Argument Retrieval from Web

Mahsa S. Shahshahani * and Jaap Kamps

University of Amsterdam, Amsterdam, The Netherlands
{m.shahshahani,kamps}@uva.nl

Abstract. We are well beyond the days of expecting search engines to help us find documents containing the answer to a question or information about a query. We expect a search engine to help us in the decision-making process. Argument retrieval task in Touché Track at CLEF2020 has been defined to address this problem. The user is looking for information about several alternatives to make a choice between them. The search engine should retrieve opinionated documents containing comparisons between the alternatives rather than documents about one option or documents including personal opinions or no suggestion at all. In this paper, we discuss argument retrieval from web documents. In order to retrieve argumentative documents from the web, we use three features (PageRank scores, domains, argumentative classifier) and try to strike a balance between them. We evaluate the method based on three dimensions: relevance, argumentativeness, and trustworthiness. Since the labeled data and final results for Toucheé Track have not been out yet, the evaluation has been done by manually labeling documents for 5 queries.

Keywords: Argument Retrieval · Web · Touché

1 Introduction

Once search engines were created to help users find to help users find the pieces of information relevant to their needs among a large amount of data. But nowadays search engines are more than that. A newer use-case for search engines is to help users in the decision-making process. In this case, the user is looking for recommendations and personal opinions to choose between some options, for example, different brands of laptops, rather than just official comparisons between their features. So, the goal of the search engine would be to retrieve web documents with an argumentative structure in which there is a discussion or personal view about the options that the user wants to choose from. For example, when the user wants to make a decision to buy a laptop, s/he does not expect to receive a ranked list of documents including comparisons between the specifications of each brand or model. Although these documents are helpful, they do not discuss disadvantages or a trade-off between features. To make the final decision, user needs to see reviews and recommendations like *"If you are a gamer, product A is not useful for you due to its low GPU, but it is the best if you are a programmer as it has a great CPU"*.

* Corresponding author

In this paper, we will study argumentative document retrieval from web pages. In lack of a large labeled corpus, we focus on unsupervised methods. We treat this problem as re-ranking rather than ranking. We assume we have an initial ranked list of documents and we use some features to re-rank these documents to make a better ranked list, putting documents with the argumentative structure on top.

Pagerank scores, sources of web documents, and argumentative classes are three features we use to re-rank the initial ranked lists. We use Clueweb12[1] as our web document source.

The rest of this paper is structured as follows. You are now reading the introduction in Section 1. Section 2 details the features we used to distinguish argumentative and non-argumentative documents, along with other features. This is followed by Section 3 detailing the experimental setup and summarizing the initial results. Finally, we end with conclusion in section 4.

2 Argument Retrieval

In this section first, the task will be defined as a re-ranking task; and then, we will discuss various features that we considered to be relevant and helpful in re-ranking documents will be discussed.

Task The goal is to re-rank documents $d_1, d_2, ..., d_n$ in response to query q, considering that these documents have already been ranked with an initial ranking model like BM25. A document should be at a higher rank if it is more relevant, more argumentative, and from a more trustworthy source.

In [6] trustworthiness has been addressed. We treat this aspect as a subjective dimension.

initial rank An initial ranking of documents based on a simple method (we chose BM25) is given and we want to re-rank them. Thus, every document has been associated with an initial rank before we perform the re-ranking. As we take the top 10 documents into account, this initial rank feature can be any number in $\{1, 2, ..., 10\}$.

Argumentative Classifier We trained a simple SVM classifier based on data from a debate corpus and a web corpus to distinguish between argumentative and non-argumentative documents. We used BERT [4] to represent arguments. As BERT model imposes a limit on the length of documents after tokenization, we use the first 512 tokens in an argument if its length exceeds this limit. In order to train the classifier, we picked a small subset of documents from each corpus. To select the documents, we submitted all 50 comparative queries in the first task of Touché shared-task in CLEF 2020 [2] to both corpora and got up to 100 documents for each query. Then, we manually removed argumentative documents from the Clueweb subset and considered the remaining documents as negative examples. All retrieved documents from args.me corpus have been considered as positive examples. The final set includes 3000 positive and 3000 negative examples. We trained the classifier on 80% of the data, and tested it on the remaining 20% of documents. It achieved 87% in terms of accuracy.

[1] https://lemurproject.org/clueweb12

Web domains Clueweb has been formed by crawling web documents with some post-filters in which pages from inappropriate websites (such as pornographic contents) and 10 percent of pages with the lowest page-rank scores have been removed. But when the user is looking for argumentative documents, documents from particular domains like Wikipedia will not be relevant since they do not present any personal opinion or advice. On the other hand, discussion forums are very helpful for what the user is looking for. We use this intuition to give a bonus to web pages from discussion forums or blogs. To do this, we define a binary feature that indicates if the source URL for a discussion contains 'forum' or 'blog' terms.

PageRank Although relevant documents are those from discussion websites they should also be trustworthy. To take this element into account, we used page rank scores to prioritize documents from more reliable sources.

Re-ranking The main goal is to re-rank documents based on defined features (initial rank, argumentativeness, domain addresses , and PageRank scores).

To generate the final ranked list, we make a heuristic ranking pipeline; First we get the initial ranked list. Second, we re-rank the list based on PageRank scores. This can result in putting a document, initially ranked very low, on top of the list. To avoid this, we limit moving documents in the ranked list to a maximum of 10 positions. Third, we re-rank the new list based on domains. To perform this step, we put the documents with positive domain feature (which means the document is taken from a blog or forum website) on top of the list. We do this for every 10 documents. Fourth, we classify the whole list using the argumentativeness classifier we have trained. We put documents with positive class on top of the list. We do this for every 20 documents and we do not move a document more than 20 positions.

This way, we reassure that we have prioritized relevance in comparison with other dimensions.

3 Experiments

In this section we will first explain experimental setup and corpora, followed by the experiments and results.

3.1 Experimental Setup

Corpora We used two corpora in our study; one argumentative corpus (args.me) and one web corpus (Clueweb12). Args.me [1] has been created by crawling 387,606 arguments from 4 debate websites to ease research in Argument Retrieval. We used the API of a publicly available search engine[2] based on Elasticsearch to retrieve pro and con arguments from this corpus[7] using BM25 model. We used this corpus for our baseline method.

Clueweb12 is a dataset made by crawling 733,019,372 documents seeded with 2,820,500 urls from Clueweb09[3]. We used a publicly available search engine [5] based on Elasticsearch to retrieve documents from Clueweb12.

[2] www.args.me

All documents have been tokenized by nltk toolkit.

PageRank scores are extracted from chatnoir search engine which has been provided by Carnegie Melon University [3].

We used pre-trained BERT-based model from Huggingface Transformers[4] framework to represent arguments for training the argumentative classifier.

All parameters for the argumentativeness classifier have been set to default values in Scikit-learn[5] library for Python.

Queries We selected 5 out of 50 comparative topics released for the second shared task in Touché track of CLEF2020 to evaluate our model (Ex. "which is better, a laptop or a desktop?").

Initial ranked list We retrieved the top 100 documents from Clueweb for each query using BM25 model.

3.2 Experiments

We evaluated the top 10 documents for each ranked list: initial ranked list, re-ranked by PageRank scores, re-ranked by domains, re-ranked by argumentativeness, and mixed model.

Evaluation results have been reported in Table 1 using NDCG@10 evaluation metric on three criteria: relevance, argumentative structure, and trustworthiness. We labeled documents using three labels: 0 for non-relevant, non-argumentative, or untrustworthy; 1 for relevant, argumentative, or trustworthy, and 2 for highly-relevant, highly-argumentative, or highly trustworthy.[6]

Baseline We retrieve documents from the argumentative corpus and expand the query with a maximum of 5 terms using the top 1000 retrieved documents. Then we use this expanded query to retrieve documents from ClueWeb.

Initial ranks Initial ranks have the most impact on putting relevant documents on top.

Pagerank score Pagerank scores impact trustworthiness by putting pages with more in-links on top.

Argumentativeness We put documents with the positive class for argumentativeness on top of the list.

Domain Blog and forum domains help to put documents from discussion websites on top. This can balance the impact of PageRank, as it tends to give a higher rank to documents from official websites.

[3] boston.lti.cs.cmu.edu/cluew eb09/wiki/tiki-index.php?page=PageRank

[4] https://github.com/huggingface/transformers

[5] https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

[6] Since we did not have official labeled data or labeling guidelines from organizers of Toucheé Track, we labeled documents based on our own guidelines.

Table 1. Results-NDCG@10 metric

Model	Relevance	Argumentativeness	Trustworthiness
initial	0.87	0.71	0.81
pagerank	0.89	0.66	0.87
domain	0.84	0.72	0.80
argumentative classifier	0.80	0.84	0.79
mixed	0.84	0.78	0.82
Baseline	0.55	0.64	0.92

Mixture Being relevant is the first condition for a desired ranked list. In addition, the second priority for a high-ranked document is having an argumentative structure and including comparisons and user reviews, as well as having a trustworthy source. . Thus we need to make sure while mixing all the features in order to re-rank the documents, we do not lose track of relevant documents in the initial ranked list. This is the reason that we started the pipeline by relevance, and limited the changes in document ranks to 10-20 positions while performing re-ranking.

Results As it has been shown in Table 1, the heuristic mixed model does not achieve the same performance as the initial ranked-list in terms of relevance, the same performance as argumentative classifier model in terms of argumentativeness, and the same performance as PageRank model in terms of trustworthiness. But, it struck a balance between all three dimensions.

4 Conclusion

In this paper, we discussed retrieving argumentative documents from the web to assist users in finding the pros and cons of the desired query. The important point in this task is to notice that the user is not only looking for relevant documents; documents including information about one or more options. We should also take argumentativeness and subjectiveness into account.

In this paper, we formulated the problem as a re-ranking task and as we do not have any training data, we treated it in an unsupervised manner.

We used a couple of simple features to re-rank documents from the web in response to an argumentative query. We showed that using a mixture of page-rank scores, web-domain addresses, and argumentative classifier leads to a better ranked list in terms of argumentativeness, relevance, and trustworthiness over the initial BM25 ranked list. PageRank scores help in the trustworthiness dimension by putting documents with more in-links on top of the ranked list. Domains and the argumentativeness classifier help in putting documents with a more argumentative and discussion-based structure in the higher ranks. After all, relevance remains the main ranking dimension: If a document is not relevant, trustworthiness or argumentativeness does not matter anymore. Thus, in the mixed model, we try to limit documents from moving too much in the ranked-list in comparison with the initial BM25 ranked-list. By forcing these limitations, we get a ranked-list with a balance between three dimensions: relevance, argumentativeness, and trustworthiness. However, we did not have the final judgments from Touché track

before submission of this paper, and the evaluations have been performed by manually labeling documents for 5 topics out of 50, and are not official results.

Acknowledgments This research was supported in part by the Netherlands Organization for Scientific Research (NWO, ACCESS project, grant # CISC.CC.016).

References

1. Ajjour, Y., Wachsmuth, H., Kiesel, J., Potthast, M., Hagen, M., Stein, B.: Data acquisition for argument search: The args.me corpus. In: Benzmüller, C., Stuckenschmidt, H. (eds.) KI 2019: Advances in Artificial Intelligence - 42nd German Conference on AI, Kassel, Germany, September 23-26, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11793, pp. 48–59. Springer (2019), https://doi.org/10.1007/978-3-030-30179-8_4
2. Bondarenko, A., Hagen, M., Potthast, M., Wachsmuth, H., Beloucif, M., Biemann, C., Panchenko, A., Stein, B.: Touché: First shared task on argument retrieval. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II. Lecture Notes in Computer Science, vol. 12036, pp. 517–523. Springer (2020), https://doi.org/10.1007/978-3-030-45442-5_67
3. Clarke, C.L.A., Craswell, N., Soboroff, I.: Overview of the TREC 2009 web track. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009. NIST Special Publication, vol. 500-278. National Institute of Standards and Technology (NIST) (2009), http://trec.nist.gov/pubs/trec18/papers/WEB09.OVERVIEW.pdf
4. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019), https://doi.org/10.18653/v1/n19-1423
5. Potthast, M., Hagen, M., Stein, B., Graßegger, J., Michel, M., Tippmann, M., Welsch, C.: Chatnoir: a search engine for the clueweb09 corpus. In: Hersh, W.R., Callan, J., Maarek, Y., Sanderson, M. (eds.) The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012. p. 1004. ACM (2012), https://doi.org/10.1145/2348283.2348429
6. Rafalak, M., Abramczuk, K., Wierzbicki, A.: Incredible: is (almost) all web content trustworthy? analysis of psychological factors related to website credibility evaluation. In: Chung, C., Broder, A.Z., Shim, K., Suel, T. (eds.) 23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume. pp. 1117–1122. ACM (2014), https://doi.org/10.1145/2567948.2578997
7. Wachsmuth, H., Potthast, M., Khatib, K.A., Ajjour, Y., Puschmann, J., Qu, J., Dorsch, J., Morari, V., Bevendorff, J., Stein, B.: Building an argument search engine for the web. In: Habernal, I., Gurevych, I., Ashley, K.D., Cardie, C., Green, N., Litman, D.J., Petasis, G., Reed, C., Slonim, N., Walker, V.R. (eds.) Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP 2017, Copenhagen, Denmark, September 8, 2017. pp. 49–59. Association for Computational Linguistics (2017), https://doi.org/10.18653/v1/w17-5106