Assignment title:

Scoping a research problem

Project title:

Filtering related news stories based on user’s interest by applying topic modelling techniques

Name: Kai Sum Yau

Student number: n10050965

Cluster: Cluster 2

Tutor name: Nishanthi Dasanayaka

Supervisor name: Aspro Yue Xu, Dakshi Kapugama Geeganage

Major: Computer Science

Research background and literature analysis

The domain of this project is information filtering of text contents. In the era of information explosion, the Internet contains an enormous amount of text contents. Filtering the required contents based on the user's interest is a challenge.

Information filtering (IF) (Belkin & Croft, 1992) is technical a type of information retrieval (IR). From 2003, most of the IF system is based on a latent Dirichlet allocation (LDA) (David, Andrew, et al., 2003). In this year, Kapugama Geeganage, Xu, et al. (2021) propose a novel approach which is Semantic-based Topic Representation using Frequent Semantic Patterns (STRuFSP). STRuFSP handles the textual data from the meaning of the words, not the traditional frequency of the word. Although this novel approach upgrades the performance of IF, it cannot handle the disambiguation and unavailable words. In this project, I will mix the existing data mining algorithms, such as classification and clustering, to offset the limitations of each algorithm. The challenge of this project is most of users have multiple topic interests and it is hard to extract multiple interests in the user's previous readings and filter the contents.

Research problem statement

In this project, I will solve the problem is the correctness of filtering the news based on the user's reading history. In IF, when the length of content is longer, the model is harder to filter (Hong & Davison, 2010). Also, filtering function is important in a website or application which stores many books, newspaper or any text contents, such as a book shop website or newspaper website. If the correctness of filtering can be improved, user can be easier to receive the new information that is highly relevant to her interest.

To address this problem, I propose creating a model that is using more than one data mining or topic modeling algorithms to train the RCV1 dataset. Then, the model extracts the user's interest and finds the relevant news in the dataset. This is achieved by testing the division and combination of difference algorithms. Finally, I will evaluate the combinations of the mixing approach.

Research Questions

1. **Which combinations of the existing classification and clustering algorithms can be information filtering approach which is suitable for filtering the large text contents?**

In 2003, LDA is published and it improves the information filtering. However, there are many limitations of the traditional LDA model. David, Andrew, et al. mentioned *"In particular, the bag-of-words assumption allow words that should be generated by the same topic to be allocated to several different topics."* Although there are many existing models which are based on the LDA, most of them still have the same limitations.

The research question can produce the direction of exploring the potential of the data mining algorithms. If the research question is answered successfully, there is a novel approach is appeared which can improve the information filtering and the limitations of traditional LDA model. Also, businesses can apply the novel approach in their recommender system to analyze their user's interest.

To answer this research question, there are three ways are recommended: (1) creating a new model and architecture like LDA, (2) improving the existing algorithm or model like four-layer hierarchical Bayesian structure (Tang, Mao, et al., 2018), (3) using more than one algorithm or models in the same time to enhance the performance. Since the limitations of time and resources, I propose the third way. Except enhancing the performance, it is possible to reduce the limitations of the algorithms.

Bowles (2015), Alsmadi and Alhami (2015) provided some useful tools and methods which can used python to apply. In Python, there are serval famous libraries, such as scikit-learn (sklearn) (scikit-learn, 2010) which is providing the tools for predictive data analysis. In these libraries, there are existing many algorithms which can be the IF approach.

The method is possible to be implemented. And I expect the output can extract user's interest in the RCV1 dataset.

2. How does the novel mixed approach improve extracting user’s interest?

Following the previous research question, after I develop a new mixed approach to extract the user’s interest, I have to evaluate the applying of the multiple algorithms to know does it improve extracting user’s interest and information filtering.

The aim of this research question is evaluating the accurate, feasibility and availability of the novel complex approach. Also, finding the limitations of applying the approach is necessary. It can help me to know what should I focus on in the future.

To answer the research question, I will follow Zheng’s (2015) guide to evaluate the novel approach. It includes the evaluation tools and the measurement units. I expect the result of the evaluation is a good accuracy score (higher than 65%) and ROC curve (the curve near the true positive rate).

References

Belkin, N. J., & Croft, W. B. (1992). Information filtering and information retrieval: two sides of the same coin?. Communications of the ACM, 35(12), 29-38.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(4-5), 993-1022.

Kapugama Geeganage, D. T., Xu, Y., & Li, Y. (2021). Semantic-based topic representation using frequent semantic patterns. Knowledge-based systems, 216, 106808. https://doi.org/10.1016/j.knosys.2021.106808

Hong, L., & Davison, B. (2010). Empirical study of topic modeling in Twitter. Proceedings of the First Workshop on social media analytics, 80-88.

Tang, Y. K., Mao, X. L., et al. (2018). Conceptualization topic modeling. Multimedia tools and applications, 77(3), 3455-3471.

Bowles, M. (2015). Machine learning in python: essential techniques for predictive analysis. Indianapolis.

Alsmadi, I., & Alhami, I. (2015). Clustering and classification of email contents. Journal of King Saud University – Computer and Information Sciences, 27(1), 46-57.

Scikit-learn. (2010). Scikit-learn about us [Invented title]. Scikit-learn. https://scikit-learn.org/stable/about.html

Zheng, A. (2015). Evaluating Machine Learning Models (1st ed.). O'Reilly Media, Inc.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 841–842. https://doi.org/10.1145/1835449.1835643

Appendix 1

Title: Latent Dirichlet Allocation

Authors: David M. Blei, Andrew Y. Ng, Michael I. Jordan

Full-text citation: Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(4-5), 993–1022.

Quality/trustworthiness: Journal of Machine Learning Research’s impact score is 9.03 in 2020. The total citation of this article is 5578.

Summary:

This article is about the Latent Dirichlet Allocation (LDA) which is a generative probabilistic mode.

Related content:

Since LDA is the most famous model in the information filtering and information retrieval, it is useful to understand the background of the research. The main findings are the structure and limitations of LDA. It helps me to answer the first research question.

Appendix 2

Title: Semantic-based topic representation using frequent semantic patterns

Authors: Dakshi T. Kapugama Geeganage, Yue Xu, Yuefeng Li

Full-text citation: Kapugama Geeganage, D. T., Xu, Y., & Li, Y. (2021). Semantic-based topic representation using frequent semantic patterns. Knowledge-Based Systems, 216, 106808. https://doi.org/10.1016/j.knosys.2021.106808

Quality/trustworthiness: Knowledge-Based Systems' impact factor is 8.038.

Summary:

This paper represents the latest method for handling the textual data. Also, it shows the comparison result between semantic based approach, pattern-based approach and word-based approach.

Related content:

The main findings are the latest situation of processing of information filtering and the novel approach of information filtering. I can understand the changing in the area of information filtering and the challenges which are still existing. It can help me to define the research problem and research questions.

Appendix 3

Title: Short text classification in twitter to improve information filtering

Authors: Bharath Sriram, David Fuhry, Engin Demir, Hakan Ferhatosmanoglu

Full-text citation: Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 841–842. https://doi.org/10.1145/1835449.1835643

Quality/trustworthiness:

Summary:

This article proposes to use a small set of domain-specific features instead of traditional classification methods.

Related content:

The main finding is the "features" approach. In the research question 1, I may use classification algorithms to combine the novel approach. The "features" approach can improve the implementation of my novel approach.