Analysis Comparison of BM11 with BM25

Alda G. M. Lumban Gaol[1 [12S18028]], Sarah H. M. Siahaan[2 [12S18032]], and Roy Gunawan Napitupulu[3 [12S18043]]

Information System, Del Institute of Technology, Indonesia
{iss18028, iss18032, iss18043}@students.del.ac.id

Abstract. Information Retrieval (IR) is a discipline that deals with the storage, organization and access of information. Today IR is a popular and ubiquitous field, so it will have a positive impact if improvements or research is carried out in this area. Based on previous research, it is stated that in the document in English, the BM25 model is superior to the BM11 model. The BM25 and BM11 models which are part of the IR approach are BIM (Binary Independence Model) which functions to determine the relevance value of a searched document based on binary weighting that is adjusted to the inputted query. This research will prove whether it is true that the BM25 model is superior to the BM11 model. The research will begin with a study of literature in scientific journals and articles according to the research case, then conduct experiments with the BM11 and BM25 models on the same two datasets, then analyze the experimental results for each model.

Keywords: Information Retrieval (IR) · BM11 · BM25.

1 Introduction

Information Retrieval (IR) is a discipline that deals with the storage, organization and access of information [1].Today, IR is one of the most popular and ubiquitous fields, so improvements in this area will have a positive impact [6]. IR is needed because it can help users obtain information that is relevant to their wishes from a large and unstructured collection of information sources [3].For example, we want to consider the need for information to find out if eating chocolate is beneficial in lowering blood pressure. We might reveal this through a search engine query: "chocolate effect pressure"; however we will evaluate the resulting document as relevant if it meets our information needs. Not only because it contains all the words in the query but it must be relevant to the information we need [1].

The Binary Independence Model (BIM) method serves to determine the relevance value of a searched document based on binary weighting that is adjusted to the inputted query [8]. BIM introduces two main assumptions which further simplify the calculation of P(D l Q, r). The assumptions are: 1) the assumption of independence, the terms in the document (and queries) are independent, the probability of the occurrence of one term in the relevant document does not

affect the possibility of the occurrence of other terms in the relevant document; and 2) only the query term determines the relevance of the document. BIM has three Extensions namely Two-Poisson model, BM11 model and BM25 model [1]. However, this study will only focus on the comparison between the BM11 and BM25 models. BM11 is a model that corrects for the weight scaling factor of the Two Poisson model to account for different document lengths and BM25 is a model that controls the amount of correction for document length with an additional parameter b. BM11 Formulas:

$$rel(D,Q) = \sum_{t \epsilon Q} (\frac{f_{t,D}\,(k+1)}{f_{t,D} + k\frac{l_d}{l_{avg}}}).w_t \tag{1}$$

BM25 Formulas:

$$rel(D,Q) = \sum_{t \epsilon Q} (\frac{f_{t,D}\,(k+1)}{f_{t,D} + k\frac{l_d}{l_{avg}} b + k(1-b)}).w_t \tag{2}$$

Where:

- l average : Average length of documents in the collection
- l Document : The length of the document D

Based on previous research, it is stated that in English documents, the BM25 model is superior to the BM11 model [10]. This research will prove whether it is true that the BM25 model is superior to the BM11 model. The research will begin with a study of literature in scientific journals and articles according to the research case, then conduct experiments with the BM11 and BM25 models on the same two datasets, then analyze the experimental results for each model.

2 Literature Review

In this study, there are two studies that serve as material for reviewing information on the comparative analysis of BM11 with BM25.

Okapi BM11 is one of the probabilistic retrieval methods introduced by Robertson and Walker and will be used in this study [9]. Previous research has been carried out to improve the performance of the conventional TF*IDF method, by using a 2-stage strategy in the retrieval process for Chinese documents [2]. The first stage uses the Okapi BM11 ranking algorithm to obtain the most relevant documents, and passage based is used in the second stage to remove irrelevant documents that have been retrieved from the first stage. From the research, a good approach was introduced in improving the conventional TF*IDF method, which, although simple, can be proven.

Robert W. P. Luk et al. conducted a study that combines the 2-poisson model, hybrid term indexing, and pseudo relevance feedback using the NTCIR-III data corpus [4]. In the 2-poisson model there are several methods, one of

[1] https://www.uni-mannheim.de/

which is the Okapi BM11. From the research conducted, it is proven that the combination of the above methods increases the mean average precision.

In 1994, S. E. Robertson and S. Walker conducted a study on the 2-Poisson Model for Probabilistic Weighted Retrieval [9]. They examined the effect of the variables k1, k2, and k3 on the results of document retrieval taken from TREC-1 and TREC-2. From the results of the experiments carried out, it is concluded that the value of k1 is 2, k2 is 1, and it will produce the best performance for the retrieval system with the Okapi BM11 and Okapi BM15 methods.

Okapi BM25 also one of the probabilistic retrieval, is a ranking system in sorting the most suitable documents based on a query. BM25 has the best formula in the best match class. The BM25 or Okapi BM25 is more effective and has higher accuracy in sorting documents based on the inputted query. In the BM25 equation, the values of k1, k2, and the value of b are parameters or constant values [1,2,3] [7].

$$BM25 = \sum_{t.Q} log10(\frac{N - nt}{nt}) - (\frac{(k1+1)fd,t}{K + fd,t}) - (\frac{k3+1)fq,t}{k3 + fq,t}) \quad (3)$$

Where:

- Q : Input user / query
- N : Number of sentences in the document
- nt : Number of term that containing the query
- fd,t: Number of term frequency
- fq,t: Number of query frequency

$$K = k1((1-b)\frac{b.dl_d}{avl}) \quad (4)$$

Where:

- dld : Number of sentences in the document
- avl : Average document length
- k1 : 1,2
- b : 0.75
- k3 : 1000

However, the BM25 score is weak for the occurrence of query terms in very long documents, and thus those very long documents can be overly penalized [5].

3 Analysis

At this stage an analysis of the BM11 and BM25 methods is carried out based on previous research and how to design and implement experiments.

subsectionResearch Stage Design The research stage design carried out in the research is as follows:

epsfig epstopdf

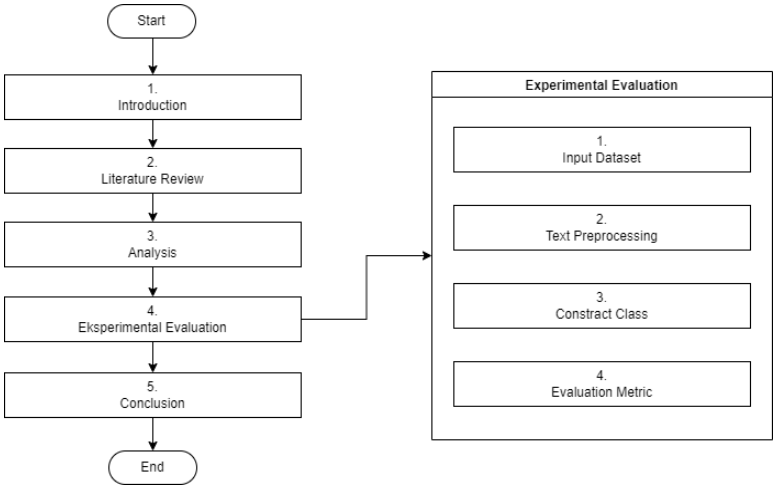

Fig. 1: Design Implementation Experiment

1. Introduction, explaining what will be the research topic. Contains an introduction to IR, BM11, BM25 and what is the problem in the research.
2. Literature Review,carried out by collecting references to be used in research in the form of scientific journals/articles, books, scientific conferences, and data sourced from the internet. Aims to obtain and collect information relevant to the research topic that will be used as material for analysis in the next stage.
3. Analysis, analyzing the BM11 and BM25 methods based on the results of previous research.
4. Experimental Evaluation, discuss the comparison of BM11 with BM25 by conducting comprehensive experiments on two real datasets: Song Lyrics and Friends Dialog.
5. Conclusion, convey the conclusions of the research results

3.1 Analysis of BM11 and BM25 Methods Based on Previous Research

Okapi BM11 and BM25 are probabilistic retrieval methods precisely part of the Two-Poisson model. BM11 is a model that corrects for the weight scaling factor of the Two Poisson model to account for different document lengths and BM25 is a model that controls the amount of correction for document length with an additional parameter b. BM25 is considered to produce better performance with higher accuracy in sorting documents based on the entered query compared to BM11. When viewed from the formula for the BM11 and BM25 methods, what makes the two methods different is that BM25 has an additional parameter in the formula, namely the value of b. The value of b is a parameter or constant value.

4 Experimental Evaluation

In this section, we discuss the comparison of BM11 with BM25 by conducting comprehensive experiments on two real datasets: Song Lyrics and Friends Dialog.

4.1 Experimental Setup

We conduct experiments over two publicly available datasets, Song Lyrics and Friends Dialog. Song Lyric dataset is a public dataset by Deep Shah and consist of song lyrics of various artists[2]. Specifically, the dataset that's going to be used for this research is Taylor Swift lyrics. Friends Dialog is a dataset in .csv format that contains text files of all scenarios as well as dialogue for each episode on FRIENDS TV Show[3]. The details of these datasets are described in Table 1 and Table 2.

Table 1: Song Lyrics Dataset

No	Attribute	Description
1	Artist	The singer name
2	Title	The song title
3	Album	Album name of the song
4	Year	Song's release year
5	Date	Song's release date
6	Lyric	Lyric of the song

Table 2: Friends Dialog Dataset

No	Attribute	Description
1	ID	ID of document
2	Episode	Episode of document
3	Dialogue	Dialogue of episode in document

Table 3 show the detail of the length of attribute from dataset that is going to be used on experiment.

[2] https://www.kaggle.com/datasets/deepshah16/song-lyrics-dataset
[3] https://github.com/abishek21/F.R.I.E.N.D.S-

Table 3: Dataset Descriptions

No	Dataset Name	Attribute	Document Average Length
1	Song Lyrics	Lyric	171.754
2	Friends Dialog	Dialogue	1434.091

We evaluate each result of two datasets by calculating the score of the predicted rating data. We define score as a ranking function used by search engines to estimate the relevance of documents to a given search query and is presented with stopword removal.

4.2 Experiment

We compared BM11 with BM25 using two datasets namely: Song Lyrics and Friends datasets. In conducting the experiment, preprocessing is carried out on the dataset that will be used. The preprocessing consists of removing nul data, removing number and whitespace, tokenization, stopwords removal, and punctuation.

We first compared the BM11 and BM25 scores. We observe that BM25 is superior to BM11 where the b value in the BM25 formula is 0.75 and k is 1.2. The comparison score table can be seen in the following table:

Table 4: BM11 on Dataset Song Lyrics

No	Score	Document
0	4.341	vintage tee brand new phone high heels on cobb...
1	1.296	justin vemon can see you standing honey with...
2	2.245	we could leave the christmas light up til jan...
3	1.301	doing good on some new shit been saying yes...
4	2.33	don like your little games don like your till...
5	1.344	betty won make assumptions about why you swit...
6	0.706	taylor swift future wanna be your end game w...
7	0.0	taylor swift promise that youre taki...
8	0.0	you are somebody taht don know but youre taki...
9	2.752	sait air and the rust on your door never need...

Table 5: BM25 on Dataset Song Lyrics

No	Score	Document
0	4.365	vintage tee brand new phone high heels on cobb...
1	1.332	justin vemon can see you standing honey with...
2	2.179	we could leave the christmas light up til jan...
3	1.298	doing good on some new shit been saying yes...
4	2.408	don like your little games don like your till...
5	1.355	betty won make assumptions about why you swit...
6	0.797	taylor swift future wanna be your end game w...
7	0	taylor swift promise that youre taki...
8	0	you are somebody taht don know but youre taki...
9	2.771	sait air and the rust on your door never need...

Table 6: BM11 on Dataset Friends Dialog

No	Score	Document
0	0.536	there s nothing to tell it s just some guy i w...
1	0.693	you guy don t understand for us kissing is as...
2	0.523	hi guys hey pheebs oh oh how d it go um not so...
3	0.632	all right phoebe if i were omnipotent for a da...
4	0.007	let it go it s not a big deal not big deal i...
5	0.549	oh look there s joey picture this is so exci...
6	1.852	hey georgeous how s it going dehydrated japanes...
7	1.852	hey georgeous how s it going dehydrated japanes...
8	0.408	terry i know i haven t worked here long bu...
9	0.391	guy there s somebody i d like you to meet wai...

Table 7: BM25 on Dataset Friends Dialog

No	Score	Document
0	0.533	there s nothing to tell it s just some guy i w...
1	0.688	you guy don t understand for us kissing is as...
2	0.523	hi guys hey pheebs oh oh how d it go um not so...
3	0.635	all right phoebe if i were omnipotent for a da...
4	0.007	let it go it s not a big deal not big deal i...
5	0.543	oh look there s joey picture this is so exci...
6	1.806	hey georgeous how s it going dehydrated japanes...
7	1.806	hey georgeous how s it going dehydrated japanes...
8	0.402	terry i know i haven t worked here long bu...
9	0.389	guy there s somebody i d like you to meet wai...

Score for Document Song Lyrics shown in the following table:

Table 8: Score on Song Lyric Dataset

Rank	Doc No	BM11	BM25
1	0	4.365	4.341
2	9	2.771	2.752
3	4	2.408	2.33
4	2	2.179	2.245
5	5	1.355	1.344
6	1	1.332	1.301
7	3	1.298	1.296
8	6	0.797	0.706
9	7	0.0	0.0
10	8	0.0	0.0

Score for Document Friends shown in the following table:

Table 9: Score on Friends Dialog Dataset

Rank	Doc No	BM25	BM11
1	6	1.806	1.852
2	7	1.806	1.852
3	1	0.688	0.693
4	3	0.635	0.632
5	5	0.543	0.549
6	0	0.533	0.536
7	2	0.523	0.523
8	8	0.402	0.408
9	9	0.389	0.391
10	4	0.007	0.007

4.3 Evaluation

Evaluation Result Based on the experiments from our research, the comparison between BM11 and BM25 shows that BM25 is not always superior to BM11. This is based on our experimental results which show that the BM11 score is higher than the BM25 score. This score indicates the relevance between the query and the document in the experiment.

This also refers to the literature review that we have done. Where it is stated that the BM25 score is weak for the occurrence of query terms in very long documents, and thus those very long documents can be overly penalized. Incidentally, the experiment we did was using long documents. Thus, BM25 is not really effective on long documents.

5 Conclusion

The BM25 and BM11 models which are part of the IR approach are BIM (Binary Independence Model) which serves to determine the relevance value of a searched document based on binary weighting that is adjusted to the inputted query. Previous research stated that the BM25 model is superior to the BM11 model and this study has proven that this statement is not always true.

Also reviewed based on the results of experiments conducted by researchers showed that the BM11 score was higher than the BM25 score. The experiment conducted by the researcher used long documents, referring to another literature review which stated that the BM25 score is weak for the occurrence of query terms in very long documents, and thus those very long documents can be overly penalized.

Thus, the researchers concluded that the BM25 is not always superior to the BM11. The BM25 is superior to the BM11 only on short documents.

References

1. Ceri, S., Bozzon, A., Brambilla, M., Valle, E.D., Fraternali, P., Quarteroni, S.: An introduction to information retrieval. In: Web information retrieval, pp. 3–11. Springer (2013)
2. yuan Chi, S., li Hsiao, C., feng Chien, L.: A practical passage-based approach for chinese document retrieval
3. Dozan, D.: Pengembangan sistem information retrieval untuk bahasa indonesia berbasis web menggunakan vector space model. Undergraduate thesis. Universitas Kristen Duta Wacana (2019)
4. Luk, R.W., et al.: Hybrid term indexing: an evaluation (2001)
5. Lv, Y., Zhai, C.: When documents are very long, bm25 fails! In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1103–1104. SIGIR '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/2009916.2010070, https://doi.org/10.1145/2009916.2010070
6. McDonell, R., Zobel, J., von Billerbeck, B.: Informativeness in information retrieval: Statistical dispersion as a measure of term specificity (2015)
7. Pardede, J., Husada, M.G., Riansyah, R.: Implementasi dan perbandingan metode okapi bm25 dan plsa pada aplikasi information retrieval (2018)
8. Purnama, F.A.: SISTEM TEMU KEMBALI INFORMASI DENGAN MENERAPKAN METODE PROBABILISTIK BINARY INDEPENDENCE MODEL (BIM). Ph.D. thesis, UNIVERSITAS ISLAM NEGERI SULTAN SYARIEF KASIM RIAU (2012)
9. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: SIGIR'94. pp. 232–241. Springer (1994)
10. Tseng, Y.H., Tsai, Y.C., Lin, C.J.: Comparison of global term expansion methods for text retrieval. In: NTCIR. Citeseer (2005)

Laporan Perbaikan Presentasi

Progress Report Proyek STBI - 01

- 12S18028 - Alda G M Lumban Gaol
- 12S18032 - Sarah H M Siahaan
- 12S18043 - Roy Gunawan

Berikut ini adalah laporan perbaikan proyek yang dikerjakan setelah dilakukannya presentasi. Mencakup:

Table 10: Perbaikan Proyek

NO	Laporan Perbaikan	Status Perbaikan
1	Penyajian hasil dalam bentuk barchart	Belum Diperbaiki
2	BM11 dan BM25 tidak bisa dicompare jika hanya dengan melalui score saja, rangking dokumen harus disesuaikan lagi terhadap query yang digunakan	Belum Diperbaiki
3	Menambahkan penjelasan query yang digunakan harus ditambahkan ke bab 4	Sudah Diperbaiki
4	Menyajikan length average doc menjadi 3 decimal	Sudah Diperbaiki