Literature draft

Student: Michael de Jong (11246618)
Supervisor: Dr. Maarten Marx

April 7, 2020

1 Problem definition draft

The main goal of the project is to make a search engine that is aware of the Jupyter notebook format. In this format, there are multiple types of text in different cells. Thus the search engine has to search the different cell types and graphs. The sub goals of the project include the creation of a test set for the search engine, the implementation of Evolutionary machine learning to be used in the ranking of the returned pages and the evaluation of the system. The search engine will be evaluated using precision, recall, F1, mean average precision and discounted cumulative gain.

2 Keywords

Search Engine, Jupyter Notebooks, Learning to Rank, Evolutionary Network

3 Relevant articles

- Grappiolo et al. 2019 builds a Jupyter notebook search engine with a different approach. This article is useful because they achieved the goal of the project with a different approach.
- Tan et al. 2020 compares state of the art neural network ranking models. An evolutionary algorithm is a neural ranking model so information of different approaches is relevant.
- Mitra and Craswell 2017 explains fundamental concepts of neural and non-neural approaches to learning vector representations of text. This, as Tan et al. 2020 but includes fundamental concepts and non neural approaches as well.
- Ibrahim and Landa-Silva 2018 is about learning to rank in information retrieval with machine learning with an evolutionary strategy. This evolutionary approach is similar to the project and thus important.
- Shen et al. 2006 contains the theory how queries can be improved for query classification. This method might be good to use to improve performance of the Jupyter search engine and a good way to do query classification.

References

[Gra+19] Corrado Grappiolo et al. "The Semantic Snake Charmer Search Engine: A Tool to Facilitate Data Science in High-tech Industry Domains". In: Mar. 2019, pp. 355–359. DOI: 10.1145/3295750.3298915.

[ILS18] Osman Ali Sadek Ibrahim and Dario Landa-Silva. "An evolutionary strategy with machine learning for learning to rank in information retrieval". In: *Soft Computing* 22.10 (2018), pp. 3171–3185.

[MC17] Bhaskar Mitra and Nick Craswell. "Neural models for information retrieval". In: *arXiv preprint arXiv:1705.01509* (2017).

[She+06] Dou Shen et al. "Query enrichment for web-query classification". In: *ACM Transactions on Information Systems (TOIS)* 24.3 (2006), pp. 320–352.

[Tan+20] Shulong Tan et al. "Fast item ranking under neural network based measures". In: *Proceedings of the 13th International Conference on Web Search and Data Mining*. 2020, pp. 591–599.