Search Halal Products Using BM25F and the Analytic Hierarchy Process

Nur Aini Rakhmawati[#1], Adnan Mauludin Fajriyadi[#2]

#*Departement of Information Systems, Institut Teknologi Sepuluh Nopember*
Jalan Raya ITS, Keputih, Sukolilo, Kota Surabaya, IndonesiaAut
[1]nur.aini@is.its.ac.id
[2]adnanfajri@gmail.com

Abstrak— **In Indonesia, the Institute of Food and Drug Administration (LPPOM) MUI is the official institution that provides information about halal products. However, the lack of information is provided on the website. Halal Nutrition Food is an application that has a function to facilitate the search for halal products that are incorporated in an integrated database in the form of Linked Open Data. To improve the searching features, we exploit BM25F. BM25F can process structured documents such as instances in RDF. The BM25F return the answer based on four fields: product name, manufacturer name, product ID and ingredient name. The weight of each field is calculated using the Analytic Hierarchy Process (AHP) where the product name gets the highest weight value. The number of keywords and the keywords occurrences influence the score of query results.**

Kata kunci— **Halal, BM25F, Linked Open Data, AHP**

I. INTRODUCTION

Indonesia is one of the largest Muslim-majority countries in the world. Percentage of Indonesian Muslims reach up to 12.7% of the world's population. Of the 205 million Indonesians, it is reported that at least 88.1 % are Muslims[1]. In Islam, a Muslim must consider a halal status of food and beverages consumed, therefore the Indonesian government established an institution to examine the status of halal food product in Indonesia. The Indonesian Institute of Food Supervision, Drugs and Cosmetic of the Indonesian Ulama Council (LPPOM MUI) is an institution under the ministry of religious affairs. The institution issues halal certificates for food, products and cosmetics. Apart from releasing halal certificates, LPPOM MUI is also responsible for providing halal products information both through the offline and online channel. Currently, LPPOM MUI provides halal product information through a website that can be accessed a website at http://halalmui.org. Users can search for a halal certificate of a product or company on the website. Furthermore, they can download the list of halal certified products in Portable Document Format (PDF). The lack of information is provided with both website and PDF file

since they only present the certificate number, the manufacturer, and the validity period of the halal product.

Therefore, Fatawi and Rakhmawati developed the Halal Nutrition Food where people can search for halal products that have been certified by the MUI[2]. The previous application exploits Linked Data[3] technology such as RDF[4] as a standard for data format and SPARQL[5] for searching any information about halal products and ingredients. In this work, we enhance the search feature in the application by using Okapi BM25F[6]. BM25F is an extension algorithm of BM25[7] which is intended for structured processing documents, such as RDF format. The BM25 algorithm lists the documents based on query terms that appear in each document collection. In this way, a user can retrieve not only a product, but he can also obtain related manufacture and ingredients based on his keywords. The contributions of this work can be listed as follows:

1. Propose BM25F for searching a halal product, manufacture and ingredients. Note that, our system does not rank a level of halal of a product
2. Develop a searching architecture for halal domain
3. Assign a list of weights for each field for searching halal stuff.

The remainder of this paper are explained as follows: Section 2 describes the related works in BM25 implementation, Section 3 explains the System architecture; Section 4 describes our methodology; Section 5 contains a discussion of the results obtained along with theoretical and practical implications, and Section 5 concludes our work.

II. RELATED WORKS

BM25 is one of the notable ranking methods in information retrieval which based on probabilistic retrieval framework [7]. The method has been implemented in multiple domains such as Indonesian Text Search Engines [8], scientific ranking [9], and mutant recovery search [10]. Regarding product search, BM25 is

used for searching a product deal [11], finding related products on Twitter [12] and E-commerce [13][14].

BM25F is also considered in semantic web data which data is stored in RDF format. The BM25F was initially implemented in the semantic web by Perez-Aguera et al [15]. After that, several similar works attempted to improve the performance of BM25F on semantic search [16][17][18]. In this paper, we exploit BM25F for searching related halal products.

Analytic Hierarchical Process (AHP)[19] was used for ranking items in multiple domains. The ranking of local governments for solid water treatment [19] and open data portal [23] is based on AHP. The performance of a set of the website in a certain domain [21] and media tourism websites [22] are compared by using AHP.

III. SYSTEM DESIGN

A. System Architecture

The web application is built with Laravel (https://laravel.com/) where MySQL (https://www.mysql.com/) as the data storage. Although the application is based on RDF, MySQL is required for storing data from an user. Product name, product code, manufacturer name and ingredient name are indexed in Apache Solr (http://lucene.apache.org/solr/). The Apache Solr receives a set of keywords as an incoming query and then returns the answer to the Web application. The detail of our system architecture is displayed in Figure 1.
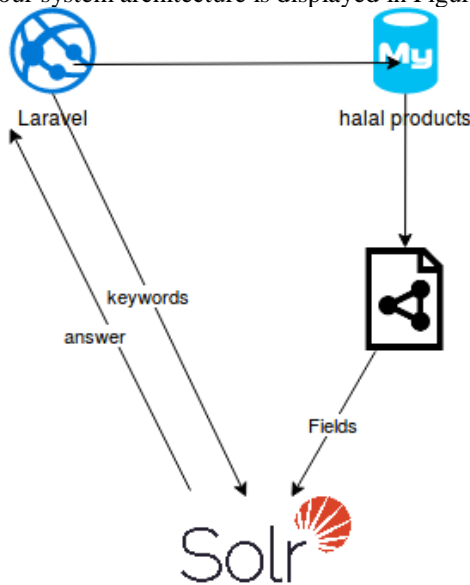

Figure 1. System architecture

B. BM25F

OKAPI (Online Keyword Access to Public Information) BM25 is a search results weighting model developed by City University London based on a basic probabilistic model that ranks documents in descending order of relevance for a document to the information needed. Okapi BM25 ranks documents based on probability and uses term frequency to rank similarity. The ranking formula on BM25F based on Pérez-Iglesias [15] is as follows:

$$R(q,d) = \sum_{t \in q} idf(t) \frac{weight(t,d)}{k_1 + weight(t,d)}$$

Where:
- R (q, d) is a query ranking q in document d (value BM25F).
- t is the term obtained from the query.
- d is a document, in RDF document means instance
- k1 is a constant that can regulate the growth of non-linear term frequencies.
- idf (t) is the idf value obtained from the following calculation:

$$idf(t) = \log \frac{N - df(t) + 0.5}{df(t) + 0.5}$$

Where:
- N is the total number of documents.
- df (t) is the number of documents that have terms t.

weight (t, d) is the weight of the term t in document d which is get from the calculation:

$$weight(t,d) = \sum_{c \in d} \frac{occurs(t,c,d).boost(c)}{((1-b_c) + b_c \frac{l_c}{avl_c}}$$

Where:
- c is a field in a document consisting of object-object predicates.
- d is a document, in RDF the document means an instance.
- Occurs(t,c,d) is the appearance of the term t in field c in the document d.
- Boost(c) is the boost factor given to field c.
- bc is a constant based on the length of field c.
- lc is the length of the field c.
- avlc is the average length from field c.

C. RDF

The Resource Description Framework (RDF) is a framework for retrieving information from data sources. These sources can be anything, including documents, people, physical objects, and abstract concepts. RDF appears now where the Web needs to be processed by the application, not just to be displayed. RDF provides a common framework for informing data so that data exchange between applications can be carried out without losing its real meaning.

D. AHP

Analytic Hierarchical Process (AHP) is a process that carries out a justification process for several criteria. AHP

was developed by Thomas Saaty in 1980 with an approach that breaks complex problems into problems that are arranged hierarchically.

AHP can be used to evaluate several criteria that can help in later decision making. In this study, the use of AHP is used to determine the boost value for each field in Solr collection The steps taken to calculate AHP are as follows:

1. Identify criteria to be used in research. In this study, there are four criteria to be calculated, namely Product Code, Product Name, Manufacturer Name, and Name of product.
2. Create a matrix of n x n, where n is the criteria to be weighted later.
3. Perform pairwise-comparison process.
4. Aggregate each column using sum operation
5. Normalize the aggregation values.

IV. METHODOLOGY

In order to determine the weight value of Product name, product code, manufacturer name and ingredient name field, we use the Analytic Hierarchical Process (AHP)[19] for calculating the rank of each field. A questionnaire was distributed to 136 respondents. The respondent must select a value from one to nine, where one value is evenly important and nine value is the most important. The questionnaire can be found at appendix. The AHP calculation generates the field weights in Table 1. Based on the questionnaire results, Consistency Ratio (CR) is 37.5% which is not acceptable since it is higher than 10% [20].

TABEL I
WEIGHT AND RANK FIELDS BEFORE ADJUSTMENT

Field	Weight	Rank
Product ID	3.4	4
Product Name	63.5	1
Manufacturer Name	9.1	3
Ingredients	24	2

Therefore, the results are adjusted to generate less CR value as seen in Table 2. After the adjustment, the CR value is getting down to 4.6%, but the rank of fields is the same as before.

TABEL III
WEIGHT AND RANK FIELDS AFTER ADJUSTMENT

Field	Weight	Rank
Product ID	4.7	4
Product Name	63.5	1
Manufacturer Name	9.8	3
Ingredients	22	2

The weight of each field will determine the answer of the query. The product name is the highest weight among the fields, followed by ingredient name, manufacturer name and product name.
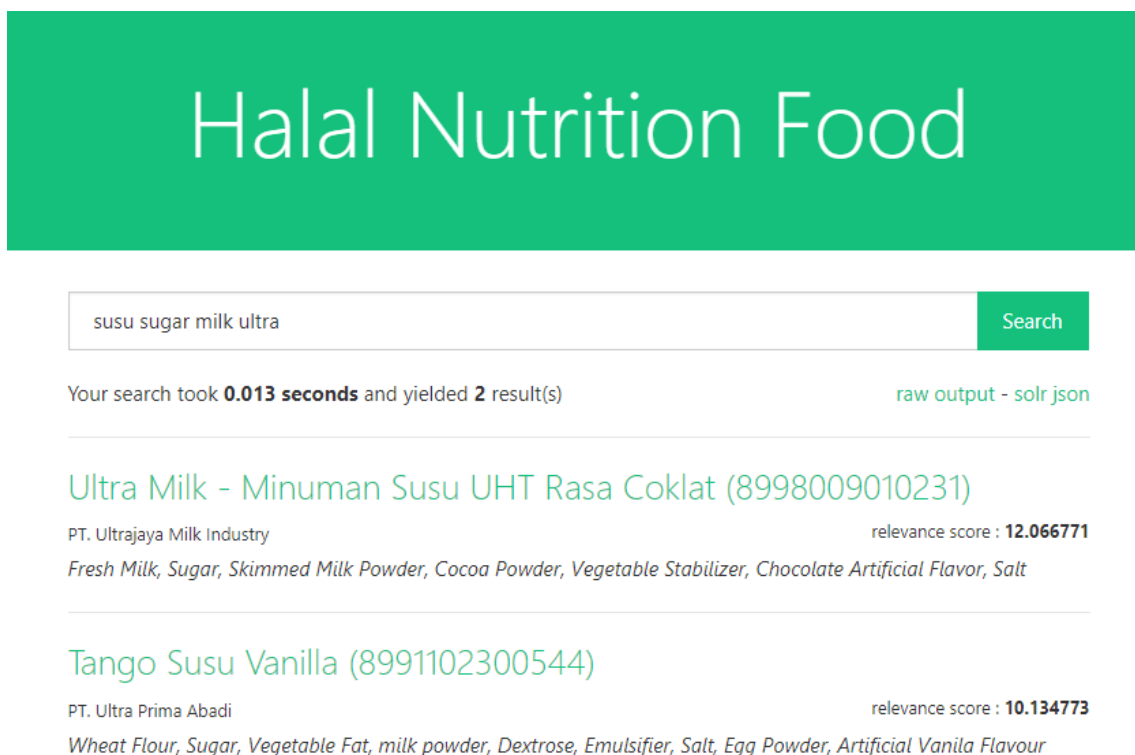
Figure 2. Screenshot of system returning the answer of a query

V. RESULTS AND DISCUSSION

Figure 2 shows our system returning the relevance products when "Susu sugar milk Ultra" (in Bahasa Indonesia) sends to the system. The keywords were distributed among three fields as seen in Table 3. Each keyword generally occurs in two fields except sugar.

TABLE III
THE DISTRIBUTION OF QUERY KEYWORDS.

Keyword	Product ID	Product Name	Manufacturer Name	Ingredients
Susu		*		*
Sugar				*
Milk		*		*
ultra		*	*	

Next, we test different numbers of keywords. Table 3 shows the rank results, keywords occurrences and BM25F scores — the more keywords delivered to the system, the higher the score of BM25F. The number of keywords occurrences has an impact on the score as well.

TABLE 4
EVALUATION RESULT FOR DIFFERENT NUMBERS OF KEYWORDS

Queries	Rank	#Keyword				BM25F
		Code	name	manufacture	ingredient	
Good	1	0	1	1	0	2.218
	2	0	1	0	0	2.192
	3	0	1	0	0	2.192
Milk Coklat	1	0	2	0	1	5.816
	2	0	1	0	2	5.648
	3	0	1	0	1	5.648
Roma Mayora Biskuit	1	0	2	1	0	8.589
	2	0	2	1	0	8.589
Susu sugar milk ultra	1	0	3	1	3	12.06
	2	0	1	1	2	10.13

VI. CONCLUSION

We have presented our system for searching a halal product using BM25F and AHP. BM25F is used for ranking relevance products based on the query keywords, while AHP determines the weight of fields. The product name is the highest weight among the fields, followed by ingredient name, manufacturer name and product name. The number of keywords and the keywords occurrences influence the score of query results.

For further research, the search query logs can be determined which fields should be given the highest weight value so that the search results are more relevant.

REFERENSI

[1] A. Indrawan, "Inilah 10 negara dengan populasi muslim terbesar di dunia," 2017, [Online]. Available: http://www.republika.co.id/berita/dunia-islam/islam-nusantara/15/05/27/noywh5-inilah-10-negara-dengan-populasi-uslim-terbesar-di-dunia

[2] J. Fatawi and N.A. Rakhawati, "Rancang bangun perangkat lunak linked open data halal dan gizi pada produk makanan dan minuman," 2016.

[3] Bizer, Christian, Tom Heath, and Tim Berners-Lee. "Linked data: The story so far." Semantic services, interoperability and web applications: emerging concepts. IGI Global, 2011. 205-227.

[4] World Wide Web Consortium. "RDF 1.1 concepts and abstract syntax." (2014).

[5] Harris, Steve, Andy Seaborne, and Eric Prud'hommeaux. "SPARQL 1.1 query language." W3C recommendation 21.10 (2013).

[6] Robertson, Stephen, Hugo Zaragoza, and Michael Taylor. "Simple BM25 extension to multiple weighted fields." Proceedings of the thirteenth ACM international conference on Information and knowledge management. ACM, 2004.

[7] Robertson, Stephen, and Hugo Zaragoza. "The probabilistic relevance framework: BM25 and beyond." Foundations and Trends® in Information Retrieval 3.4 (2009): 333-389.

[8] Saputra, T. (2013) 'Tuning Parameter dalam Fungsi Okapi BM25 pada Mesin Pencari Teks Bahasa Indonesia'. Available at: http://repository.ipb.ac.id/xmlui/handle/123456789/65325 (Accessed: 21 August 2018).

[9] Aberer, K., Boyarsky, A., Cudré-Mauroux, P., Demartini, G., & Ruchayskiy, O. (2011). Sciencewise: A web-based interactive semantic platform for scientific collaboration. In 10th International Semantic Web Conference (ISWC 2011-Demo), Bonn, Germany.

[10] Dorff, K.C., Wood, M.J. and Campagne, F., 2006, November. Twease at TREC 2006: Breaking and fixing BM25 scoring with query expansion, a biologically inspired double mutant recovery experiment. In TREC.

[11] Blanco, Roi, Michael Matthews, and Peter Mika. "Ranking of daily deals with concept expansion." Information Processing & Management 51.4 (2015): 359-372

[12] Esparza, Sandra Garcia, Michael P. O'Mahony, and Barry Smyth. "Mining the real-time web: a novel approach to product recommendation." Knowledge-Based Systems 29 (2012): 3-11.

[13] Jabeur, L. B. et al. (2016) 'A Product Feature-Based User-Centric Ranking Model for E-Commerce Search', in Fuhr, N. et al. (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer International Publishing (Lecture Notes in Computer Science), pp. 174-186.

[14] Petteri Nurmi, Antti Salovaara, Andreas Forsblom, Fabian Bohnert, and Patrik Floréen. 2014. PromotionRank: Ranking and Recommending Grocery Product Promotions Using Personal Shopping Lists. ACM Trans. Interact. Intell. Syst. 4, 1, Article 1 (April 2014), 23 pages. DOI: https://doi.org/10.1145/2584249

[15] J. R. Perez-Ag?uera, J. Arroyo, J. Greenberg, J. P. Iglesias, and V. Fresno, "Using bm25f for semantic search," in Proceedings of the 3rd international semantic search workshop . ACM, 2010, p. 2.

[16] R. Delbru, N. A. Rakhmawati, and G. Tummarello, "Sindice at semsearch 2010," in Proceedings of the 19th International World Wide Web Conference, Raleigh, North Carolina, USA . Citeseer, 2010.

[17] Blanco, R., Mika, P., & Vigna, S. (2011, October). Effective and efficient entity search in RDF data. In International Semantic Web Conference (pp. 83-97). Springer, Berlin, Heidelberg.

[18] Zhiltsov, Nikita, and Eugene Agichtein. "Improving entity search over linked data by modeling latent semantics." Proceedings of the 22nd ACM international conference on Information & Knowledge Management. ACM, 2013.

[19] Antonopoulos, I. S., Perkoulidis, G., Logothetis, D., & Karkanias, C. (2014). Ranking municipal solid waste treatment alternatives considering sustainability criteria using the analytical hierarchical process tool. Resources, Conservation and Recycling, 86, 149-159.

[20] Mu, E. and Pereyra-Rojas, M., 2017. Understanding the analytic Hierarchy process. In Practical Decision Making (pp. 7-22). Springer, Cham.

[21] Rakhmawati, Nur Aini, et al. "A performance evaluation for assessing registered websites." Procedia Computer Science 124 (2017): 714-720.

[22] Roy, Jagannath, et al. "Evaluation and selection of medical tourism sites: A rough analytic hierarchy process based multi?attributive border approximation area comparison approach." Expert Systems 35.1 (2018): e12232

[23] Kubler S, Robert J, Neumaier S, Umbrich J, Le Traon Y. Comparison of metadata quality in open data portals using the Analytic Hierarchy Process. Government Information Quarterly. 2018 Jan 1;35(1):13-29.