CS512 Project Proposal: Leveraging Document Co-Clicks To Infer Similarities

Daniel Campos
Revanth Reddy
Shweta Garg
dcampos3,revanth3,shwetag2@illinois.edu

Abstract

In the last few decades search engines like Google, Bing, Baidu, and Yandex have become the primary way that people around the world interact with information. Their constant and diverse usage has made these engines ideal sources for training data like document co-clicks. Using the ORCAS dataset we seek to explore the intersection of Data Mining and Deep Learning and answer the question: Can document co-clicks be used to learn similarity between concepts?

CCS Concepts: • **Information systems** → **Query reformulation**; **Query log analysis**; **Data stream mining**.

Keywords: datasets, neural networks, information retrieval, data mining, query rewriting, synonym prediction

ACM Reference Format:
Daniel Campos, Revanth Reddy, and Shweta Garg. 2021. CS512 Project Proposal: Leveraging Document Co-Clicks To Infer Similarities. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Being able to understand the similarity between words is a core component of many modern computational systems. From a search engine, which needs to understand that the user queries *How tall is tom cruise?* and *Tom cruise height?* have the same intent. To an e-commerce website that can use similarity between *car parts* and *auto parts* to provide similar product recommendations being able to recognize and understand similarity between words is crucial. Supervised methods tend to rely on carefully constructed synonym dictionaries or labeled datasets. Unsupervised methods like contextual [11] and static word embeddings [24] provide a scalable and highly accurate notion of synonym search using vector similarity but once again rely on a large and diverse corpus for pre-training. Whether labeled or unlabeled, low-resource languages do not always have the luxury of large data sources. A learning methods that find alternative methods of dataset construction will prove useful to scale nlp systems.

Modern commercial search engines have hundreds of millions of daily active users around the world. While users may speak different languages and search for different document, the use of search engines as a source of knowledge brings all users together. Every time a user issues a query and engages with a document they are providing a signal to the search engine. User clicks have long been used to improve information retrieval systems [7] but because of the sensitive nature of users queries, few datasets have been released publicly and there has been even less public research into these datasets.

Building on the Question answering and Information Retrieval benchmark, MSMARCO [3], the ORCAS [8] dataset features the largest publicly accessible document click dataset. It features 18.4 million document URL clicks on 10.4 million queries and 1.4 million documents which were extracted from the logs of a commercial search engine. This dataset captures both the similarity and difference of user searches. For example, the query *pandas* features clicks on documents relating to the python programming library and the panda animal yet the query *panda* features clicks on documents related to the restaurant chain panda express and the panda animal.

Using the ORCAS dataset, we seek to explore a broad array of data mining techniques to produce novel training data to be used for transfer learning in tasks like query rewriting and synonym prediction. We will explore the usage dynamic network alignment, topic clustering and link prediction to produce training data which we will then use to explore its effects on the aforementioned tasks. Our goal is to explore how well click based data can replace traditional training corpora.

2 Main Ideas

The main focus of our work is to explore what kind of value we can extract out of the ORCAS dataset with regards to

Conference'17, July 2017, Washington, DC, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

using it as distant supervision data. We believe that if our efforts are successful this method can be explored for low resource languages.

2.1 Dataset Mining

The focus of our work is to start with general exploration and processing of the ORCAS dataset. We will explore the application of traditional data mining algorithms to explore what signals we can extract. We seek to create a set of signals which we can be used in our downstream experiments to prove or disprove our hypothesis. Some of the methods we plan to explore include: phrase based clustering, click based query clustering, n-gram association analysis, etc.

2.2 Query Synonym Prediction

Using the ORCAS dataset, we formulate the query synonym prediction task as a task where the goal is to identifying the word synonyms in queries. These synonyms can then be used to identify similar queries efficiently. // In this work, we intend to use contextual and non contextual word representations to explore how these methods can represent n-gram and query similarity, and leverage them to build our own knowledge base of query synonyms. We plan to evaluate our approach by predicting the degree of connection between queries (measured by proximity in click graph) and between n-gram terms. Another way to evaluate our mined synonyms would be to compare them against those from English dictionaries such as Merriam-Webster dictionary. We also plan to perform a qualitative evaluation by getting a small portion of the generated query synonyms analyzed by human evaluators.

To further explore how such user logs can help learn signals about query similarities, we plan to leverage the query co-clicks from ORCAS in a transfer learning setting for a related task of Quora duplicate question detection[1]. We will investigate how well a model trained on the query co-clicks from ORCAS performs on the Quora dataset and will also explore into using the co-clicks data to come up with simpler methods that can match the performance of more complex methods [1, 5, 6, 10].

2.3 Query Rewriting

Query rewriting is the process of automatically expanding a search query to better understand the user's intent. Query rewriting is typically used for improving the recall of IR systems, to retrieve a larger set of relevant results. We propose to use queries that have the same co-clicks in ORCAS as parallel data to train a sequence-to-sequence model to do query re-writing. Specifically, we intend to finetune a pre-trained encoder-decoder model [22, 27] using these similar query pairs. Then, we plan to explore how our query re-writing approach can be used to improve established ranking models like BM25 [29] and neural IR models [20, 21] on popular IR benchmarks like MS MARCO [3].

Another potential direction to explore is converting keyword based queries to natural language queries. Current neural network based IR systems, which are built for semantic search, might fall short against keyword-based queries. Hence, converting such queries into natural language would not just improve IR performance but also help understand searcher intent and query context, which are critical in query understanding. To achieve this, we plan to create parallel data from co-clicks, with smaller queries considered to be keyword-based and those with more words categorized as natural language queries. We will evaluate our approach using the TREC Web track[2], which contains both keyword-based and natural language questions with information seeking behaviors that are common in web search.

3 Preliminary Plan

Our plan essentially has two stages: data exploration and transfer learning.

3.1 Milestones

1. Project Scoping(Feb 25th): Provide a scope of work to be attempted and problem framing.
2. Data Exploration and Clustering(March 15th): Application of Data Mining algorithms to visualize and explore clusters both of query terms and documents.
3. Transfer Learning Labels(March 28): Using finding from data exploration and clustering we will create processed data to be used with our transfer tasks.
4. Synonym Evaluation dataset and Baselines(March 28th): Finalize query synonym prediction task and produce baselines that do not leverage click data.
5. Query Rewriting baseline(March 28): Finalize query rewriting task and produce baselines that do not leverage click data.
6. Midterm Report(March 30): Discuss progress and learning.
7. Experiments across tasks(April 15th): Initial results using transfer data on benchmark tasks. Use results to go back and tweak mining methods.
8. Experiments across tasks v2(April 30): Updated results using improved data.
9. Final report (May 5th)

3.2 Roles

- **Daniel Campos**: Application of Data Mining techniques to dataset for exploration and transfer label creation, report writing, IR experiments
- **Revanth Reddy**: Query rewriting baseline, experiments and tweaking.

[1] Quora Duplicate Question Detection

[2] https://trec.nist.gov/data/webmain.html

- **Shweta Garg**: Query Synonym Prediction task analysis, survey on existing baselines and their implementation, experiments and tweaking.

4 Related Work

We break our relevant reading into 3 sections: network mining, query rewriting, and synonym detection.

4.1 Network and Click Mining

Network Mining has been extensively used in various domains such as social networks and biomedical domain for performing some downstream analytic tasks such as recommender systems, link prediction, node classification etc [13, 14]. One popular sub-branch of Network Mining is Click Mining that uses user clicks, co-clicks and query logs for mining interesting patterns. Since user click logs are private hence collection of such a dataset becomes sensitive therefore we do not have many publicly available datasets for this. [17] introduce a new dataset - Clickage - for click mining using user search logs. ORCAS [8] dataset features the largest publicly accessible document click dataset. Several papers are research of the area of Click Mining. [4] makes use of a user's immediate and preceding queries in the search query log for a context-aware query suggestion. Some others [7, 19, 23] also make use of Click Mining of Query logs for basic Information Retrieval tasks.

4.2 Query Rewriting

Query rewriting has always been an important problem in information retrieval. One approach [30] is to expand the query using the top ranked documents from the original query. Later approaches [2, 9] rely on user generated data and focus on using user query logs to generate expansions via probabilistic frameworks. In constrast, we plan to leverage pre-trained language models to learn the distribution over queries with co-clicks and do query rewriting in a generative fashion.

4.3 Synonym Prediction

HolisticOpt [15] aligns with our thought process as it makes use of query log clicks and web table attribute name cooccurrences to find attribute synonyms that can be used to boost the performance of search engines. Fei et al. propose a synonym prediction approach for the medical domain by using a multi-task model with a hierarchical term task relationship to learn entity embeddings.

Transfer learning is an active area of research in machine learning. Recent approaches have shown transfer learning via distant supervision to be effective in the domains of relation extraction [18, 25], multilingual models [16], domain adaptation [31], question answering [28] and information retrieval [26].

References

[1] K Abishek, Basuthkar Rajaram Hariharan, and C Valliyammai. 2019. An enhanced deep learning model for duplicate question pairs recognition. In *Soft Computing in Data Analytics*. Springer, 769–777.

[2] Ioannis Antonellis, Hector Garcia-Molina, and Chi-Chao Chang. 2008. Simrank++ query rewriting through link analysis of the clickgraph (poster). In *Proceedings of the 17th international conference on World Wide Web*. 1177–1178.

[3] Daniel Fernando Campos, T. Nguyen, M. Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, L. Deng, and Bhaskar Mitra. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *ArXiv* abs/1611.09268 (2016).

[4] Huanhuan Cao, D. Jiang, J. Pei, Qi He, Zhen Liao, E. Chen, and Hang Li. 2008. Context-aware query suggestion by mining click-through and session data. In *KDD*.

[5] Andreas Chandra and Ruben Stefanus. 2020. Experiments on Paraphrase Identification Using Quora Question Pairs Dataset. *arXiv preprint arXiv:2006.02648* (2020).

[6] Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2018. Quora question pairs. *University of Waterloo* (2018).

[7] A. Chuklin, P. Serdyukov, and M. Rijke. 2013. Click model-based information retrieval metrics. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (2013).

[8] Nick Craswell, Daniel Fernando Campos, Bhaskar Mitra, E. Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 20 Million Clicked Query-Document Pairs for Analyzing Search. *Proceedings of the 29th ACM International Conference on Information Knowledge Management* (2020).

[9] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. 2002. Probabilistic query expansion using query logs. In *Proceedings of the 11th international conference on World Wide Web*. 325–332.

[10] Elkhan Dadashov, Sukolsak Saksuwong, and Katherine Yu. [n.d.]. Quora question duplication.

[11] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[12] Hongliang Fei, Shulong Tan, and Ping Li. 2019. Hierarchical Multi-Task Word Embedding Learning for Synonym Prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 834–842. https://doi.org/10.1145/3292500.3330914

[13] Philippe Fournier-Viger, J. Lin, R. Kiran, Y. Koh, and R. Thomas. 2017. A Survey of Sequential Pattern Mining.

[14] Sayantani Ghosh, Sudipta Roy, and S. Bandyopadhyay. 2012. A tutorial review on Text Mining Algorithms.

[15] Yeye He, Kaushik Chakrabarti, Tao Cheng, and Tomasz Tylenda. 2016. Automatic discovery of attribute synonyms using query logs and table corpora. In *Proceedings of the 25th International Conference on World Wide Web*. 1429–1439.

[16] Michael A. Hedderich, D. Adelani, D. Zhu, J. Alabi, Udia Markus, and D. Klakow. 2020. Transfer Learning and Distant Supervision for Multilingual Transformer Models: A Study on African Languages. *ArXiv* abs/2010.03179 (2020).

[17] Xian-Sheng Hua, Linjun Yang, Jingdong Wang, Jing Wang, M. Ye, Kuansan Wang, Y. Rui, and J. Li. 2013. Clickage: towards bridging semantic and intent gaps via mining click logs of search engines. In *MM '13*.

[18] Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions. In *AAAI*.

[19] Emilia Kacprzak, Laura M. Koesten, L. Ibáñez, E. Simperl, and J. Tennison. 2017. A Query Log Analysis of Dataset Search. In *ICWE*.

[20] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6769–6781.

[21] Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 39–48.

[22] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.

[23] W. Li and G. Jones. 2013. Enhanced Information Retrieval by Exploiting Recommender Techniques in Cluster-Based Link Analysis. In *ICTIR*.

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, G. S. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *ArXiv* abs/1310.4546 (2013).

[25] M. Mintz, Steven Bills, R. Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL/IJCNLP*.

[26] Bhaskar Mitra. 2020. Neural Methods for Effective, Efficient, and Exposure-Aware Information Retrieval. *ArXiv* abs/2012.11685 (2020).

[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21 (2020), 1–67.

[28] Revanth Gangi Reddy, Bhavani Iyer, Md Arafat Sultan, Rong Zhang, Avi Sil, Vittorio Castelli, Radu Florian, and Salim Roukos. 2020. End-to-End QA on COVID-19: Domain Adaptation with Synthetic Training. *arXiv preprint arXiv:2012.01414* (2020).

[29] Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

[30] Jinxi Xu and W Bruce Croft. 1996. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. 4–11.

[31] Rong Zhang, Revanth Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avirup Sil, and Todd Ward. 2020. Multi-Stage Pretraining for Low-Resource Domain Adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5461–5468.