International Journal on Advanced Science Engineering Information Technology

Vol.8 (2018) No. 4-2
ISSN: 2088-5334

Auto Halal Detection Products Based on Euclidian Distance and Cosine Similarity

Nur Aini Rakhmawati[#], Azmi Adi Firmansyah[#], Pradita Maulidya Effendi[#], Rosyid Abdillah[#], Taufiq Agung Cahyono[#]

[#]*Department of Information Systems, Faculty of Information and Communication Technology, Institut Teknologi Sepuluh Nopember Surabaya, Indonesia*

E-mail: azmiadifirmansyah@gmail.com, praditame@gmail.com, rosyid.17052@mhs.its.ac.id, taufiqcahyono.17052@mhs.its.ac.id
Corresponding author e-mail: nur.aini@is.its.ac.id

***Abstract*— Although Indonesia is the world the world's most populous Muslim-majority country, the number of halal-certified products in Indonesia is only 20% of the products on the Indonesian market. Halal certification is voluntary as such there are many food products which are halal but are not certified as halal. In principle, these food products may have similar halal ingredients with halal-certified products. In this study, we build a system that can compare products that have not been certified halal with halal certified products based on its ingredients. The food products are collected from Open Food Facts, Institute For Foods, Drugs, And Cosmetics Indonesian Council Of Ulama (LPPOM MUI) and our halal system. As of this paper writing, the halal-certified products are obtained from LPPOM MUI. The system uses the Euclidean Distance and Cosine Similarity that generate top-5 similar products. Those two similarity calculations are based on Term Frequency-Inverse Entity Frequency weighting function. The weighting function calculates the frequency of a term on a product name and ingredients. If a similarity value of a product with no halal certification and a halal-certified product is higher than 75%, then the former could be indicated as a halal product. In the end, the system can give a recommendation of unknown products from a related pool of halal-certified products based on similarity of product composition. Cosine similarity accuracy is higher than Euclidean Distance and MoreLikeThis accuracy. Cosine similarity gets the highest precision because the cosine similarity is based on the vector angle of the term in a product.**

***Keywords*— halal; ingredients; euclidean distance; cosine similarity.**

I. INTRODUCTION

Based on the population census in 2010, the population in Indonesia is 207.176.162 million people, where 87.2% of them are Moslems [1]-[3]. To assist the Muslim needs, Institute For Foods, Drugs, And Cosmetics Indonesian Council Of Ulama (LPPOM MUI), an authorised institution to supervise halal food products in Indonesia. LPPOM MUI was established by the Indonesian Ulama Council and the authority to conduct audits and issue halal certificates [4] to industries such as food, medicine, and cosmetics, Slaughter House and restaurant/catering/ kitchen [5].

According to the LPPOM MUI report [6], the number of halal-certified products is increasing. However, the number of halal-certified products in Indonesia is only 20% of the products on Indonesia market because it is not mandatory for a company to apply for halal certificates for their products [7]. A product which is not halal certified could not be ascertained to be a haram product, since the composition of the product may have similarities to a halal-certified product. Therefore, we build a system that compares products that have not been certified against similar products that have been halal certified based on the composition of the ingredients.

To date, there are over 200 halal certification organisations in the world, but a Muslim might be challenging to distinguish halal products especially in non-Muslim countries [8]. Our system allows users to survey, check and ensure the food products based on their ingredients in comparison to Halal certified products. It is important to note that the limitation of this work is that we only focus on the components of a product which our system does not include a food processing system. Open Food Facts (https://world.openfoodfacts.org/), a crowdsourcing food database system provides some of the products that are labelled as a halal product. However, the label is uncertain because it does not come from halal certification organisations and everyone is free to claim a halal status of a product without supervision.

Our system uses Euclidean distance and cosine similarity method between two products based on the similarity level by making the composition of the product as a comparison in

assessing whether or not the product is halal. Concerning finding the similarity of products, Cosine similarity is expected to increase the value of accuracy [9]. Some previous studies have used cosine similarity and Euclidean distance to perform groupings of vector data from the reduction results with Principal Component Analysis (PCA) since the changes in the average correction of the original data may alter the location of the document [10]. Another study [11] successfully improved classification accuracy using Euclidean distance on the authenticity of tea samples.

Our main contributions are explained as follows:

- Proposing two similarity methods to detect the halal status of products with no halal certification
- Providing a system that can display the top five of similar products that are already being halal certified concerning the product in question

This paper is structured as follows: Section II describes two of our similarity methods and our existing system. The methodology used in this system is explained in Section III. The results and conclusion can be found in Section IV and Section V respectively.

II. Material and Method

Briefly, Fig. 1 describes our system. All products are compared to each other. If a similarity value of a product with no halal certification and a halal-certified product is higher than 75%, then the former could be indicated as a halal product.

Our methodology consists of four steps, namely data collection, data indexing, TF-IEF weighting terms, and Euclidean distance and cosine similarity calculations.

A. Data collection

Our data were obtained from three sources: 1) crowdsourcing using a product input feature in our website http://halal.addi.is.its.ac.id; 2) web scrapping LPPOM MUI website http://www.halalmui.org/, and 3) Open food facts (https://world.openfoodfacts.org/).

http://www.halalmui.org/ only provides the name of products, the name of manufactures, and halal certificate expiry date. https://world.openfoodfacts.org/ is an initiative for establishing a collaborative, free and open database of food products from around the world. Open Food Facts provides information about food products including food, manufacture, ingredients and nutrition.

In 2016, Halal Nutrition Food framework, http://halal.addi.is.its.ac.id was developed by Fatawi and Rakhmawati [14]. The system facilitates users to enter the halal products, search for halal products, and find more information by exploiting Linked Data technologies [15].

B. Data Indexing

The collected data were then indexed into the Apache Lucene (https://lucene.apache.org/) to store the terms of labels and ingredients from the document into the Lucene index. Before product data was indexed, data must be pre-processed to accelerate the weighting and calculation of similarity between products. The initial pre-processing was tokenisation, which is cutting the sentence into its constituent words, called tokens, based on spaces and punctuation. This token was later indexed into Apache Lucene. Examples of indexing of SampleP Mie Goreng and SampleQ Rasa Soto Ayam can be seen in Fig. 2 and 3.

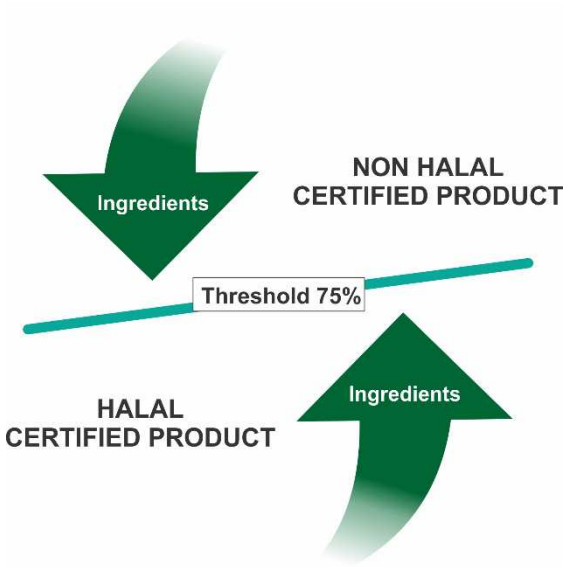

Fig. 1 Prediction system comparing products with no halal certification and halal certified products

C. TF-IEF Weighting

Delbru presents weighting for Linked Data search which is called the Term Frequency-Inverse Entity Frequency or TF-IEF [16]. This weighting measures the importance of a term across entities in the collection. In our case, product and ingredient are an entity. For instance, p_i is an ingredient in product p, the weight of p_i in product p is

$$w(p_i, p) = \frac{f(p_i,p)}{|p|} \log \frac{N}{1+ f(p_i,p)} \quad (1)$$

Where $f(p_i, p)$ is the number of occurrences of the ingredient p_i , $|p|$ specifies the number of ingredients in the product p, and N is the number of products in the dataset.

The product along with ingredients was weighted using TF-IEF[16]. The weighting of the TF-IEF compared the composition of each product with the composition of other products that have been stored. Apache Lucene reads the previous index before calculating the weight of terms.

Table I and II are two TF-IEF weighting examples for the *SampleP Mie Goreng* (p) and the *SampleQ Rasa Soto Ayam* (q) products.

Doc: SampleP *Mie Goreng*

Label:
SampleP, *mie, goreng*

containsIngredient:
vegetable,oil,salt,wheat,flour,sewwt,soy,chilli ,sauce

Fig. 2 Indexing of *SampleP Mie Goreng* in Apache Lucene

Doc: SampleQ *Rasa Soto Ayam*

Label:
sampleQ, *Soto, rasa, ayam*

containsIngredient:
Sugar, Salt, Vegetable Oil, Wheat Flour, Tapioca Starch, Acidity Regulator, Artificial Chicken Flavour, Spring Onion, Garlic Powder, Onion Powder, Pepper Powder, Celery Powder, Chilli Powder, Spice, Tartrazine, Monosodium Glutamate

Fig. 3 Indexing of *SampleQ Rasa Soto Ayam* in Apache Lucene

TABLE I
WEIGHTING TF-IEF *SAMPLEP MIE GORENG* PRODUCTS

Term	$f(p_i,p)$	$w(p_i,p)$
Vegetable	1	0.189
Oil	1	0.189
Salt	1	0.189
Sauce	2	0.34
…	…	…
Chilli	1	0.189

TABLE II
WEIGHTING TF-IDF SAMPLEQ PRODUCT RASA SOTO CHICKEN

Term	$f(q_i,q)$	$w(q_i,q)$
Sugar	1	0.155
Salt	1	0.155
Wheat	1	0.155
Butylated	2	0.155
…	…	…
Glutamate	1	0.155

D. Calculation of Euclidean Distance and Cosine Similarity

Euclidean Distance is a type of distance measurement and is the most commonly used cluster analysis to measure the distance from a data object to a cluster centre. Euclidean distance is a geometric distance between two data objects. The closer the distance between the two objects, the more identical they are [12]. Similarity on Euclidean distance has a range of values from 0 to 1. Zero (0) value states that the two products do not have any resemblance, while one (1) means that the two products are identical.

Given p as the first product and q as the second product, the Euclidian distance formula for finding the similarity between product p and q can be defined as follows:

$$sim(p,q) = \frac{1}{dist(p,q)+1} \quad (2)$$

Where *dist(p,q)* can be formulated as follows:

$$dist(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (3)$$

Cosine method similarity is a method used to calculate the similarity (degree of similarity) between two objects. In general, the calculation of this method is based on the vector space similarity measure. The similarity between two objects is expressed in the two vectors by using keywords (keywords) of a document as size [13]. Vector representation is used to facilitate calculations of long documents. Similar to Euclidean distance, Cosine similarity has a range of values from 0 to 1.

Given p as the first product and q as the second product, the Cosine similarity formula for finding the similarity between two products can be defined as follows:

$$sim(p,q) = \frac{\sum_{i=1}^{n} w(p_i,p)w(q_i,q)}{\sqrt{\sum_{i=1}^{n} w(p_i,p)^2}\sqrt{\sum_{i=1}^{n} w(q_i,q)^2}} \quad (4)$$

Where $w(p_i,p)$ and $w(q_i,q)$ are the weighting function for each ingredient in product p and q.

The Euclidean distance calculation takes into account the distance between two points in the Euclidean space. The first TF-IEF document was calculated to represent it as a point in the Euclidean space.

Table III shows the steps in calculating Euclidean distance between *SampleP Mie Goreng* product (p) and *SampleQ Rasa Soto Ayam* (q).

TABLE III
EUCLIDEAN DISTANCE CALCULATION

Term	$w(p_i,p)$	$w(q_i,q)$	$(w(p_i,p) - w(q_i,q))^2$
Vegetable	0.189	0.155	0.0012
Oil	0.189	0.155	0.0012
Salt	0.189	0.155	0.0012
Wheat	0.189	0.155	0.0012
…	…	…	…
Chilli	0.000	0.000	0.0357
Total			0.3727
SQRT			0.6105

Based on the similarity formula on Euclidean distance, the similarity score between *SampleP Mie Goreng* and *SampleQ Rasa Soto Ayam* is:

$$sim(p,q) = \frac{1}{0.61+1} = 0.62 \quad (5)$$

Calculation of cosine similarity also uses TF-IEF weighting terms. Table IV shows the calculation steps in cosine similarity *SampleP Mie Goreng product* (p) and *SampleQ Rasa Soto Ayam* (q).

TABLE IV
COSINE SIMILARITY BETWEEN SAMPLEP *MIE GORENG* AND SAMPLEQ *RASA SOTO AYAM.*

Term	$w(p_i,p)$	$w(q_i,q)$	$w(p_i,p)\,w(q_i,q)$	$w(p_i,p)^2$	$w(q_i,q)^2$
Vegetable	0.189	0.155	0.029	0.036	0.024
Oil	0.189	0.155	0.029	0.036	0.024
Salt	0.189	0.155	0.029	0.036	0.024
Wheat	0.189	0.155	0.029	0.036	0.024
…	…	…	…	…	…
Chili	0.189	0.000	0.000	0.036	0.000
Total			0.146	0.401	0.264
SQRT				0.634	0.514

Cosine similarity score weights between *SampleP Mie Goreng* products and *SampleQ Rasa Soto Ayam* can be calculated as follows:

$$sim(p,q) = \frac{0.146}{0.634\ x\ 0.514} = 0.45 \quad (6)$$

III. Results and Discussion

Based on Euclidean Distance and Cosine Similarity testing on the Halal Nutrition Food application feature, an exemplary search result is shown in Fig. 4.

Each search result displayed a list of other products related to the product being searched. List of related products with *SampleQ Mie Goreng* are depicted in Table V and Table VI.

Based on the products list in Tables 5 and 6, it can be seen that the related products are Instant Noodle products. Therefore, the result of using Euclidean Distance and Cosine Similarity on *SampleP Mie Goreng* is relevant. Meanwhile, Fig. 5 displays a notification of a product (*Chocolate Sample*) that does not have a halal certificate, and does not contain additives in the haram category, has a similarity of more than 75% based on Euclidean distance or cosine similarity.

TABLE V
List of Products Related to SampleQ *Mie Goreng* Using Cosine Similarity

Produk	*Cosine Similarity*
Sample *Mie Goreng Rasa Sate*	83,9167%
Sample *Mie Rasa Kari Ayam*	74,8913%
Sample *Mie Rasa Ayam Special*	74,8193%
Sample *Mie Rasa Ayam Bawang*	74,5665%
Sample *Mie Goreng Rendang*	74,4058%

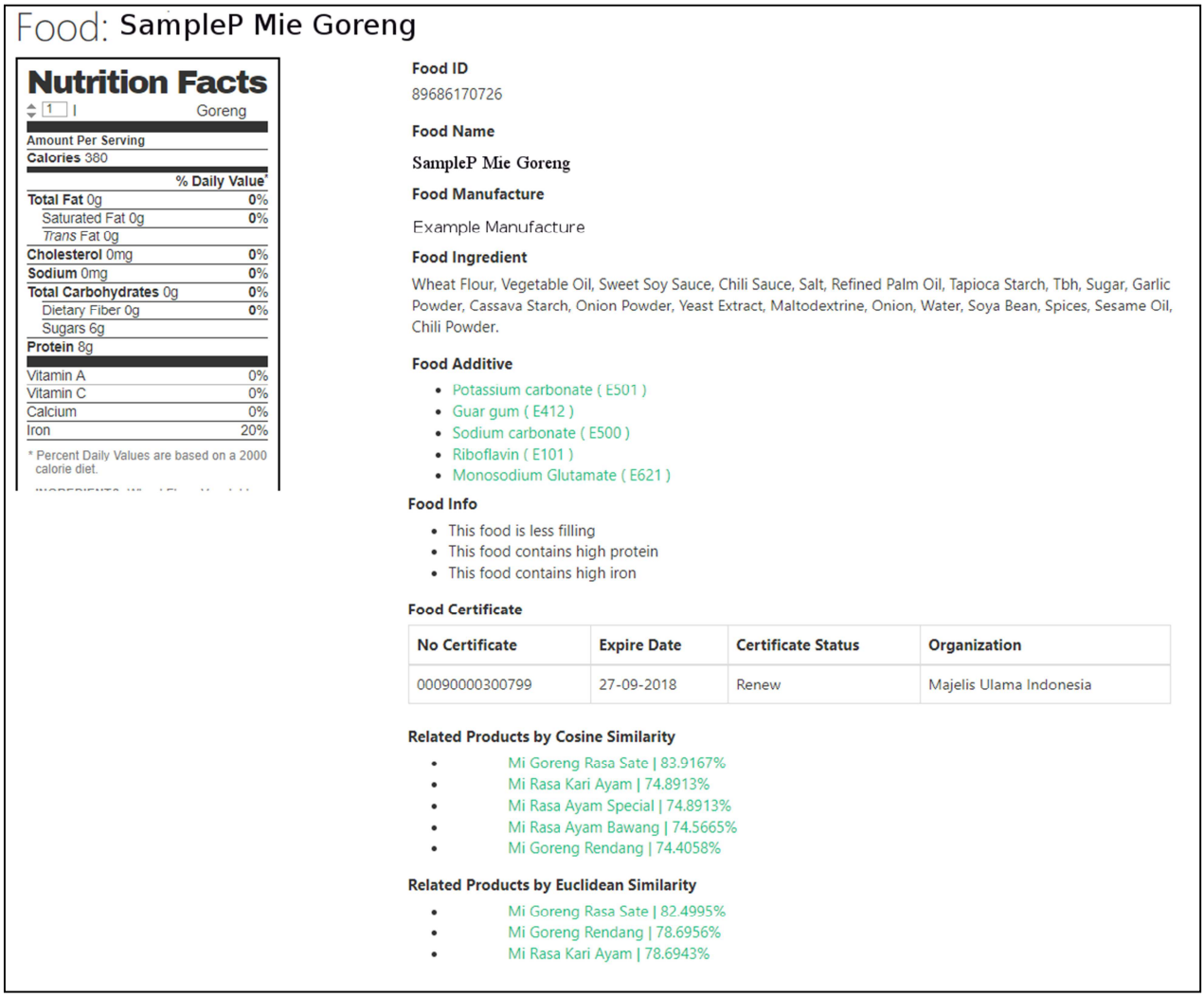

Fig. 4 Examples of related product search results on Halal Nutrition Food Applications Using Euclidean Distance and Cosine Similarity

TABLE VI
List of Products Related to SampleQ *Mie Goreng* Using Euclidean Distance

Produk	*Euclidean Distance*
Sample *Mie Goreng Rasa Sate*	82,4995%
Sample *Mie Goreng Rendang*	78,6956%
Sample *Mie Rasa Kari Ayam*	78,6943%

To test the relevance of the related product shown, precision testing of the Euclidean distance and cosine similarity algorithms was performed. Besides, we also examined the significance of the associated products by using MoreLikeThis from Apache Lucene. Precision results of several products can be seen in Table VII.

Further research requires more samples and evaluations to investigate the reliability and reproducibility of results. Before weighing the term, it is suggested to pre-process the document using a similarity string algorithm, such as Levenshtein [17] or Jaccard [18] to avoid in case of typing errors.

TABLE VII
PRECISION SCORE COMPARISON USING EUCLIDEAN DISTANCE, COSINE SIMILARITY, AND MORELIKETHIS

Products	*Cosine Similarity*	*Euclidean Distance*	MoreLike This
Sample Milk Plain	100%	80%	100%
Sample *Rasa Kacang Hijau*	40%	40%	40%
Sample Original	100%	80%	80%
SampleQ *Mie Goreng*	100%	100%	100%
Sample Wafer Roll	80%	60%	80%
Average Precision	**84%**	**72%**	**80%**

Table VII shows that the precision of related products searches using Euclidean distance, Cosine similarity and MoreLikeThis are 72%, 84%, and 80% respectively. Precision score using Euclidean distance yields the smallest value because Euclidean distance measures the distance between two points in the Euclidean space that is influenced by term weight. Cosine similarity gets the highest precision because the cosine similarity is based on the vector angle of the term in a document.

IV. CONCLUSION

The related product search system in the Halal Nutrition Food was developed. The relevant product search feature displays top five products that are on the product detail page. Cosine similarity surpasses the performance of Euclidean Distance and MoreLikeThis from Apache Lucene; and as such is recommended for further use in the product search system. It is noteworthy that the findings of this work do not rule out the needs of halal certification process which include documents and site audits; covering all aspects of food processing from ingredients, processing line and logistics. Nevertheless, the system offers an opportunity for users to survey, check and ensure the food products based on their elements in comparison to Halal certified products. In the future, the system could be further refined to include the standard requirements for halal certification.

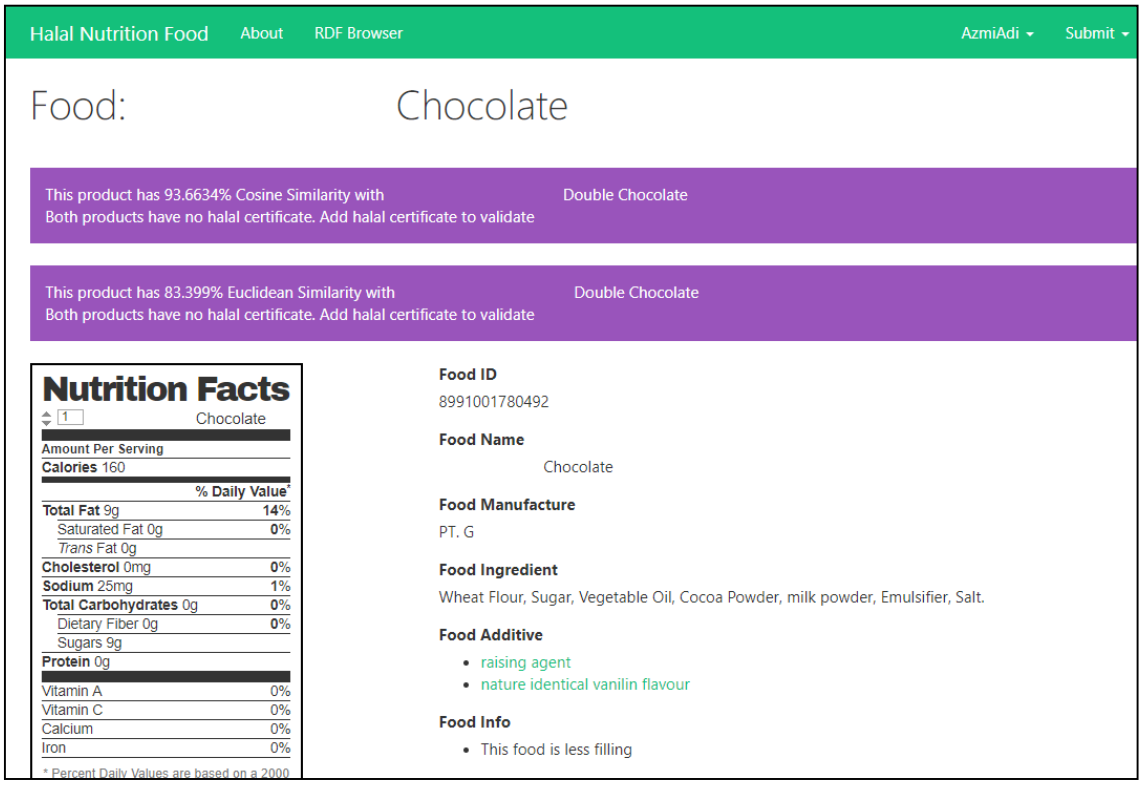

Fig. 5 Notification of product (*chocolate sample*) that does not have a halal certificate, and does not contain additives in the haram category, has a similarity of more than 75% based on Euclidean distance or cosine similarity.

ACKNOWLEDGMENT

This research was supported by funding from Lembaga Penelitian dan Pengabdian kepada Masyarakat, Institut Teknologi Sepuluh Nopember Surabaya (LPPM - ITS) and Kementerian Riset, Teknologi, dan Pendidikan Tinggi (or Ministry of Higher Education Indonesia) with the scheme of Department Research and the grant number: 1156/PKS/ITS/2017.

REFERENCES

[1] Badan Pusat Statistik Indonesia. (2010). *"Jumlah dan Distribusi Penduduk"*. [Online]. Available: http://sp2010.bps.go.id/index.php/site/index

[2] Santoso, A. B. (2015). "Berdasar Survei Ini, Pertambahan Penduduk Kristen di Indonesia Lebih Cepat Dibanding Muslim. [Online]. Available:http://www.tribunnews.com/internasional/2015/04/05/berdasar survei-ini-pertambahan-penduduk-kristen-di-indonesia-lebih-cepat-dibanding-muslim

[3] Tomoutou. (2017). “Jumlah Penganut Agama di Indonesia Tiap Provinsi”.[Online].Available: http://tumoutounews.com/2017/11/08/jumlah-penganut-agama-di-indonesia-tiap-provinsi/
[4] Jati, S. “Sertifikasi Halal MUI”. Jakarta: Majelis Ulama Indonesia. 2017
[5] LPPOM MUI. (2017). “Prosedur Sertifikasi Halal MUI”. [Online]. Available: http://www.halalmui.org/mui14/index.php/main/go_to_section/56/162/page/1
[6] LPPOM MUI. (2017). “Statistik Sertifikasi Halal Indonesia”. [Online]. Available: http://www.halalmui.org/mui14/index.php/main/go_to_section/59/1368/page/1
[7] Pratama, A. F. (2014). “Produk Bersertifikasi Halal di Indonesia Baru 20 Persen, Malaysia Sudah 90 Persen”. [Online]. Available: http://www.tribunnews.com/nasional/2014/03/07/produk-bersertifikasi-halal-di-indonesia-baru-20-persen-malaysia-sudah-90-persen
[8] B, Ali, and J.M. Regenstein, “ Halal food certification challenges and their implications for Muslim Societies Worldwide,” *Electronic Turkish Studies*, vol. 9, pp. 111-130, Nov. 2014.
[9] A. Y. Rofiqi, “Clustering Berita Olahraga Berbahasa Indonesia Menggunakan Metode K-Medoid Bersyarat” SimanteC Journal, vol. 6 no 1. pp. 25-32.], June.2017.
[10] T. Korenius, J, Laurikkala, M. Juhola, “On principal component analysis, cosine and Euclidean measures in information retrieval,” Information Sciences, vol. 177, pp. 4893-905, Nov 2007.
[11] W.He et al. “Validation of origins of tea samples using partial least squares analysis and Euclidean distance method with near-infrared spectroscopy data,” Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, vol 86, pp. 399–404, Feb. 2012.
[12] C. Bizer, “The emerging web of linked data,” IEEE Intelligent Systems, vol. 24, pp.87-92, Oct. .2009
[13] G.A.Pradnyana, ER, N. A. S. “Perancangan dan mplementasi Automated Document Integration dengan menggunakan Algoritma Complete Linkage Agglomerative Hierarchical Clustering”. Jurnal Ilmu Komputer, vol. 5, 2012.
[14] Fatawi, J, N.A Rakhmawati, “Rancang bangun perangkat lunak linked open data halal dan gizi pada produk makanan dan minuman,” 2016.
[15] C. Bizer, T. Heath, T.Berners-Lee. "*Linked data: The story so far*." In Semantic services, interoperability and web applications: emerging concepts, pp. 205-227. IGI Global, 2011.
[16] R. Delbru, et al. “Searching web data: an entity retrieval model”. Digital Enterprise Research Institute, National University of Ireland, Galway, 2010
[17] V.I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,”. Soviet physics doklady, vol. 10, no. 8, pp. 707-710, Feb.1996.
[18] S. Niwattanakul, J. Singthongchai, E. Naenudorn, S. Wanapu, S, “Using of Jaccard coefficient for keywords similarity,” in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, no. 6. Marc. 2013.