CSE 535 - Information Retrieval
Project 3 Report

Ajith Kumar Natarajan
UBIT name: ajithkum
UB Person number: 50318505
ajithkum@buffalo.edu

Abstract

In this project I have implemented three different types of Information Retrieval models to study their practical application. All the three types of implementation were *probablistic models*. Specifically, they are:

- Best Matching 25 (BM25)
- DFR (Divergence From Randomness)
- Language Model (LM)

Mean Average Precision (MAP) was the selected method of evaluation measure. The formula to calculate MAP is:

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP(q)}}{Q}$$

Figure 1: MAP evaluation formula

Here *AveP* stands for average precision and *Q* is the number of queries.
The MAP values obtained for different methods were:

MAP values	
BM25	0.7095
DFR	0.7022
LM	0.7056

1 Introduction

In this project, Twitter dataset consisting of tweets in three different languages (English, Russian, German) have been taken to implement the IR model. Separate cores were created in Solr with respective similarity factories. The dataset was indexed on on each of the three cores. The tweets returned for each query from the different model types are evaluated using trec_eval, the software used in Information Retrieval conferences, by comparing the result returned with the ground truth.

2 Information Retrieval models

There are different types of Information Retrieval (IR) models. Each IR model uses a different strategy to evaluate and retrieve relevant documents. The models are categorized as follows:

- Similarity-Based Models
- Algebraic Model - Eg: Vector Space Model
- Set-Based - Eg: Boolean model, extended Boolean model
- Probabilistic Relevance Model - Eg: BM25, DFR, LM
- Query Likelihood Model

Here is a small description of the types of model implemented in this work.

3 Probabilistic models

This type of model casts relevance of the returning values as a probability problem. According to probabilistic models relevance score is reflected by the probability a user will consider the result relevant.

3.1 Best Matching 25 (BM25)

BM25 improves upon TF-IDF. It ranks the document set based on the query terms appearing in document dataset, regardless of their proximity within the document.

Given a query Q, containing keywords *q1,...,qn*, the BM25 score of a document D is:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)},$$

Figure 2: BM25 formula

where *f(qi,D)* is *qi*'s term frequency in the document D, *D* is the length of the document D in words, and *avgdl* is the average document length.

3.2 Divergence from Randomness (DFR)

It is used to verify the amount of relevant information present in the document. It uses term weight as standard, which is calculated using divergence between the term distribution produced by models like Bose-Einstein and the actual term distributions.

3.3 Language Models

Language model approach to IR directly models the idea that a document is relevant to a query if the document model is likely to generate the query. This will in turn happen if the document contains the query words often. It is a vicious circle definition that makes clear exploitation of the logic.

4 Procedure of implementation

- Three different cores were created
- The similiarity functions for each of the cores were defined
- The dataset was posted to each of the Solr core
- MAP value was calculated to evaluate the IR model
- Parameters were tuned and the MAP value was calculated repeatedly to observe the change in the value of MAP

5 Solr configuration

The similarity class was configured globally for all field types by using the class **solr.SchemaSimilarityFactory** which accepts a parameter **defaultSimFromFieldType** specifying the default field on which the similarity function is applied. The configuration is a shown as below.

5.1 BM25

The similarity class used for BM25 is solr.BM25SimilarityFactory. The formula used for BM was already provided.

```
<similarity class="solr.SchemaSimilarityFactory">
  <str name="defaultSimFromFieldType">text_en</str>
</similarity>
```

```
<similarity class="solr.BM25SimilarityFactory">
  <str name="b">0.5</str>
  <str name="k1">1.2</str>
</similarity>
```

The maximum value was obtained when the k1 parameter was set to 1.2 and b was set to 0.5. The MAP value obtained was 0.7095.

k1	b	MAP
0	0	0.6969
0.5	0.5	0.7001
1	0.7	0.7002
1	0.7	0.7002
1.2	0.5	0.7095

It was observed that the relationship between the k1, b and the MAP score is not a linear relation.

5.2 DFR

For DFR, as per the project requirements the parameters were set to “BasicModelG” plus “Bernoulli” first normalization plus “H2” second normalization.

```
<similarity class="solr.SchemaSimilarityFactory">
  <str name="defaultSimFromFieldType">text_dfr</str>
</similarity>
```

```
<similarity class="solr.DFRSimilarityFactory">
  <str name="c">7.0</str>
  <str name="normalization">H2</str>
  <str name="afterEffect">B</str>
  <str name="basicModel">G</str>
</similarity>
```

The maximum MAP value obtained was 0.7022.

5.3 LM

The language model that was used here is Dirichlet smoothing.

```
<similarity class="solr.LMDirichletSimilarityFactory">
  <str name="defaultSimFromFieldType">text_lm</str>
</similarity>
```

The MAP value achieved was 0.7056.

6 Improving MAP score

The MAP score improvement was done by two procedures:

- Query boosting was implemented. Though this improved the performance, it was not very evident.
- Parameter tuning. Various parameters of different IR model typed were varied.

7 Query parser

The edismax query parser has been implemented here. Various parameters experimented are query fields (qf) and phrase fields (pf), query size (qs), phrase size (ps), pair of words (pf2) and word triplets (pf3).

```
<str name="defType">edismax</str>
<str name="qf">text_en^2 text_ru^2 text_de^2</str>
<int name="qs">10</int>
<str name="pf">text_en^2 text_ru^2 text_de^2</str>
<int name="ps">10</int>
<str name="pf2">text_en^10 text_ru^10 text_de^10</str>
<int name="ps2">5</int>
<str name="pf3">text_en^10 text_ru^10 text_de^10</str>
<int name="ps3">10</int>
```

8 Outputs

The following were the MAP values obtained for different model types.

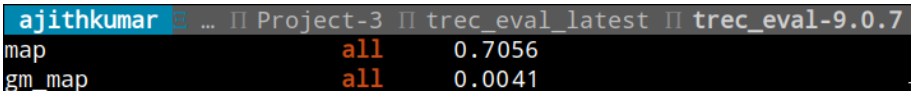
8.1 BM25

```
ajithkumar … Project-3 trec_eval_latest trec_eval-9.0.7
map                     all     0.7095
gm_map                  all     0.0063
```

8.2 DFR

```
ajithkumar … Project-3 trec_eval_latest trec_eval-9.0.7
map                     all     0.7022
gm_map                  all     0.0076
```

8.3 LM



References

[1] Introduction to information retrieval by Christopher D. Manning, Hinrich Schütze, and Prabhakar Raghavan

[2] Stack Overflow

[3] Solr documentation