Automatic Timeline Generation using Learning To Rank

Anonymous ACL submission

1 Problem Definition

For a better understanding of world affairs, a simple ordering of related articles as they are published in newspapers or online media is beneficial for readers. The task is to arrange news articles related to a specific event in chronological order automatically.

2 Background

Chronological order or Timeline is an effective and straightforward way to understand the actions over time belonging to any event. It gives a complete picture of the event, actors, temporal data; situations changed over time with actions and the outcome. chronological orders not only help in better understanding, but they are helpful in governance and legal matters. Our solution automates the creation of chronological orders in an efficient way when a large number of articles are provided.

Learning to Rank(LTR) is a well-studied concept in the field of Information retrieval and web search engines. We have mapped the ‘Ranking’ with ‘Sorting’ in our proposed solutions.

3 Dataset

Collection of 396 news articles related to Nirbhaya rape case were scrapped. This corpus is divided into two sets.

Set 1 : Articles text files with title as the name of the news headline.

Set 2 : News publishing date mapped against the article’s name.

We have performed web scraping for gathering information from the Hindustan Times website. Following steps are performed for the same:

Step 1: Inspect Nirbhaya Case news articles from Hindustan Times home page.

https://www.hindustantimes.com/search?q=nirbhaya

Step 2: Understand the HTML/CSS source code of the web pages using BeautifulSoup and requests. “Requests” is used to get the site’s HTML code into a python script. Beautiful Soup is a Python library for parsing structured data. It allows interacting with HTML in a similar way to how one would associate with a web page using developer tools. Developer tools can help to understand the structure of a website. All modern browsers come with developer tools installed. Developer tools allow to interactively explore the site’s DOM to understand better the source that we’re working with. To dig into the page’s DOM, select the Elements tab in developer tools. We’ll see a structure with clickable HTML elements.

Step 3: Save the article body and dates (for the ground truth purpose) in the separate files with the file name as the name of the article.

Step 4:Since not all the articles are related to the Nirbhaya rape case specifically. Therefore, we have filtered out related articles manually. After this step, the size of the final data set becomes 128 articles.

4 Literature Review

Ranking is a well-known deliverable of any information retrieval system. There exist conventional methods of ranking in IR which can be categorized between query dependent and query independent approaches. The problems with approaches are that the evaluation metrics are not differentiable with respect to parameters, parameter tuning is a difficult task for good performance of ranking model and combination of such models to achieve efficiency is hard. Machine learning can help in all the above issues and with the advancement of deep learning models, it gives good results.

There is a learning system to which we fed the labelled data and it is trained. When the IR model gives to be Top K documents in some order, the inference part of the framework i.e. the Ranking system gives the output of ranked documents which tries to approximates the actual order should be.

There are three approaches to model the learning phase which are: the pointwise approach, the pairwise approach and the listwise approach. Thus, they imply different input and output spaces with different loss functions.

4.1 Pointwise Approach

The input contains the feature vector of each document and output contains the relevance degree of each document. The loss function decides the degree to which the model has predicted the output in reference to the ground truth. Thus, various regression machine learning models can be used directly.

4.2 Pairwise Approach

The input space is defined by the two feature vectors of two documents and output are the explaining which document is better in terms of ranking. The goal is to reduce the number of inversions in the ranking. This can easily be achieved by classification machine learning models for binary classification problem.

4.3 Listwise Approach

The input space contains all the documents at one place and output space has two categories. Some algorithms give output as degree of relevance of each document when all documents were taken together where some algorithms output is the ranking of the documents. Thus these model try to directly optimise the value of evaluation metrics

There are different kind of evaluation metrics are used for Learning to Rank concept such as Mean average precision, NDCG and DCG, precision@N etc.

4.4 RNN for Text Processing

The study suggests that Recurrent Neural Networks (RNN) are doing an excellent job for the text analysis tasks like text classification, prediction of next sentence for a paragraph, etc.RNN are designed to make use of sequential data when the current step has some relation with the preceding steps. The study suggests that Recurrent Neural Networks (RNN) are doing an excellent job for the text analysis tasks like text classification, prediction of next sentence for a paragraph, etc.RNN is designed to make use of sequential data when the current step has some relation with the preceding steps. This property of RNN makes it worthy of dealing with natural language processing, time-series data etc.We can further train our RNN model to perform advanced tasks like language modeling, image captioning etc.

4.5 Loss Functions

There are two categories of loss functions—first, direct optimization of metrics like NDCG, MAP such as softNDCG in SoftRank, AdaRank-MAP. Second, proposing new metrics that directly identify the characteristics of the problem statement such as cross-entropy loss in ListNet, cosine loss in cosineRanking, log-likelihood loss in ListMLE.

5 Proposed solution

A feature vector corresponding to each document in the dataset is obtained using Doc2Vec. Date from the metadata is used to obtain the true order of the articles.

5.1 Proposed Solution with Pairwise approach

A feature vector corresponding to each document in the dataset is obtained using Doc2Vec. Date from the metadata is used to obtain the true order of the articles. The training data format is as follow:

A pair of feature vectors of documents di, dj is passed as input. The target label is 0 or 1 according to the true order of di and dj. Hence, a pair of documents constitutes one data point. The dataset consists of 34 articles. The dataset is divided into training and testing using a 75:25 train test split. This means there are 300 train data points and 36 test data points. Finally, the 9 articles in test set are sorted using the trained model as comparator. The predicted permutation is evaluated against the true permutation using NDCG.

5.2 Proposed Solution with List wise approach

In this approach, we passed the input document vectors (d1 ,d2,d3....dn) of the training points one by one to the RNN model. The RNN model generates the corresponding output vectors (o1,o2,o3......on) and the updated hidden unit. This hidden state is again fed into the model with the next input document vector. Once all the input document vectors are passed into the model, the final hidden unit has seen all the training document vectors.This means that it has learned the semantics of the articles. This hidden unit is in vector form, which is appended to each output vector (oi) and further fed into multi-layer perceptron (MLP). MLP generates a unique score (vi) for each vector. Now, these output vectors are sorted based on this unique score (vi). This sorted ordering is compared with the actual ordering (chronological ordering) of the training articles to compute mean squared error loss.This loss is used to update the weights of RNN during backpropagation.This step is repeated several times until the loss does not decrease further.

Once the model is trained ,it can now be used to find the chronological ordering of the news articles.

5.3 Evaluation measure

In NDCG, the relevance scores are assigned as follows: Let D = [d1, d2,..., dk] be a set of documents in true order. The corresponding relevance scores of the documents are [d1=k, d2=k-1, . . . , dk=1]. 4.3 Work ahead: We plan to use more sophisticated ML classification models e.g. Neural networks to approximate the comparator function which is used in sorting the documents. Since Listwise approach of LTR is considered to be better than the other two for ranking, We will try to understand the listwise approach and see its feasibility in the given problem statement.

6 Establishing Baseline

As discussed, there are three approaches in the LTR domain: Pointwise, Pairwise and Listwise approach. From thorough research, we have found that the Baseline for our problem statement does not exist. So we created our own baseline. For creating a baseline we have used a pairwise approach. We trained a Machine Learning classification model, specifically Logistic Regression to approximate the comparator function. The accuracy of the logistic model-based comparator is as follow: Train accuracy: 93.33Test accuracy: 50Using this comparator, the test articles are sorted. The NDCG evaluation metric is used to quantify the sorting performance. With the test set of 9 articles, the NDCG obtained is 0.7079.

7 Result and Inferences

- In the pairwise approach, we have tried different machine learning models to see how good they perform on test data and actual ranking task. In accuracy graphs below, the random forest seems promising in the training phase, but it suffers from overfitting. Overall, the K-Nearest Neighbour classifier is best by giving 61% test accuracy. The NDCG remains the same at 0.7079 because model learning is comparable in terms of test accuracy, which ranges from 44% to 61%.

- For evaluating list wise approach we again used NDCG measure.We computed the median of all the NDCG scores of test points and achieved score of 0.62

- In the proposed pairwise approach, the comparator takes into consideration only 2 documents at a time. This approach does not take into account the overall context of the set of documents. In contrast, the proposed list wise approach processes all the documents of the set one by one and computes the hidden states of the RNN. the final hidden state carries the overall context of the set of documents, which is used during the computation of the scores of the documents. This way since the proposed list wise approach considers the overall context of the documents for their ordering, it gives better results than the proposed pairwise approach.

- When implementing List wise approach using RNN , the selection of appropriate loss function is the major concern.The selection of loss function highly varies from case to case.

- There was no proper relation between the accuracy of model and size of the training document vectors list.Therefore , we have fed training document vectors list of random sizes.

References

[1] Liu, Tie-Yan. ”Learning to rank for information retrieval.” Foundations and Trends® in Information Retrieval 3.3 (2009): 225-331.

[2] Dong, Xishuang, et al. ”An overview of learning to rank for information retrieval.” 2009 WRI World Congress on Computer Science and Information Engineering. Vol. 3. IEEE, 2009.

[3] Ai, Qingyao, et al. ”Learning a deep listwise context model for ranking refinement.” The 41st International ACM SIGIR Conference on Research Development in Information Retrieval. 2018.

[4] Taylor, Michael, et al. ”Softrank: optimizing nonsmooth rank metrics.” Proceedings of the 2008 International Conference on Web Search and Data Mining. 2008.

[5] Cao, Zhe, et al. ”Learning to rank: from pairwise approach to listwise approach.” Proceedings of the 24th international conference on Machine learning. 2007.

[6] Dinu, Liviu P., and Radu-Tudor Ionescu. ”A rank-based approach of cosine similarity with applications in automatic classification.” 2012 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing. IEEE, 2012.

[7] Xia, Fen, et al. ”Listwise approach to learning to rank: theory and algorithm.” Proceedings of the 25th international conference on Machine learning. 2008.