BITS ID : 2019HC04578

Title of the paper : Improving Ad Hoc Retrieval With Bag Of Entities

Authors :

Gustavo Gonçalves
João Magalhães
NOVA LINCS
Universidade NOVA de Lisboa
Caparica, Portugal

Chenyan Xiong
Jamie Callan
Language and Technology Institute
Carnegie Mellon University
Pittsburgh, United States of America

Published in Text REtrieval Conference (TREC) - 2018

Summary

Problem addressed : Improving feature-based learning-to-rank (LTR) search engine performance using entity information.

The retrieval accuracy can be improved in Web, medical and academic search by combining the traditional Bag-of-Words (BoW) and entity-oriented representations.

Taking the News industry, which mainly focuses on entities such as people, companies and countries etc., “Washington Post” news dataset was selected for this core work along with some text scraped from social media posts that was embedded in the news.

Using entities to bring the semantically similar concepts closer in an entity space which can help in improving the ad-hoc retrieval.

Contributions:

There were 2 hypotheses proposed, to improve the retrieval results.

1) Using entities along with the BoW representation.
2) Comparing the BoW model and entity representation in dealing with short fields of text (e.g., titles).

There were many approaches explored to make most use of the added value of the entities to improve the retrieval results.
(i) in-performing query expansion with respect to the added entity values.
(ii) Using Latent entity space, space where the query and a document relevancy is calculated and projected.

Bag-of-Entities (BoE) model introduced by Xiong, focuses on transposing the documents to an entity space which is defined for both queries and documents, by entity linked systems.

3 different entity tagging systems were compared “Freebase”, “Wikipedia” and “TagMe”, out of which “TagMe” was selected as it provides a balance between recall and precision.

LTR uses machine learning to identify the textual features (document length, TF , IDF) of a document to be used to optimize the metric such as Mean Average Precision (MAP). The LTR systems are used to set up standards for an information retrieval system, which are used as baseline to evaluate the new search systems.

Based on the preview of the given dataset, we could see most of the documents in the collection are from blog posts, 65% of documents are dated between 2014 and 2016, also highlighting the missing dates on blogs.The news data was indexed by separating each document in its title and body, using virtual document model, considering the data from social media posts, multimedia captions etc.,

Approach :

The entity representation was created for both corpus and the queries, where there was a dedicated field for “title” entity and for the “body” entity. The LTR architecture was employed by re-ranking the results provided by BM 25 (Best Matching) rank of 1000 documents.

1) **Entity Linking:**

“TagMe” entity tagging system was used to link against the WIkipedia dump. The idea behind this inking is to link the similar concepts to the same entity representation, reducing the dispersion between semantic synonyms. This results in exact match between queries and document terms by entity representation or ID.

“TagMe” aims for balance between precision and recall, causing some incorrectly linked entities. However, these incorrect links are not a major issue due to TagMe’s consistency, thus still approaching the terms in the entity space.

This situation will harm our results, which is one of the limitations of our work.

2) **Feature Selection and Re-ranking**

The evaluated runs had LTR models trained on 150 queries from Wall Street Journal collection from 1987 to 1992. Training was with 10-fold cv to reduce the bias and reduce overfitting using the identified combination of features. The BM25 run was re ranked with default parameters using the trained models on Washington Post collection.

With each run the retrieval accuracy was evaluated for various documents and queries for using BoE versus BoW approaches and the short and long textual elements such as title and document's body.

Results :

BM25-b-BoW was the baseline (values calculated MAP, nDCG@1000 and P@10) proposed to be compared with other representations.
The entity representations (t-BoE, b-BoE and bt-BoE) runs were executed, however, they were not able to beat the BM25 baseline, However, when used in combination with BoW representation (bt-BoWBoE) it substantially improves the BoW LTR runs, bt-BoW and b-BoW.

In general, combining both the title and body textual representation the system's performance improves for both BoW and BoE, which is also true when the title field is removed from models bt-Bow and bt-BoWBoE.

Conclusion :

Using entities alone is not enough to beat the baselines,however, when used along with BoW , the system's performance is increased. Use of title and body together also helps in system's performance improvement. In the short textuals , the experiments looked promising with entity representation. Hence entity representation is a promising tool to be researched further.