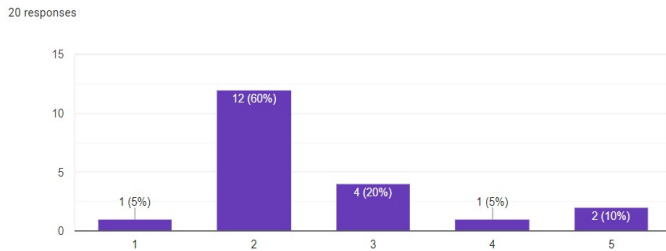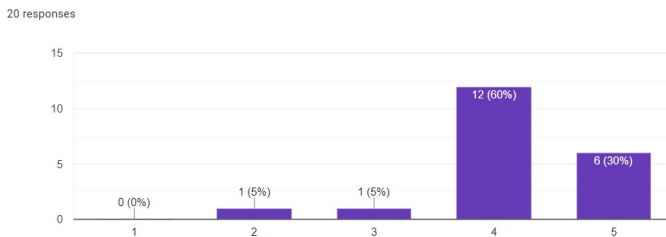
Fig. 9. Word Frequency Chart

Fig. 10. TextRank Chart

TextRank algorithm was rated as 4 out of 5 from 60% from people who read it and 30% were rating 5 out of 5 as a summary as shown in Fig. 10. While Word Frequency algorithm was having bad rates from the reader as 60% were rating its summary as 2 out of 5 which is of course a bad percentage as shown in Fig. 9. These results ensure that TextRank has more efficient results in summarization.

To further emphasize that, the output summary from the TextRank algorithm was revised by an expert besides the normal readers rating. The feedback considers that the output summary fulfills all the requirements of a good summary such as:

- The length is about 10% of the original.

- Short paragraphs that contain the key ideas and to the point.

- Could be read and clearly understood without referring to the original article.

- Used the appropriate language just like that used in the original article.

V. Discussion

The results of our proposed system presented a very good and understandable summary, as this information was ensured by an expert in this type of fields. TextRank summary was more acceptable in the survey results according to readers perspective and that's an indication that user is more comfortable with reading the summary after applying the experiment. Also, the system fully works online so any type of aggregated articles will be on the spot and will be chosen from trusted and determined sources.

VI. Conclusion

After testing two summarization algorithms, TextRank algorithm was the chosen approach to be applied in the summarization system over Word Frequency algorithm. The reason behind applying TextRank algorithm was simply that TextRank gives more efficient summary for the reader. The system will generate output summary from online sources that contains key ideas to the certain article topic and it could be understood without referring to the original article.

Acknowledgment

References

[1] S. Bergamaschi, F. Guerra, M. Orsini, C. Sartori, and M. Vincini, "Relevant news: a semantic news feed aggregator," in *Semantic Web Applications and Perspectives*, vol. 314. Giovanni Semeraro, Eugenio Di Sciascio, Christian Morbidoni, Heiko Stoemer, 2007, pp. 150–159.

[2] S. Chowdhury and M. Landoni, "News aggregator services: user expectations and experience," *Online Information Review*, vol. 30, no. 2, pp. 100–115, 2006.

[3] R. Bahana, R. Adinugroho, F. L. Gaol, A. Trisetyarso, B. S. Abbas, and W. Suparta, "Web crawler and back-end for news aggregator system (noox project)," in *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*. IEEE, 2017, pp. 56–61.

[4] G. Paliouras, M. Alexandros, C. Ntoutsis, A. Alexopoulos, and C. Skourlas, "Pns: Personalized multi-source news delivery," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2006, pp. 1152–1161.

[5] D.-S. Jeon and N. N. Esfahani, "News aggregators and competition among newspapers in the internet (preliminary and incomplete)," 2012.

[6] N. Diakopoulos, M. De Choudhury, and M. Naaman, "Finding and assessing social media information sources in the context of journalism," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2012, pp. 2451–2460.

[7] M. Aniche, C. Treude, I. Steinmacher, I. Wiese, G. Pinto, M.-A. Storey, and M. A. Gerosa, "How modern news aggregators help development communities shape and share knowledge," in *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*. IEEE, 2018, pp. 499–510.

[8] K. Lerman, "Social information processing in news aggregation," *IEEE Internet Computing*, vol. 11, no. 6, pp. 16–28, 2007.

[9] 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.

[10] K. A. Isbell, "The rise of the news aggregator: Legal implications and best practices," *Berkman Center Research Publication*, no. 2010-10, 2010.

[11] C. Grozea, D.-C. Cercel, C. Onose, and S. Trausan-Matu, "Atlas: News aggregation service," in *2017 16th RoEduNet Conference: Networking in Education and Research (RoEduNet)*. IEEE, 2017, pp. 1–6.

[12] O. Oechslein, M. Haim, A. Graefe, T. Hess, H.-B. Brosius, and A. Koslow, "The digitization of news aggregation: Experimental evidence on intention to use and willingness to pay for personalized news aggregators," in *2015 48th Hawaii International Conference on System Sciences*. IEEE, 2015, pp. 4181–4190.

[13] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," *Advances in automatic text summarization*, pp. 111–121, 1999.

[14] G. Rossiello, P. Basile, and G. Semeraro, "Centroid-based text summarization through compositionality of word embeddings," in *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, 2017, pp. 12–21.

[15] Y. Li and B. Merialdo, "Multi-video summarization based on video-mmr," in *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. IEEE, 2010, pp. 1–4.

[16] I. Mani, *Automatic summarization*. John Benjamins Publishing, 2001, vol. 3.

[17] O. Tas and F. Kiyani, "A survey automatic text summarization," *PressAcademia Procedia*, vol. 5, no. 1, pp. 205–213, 2007.

[18] F. Hamborg, N. Meuschke, and B. 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.

[19] G. Paliouras, A. Mouzakidis, V. Moustakas, and C. Skourlas, *PNS: A Personalized News Aggregator on the Web*, 01 1970, vol. 104, pp. 175–197.

[20] B. Balcerzak, W. Jaworski, and A. Wierzbicki, "Application of textrank algorithm for credibility assessment," in *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, vol. 1. IEEE, 2014, pp. 451–454.

[21] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[22] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in *2016 4th International Conference on Cyber and IT Service Management*. IEEE, 2016, pp. 1–6.

[23] P. Chen, H. Xie, S. Maslov, and S. Redner, "Finding scientific gems with google's pagerank algorithm," *Journal of Informetrics*, vol. 1, no. 1, pp. 8–15, 2007.

[24] S. Liu, K. Huang, and J. Chai, "Research of news text with word frequency statistics and user information," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*. IEEE, 2017, pp. 2633–2637.