Research Proposal – Produce high quality of content searching

At the time of big data booming, the time and cost spend of data exploratory may increase sharply. The need of development on an efficient search visualization is obvious. For the reason, there is a large need of the text summarization incorporate with the search engine system. In this proposal, I would like to research on search engine and text summarization techniques.

Background

Search Engine is a kind of software system designed to search information from world wide web based on users input. Most of the search engine nowadays capable to accept fuzzy input, that is, the search result is based on the proximity of the input. As the success of the web search engine. Some *search framework* also rolled out for different application has already proposed for commercial use, such as *Elastic Search*[i], a Lucene-based NOSQL framework and *Apache Solr*[ii], a document compatible enterprise search platform. These search engines also support distributed deployment, facilities the multiple search task in different node.

Applications such as DuckDuckGo provides InstantAnswerAPI[iii] for instant answer system for a particular type of question, topic summaries, categories and disambiguation. The system accepts abstract type of input and the input could relate to a topic, text or a source. The source of the come from more than 100 of source (including Wiki Open Search, Wikia, CrunchBase, etc). Second, Google news[iv], an instant and continuous news aggregator which earns billions of dollars annually. The system categorizes the news from different publisher and newspaper into different areas. For instance, technology, international, sports, etc. Besides, the application also provides recommendation system to specific area of interest from users. See more for the recent research (Flora et al. 2017).

The development of the GETA (General Engine for Transposable Association)[v] provide a versatile of application in terms of associative search. Some existing application (e.g. WebcatPlus, IMAGINE Book Search) show good result on the related word suggestion, search recommendation. GETA uses *WAM (Word Article Matrix)* to indicate the occurrence of words in each article and help relate the association between them. In the meanwhile, it is also possible to investigate other method on the associative search. In order to indicate the importance of certain keywords in the document the engine uses a common approach *TD-IDF (Term Frequency-Inverse document frequency)* to reflect score of the keywords.

Text Summarization is a wide range of application including news article, public opinion and scientific papers. In current research, *Encoder-decoder framework* (Cho et al., 2015), is one of the popular approaches for the summary generation. The method is possible to be implemented by Recurrent Neural Network (RNN) and Convolutional Neural Network. More precisely, some researches have shown CNN were able to perform the task in parallel, which apparently improve the efficiency.

Problem Statement

My main focus of research is to improve the performance of the search engine and the text summarization.

i) Searching. As mentioned, GETA uses WAM to associate the relation between documents. Besides using keyword, it is possible that the meta data of document could be used for the association. For the simplest case, the search engine should be able to relate the document written be same author. Besides, the authority of every single document may not the same. And the system should provide a suitable ranking strategy to those candidates and select the most suitable result.

On the other hand, the search engine may not reflect the actual impact of the results. For example, some document may not have high proximity to the keyword, but referenced by many highly related documents, should also related to the search result.

Ii) Summarization To perform a summarization task, it may require paraphrase the paragraph, and the keywords may not necessarily repeated. As what GETA proposed, a single keyword could have multiple of analogous keywords, and then search the related document using those keywords. In the same way, the summarization should generate the most suitable sentence in the same manner.

Related Works

Researches Some applications (Eniat et al. 2000; Delort et al. 2003) work with search engine and provide web summarization and relate the topic main as much as possible. The research improves the search performance by counting the frequency the page pointing to each web page. The authors constructed a new frameworks for the on scientific paper summarizations (Michihiro et al. 2019), which uses Long Term Short Term Memory (LTSM) for the encoding task and Graph convolutional network (GCN) for the new model were trained by the corpus it proposed and have shown improvement on the ROUGE (2F & 3F) and SU4-F score. In the research, a graph was constructed to relate the input sentence of the articles via edges. The relationship between sentence were evaluated using TF-IDF cosine similarity.

Methodology

Meanwhile to provide an improved searching service, we would also like to introduce a slight text summarization to end users. For instance, some prevailing search engine (e.g. Google, DuckDuckGo, Bing) will give a related definition or disambiguation response to the search question. In short, the research contains two goals: search engine and text summarization. It is desirable to investigate other possible technique search.

i) Searching

To overcome the impact reflection problem, we would like to investigate and develop an approach to recognize the actual impact of the document. We would develop a hybrid model of searching, which perform weighted justification for the impact reflection on both its own content and the impact on specific community. For example, the search engine will perform the. Then perform a weighted sum approach and rank those results.

ii) Text Summarization

Data Preprocessing. In order to overcome the drawback of impact reflection. We would like to identify the most similar terms (e.g. TF-IDF cosine similarity). And then the research will evaluate the authority of each reference and rank them in weighted approach. Finally associate the documents using graph-approach.

Summary Generation. We would like to develop a new unsupervised technique in topic representation approaches. For instance, some well-known approaches include Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) and Bayesian Topic Model, which have been widely used in multi-document summarization and query-focused summarization.

Evaluation

Two area will be examined: searching and summarization. For searching, we would like to evaluate the performance and the accuracy of the search engine. For the text summarization, the accuracy and the quality of result will be tested. Our result will be compared with the other existing and latest platform, for instance, DuckDuckGo InstantAnswerAPI. The original GETA system will be treated as baseline to indicate the performance improvement.

Human Evaluation. We will invite around 5-10 people to evaluation. They will evaluate the instance in their searching and summarization performance. Result will be scored between 1-5, the testing instance will be compared with the average score.

Automatic Indicator. The testing data will be at least 10 million of documents. The performance of the text summarization could be benchmarked with commonly used automatic indicators, ROUGE (2F & 3F) and SUR4-F.

Conclusion

In this research, we will investigate the performance and conduct improvement on the GETA. Apart from that, we would also like to propose a novel approach for search engine and text summarization. During the research, I would like to have a further understanding of Natural Language Processing techniques, specifically in the field of information retrieval approaches.

Reference

Amato, F., Moscato, V., Picariello, A., Sperlí, G., D'Acierno, A., & Penta, A. (2017). Semantic summarization of web news. Encyclopedia with Semantic Computing and Robotic Intelligence, 1(01), 1630006.

Cho, K., Courville, A., and Bengio, Y. (2015). Describing multimedia content using attention-based encoder–decoder networks. arXiv preprint arXiv:1507.01053.

Einat Amitay and Cécile Paris. 2000. Automatically summarising web sites: isthere a way around it?. In Proceedings of the ninth international conference onInformation and knowledge management. ACM, 173–179.

J-Y Delort, Bernadette Bouchon-Meunier, and Maria Rifqi. 2003. Enhancedweb document summarization using hyperlinks. In Proceedings of the fourteenthACM conference on Hypertext and hypermedia. ACM, 208–215.

[i] https://www.elastic.co/products/elasticsearch

[ii] http://lucene.apache.org/solr

[iii] https://duckduckgo.com/api

[iv] https://news.google.com

[v] https://geta.ex.nii.ac.jp/e/