TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 18, No. 3, June 2020, pp. 1268~1274
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v18i3.14028

Design of optimal search engine using text summarization through artificial intelligence techniques

Kaushik Sekaran[1], P. Chandana[2], J. Rethna Virgil Jeny[3], Maytham N. Meqdad[4], Seifedine Kadry[5]
[1,2,3]Department of Computer Science and Engineering, Vignan Institute of Technology and Science, India
[4]Al-Mustaqbal University College, Iraq
[5]Department of Mathematics and Computer Sciecne, Faculty of Science, Beirut Arab Univeristy, Lebanon

Article Info

Article history:

Received Sep 2, 2019
Revised Jan 18, 2020
Accepted Feb 23, 2020

Keywords:

Artificial intelligence
Bot creation
Natural language processing
Search engine
Text summarization

ABSTRACT

Natural language processing is the trending topic in the latest research areas, which allows the developers to create the human-computer interactions to come into existence. The natural language processing is an integration of artificial intelligence, computer science and computer linguistics. The research towards natural Language Processing is focused on creating innovations towards creating the devices or machines which operates basing on the single command of a human. It allows various Bot creations to innovate the instructions from the mobile devices to control the physical devices by allowing the speech-tagging. In our paper, we design a search engine which not only displays the data according to user query but also performs the detailed display of the content or topic user is interested for using the summarization concept. We find the designed search engine is having optimal response time for the user queries by analyzing with number of transactions as inputs. Also, the result findings in the performance analysis show that the text summarization method has been an efficient way for improving the response time in the search engine optimizations.

Corresponding Author:

Seifedine Kadry,
Department of Mathematics and Computer Sciecne, Faculty of Science,
Beirut Arab Univeristy, Beirut, Lebanon.
Email: s.kadry@bau.edu.lb

1. INTRODUCTION

Natural language processing (NLP) [1-3] is the trending area of research which allows machines to understand humans in a smarter way. It allows computers to analyze, understand and derive the meaning from human instructions. Using natural language processing the programmers could organize and design the knowledge based representations to perform the tasks like summarization, sentiment analysis, speech recognition and topic segmentation. The natural language processing considers the hierarchy of the language starting with sentence then phrases and finally sentences. By analyzing the language with their meanings, NLP processes the sentences by correcting the grammar, speech to text conversion by automatically translating the languages.

Natural language processing is used to analyze text and allows machine to understand the instructions or commands given by humans in the form of speech. The natural language processing is generally used for automatic machine answering, machine translation [4] and text mining. It is considered as the toughest research area to analyze the human speech into text form inspite of their pronunciations and

need to interpret their meanings and understand the associations between words to create a meaning. Therefore, it can be a said that this is the toughest research area for the programmers to design.

The summarization [5, 6] technique is the idea of data mining, which focus on the centralizing the idea of topic by considering the entire text in the document given as an input. It conveys important information in the original text in summarized form. Automatic text summarization produces the shorter view of text with the semantics which reduces the reading time. The classification of text summarization includes Informative and Indicative. Indicative presents the main idea of text to the user which covers 10 percent of the main text. Informative scheme presents the concise information of the main text which covers 30 percent of the original text.

The Search engines play a vital role in a human life to obtain the information in a fraction of seconds, for different users queries, there are famous search engines like Google, Yahoo, Bing, Ask etc which performs the task of search operations with respect to the user query and displays the relevant content within a fraction of seconds. The search engines are designed to display the contents for the query in the format of title, URL and description format which consist of subscripted text with read more option in the description part which allows the user to read or know the information about that content after clicking on that URL. In our paper, we make an attempt to design a search engine which provides the summarized content about the data we are looking instead of clicking on URL to know the information about selected result.

2. RELATED WORK

Atif Khan et al., [7] presents a semantic graph approach with improved ranking algorithm for abstractive summarization of multi-documents. The predicate argument structures are the graph nodes constructed from the source documents collected from multiple sources where the predicate argument structures (PASs) can be referred as the semantic structure of sentence, which is automatically identified by using semantic role labeling; while graph edges represent similarity weight, which is computed from PASs semantic similarity. He conducted experiment using DUC-2002, a standard dataset for document summarization. Experimental results present the superior performance than other summarization approaches.

Samrat Babar [8], proposed an automatic summarization text for absorbing the relevant content from the number of documents available. The author discussed about the importance of automatic summarization with the basic definitions of text summarization. He discussed about the various research areas for considering the automatic summarization like machine learning, Natural language processing. The author explained the important extraction and differences between both extractive and abstractive summarizations with two groups of text summarization namely indicative and inductive summarizations [9].

Deepali K. Gaikwad, et al., [10], proposed the importance of text summarization [11] as a branch of natural language processing with the abstract presentation of information available in the internet. They presented about the details of both the extractive and abstractive approaches along with the techniques used, its performance achieved, along with advantages and disadvantages of each approach. They presented all the details of both the extractive and abstractive approaches along with the techniques used, its performance achieved, along with advantages and disadvantages of each approach. Text summarization has its importance in both commercial as well as research community. As abstractive summarization requires more learning and reasoning, it is bit complex then extractive approach but, abstractive summarization provides more meaningful and appropriate summary compare to extractive. They presented various types of text summarization techniques with various forms of approaches.

Rashmi Kurmi, et al., [12] implemented a method to reduce cost and time. The method works on the principal of maximal marginal significance between word and sentence. The maximal marginal significanceis decided by the unit step function used which contains database with of useless words or words which can't impact the meaning of document are maintained. The input document is traversed and words containing in the database are eliminated from initial position of the sentence to the end.

Luciano Cabral el al., [13] proposed an automatic summarization method which displays the summaries of news pages on Android-enabled mobile devices to the different forms of users. The method contain two basic approaches where the first approach preprocesses web pages by reformatting or adapting them to a more appropriate way of viewing on small screens, without altering the original content. Second approach presents the most important and relevant content of a page to the user, with respect to the need for grasping the basic information.

3. ARCHITECTURE

The Architecture for the proposed framework is given in Figure 1. The architecture of the proposed work start with a graphical user interface design which allows user to input his query as in a general search

engine, where the search program fetches and extracts data from multiple search engines like Google, Bing and Yahoo using the spider and robot programs for the given user query and stores in the database. The extractor program in the GUI performs the automatic keyword extraction [14-17] for the obtained database to present the visibility of frequent search terms for the user which allows user to reframe the query he would like to request. The obtained data is presented in the form of title, URL and description format as in general search engines. The display of results is presented with the link to view the overview of content which presents the summarized text of the selected result without navigating to the page of the URL. The content summarization is performed by the summarization technique [18-22] of natural language processing [23-25]. In future, the scope of the NLP could be extended towards cloud based [26-30] processing of AI techniques.

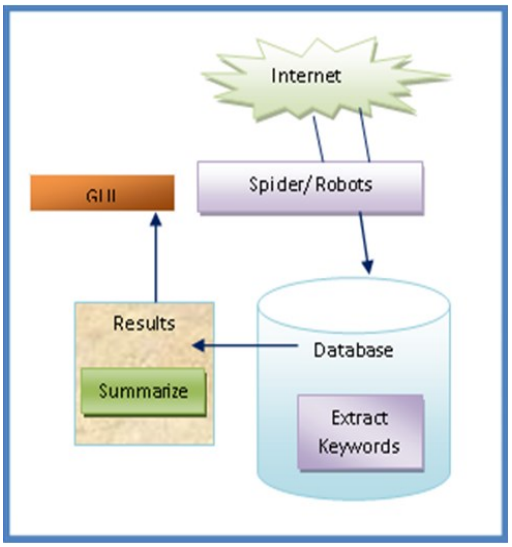

Figure 1. The framework for the search engine design

4. ALGORITHM

Input: D Database = $\{T1, T2 \ldots Tn\} \forall\, Ti\, \{i = 1, 2 \ldots n\}$
Output: Result Display with Summarized Text
Parameters: Swi = Array of Stop Words
attsi = Description attribute
KWi = Words stemmed from
Description attribute
WFi = Word frequencies after
stemming
Sti = Summarized text of
selected result
Method:
- Consider the list of stop words to be removed from description attribute Swi= {α};
- Perform the following operations, for each tuple Ti in Database D,
Consider the Description attributes in D and name it as 'attsi'
- Stem the stop words from the text of description attribute and separate the keywords as follows: Wk=Separate (attsi, Swi);
- The word frequency (freq (w)), word degree (deg (w)) are considered for calculating the word score WS that is, the ratio of degree to frequency (deg(w)/freq(w)).
a. Wordf(word)=words(D,attsi,WK)
b. Wordd(word)=words(D, attsi)
c. WS= Wordf/Wordd
- The frequency (f) is considered as the highest score received for the words after stemming the description attribute. f=highest_score (WS)
- Generate the list of words extracted along with frequency in descending order.
- Select the result for display, feed the url to summarizer to generate the overview of the text.
Result= Summarize(url)
(Or) Res=Sti(attsi)

5. RESULTS

Lot of API's and tools [20] there to analyze the current texts and also the contextual databases, yet another significant achievement in the field of NLP, intellexer API is one of the prominent platform for getting the solutions on text mining. It could be easily incorporated with the programming languages like python, php and java which help developers to present the innovative and instant solutions for the information search, extraction and semantic modeling. We have analyzed using this API and came up with the results as shown in Figure 2. The proposed work is presented with the results as follows:

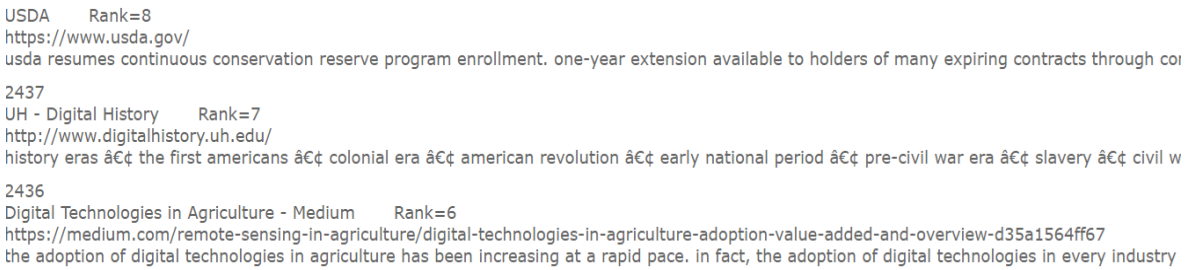

Figure 2. The results for the input query is collected from search engines like Google, Yahoo and Bing using the crawler programs

The user interface is designed as a simple search engine which collects data using web crawlers and fetches information from the popular search engines like Google, Yahoo, Bing search engines in the format of <Title, Url, Description> and stored in the database in the order they have fetched. The data scraped from the database may contain some missing, irrelevant values and therefore it is filtered and pruned from the database which contains only the relevant information and maintained with the index as shown in Figure 3. The data collected from the crawler program is pruned and description part collected is further splitted into set of keywords for propagating general search terms [17] to provide convenience for the user, while searching for the content. The results are presented to the user in a tabular form where the user selects the results to get overview of the content he desired to view as shown in Figure 4. The generalized summary of the result selected by the user is presented as a document summary which contains overview of the topic on which their website is designed about. It presents the overview information for which the user has selected as shown in Figure 5.

levels	1
cover	1
brand	1
imaginable	1
houses brilliant selfie cameras	15
click perfect selfies	9
vivo mobile phones	9
vivo mobiles falls	9
android based smartphones	9
brilliant camera	5
totally worth	4
sleek design	4
device make	4
features packed	4
price range	4
low cost	4
money	1
rs	1
modern shopping centre	8
good soft play	7
soft play	5
kids whilst	4

Figure 3. The Results fetched using crawler program splits the entire text in the description part into set of keywords

2297	organic farming	organic farming is a method of crop and livestock production that involves much more than choosing not to use pesticides, fertilizers, genetically modified organisms, antibiotics and growth hormones.	Select
2298	organic farming	organic farming is a production system which avoids or largely excludes the use of synthetically compounded fertilizers, pesticides, growth regulators, genetically modified organisms and livestock food additives.	Select
2299	organic farming	18 feb 2017 ... organic farming is done to release nutrients to the crops for increased sustainable production in an eco-friendly andpollution-free environment.	Select
2300	organic farming	international federation of organic agriculture movements (ifoam), an international organization established in 1972 for organic farming organizations defines ...	Select
2301	organic farming	put simply, organic farming is an agricultural system that seeks to provide you, the consumer, with fresh, tasty and authentic food while respecting natural ...	Select
2302	organic farming	organic farming - agriculture and rural development.	Select
2303	organic farming	no matter how much a man progresses, agriculture is an occupation that was, is and will always be undertaken since it suffices one of the most important basic ...	Select
2304	organic farming	many organic farmers, including wende elliott and joe rude of colo, iowa, view organic production as a means to work with the environment and maintain the ...	Select
2352	organic farming	organic farming is an alternative agricultural system which originated early in the 20th century in reaction to rapidly changing farming practices. organic farming continues to be developed by various organic agriculture organizations today.	Select

Figure 4. Display of results presenting or the user to select the summarized text

Document structure: *NewsArticle*

Document topics:

Document summary:

The site uses cookies for analytics, personalized content and ads.
By continuing to browse this site, you agree to this use.
Learn more
Digital Agriculture: Farmers in India are using AI to increase crop yields
By
Microsoft News Center India
7 November, 2017
The fields had been freshly plowed.
The furrows ran straight and deep.
Thousands of farmers across Andhra Pradesh (AP) and Karnataka waited to get a text message before they sowed the seeds.
The SMS, which was delivered in Telugu and Kannada, their native languages, told them when to sow their groundnut crops.
In a few dozen villages in Telengana, Maharashtra and Madhya Pradesh, farmers are receiving automated voice calls that tell them whether their cotton crops are at risk of a pest attack, based on weather conditions and crop stage.
Meanwhile in Karnataka, the state government can get price forecasts for essential commodities such as tur (split red gram) three months in advance for planning for the Minimum Support Price (MSP).
Welcome to digital agriculture, where technologies such as Artificial Intelligence (AI), Cloud Machine Learning, Satellite Imagery and advanced analytics are empowering small-holder farmers to increase their income through higher crop yield and greater price control.

Figure 5. Generates the summary of the content for the result selected by the user

6. PERFORMANCE ANALYSIS

The performance evaluation of the search Engine is assessed by considering the database size, execution time and response time for the number queries executed where the response time is directly proportional to the network latency and bandwidth as shown in Figure 6.

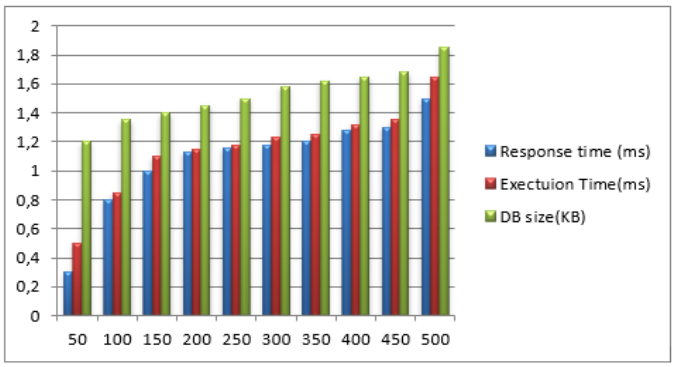

Figure 6. Performance evaluation of search engine

7. CONCLUSION

The search engines play a vital role in information retrieval process for any format of queries presented by user. Thus it is crucial step in a search engine design to interpret the query and should present the results in an effective manner so that user should visualize a great look and feel about the interface. In order to provide more flexible usage of search engine, the search results are further elaborated by providing the small summary content of the result they are looking by clicking on the select option for each result displayed rather than navigating to the webpage by clicking the url in the result page. In our paper, we proposed the framework for designing the search engine using automatic keyword extraction and summarization techniques. The performance of our search engine is evaluated with respect to the database size, Response time and Execution time/ throughput time.

REFERENCES

[1] Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, no. 12, pp. 2493-2537, March 2011.

[2] Grosz, Barbara J., Karen Sparck Jones, and Bonnie Lynn Webber, "Readings in natural language processing," *Morgan Kaufmann,* August 1986.

[3] Chowdhury, Gobinda G., "Natural language processing," *Annual review of information science and technology,* vol. 37, no. 1, pp. 51-89, 2003.

[4] Sarkar, Kamal, Mita Nasipuri, and Suranjan Ghose, "Using machine learning for medical document summarization," *International Journal of Database Theory and Application*, vol. 4, no. 1, pp. 31-48, March 2011.

[5] PadmaLahari, E., D. V. N. Siva Kumar, and Shiva Prasad, "Automatic text summarization with statistical and linguistic features using successive thresholds," *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, Ramanathapuram, pp. 1519-1524, 2014.

[6] Saranyamol, C. S., and L. Sindhu, "A survey on automatic text summarization," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 6, pp. 7889-7893, 2014.

[7] Cabral, Luciano, Rinaldo Lima, Rafael Lins, Manoel Neto, Rafael Ferreira, Steven Simske, and Marcelo Riss, "Automatic Summarization of News Articles in Mobile Devices," *2015 Fourteenth Mexican International Conference on Artificial Intelligence (MICAI)*, Cuernavaca, pp. 8-13, 2015.

[8] Imam, Ibrahim, Nihal Nounou, Alaa Hamouda, Hebat Allah, and Abdul Khalek, "Query Based Arabic Text Summarization," *International Journal of Computer Science and Technology*, vol. 4, no. 2, pp. 35-39, June 2013.

[9] Reeve, Lawrence H., Hyoil Han, Saya V. Nagori, Jonathan C. Yang, Tamara A. Schwimmer, and Ari D. Brooks, "Concept frequency distribution in biomedical text summarization," *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 604-611, 2006.

[10] Khan, Atif, and Naomie Salim, "A review on abstractive summarization methods," *Journal of Theoretical and Applied Information Technology*, vol. 59, no. 1, pp. 64-72, January 2014.

[11] Gaikwad, Deepali K., and C. Namrata Mahender, "A review paper on text summarization," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5, no. 3, pp. 154-160, March 2016.

[12] Vishal Gupta, "A Survey of Recent Keywords and Topic Extraction Systems for Indian Languages," *International Journal of Engineering Trends and Technology (IJETT)*, vol. 6, no. 6, pp. 340-343, December 2013.

[13] Gupta V. And Lehal G. S., "A Survey of Text Summarization Extractive Techniques," *Journal of Emerging Technologies in Web Intelligence*, vol. 2, no. 3, pp. 258-268, August 2010.

[14] Cabral, Luciano, Rinaldo Lima, Rafael Lins, Manoel Neto, Rafael Ferreira, Steven Simske, and Marcelo Riss, "Automatic Summarization of News Articles in Mobile Devices," *2015 Fourteenth Mexican International Conference on Artificial Intelligence (MICAI)*, Cuernavaca, pp. 8-13, 2015.

[15] Gupta Vishal, "A Survey of Text Summarizers for Indian Languages and Comparison of their Performance" *Journal of Emerging Technologies in Web Intelligence*, vol. 5, no. 4, pp. 361-366, November 2013.

[16] Young, Tom, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55-75, Aug. 2018.

[17] Viswanath Meghana, "Thesis: Ontology-Based Automatic Text Summarization," M. Sc Thesis, Vishweshwaraiah Institute of Technology, India, 2009.

[18] Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky, "The Stanford CoreNLP natural language processing toolkit," In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55-60, June 2014.

[19] Khan, Atif, Naomie Salim, Haleem Farman, Murad Khan, Bilal Jan, Awais Ahmad, Imran Ahmed, and Anand Paul, "Abstractive text summarization based on improved semantic graph approach," *International Journal of Parallel Programming*, vol. 46, no. 5, pp. 992-1016, February 2018.

[20] Babar, Samrat, and M. Tech-Cse, "Text summarization: An overview," Reseach Gate, 2013. [online]. Available from: https://www. researchgate. net/publication/25794 7528_Text_SummarizationAn_Overview.

[21] Manne, Suneetha, Zaheer Parvez Shaik Mohd, and S. Sameen Fatima, "Extraction based automatic text summarization system with HMM tagger," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 1, no. 3, pp. 118-123, July 2011.

[22] Liu, Peter J., Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer, "Generating wikipedia by summarizing long sequences," c*onference paper at ICLR 2018*, January 2018.

[23] Khairuddin Khalid, Azah Mohamed, Ramizi Mohamed, Hussain Shareef, "Performance Comparison of Artificial Intelligence Techniques for Non-intrusive Electrical Load Monitoring," *Bulletin of Electrical Engineering and Informatics*, vol. 7, no. 2, pp. 143-152, June 2018.
[24] Vinnarasu A., Deepa V. Jose, "Speech to text conversion and summarization for effective nderstanding and documentation," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 5, pp. 3642-3648, October 2019.
[25] Husni Thamrin, Gunawan Ariyanto, Irma Yuliana, Wawan Joko Pranoto, "Crowdsourcing in developing repository of phrase definition in Bahasa Indonesia," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 17, no. 5, pp. 2321-2326, October 2019.
[26] Sekaran Kaushik and P. Venkata Krishna, "Big Cloud: a hybrid cloud model for secure data storage through cloud space," *International Journal of Advanced Intelligence Paradigms*, vol. 8, no. 2, pp. 229-241, January 2016.
[27] Sekaran, K., & Krishna, P. V., "Cross region load balancing of tasks using region-based rerouting of loads in cloud computing environment," *International Journal of Advanced Intelligence Paradigms*, vol. 9, no. 5-6, pp. 589-603, January 2017.
[28] Sekaran, K., Khan, M. S., Patan, R., Gandomi, A. H., Krishna, P. V., & Kallam, S., "Improving the Response Time of M-Learning and Cloud Computing Environments Using a Dominant Firefly Approach," in *IEEE Access*, vol. 7, pp. 30203-30212, 2019.
[29] Kaushik, S., Singh, S., & Pathan, R. K., "Design of novel cloud architecture for energy aware cost computation in cloud computing environment," *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*, Vellore, pp. 1-6, 2017.
[30] Sekaran, K., Vikram, G. R., Chowdar, B. V., & Raju, U. N. P., "Combating Distributed Denial of Service Attacks Using Load Balanced Hadoop Clustering in Cloud Computing Environment," In *Proceedings of the 2nd International Conference on Digital Technology in Education*, pp. 77-81, October 2018.