Bachelor Informatica

University of Amsterdam

Creating an experimental environment for participating in TREC Fair Ranking Track

René Emmaneel

December 9, 2020

Supervisor(s): Fatemeh Sarvi

Signed:

Abstract

Ranking systems based on queries are used everywhere in the modern online world, from Google search results to jobs and job applicants. To create a ranking of items based on a query, a learning to rank (LTR) algorithm is used to create a model for the ranking of items. Traditional learning to rank algorithms maximize the utility of the rankings for the user, however increasingly information access systems care about the utility for both the user and producer of the items. The TREC Fair Ranking Track is an annual academic search task with the stated goal of developing a benchmark for evaluating retrieval systems in terms of fairness. Based on the track definition we create a framework with the goal of creating a robust environment for participating in the fair ranking track. We implement two learning to rank algorithms utilizing the framework. The first algorithm we implement is the rank support vector machine algorithm. The second algorithm is the FAIR-PG-RANK algorithm. We discuss the performance of the implemented algorithms based on both the utility and fairness metrics as defined by the TREC Fair Ranking Track, and compare it to existing work. We found that the algorithms performed better/equal/worse than existing implementations in both the utility and fairness. Finally, we discuss future work on both the framework and on implementations of fair LTR algorithms.

Contents

CHAPTER 1

Introduction

Ranking systems allow items to be seen by the users, where the exposure each item gets is significantly determined by the position in the ranking. Traditionally a ranking system would order items based on their relevance so that the system is of maximum utility to the user. However, modern information access systems often influence both the consumer and the producer of the content that they serve. Examples are hiring platforms where it is important that the system has a high utility for both the employer and the job-seeker. But also more traditional environments such as music, book or video recommendations can be considered two-sided because the indirect matching of users to content creators.

The ranking system allows exposure to the producers, and for various applications it is important to fairly give exposure to various groups. While it is unlikely that a universal definition of fairness can be created, there are various attempts to measure unfairness and devise various fair ranking algorithms to ensure that the received attention of a given subject is approximately equal to its deserved attention [1].

To evaluate different fair ranking algorithms, the TREC Fair Ranking Track was created. The stated goal was to develop a benchmark for evaluating retrieval systems in terms of fairness, as well as releasing a dataset for benchmarking fair ranking algorithms. The TREC 2019 Fair Ranking Track was an academic search task, where a set of academic article abstract and queries were submitted to an adademic search engine. The central goal is to provide fair exposure to different groups of authors, where the group definitions can be arbitrary, while keeping the utility of the search results high. Various sets of data has been provided, as well as a detailed evaluation protocol. The data of the 2019 track includes the Semantic Scolar Open Corpus, which consists of data describing various papers. It also includes query data, and query sequences. The metrics and data used by the TREC Fair Ranking Track will be discussed in chapter 3.

There are two main ways we will consider for creating a fair ranking algorithm to participate in the Fair Ranking Track. The first way revolves around first creating a traditional LTR algorithm using a rank support vector machine. This is a machine learning algorithm to classify, given a pair of documents, which of the two is more relevant [2]. We use the RankSVM and a simple adjust for exposure algorithm to transform the LTR algorithm into a learning to fairly rank algorithm [3]. Our second approach will incorporate the goal of fairness directly in the machine learning algorithm by defining fair ranking policies and searching the space of fair ranking policies via a policy-gradient approach [4]. The implementation of these two algorithms will be used to answer the following sub research question: *Can we implement a fair learning to rank algorithm for participating in the TREC Fair Ranking Track.* We discuss previous work on fair LTR algorithms in chapter 2. In chapter 4 we will provide a detailed explanation of the algorithms we use.

To participate in the TREC fair ranking track, an environment needs to be build for the entire process. This process includes handling of the input data, implementing the evaluation algorithm, testing of the framework and implementing LTR machine learning algorithms. To design and implement the framework the following research question will be answered: *How can a modular framework be constructed to create an experimental environment for participating in the TREC fair ranking track?*.

To answer the main research questions we will give an overview of the components needed for a framework, and the implementation details of the framework, in chapter 4. In chapter 5 we will provide results of the implemented algorithms and of the framework itself, and in chapter 6 we provide a conclusion and discussion.

CHAPTER 2

Related work

We discuss both work related to fair LTR algorithms in general, and work related to the TREC Fair Ranking Track.

CHAPTER 3

Theoretical background

3.1 Ethical aspect

In this section we will discuss the ethical aspect of the need of fair LTR algorithms, as well as the purpose of the TREC Fair Ranking Track.

3.2 Overview of the TREC Fair Ranking Track

In this section we will give an overview of the data provided by the TREC Fair Ranking Track, as well as the metrics used by the track to score the algorithms on fairness and utility. Also this part should be used to talk more about the TREC Fair Ranking Track.

3.2.1 Semantic Scholar Open Research Corpus

The Sementic Scolar Open Research Corpus is a large archive containing metadata of research papers [5]. It contains 47GB data in JSON format, containing information about papers in many different disciplines. A large dataset is needed for sufficient training sample size. The following data is available for most papers.

- **S2 Paper ID**
- **DOI**
- **Title**
- **Year**
- **Abstract**
- **Venue name**
- **Authors** (resolved to author IDs)
- **Inbound citations** (resolved to S2 paper IDs)
- **outbound citations** (resolved to S2 paper IDs)

For the 2019 TREC Fair ranking Track there was no semantic corpus subset available, however a training set was provided. Using this training set we extract the semantic corpus for the used documents using the available API, which gives the data for a specific document given the S2 Paper ID.

Using this data, we can extract the corpus into smaller csv files containing the metadata of each paper and author. The TREC fair ranking track provided the following 3 csv files, however

we also created functionality to extract the following information ourself given the Sementic Scolar Open Research Corpus. The paper metadata is as follows.

paper_metadata.csv:

- **S2 Paper ID**
- **Title**
- **Year**
- **Venue**
- **Number of inbound citations**

Using the same data, we can also extract metadata information for each author of the papers. During the extraction, we count for each author the amount of papers they worked on and the amount of citations each papers has. We use this information for calculating the i10 and H_index scores, as explained below.

author_metadata.csv:

- **Author ID**
- **Name**
- **Citation count:** sum of all citations of each published paper
- **Paper count**
- **i10:** amount of publications with atleast 10 citations
- **H_index:** The maximum value of h such that the given author has published h papers that have each been cited at least h times.
- **H_class:** The class is 'H' if the author has a H_index of atleast 10, and the class is 'L' if the H_index is less than 10

To combine the papers and authors, we need a third linker file, which combined the paper ID with each corresponding author ID, and the position in which they were attributed.

- **S2 Paper ID**
- **Author ID**
- **Position**

3.2.2 Measuring fairness

Fairness can be subjective and differ for various use cases, and as such the TREC fair ranking track has provided a fairness definition for the acadamic search task. The goal of the fair ranking track is to provide fair exposure to different groups of authors, where the group definition may be arbitrary. Before describing the fairness definition we have to describe the way exposure can influence the discoverability for a certain group, by defining formulas for both the exposure and the relevance an author has.

The exposure an author has is related to the positions of the papers he contributed to, where the first ranked paper has the highest exposure and lower ranked papers have lower exposure. The formula used to measure exposure is the Expected Reciprocal Rank metric [6]. This metric describes the idea that a user is more likely to stop browsing if they have seen a highly relevant document. The user will look at each document in order, and will have a chance to stop browsing depending on the relevance of the document. It also implements an abandonment probability γ, which represent the chance the user will examine the next document, meaning that there is a $1 - \gamma$ chance of abandoning the search. The exposure of the document on place i given the Expected Reciprocal Rank metric is as follows.

$$e_i = \gamma^{i-1} \prod_{j=1}^{i-1} (1 - f(r_j))$$

where $f(r_j)$ is a function to transform the relevance of document d_j into a probability of stopping the search after seeing the document. For the fair ranking track the relevance is either 0 or 1, and the function is given as $f(r_d) = 0.7 \cdot r_d$. The continuation probabilty is given as $\gamma = 0.5$.

Because the task is to provide fair exposure to authors instead of documents, we have to calculate the cumulative exposure of the documents which each author has contributed to. Given a ranking of documents π, we can calculate the exposure of author a as follows.

$$e_a^\pi = \sum_{i=1}^{n} (e_i) \cdot I(\pi_i \in \mathcal{D}_a)$$

Where π_i is the document on position i, and $\mathcal{D}_a$ is the documents which include a as an author. Because the final result of the track consist of a series of different rankings, we can calculate the amoritezed exposure of author a as,

$$e_a = \sum_{\pi \in \Pi} e_a^\pi$$

where Π is the sequence of all rankings.

While the exposure of an author is dependent on the ranking of the documents, the relevance of the author is dependent on the relevance of the individual documents the author worked on. The author relevance is simply the sum of the stop probability for each document where a is an author

$$r_a^\pi = \sum_{d \in \mathcal{D}_a} f(r_d)$$

And the amortized relevance for an author is defined as

$$r_a = \sum_{\pi \in \Pi} r_a^\pi$$

We can assume that each author is part of exactly one group. The goal of the fair ranking track to provide fair exposure for each group relatively to the relevance for each group. Let $\mathcal{G}$ be the set of all groups and $\mathcal{A}_g$ be the set of all authors in group g. The group exposure and relevance are defined as,

$$\mathcal{E}_g = \frac{\sum_{a \in \mathcal{A}_g} e_a}{\sum_{g' \in \mathcal{G}} \sum_{a \in \mathcal{A}_{g'}} e_a}$$

$$\mathcal{R}_g = \frac{\sum_{a \in \mathcal{A}_g} r_a}{\sum_{g' \in \mathcal{G}} \sum_{a \in \mathcal{A}_{g'}} r_a}$$

Note that $\sum_{g \in \mathcal{G}} \mathcal{E}_g = 1$ and $\sum_{g \in \mathcal{G}} \mathcal{R}_g = 1$. We can use these metrics to calculate the deviation between the exposure and relevance for each group.

$$\Delta_g = \mathcal{E}_g - \mathcal{R}_g$$

Groups should receive exposure proportional to relevance, meaning that a perfectly fair model provides a deviation of zero for each group. We can compute the fair exposure using the square norm.

$$\Delta = \sqrt{\sum_{g \in \mathcal{G}} \Delta_g^2}$$

We measure the relevance of a ranking using the Expected Reciprocal Rank metric, which is the same metric as used in the fairness measurment. Recall that the exposure of document i is defined as,

$$e_i = \gamma^{i-1} \prod_{j=1}^{i-1} (1 - f(r_j))$$

We multiply the exposure of each document with the stop probability to calculate the total utility of a ranking π as follows,

$$u^{\pi} = \sum_{i=1}^{n} e_i \cdot f(r_i)$$

The average utility of all rankings, $U = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} u^{\pi}$, will be the final relevance metric.

CHAPTER 4

Method

In this chapter we will discuss the various needed aspects for the framework, and the LTR algorithms.

4.1 Framework

In this section we will discuss the framework.

4.2 Learning to rank algorithms

Before looking at a way to create a model for learning to fairly rank, we will first propose a model for a standard LTR model on the TREC fair ranking track data. LTR, when applied to document retrieval, is a task as follows. In training a set of queries and documents for each query is provided, as is the relevance judgement for each document. Using this data, a ranking function is created such that the model can predict a ranked list of documents for a query. The two major approaches for creating a ranking function is either a non-learning approach or a LTR approach. Learning approaches for information retrieval (IR) have been widely studied and shown to be useful for ranking articles. [7]

A LTR framework is systematically represented in Figure 4.2. As can be seen in the figure, a training set is needed to train the ranking model. The training data consist typically of n training queries, and m documents for each query. Each document is represented by a feature vector, which is a list of values calculated before training based on the document and query. Each query also has a relevance judgement y, which can for example be the order in which the documents are ranked, or can be a vector representing the relevance for each document based on the query. After training a model has been created, which can be used to rank a set of given documents based on the query.

There are three main LTR algorithms. The pointwise approach, which learns a function for directly predicting the relevance for each document for a query. The pairwise approach, which learns a function that takes in a pair of documents and a query and tries to predict the document that is most relevant for the given query. Lastly there is the listwise approach, which learns a function to directly rank a set of input documents based on a query.

In this thesis we have decided to create both pairwise and listwise ranking algorithm. This decision was made based on the training data available in the TREC fair ranking track. As described in chapter 3, the training data consist of queries and a certain amount of documents per query, with a boolean relevance score for each document. Because of the boolean relevance score, the pointwise approach would not result in a good ranking model. We will first discuss the various features extracted for this model, after which we will discuss the two learning algorithms itself.

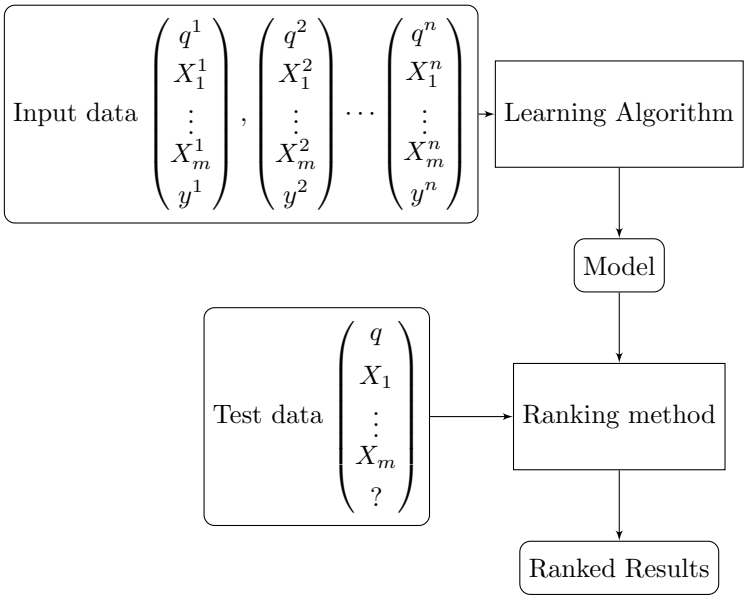

Figure 4.1: LTR framework

ID	Feature Description
1-3	TF in title, venue, abstract
4-6	IDF in title, venue, abstract
7-9	TF_IDF in title, venue, abstract
10-12	BM25 in title, venue, abstract
13-21	LMIR features in title, venue, abstract
22-24	length of title, venue, abstract
25	amount of in citations
26	amount of out citations
27	year

Table 4.1: Features used in the rankSVM model

4.2.1 Feature extraction

In LTR, each query-document pair is represented by a feature vector, and each feature gives an indication to the relevance of the document to the query. For example one feature could be the frequency of the query words in the abstract of the document. In this section we will discuss the various types of features, and which features we have used in constructing the feature vectors for the LTR framework.

Features can be categorized in 3 distinct categories. The query-document category means that the feature is dependant on both the feature and document, examples are term frequency and BM25. The other two categories are only dependant on the query or the document respectively.

The different features as used in the rankSVM model for the TREC fair ranking track are described in table 4.1. The explanation about these features are listed below.

- Features 1-3 are the term frequency (TF) features. The features will count the amount of occurences of each term in the query in the respective field on the document.
- Features 4-6 are the inverse document frequency (IDF) features. The inverse document frequency function is a measure about how much information a given word provides. The function is defined as follows,

$$IDF(t) = \log \frac{N}{n(t)}$$

with N being the total number of documents, and n(t) being the amount of documents with the term t.

- Features 7-9 are the term frequency - inverse document frequency (TF-IDF) features. For every term of the query, the TF and IDF of the specific term is multiplied. This feature is correlated by a high frequency of a term, which has a low document frequency.

- Features (G-H) are the BM25 features. This is a TF-IDF-like retrieval function, first developed for TREC-3 [8]. The function used is defined as follows.

$$BM25(q,d) = \sum_{q_i \in q \cap d} IDF(q_i) \cdot \frac{TF(q_i,d) \cdot (k+1)}{TF(q_i,d) + k \cdot (1 - b + b \cdot \frac{length(d)}{avglength()})}$$

  where length(d) is the amount of words in document d, and avglength() is the average amount of words in all documents. k and b are free variables, which can be set to various values depending on the dataset, but are set to k=1.2 and b=0.75, which experiments have shown to be reasonable values [9].

- Features 13-21 are the language model information retrieval (LMIR) features. LMIR is a statistical language model, with the goal of predicting the probability of the document's language model generating the terms of the query [10]. The way LMIR aims to find this probability is by combining the various probabilities of a certain term in the query being generated by the document's model ($P(t \mid d)$), which gives the equation as follows.

$$P(q \mid d) = \prod_{i=1}^{M} P(t_i \mid d)$$

  where $t_1, \cdots, t_M$ are the terms in query q and $P(t \mid d)$ is the document's language model. There are many variants of LMIR, which mainly differ in the use of the document's language model, of which we use three different methods as described by Zhai and Lafferty [11].

- Features 13-15 are the Jelinek Mercer LMIR features. The language model is based on a combination of the percentage of occurence of a term in the given document, as well as the percentage of occurence of a term in all documents combined. The document's language model can then be constructed as follows.

$$P(t_i \mid d) = (1 - \lambda)\frac{TF(t_i,d)}{LEN(d)} + \lambda\frac{TOT\_TF(t_i,C)}{TOT\_LEN(C)}$$

  where $TOT\_TF(t_i,C)$ is the term frequency of term i in the entire corpus, $TOT\_LEN(C)$ is the total amount of terms in the corpus and λ is the smoothing factor, which is set to 0.1 in our framework.

- Features 16-18 are the Dirichlet LMIR features.

$$P(t_i \mid d) = \frac{TF(t_i,d) + \mu TOT\_TF(t_i,C)}{TOT\_LEN(C) + \mu}$$

  with μ is 2000

- Features 19-21 are the Absolute discount LMIR features.

$$P(t_i \mid d) = \frac{max(TF(t_i,d) - \delta, 0)}{TOT\_TF(t_i,C)} + \sigma\frac{TOT\_TF(t_i,C)}{TOT\_LEN(C)}, \text{with } \sigma = \delta\frac{UNIQUE(d)}{LEN(d)}$$

  with μ is 0.7

- Features 22-27 are features various other features, including the length in words of various parts of the documents, as well as the amount of citations a documents has and the year of publication.

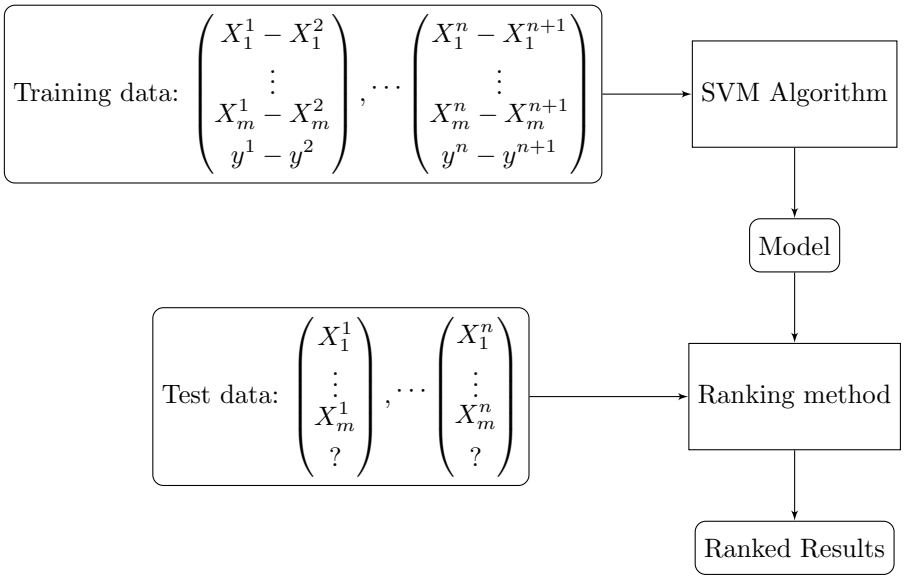

Figure 4.2: Learning to rank framework

4.2.2 Ranking Support Vector Machine

Support Vector Machine (SVM) algorithm is a supervised learning model with the goal of categorizing data represented as points in space [12]. Given a set of training examples, each being labeled in one of the two possible categories, the SVM algorithm builds a model to categorize new examples in one of the two categories. The training examples are represented as a combination of a label and a feature vector, and the model output is represented as the maximum-margin hyperplane that divides the points of one category and the other, so that the distance from the hyperplane and the closest point from either group is maximized. Utilizing the SVM algorithm to categorize documents as relevant or not has been used to create information retrieval models before [13]. However the main goal of a ranking system is to predict the relative order of documents based on a query, while the above described pointwise approach merely learns the degree to which a document is relevant based on the query. We instead use an algorithm that directly takes the goal of ordering more relevant documents higher into account.

Ranking SVM is a variant of the above described SVM algorithm. The goal of the algorithm is to learn a model which, given two documents and a query, learns which document is more relevant [2]. The main benefit of the pairwise approach is that creating a model directly creates a way to sort documents by relevance utilizing the model.

4.2.3 Policy Learning for Fairness in Ranking

In this subsection we will discuss the implementation of FAIR-PG-RANK as described by Singh, Joachims [4]

CHAPTER 5

Experiments

In this chapter we will show the results of the framework and the implemented LTR algorithms

CHAPTER 6

Conclusions

In this chapter we will reflect on the previous chapter

Bibliography

[1] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. *CoRR*, abs/1805.01788, 2018.

[2] Thorsten Joachims. Optimizing search engines using clickthrough data. KDD '02, page 133–142, New York, NY, USA, 2002. Association for Computing Machinery.

[3] Meng Wang, Haopeng Zhang, Fuhuai Liang, Bin Feng, and Di Zhao. ICT at TREC 2019: Fair ranking track. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019*, volume 1250 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2019.

[4] Ashudeep Singh and Thorsten Joachims. Policy learning for fairness in ranking, 2019.

[5] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. Construction of the literature graph in semantic scholar. In *NAACL*, 2018.

[6] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 621–630, New York, NY, USA, 2009. Association for Computing Machinery.

[7] Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. Letor: A benchmark collection for research on learning to rank for information retrieval. 2016.

[8] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. pages 109–126, 1996.

[9] Hinrich Schütze Christopher D. Manning, Prabhakar Raghavan. An introduction to information retrieval. *Cambridge University Press*, page 270, 2009.

[10] Tao Tao, Xuanhui Wang, Qiaozhu Mei, and ChengXiang Zhai. Language model information retrieval with document expansion. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 407–414, New York City, USA, June 2006. Association for Computational Linguistics.

[11] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 334–342, New York, NY, USA, 2001. Association for Computing Machinery.

[12] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[13] Ramesh Nallapati. Discriminative models for information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, page 64–71, New York, NY, USA, 2004. Association for Computing Machinery.