"Haberim Var" – Turkish News Retrieval System

Erol ÖZKAN
Hacettepe University
Computer Engineering Department
Ankara, Turkey
erolozkan@outlook.com

***Abstract*—Thanks to the development of internet technologies, we are provided with a great amount of information. However, since this information is so vast, the inability to process this huge amount of data becomes more and more clear. Online news retrieval also suffers from these problems. The data in this domain is unstructured and there is huge amount of data. In this paper, we design and implement a Turkish news retrieval system called "Haberim Var". The goal of this system is to accurately retrieve the latest Turkish news articles. We build our own news dataset using well known Turkish news sources. We implement text acquisition, index creation, query-engine, user interface and evaluation components. In order to evaluate our system, we use MRR metric. Since this dataset doesn't contain evaluation data, article titles are used as pseudo-queries. Results show that our system gives satisfying results in terms of both effectiveness and efficiency. Using "Haberim Var", users are able to find the Turkish news articles they are looking for.**

***Index Terms*—Information Retrieval, Turkish News Dataset, Text Acquisition, Index Creation, Query-Engine, User Interface, Evaluation, MRR...**

I. Introduction

As information on web increases, it is getting more difficult for users to find the information they are looking for. In this paper, we design and implement a Turkish news retrieval system called "Haberim Var". The goal of this project is to accurately retrieve the latest news articles from well-known Turkish news sources. We believe many existing retrieval systems already use these techniques [1]. However, there has been little research on retrieval of Turkish news. Turkish language possesses specific challenges. Furthermore, news retrieval is still far from a solved problem and there is still room for improvements [2].

There are some applications such as Bundle [3], MSN [4], Google News [5] and Yandex News [6] that provide their users with some kind news information. However; Bundle and MSN don't provides any search functionality. Yandex News is not available in Turkish. Finally, although, Google News provides specific news search capabilities, it is not national and it is not totally accurate for Turkish.

We aim to retrieve the latest Turkish news articles. To achieve this, we design and implement several components including; text acquisition, index creation, query-engine, user interface and evaluation. These components are utilized at different stages of our retrieval process.

- **Text acquisition** identifies and downloads the latest news articles from the news sources,
- **Index creation** transform documents into searchable index terms.
- **Query-engine** generates a ranked list of documents, given a single query.
- **User interaction** provides simple and elegant interfaces for users.
- **Evaluation** measures the system performance.

We implement Bag of Words and BM25 retrieval models. We use MRR metric to evaluate our system. Since our dataset doesn't contain evaluation data, we use article titles are as pseudo queries. Results show that BM25 gives better results in terms of both effectiveness and efficiency than Bag of Words approach.

The remainder of the paper is organized as follows. In Section II, we explain the problem definition. In Section III, we survey the similar retrieval systems. Our architecture is detailed in Section IV. In Section V, we give details about our dataset. In Section VI, we explain the evaluation strategy and we present our results. Finally, we conclude our work in Section VIII.

II. PROBLEM DEFINITION

In modern information retrieval systems, in order to retrieve the documents, a relevance score is computed according to a statistical procedure. Then, top ranking documents are retrieved and presented to user.

There has been many different approaches for retrieval of the documents. One approach has been to calculate a similarity score between the query and the documents. The BM25 weighting model is assumed to be one of the most successful of this approach. In this retrieval model, the relevance score is computed according to the following equation.

$$s_t(q,d) = w_q \frac{(k_1+1)tf}{k_1\left((1-b)+b\frac{dl}{avg\_dl}\right)+tf} w_{IDF}$$

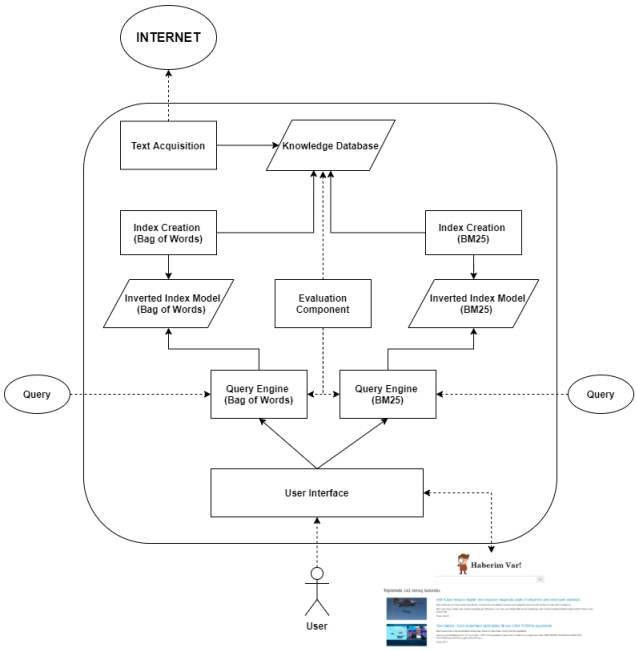
In this equation, the aim is to maximize the $s_t(q,d)$ function which represents the term-document similarity. q is the query and d is the document. tf and w_{IDF} are the document term frequency and the IDF factor. dl and avg_{dl} are the document length and average document length. w_q is query-only weight. Finally, b and k1 are the parameters.

III. RELATED WORK

Information Retrieval (IR) applied to news has been a popular research area. There have been many different approaches [1]. Catena et al. [7] introduce a variant of the well-known BM25 model, called BM25 Passage (BM25P). BM25P is specifically build to improve the effectiveness of a retrieval system in specifically news domain. They point out that most relevant terms are primarily concentrated at the beginning and at the end of news articles. Using this idea, they calculate the relevance score according to linear combination of term frequencies of each passage in the articles. They show that their method gives better results than BM25 in term of both NDCG and MRR.

IV. ARCHITECTURE

In this section we describe the architecture of our system. The architecture of “Haberim Var” is shown in Figure 1.



“Haberim Var” is composed of six main components. These components are text acquisition, knowledge database, index creation, query engine, user interaction and evaluation. Furthermore, we implement two different retrieval models. These models are Bag of Words approach and BM25 weighting model.

A. Text Acquisition

The text acquisition component of “Haberim Var” is responsible for identifying and downloading the latest news sources from the news sources. It crawls the web and downloads the latest news articles. It builds the knowledge database that contains the news articles and their metadata. Text acquisition component should be executed on regular basis to update the knowledge database. Otherwise news collection of “Haberim Var” may become stale. Furthermore, text acquisition component is composed of two sub-components. These are crawler and converter.

1) Crawler: Crawler is responsible for identifying the news articles on web. Given a news source URL, it follows the paths and discovers the news articles. If, it classifies these sources as news articles, it saves them into knowledge database.

2) Converter: "Haberim Var" require documents to be in the form of plain text. However, the documents found by the crawler is usually in the form of HTML. Hence, converter is responsible for converting HTML pages into plain text format.

B. Knowledge Database

The knowledge database is the component where all documents and their metadata are stored. In "Haberim Var", documents and their metadata are stored in the disk as JSON files.

C. Index Creation

Given the large number of documents, index creation component is responsible for transforming documents into searchable index terms. In this process, statistical information about the words within the documents are extracted. This information is later used by the query-engine component to rank the documents. The output of index creation component is a searchable file that is stored on disk. Since, large numbers of news articles are stored and indexed in "Haberim Var", index creation component needs to be both effective and efficient.

Index creation component is composed of four different sub-components. These are parser, stopping, stemming and inversion.

1) Parser: Parser is responsible for creating a sequence of tokens from the news articles. In "Haberim Var" tokens correspond to single words.

2) Stopping: The stopping is responsible for removing common words from the text. The words contained in all documents do not give much information about topics covered in the news articles. Hence, removing them usually has no or little effect. In "Haberim Var", most common Turkish words are removed in order to reduce the size of our indexes.

3) Stemming: Stemming is responsible for grouping words that are derived from the same stem. For example; “araba”, “arabalar”, and “arabam” are all converted into same stem "araba". By replacing these words with their stem, The likelihood of matching queries with their relevant documents is increased.

4) Inversion: The inversion is responsible for changing document-term information into term-document information. This type of conversation is required for system to be efficient.

D. Query-Engine

The query-engine component is responsible for generating a ranked list of documents. Given the query, it returns the list of relevant news articles and their scores. This process is also called ranking. It must be both effective and efficient. Similar to index creation; parser, stopping and stemming subcomponents are executed to make the document terms and query terms in the same format.

E. User Interaction

The user interaction component provides simple and elegant user interface to users to type their query and see the top ranking documents. It accepts a query from the user, and returns the top ranked news articles from the query-engine.

User interaction is composed of two sub-components. These are the query-interface and the results-interface.

1) The Query-Interface: The query-interface provides an interface for users to type their query. This interface is shown in the Figure 2.

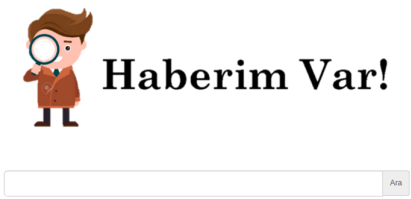


Fig. 1. The Query-Interface

Users type their query in the input field. When they click the search button, the query is sent to the query-engine and the results-interface is shown.

2) The Results-Interface: The results-interface is responsible for displaying the retrieved documents received from the query-engine. This interface is shown in the Figure 3.

Fig. 2. The Results-Interface

Top ranked news articles, their images and their scores are presented to the user. Also, some snippets are generated in order to give more information about the articles.

V. DATASET

There are many English news retrieval datasets such as; Aquaint [8], Signal [9] and RCV1 [10]. Aquaint consists of newswire text data in English, drawn from three sources: the Xinhua News Service, the New York Times News Service, and the Associated Press Worldstream News Service. Signal is released to facilitate conducting research on news articles. It is intended to serve the community for research on news retrieval in general. Finally, RCV1 is Reuters's news dataset and it contains news stories for use in research and development of in natural language processing, information retrieval, and machine learning systems.

Although there are several English news retrieval datasets, there aren't many Turkish news retrieval datasets. For this reason, we build our own Turkish news dataset using several Turkish news sources. The statistics about our Turkish news dataset are shown in the Table 1.

TABLE I
STATISTICS ABOUT TURKISH NEWS DATASET

Source	Number of Documents	Average Query Len.	Average Article Length
Haberturk	165	7.93	382.18
Milliyet	453	7.06	431.66
Hurriyet	523	6.99	353.42
Ntv	57	7.78	385.84
Sputniknews (tr)	593	9.13	269.32
Cnnturk	280	6.55	298.53
TOTAL	**2071**	**7.57**	**353.49**

We download around 2.000 articles from six different Turkish news sources. The average query (title) length and the average article length of our dataset are also presented in the table.

VI. EVALUATION

A retrieval system's performance can be measured in terms of effectiveness and efficiency [1]. A system is effective if it is able to retrieve the most relevant set of documents. And, a system is efficient if documents are retrieved as quickly as possible. We evaluate our system in terms of both effectiveness and efficiency.

A. Effectiveness

We use MRR metric to evaluate our system in terms of effectiveness. The Reciprocal Rank (RR) in information retrieval calculates the reciprocal rank at which the first relevant document is retrieved [11]. RR is 1 if the first relevant document is retrieved at first rank. RR is 0.5 if the first relevant document is retrieved at the second rank, and so on. When averaged across all of the queries, this measure is called the Mean Reciprocal Rank (MRR). MRR score is computed according to the following equation.

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i}$$

In this equation, $\mathrm{rank_i}$ refers to the rank position of the first relevant document. And Q refers to the queries. The mean of RR scores is calculated.

Since our dataset do not provide any evaluation data, we use same methodology used by Catena et al. [7]. We use article titles as queries. We assume that, there is only one relevant news article for each query and it is the article the title belongs to. All other articles are considered to be non-relevant. Effectiveness scores of our Bag of Words and BM25 approaches are shown in the Table 2.

TABLE II
EFFECTIVENESS SCORES FOR BAG OF WORDS AND BM25 METHODS

Source	Bag of Words	BM25
Haberturk	0.34	0.61
Milliyet	0.46	0.66
Hurriyet	0.42	0.65
Ntv	0.54	0.74
Sputniknews (tr)	0.52	0.87
Cnnturk	0.41	0.63
TOTAL	**0.45**	**0.71**

The table shows that BM25 weighting model gives almost 2 times better results than Bag of Words approach. Also, it is clear that scores are similar for different news sources.

B. Efficiency

The efficiency of a retrieval system can be measured by monitoring the system performance. A variety of measurements are used, such as response time and throughput [1]. We evaluate the efficiency of our system by measuring the index creation and the retrieval times. Measured times of "Haberim Var" are shown in the Table 3.

TABLE III
EFFICIENCY SCORES FOR BAG OF WORDS AND BM25 METHODS

Process	Bag of Words	BM25
Index Creation	10 min. 35 sec.	0 min. 10 sec.
Retrieval	17 min. 39 sec.	01 min. 33 sec.

Retrieval times are measured using every article as a query. Table shows that BM25 weighting model is way more efficient than Bag of Words approach in terms of both index creation and retrieval. Bag of Words approach takes too much time for both of these processes.

VII. CONCLUSION

There is an excessive amount text on web. And people, somehow, need to find the information they are looking for. This raises the question of how to best retrieve the best news document set for a given query. Many existing methods are developed to solve this problem. In this paper, we design and implement a Turkish news retrieval system. We aim to accurately retrieve the latest Turkish news articles. We build our own news dataset and implement text acquisition, index creation, query-engine, user interface and evaluation components. In order to evaluate our system, we use MRR metric and use article titles as pseudo-queries. Results show that BM25 weighting model gives better results than Bag of Words approach. Using "Haberim Var", users are able to find the latest Turkish news articles they are looking for.

REFERENCES

[1] B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*, 1st ed. USA: Addison-Wesley Publishing Company, 2009.

[2] "Newsir16," Available at: https://research.signal-ai.com/newsir16/index.html, Last visit: 27/05/2020.

[3] "Bundle haber," Available at: https://www.bundlehaber.com/, Last visit: 27/05/2020.

[4] "Msn," Available at: https://www.msn.com/tr-tr/, Last visit: 27/05/2020.

[5] "Google news," Available at: https://news.google.com/, Last visit: 27/05/2020.

[6] "Yandex news," Available at: https://yandex.ru/news/, Last visit: 27/05/2020.

[7] M. Catena, O. Frieder, C. I. Muntean, F. M. Nardini, R. Perego, and N. Tonellotto, "Enhanced news retrieval: Passages lead the way!" in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1269–1272. [Online]. Available: https://doi.org/10.1145/3331184.3331373

[8] "Aquaint dataset," Available at: https://catalog.ldc.upenn.edu/LDC2002T31, Last visit: 27/05/2020.

[9] "Signal dataset," Available at: https://research.signal-ai.com/newsir16/signal-dataset.html, Last visit: 27/05/2020.

[10] "Rcv1 dataset," Available at: https://trec.nist.gov/data/reuters/reuters.html, Last visit: 27/05/2020.

[11] N. Craswell, *Mean Reciprocal Rank*. Boston, MA: Springer US, 2009, pp. 1703–1703. [Online]. Available: https://doi.org/10.1007/978-0-387-39940-9$_4$88