Evaluation of IR Models

Parth Shah
SEAS, University at Buffalo
Buffalo, NY, 41214
Person Number: 50291125
parthnay@buffalo.edu

Abstract

The goal of this project is to implement various IR models, evaluate the IR system and improve the search result based on your understanding of the models, the implementation and the evaluation.

1 Introduction

This project deals with the implementation of IR models such as Vector Space Model, BM25 model, DFR models based on Solr using twitter data from train.json and the results are evaluated using TREC_eval program. We are given 15 training queries and 5 testing queries in languages – German , English and Russian. The key concept of this project is to improve the performance of IR system by considering primarily MAP(mean average precision)as an evaluation measure. The three models to be tested are

1. Best Matching (BM25)
2. Vector Space Model (VSM)
3. Divergence From Randomness (DFR)

2 Dataset

We are given twitter data in three languages - English, German and Russian, 15 sample queries and the corresponding relevance judgement in files named queries.txt and qrel.txt.
The data given is Twitter data saved in json format, train.json. Three languages are included - English (text_en), German (text_de) and Russian (text_ru).
train.json: This file contains the tweets with some fields extracted from raw data.
Sample tweet format is as follows:

"lang": ,
"id": ,
"text_de": ,
"text_en": ,
"text_ru": ,
"tweet_urls": [ ],
"tweet_hashtags": []

3 Best Matching 25 (BM25)

The Okapi BM25 model is a probabilistic information retrieval model which was originally designed for short-length documents. In Solr, the similarity class for this is solr.BM25SimilarityFactory. Made core train_newcore for implementing BM model. Added similarity field in the schema.xml. Following are the screenshots of similarity class for default

Boolean Model and corresponding
MAP value.
The BM25 Model comes with 2 parameters -

1. k1 - Controls non-linear term frequency normalization (saturation).
2. b - Controls to what degree document length normalizes values.

On tweaking both the values of k1 and b, the results chosen were k1 = 1.2 and b = 0.75.
The default MAP values for BM25 for the test queries were -

```
map                     001 0.3433
map                     002 0.4202
map                     003 0.5729
map                     004 0.5724
map                     005 0.5000
map                     006 0.4991
map                     007 0.8333
map                     008 1.0000
map                     009 1.0000
map                     010 1.0000
map                     011 1.0000
map                     012 0.6616
map                     013 0.1041
map                     014 0.6386
map                     015 0.8667
map                     all 0.6675
```

4 Vector Space Model (VSM)

In Vector Space Model(VSM), the queries and documents are represented as vectored quantities in multi-dimensional space where each index item is a dimension and weights are TF-IDF values. In Solr, the corresponding similarity class is
solr.ClassicSimilarity.There are no parameters to tweak in a Vector Space Model.
The default MAP Values for VSM for the test queries were -

```
map                     001 0.3403
map                     002 0.4011
map                     003 0.5729
map                     004 0.5724
map                     005 0.5000
map                     006 0.5257
map                     007 1.0000
map                     008 1.0000
map                     009 1.0000
map                     010 1.0000
map                     011 1.0000
map                     012 0.4615
map                     013 0.1098
map                     014 0.7028
map                     015 0.7721
map                     all 0.6639
```

5 Divergence From Randomness (DFR)

In Divergence from randomness model, the term-weight is inversely related to the probability of term-frequency within the document obtained by a model of randomness.
In Solr, the similarity class for this is given by solr.DFRSimilarityFactory.
It one type of probabilistic model. It is basically used to test the amount of information carried in the documents. It is based on Harter's 2-Poisson indexing-model. The 2-Poisson model has a hypothesis that the level of the documents is related to a set of documents which contains words occur relatively greater than the rest of the documents.
DFR modes provides 3 scope for tweaking -

1. BasicModel: Basic model of information content
2. AfterEffect: First normalization of information gain
3. Normalization: Second (length) normalization

As suggested by project documentation, the parameters set were “BasicModelG : Geometric approximation of Bose-Einstein” plus “Bernoulli : Ratio of two Bernoulli processes” first normalization plus “H2 : Term frequency density inversely related to length” second normalization.
The default MAP Values for VSM for the test queries were -

```
map                   001 0.3722
map                   002 0.4160
map                   003 0.5471
map                   004 0.5484
map                   005 0.5000
map                   006 0.5065
map                   007 0.8333
map                   008 1.0000
map                   009 1.0000
map                   010 1.0000
map                   011 1.0000
map                   012 0.7495
map                   013 0.1041
map                   014 0.6386
map                   015 0.8667
map                   all 0.6722
```

Thus the following were the default values as obtained from implementing the basic default models for BM25, VSM and DFR.

	BM25	VSM	DFR
MAP Values	0.6675	0.6639	0.6722

6 Improving MAP Values

1. Model Specific Tuning

We have parameters we can change for BM25 Model like k1 and b, Using different values for k1 and b the different MAP values were obtained. Comparing these MAP Values the best pair was chosen to be k1 = 1.3 and b = 0.75
The changes in MAP values with these paramet changes were -

	Default	Model Specific
BM25	0.6675	0.6705(k1=1.2 & b=0.75)
VSM	0.6639	-
DFR	0.6722	-

2. Query Translation

We found that SOLR was not returning any Non-English documents for English Queries and Similar Observation was seen for other languages as well. Logically understanding this ambiguity, a Russian Twitter user is expected to write same content/Information in Russian language and hence, our implementation should search for same Information in Russian language as well. Additionally, we realized it's not Just sufficient to search blindly with all above translated queries but institutively it makes sense to search Russian query in Russian field, English in English and German and German. Therefore, we combine this information in Implementation

Thus altering a query which was given as

001 Russia's intervention in Syria

To

001 text_en:(Russia's intervention in Syria) OR text_ru:(Вмешательство России в Сирии OR Russia's intervention in Syria) OR text_de:(Russlands Intervention in Syrien OR Russia's intervention in Syria)

To return the relevant documents for the query in the same language as that of the query.

This increased the MAP Values as follows

	Default	Model Specific	Query Translation
BM25	0.6675	0.6705(k1=1.2 & b=0.75)	0.7116
VSM	0.6639	-	0.7219
DFR	0.6722	-	0.7162

3. Synonym

We added synonyms for words that occur in the queries to thus retrieve more queries and set the relevance score accordingly.

On adding synonyms to the to synonyms.txt found in core/conf/synonyms.txt , we find that the scores for models improved to an extent except for VSM.

The MAP values for VSM were actually reduced by use of Synonyms thus we did not use synonyms for VSM Model.

The increased MAP Values are as follows

	Default	Model Specific	Query Translation	Synonyms
BM25	0.6675	0.6705(k1=1.2 & b=0.75)	0.7116	0.7224
VSM	0.6639	-	0.7219	0.7005
DFR	0.6722	-	0.7162	0.7338

4. Query Boosting

A query searched in particular language is more likely to be relevant to documents in that language. Hence, its becomes imperative to set weight of test field mode if Query is in language sa as that text field

In this technique, as shown below we extract the language and then we set weight of text field with that language higher than other text fields.

The increased MAP Values are as follows

	Default	Model Specific	Query Translation	Synonyms	Query Boosting
BM25	0.6675	0.6705(k1=1.2 & b=0.75)	0.7116	0.7224	0.7245
VSM	0.6639	-	0.7219	0.7005	0.6611
DFR	0.6722	-	0.7162	0.7338	0.7375

7 Observation

We tweaked the various IR models to obtain the following MAP Values

BM25 Gave the best MAP Values

.

	Default	Model Specific	Query Translation	Synonyms	Query Boosting	Final MAP
BM25	0.6675	0.6705(k1=1.2 & b=0.75)	0.7116	0.7224	0.7245	0.7245
VSM	0.6639	-	0.7219	0.7005	0.6611	0.7219
DFR	0.6722	-	0.7162	0.7338	0.7375	0.7375

Acknowledgments

This report was prepared in accordance to the project3_Evolution_of_IR_Models.pdf provided, BM25, VSM and DFR documentation available on Wikipedia and LearnInformationRetreival.com

References

[1] https://en.wikipedia.org/wiki/Okapi_BM25

[2] https://en.wikipedia.org/wiki/Vector_space_model

[3] https://lucene.apache.org/core/4_4_0/core/org/apache/lucene/search/similarities/DFRSimilarity.html1

[4] https://en.wikipedia.org/wiki/Divergence-from-randomness_model