JIKI. 8/1 (2015), 45-52
DOI: http://dx.doi.org/10.21609/jiki.v8i1.283

PREFERENCE BASED TERM WEIGHTING FOR ARABIC *FIQH* DOCUMENT RANKING

Khadijah Fahmi Hayati Holle, Agus Zainal Arifin, and Diana Purwitasari

Informatics Engineering Department, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember, Jalan Raya ITS, Surabaya, 60111, Indonesia

E-mail: khadijah.holle13@mhs.if.its.ac.id, agusza@cs.its.ac.id, diana@if.its.ac.id

Abstract

In document retrieval, besides the suitability of query with search results, there is also a subjective user assessment that is expected to be a deciding factor in document ranking. This preference aspect is referred at the *fiqh* document searching. People tend to prefer on certain *fiqh* methodology without rejecting other *fiqh* methodologies. It is necessary to investigate preference factor in addition to the relevance factor in the document ranking. Therefore, this research proposed a method of term weighting based on preference to rank documents according to user preference. The proposed method is also combined with term weighting based on documents index and books index so it sees relevance and preference aspect. The proposed method is Inverse Preference Frequency with α value (IPF_α). In this method, we calculate preference value by IPF term weighting. Then, the preference values of terms that is equal with the query are multiplied by α. IPF_α combined with the existing weighting methods become TF.IDF.IBF.IPF_α. Experiment of the proposed method uses dataset of several Arabic *fiqh* documents. Evaluation uses recall, precision, and f-measure calculations. Proposed term weighting method is obtained to rank the document in the right order according to user preference. It is shown from the result with recall value reach 75%, precision 100%, and F-measure 85.7% respectively.

Keywords: *document ranking, document retrieval, preference, term weighting, IPF_α*

Abstrak

Dalam pencarian, selain kesesuaian *query* dengan hasil pencarian, terdapat penilaian subjektif pengguna yang diharapkan menjadi faktor penentu dalam perangkingan dokumen. Aspek preferensi tersebut tampak pada pencarian dokumen fiqih. Seseorang cenderung mengutamakan metodologi fiqih tertentu meskipun tidak mengabaikan pendapat metodologi fiqih lain. Faktor preferensi menjadi hal yang diperlukan selain relevansi dalam perangkingan dokumen. Oleh karena itu, pada penelitian ini diajukan metode pembobotan kata berbasis preferensi untuk merangkingkan dokumen sesuai dengan preferensi pengguna. Metode yang diajukan digabungkan dengan pembobotan kata berbasis indeks dokumen dan buku sehingga mampu memperhatikan aspek kesesuaian (*relevance*) dan keutamaan (*preference*). Metode pembobotan yang diusulkan disebut dengan *Invers Preference Frequency with α value* (IPF_α). Langkah pembobotan yang diusulkan yaitu dengan perhitungan nilai preferensi *term* dengan pembobotan IPF. Kemudian nilai preferensi dari *term* dokumen yang sama dengan *term query* dikalikan dengan $\boldsymbol{\alpha}$ sebagai penguat. IPF_α digabungkan dengan metode pembobotan yang telah ada menjadi TF.IDF.IBF.IPF_α. Pengujian metode yang diusulkan menggunakan dataset dari beberapa dokumen fiqih berbahasa Arab. Evaluasi menggunakan perhitungan *recall*, *precision*, dan *F-measure*. Hasil uji coba menunjukkan bahwa dengan pembobotan TF.IDF.IBF.IPF_α diperoleh perangkingan dokumen dengan urutan yang tepat dan sesuai dengan preferensi pengguna. Hal ini ditunjukkan dengan nilai maksimal *recall* mencapai 75%, *precision* 100%, dan *F-measure* 85.7%.

Kata Kunci: *perangkingan dokumen, temu kembali dokumen, preferensi, pembobotan kata, IPF_α*

1. Introduction

Document ranking is common discussion in information retrieval research [1]. Harrag et al. was using vector space model to perform Arabic document ranking [2]. In the vector space model (VSM), the text content is represented as a vector in the term space. In general, text to vector representation can be classified into two tasks: indexing and term weighting.

The most popular term weighting method is TF.IDF weighting [3,4]. TF.IDF weighting is combining term frequency (TF) and inverse document frequency (IDF). TF measures the density of a term in a document and IDF provides the lowest score of those terms that appear in multiple docu-

ments; because of this, the TF.IDF score gives positive discrimination to rare terms and is biased against frequent terms.

However, by TF.IDF weighting, information about a set of documents in a certain class is not calculated. Therefore, Fuji et al. [5] proposed class indexing based term weighting. They implement a class indexing based TF.IDF.ICF weighting method in which inverse class frequency (ICF) is incorporated. They also implement the inverse class space density frequency (ICSdF). This method was proved to have higher precision and recall than those on TF.IDF weighting method.

In addition, Fauzi [6] proposed book indexing based term weighting for Arabic document ranking called inverse book frequency (IBF). IBF is used to improve the performance of document ranking that have hierarchy in the form of books that have many pages. IBF calculation is multiplied by the previous method and becomes TF.IDF. ICF.IBF weighting method. TF.IDF.ICF.IBF weighting method proven can be applied to Arabic document ranking and has a better performance than the previous method.

Those existing term weighting methods rank the documents based on relevance of documents. Nevertheless, beside relevance factor it is necessary to investigate preference factor. Many researches implement preference in the proposed method. User preference is useful in recommender system [7,8], query enrichment [9,10], data ranking [11], and others. Chulyadyo et al. [7] introduced approach of constructing a personalized recommendation model from Probabilistic Relational Model with the help of users' preferences over their search criteria. De Amo et al. [12] focused on preference elicitation and proposed an automatic preference elicitation method based on mining techniques by capturing implicit user's choices. Preference elicitation also can be done by using a query interface where users are asked to express their preferences [13].

In the document ranking, besides the relevancy of the query with the search results, there is user subjective assessment that expected to be the deciding factor. This preference aspect referred at the *fiqh* document searching. *Fiqh* document can be grouped in a particular *mazhab*. *Mazhab* is paradigm or *fiqh* methodology that has been a factor in decision-making [14]. Several famous *fiqh* methodologies are *Syafi'iyah*, *Hanabilah*, *Hanafiyah*, and *Malikiyah*. People tend to prefer on certain *fiqh* methodology without reject other *fiqh* methodologies.

Therefore, this research proposed a method of term weighting based on preference to rank document according to user preference. This preference-based term weighting is called Inverse Preference Frequency with α value (IPF_α). The IPF function assigns the lowest score to those terms that appear in multiple candidate user preference. Constant α is a multiplier that will increase the weights according to user preference. Proposed method combined with existing method to be TF. IDF.IBF.IPF_α term weighting method. This method is implemented to rank Arabic *fiqh* document as document that have a preference aspect. It also can be implemented to rank other document that have a preference aspect. This method combined relevance and preference that will rank the document in the correct order, that relevant to query according to user preference.

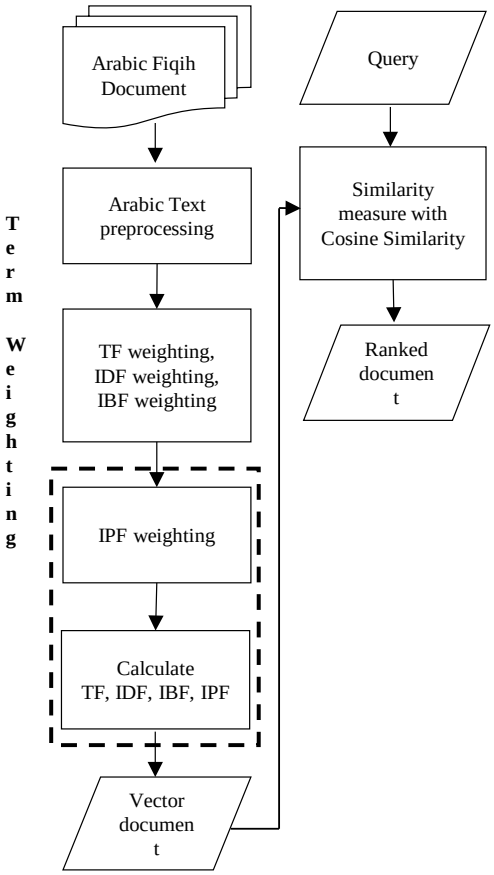


Figure 1. Phase of proposed method.

2. Methods

This paper proposed term weighting method that combine relevance aspect and preference aspect to rank Arabic *fiqh* document. Phase of proposed method include preprocessing and similarity measure are shown in Figure 1.

Dataset

To implement proposed method, we can use types of documents that have a preference aspect as data

in the experiment. In this reseach the dataset is Arabic *fiqh* document because it one of the types of documents that have a preference aspect. A document is each page of the book. The *fiqh* books are books that have been categorized into four *fiqh* methodologies that are *Hanafiyah*, *Malikiyah*, *Syafi'iyah*, and *Hanabilah*. For each *fiqh* methodology taken three books and for each book taken 10 documents. So the number of book and document that we used in this research are 12 *fiqh* books and 120 documents. Data taken from the Arabic e-book (http://shamela.ws/).

Preprocessing

Whole of collection documents through the preprocessing phase. The preprocessing consists of several stages, which are tokenizing, filtering, stopword removal, and stemming [15]. Token-izing is done to token entire contents of the document into a set of single word. In filtering stage, vowel *(harokat)* and punctuation are remo-ved. Words that frequently appear in the document but did not have significant value in a document are removed in stopword removal stage. Stopword list is taken from the website http://-Arabicstemmer.codeplex.com/. The last stage of preprocessing is stemming. Stemming stage used to obtain the root of each word by finding the base word and removing affix and suffix. To implement this process we used Light Stemmer [16-17] which are in lucene library (http://lucene.apache.-org/).

Term Weighting

In term weighting phase we calculate TF (term frequency), IDF (inverse document frequency), and IBF (inverse book frequency) of each term in the whole documents. Then the calculation of IPF (inverse preference frequency) also calculated for each term in the whole documents. For term document which equal with term query, IPF weight of the term is multiplied by the value of α to obtain the IPF_α (inverse preference frequency with α value). Then TF.IDF.IBF.IPF_α can be calculated. Here is the formula for each term weighting.

Term frequency is the simplest method in term weighting [3-4]. Each term is assumed have significant level in accordance with the number of occurrences of the term in document. Weight of term *t* in document *d* as defined in equation(1).

$$W_{TF}(t_i, d_j) = f(t_i, d_j) \tag{1}$$

where $f(t_i, d_j)$ denotes number of term t_i in document d_j.

IDF is calculated to determine the significance of each term in differentiating one document to the others [3,4]. Term that is rarely appears in the document is very valuable. Significance of each term is assumed have the opposite proportion with the number of documents that contain the term. The common term weighting as defined in equation(2).

$$W_{IDF}(t_i, d_j) = 1 + log\left(\frac{D}{d_{(t_i)}}\right) \tag{2}$$

where *D* denotes the total number of documents in the collection and $d(t_i)$ is the number of documents in the collection in which term t_i occurs at least once.

IBF calculates occurrence the term on the set of books [6]. Term that rarely appears in the collection of books is very valuable. Significance of each term is assumed have the opposite proportion with the number of books that contain the term. IBF calculation can be adopted from the IDF as defined in equation(3).

$$W_{IBF}(t_i, b_k) = 1 + log\left(\frac{B}{b_{(t_i)}}\right) \tag{3}$$

where *B* denotes the total number of books in the collection and $b(t_i)$ is the number of books in the collection in which term t_i occurs at least once.

IPF calculates occurrence the term in the group of candidate user preference. In this research, candidate user preference is four *fiqh* methodologies. Term that rarely appears on the fiqh methodology is very valuable.

Significance of each term is assumed have the opposite proportion with the number of *fiqh* methodologies that contain the term. As with IBF, IPF calculation can be adopted from the IDF as defined in equation(4). While IPF_α obtained by multiplying IPF with value of α for term that equal with user query, IPF_α as de-fined in equation(5). In Arabic document *fiqh*, do-cument of each book have been categorized in particular *fiqh* methodology. So, we can define particular document categorize as certain user pre-ference according to chosen *fiqh* methodology.

$$W_{IPF}(t_i, p_l) = 1 + log\left(\frac{P}{p_{(t_i)}}\right) \tag{4}$$

$$W_{IPF_\alpha}(t_i, d_j, p_l) = \begin{cases} \left(1 + log\left(\frac{P}{p_{(t_i)}}\right)\right) \times \left(\frac{\alpha}{2} + 0.5\right), d_j \in UP, t_i \in Q \\ \left(1 + log\left(\frac{P}{p_{(t_i)}}\right)\right) \times \left(1 - \left(\frac{\alpha}{2} + 0.5\right)\right), d_j \notin UP, t_i \in Q \\ \left(1 + log\left(\frac{P}{p_{(t_i)}}\right)\right), t_i \notin Q \end{cases} \tag{5}$$

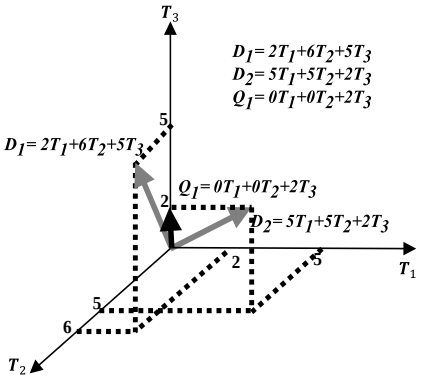


Figure 2. Document vector and query vector.

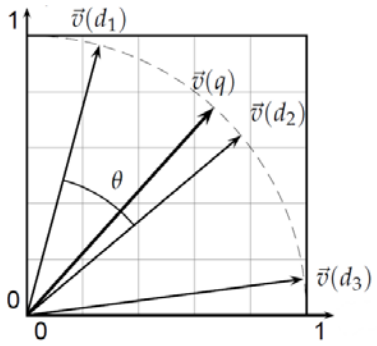


Figure 3. Representation of cosine similarity.

where P denotes the total number of *fiqh* methodologies in the collection, $p(t_i)$ is the number of *fiqh* methodologies in the collection in which term t_i occurs at least once, α denotes degree of preference it between 0 to 1, this value is given value by user according to the level of user preference. When α=0 that indicates no one in priority, while α=1 indicates only priority to user preference. UP is user preference and Q denotes query.

The idea of IPF_α is to close the document vector to the query vector for document that includes chosen group preference options. It is done by increase weight of the term in document that include to chosen group as user preference and decrease weight of the term in other document that not include to chosen group. Therefore, such document including chosen group can be in higher position in document ranking than the other positions.

Preference value of document that include chosen group preference options is inversely proportional to the other document that not include user preference. So, we can use α as multiplier for preference value of document that include particular user preference and $1-\alpha$ as multiplier for preference value other document. It works when α=1. But, when α=0, preference value of document that not include particular user preference it will have a higher value than particular document in user preference. So, we need to divide α by 2 and add by 0.5. Therefore, when α=0, term weight will be balanced between document include user preference and other document which are not included. It indicates no one in priority.

The IPF weight multiplied by $\alpha/2+0.5$ when document include to chosen group as user preference (UP) and the term document equal with term query (Q). So the greater value of α, the term will be given higher weight. Conversely, when the document is not included in the category of user preference and term document equals with term query, the IPF weight multiplied by $1-(\alpha/2+0.5)$. So the greater value of α, the term will be given lower weight. Meanwhile, if the term document is not equal with term query, weight IPF is only given without value α as a multiplier.

Combinations of each weight give various term weighting methods. In this research, we used $TF.IDF.IBF.IPF_\alpha$ term weighting method by multiplying each weights as defined in equation(6).

$$W_{TF.IDF.IBF.IPF_\alpha}(t_i,d_j,b_k,p_l) = W_{TF.IDF}(t_i,d_j)\times W_{IBF}(t_i,b_k)\times W_{IPF_\alpha}(t_i,d_j,p_l) \quad (6)$$

Vector Space Model (VSM)

Vector space model represents a document or query as a vector in a space terms [18]. This space has a dimension as the number of terms. In other words, to document that have the N terms is required N dimensions. Term in vector space model is taken from the entire unique terms in the whole documents. Each vector is represented according to the weight of the existing term. The existing term denotes the coordinate axes according to the number of term. While vector is a point whose position is based on the values of weight term being coordinate axes. Document vector representation shown in Figure 2.

Similarity Measure

One of the popular text similarity measure is the cosine similarity [19]. This measure calculate the cosine value of the angle between two vectors. In Figure 3, there are three vectors of documents d1, d2 and d3 and the query vector q. cosine similarity calculate the cosine value θ of the query and three other documents. This value indicates the degree of similarity with the query document.

Cosine similarity value is between 0 and 1, where 0 indicates that the document did not match at all, and 1 indicates that between the query and the document completely match. Cosine similarity measurement is defined as equation(7).

TABLE 1
Confusion Matriks

		Manual Document retrieval	
		Relevant according to user Preference	Relevant not according to user Preference
Document retrieval using TF.IDF.IBF.IPF$_\alpha$	Retrieved	*True Positive (tp)*	*False Positive (fp)*
	Not Retrieved	*True Negative (tn)*	*False Negative (fn)*

$$\cos(q, d_j) = \frac{\sum_{t_i}[w(t_i, q)] \cdot [w(t_i, d_j]}{\sqrt{\sum|w(q)|^2} \cdot \sqrt{\sum|w(d_j)|^2}} \quad (7)$$

where *cos(q,d$_j$)* denotes cosine value between query and document *j*, $\boldsymbol{w(t_i, q)}$ denotes weight TF. IDF.IBF.IPF for term t_i in query, $\boldsymbol{w\,(t_i, d_j)}$ denotes weight TF.IDF.IBF.IPF$_\alpha$ according to equation (6) for t_i in document *j*, $\sqrt{\sum|\boldsymbol{w(q)}|^2}$ and $\sqrt{\sum|\boldsymbol{w(d_j)}|^2}$ each is denotes vector length of query and vector length of document *j*. For the example, vector length of document are $\sqrt{\sum|\boldsymbol{w(d_j)}|^2}$ = (TF.IDF.IBF. IPF$_{\alpha t1}{}^2$ + TF.IDF.-IBF.IPF$_\alpha t_2{}^2$ + TF.IDF.IBF.IPF$_\alpha t_3{}^2$ +...+ TF.IDF.-IBF.IPF$_\alpha t_i{}^2$)$^{1/2}$, where TF.IDF.IBF. IPF$_\alpha t_i$ are weight of term t_i in document d_j vector.

3. Results and Analysis

To evaluate our proposed method, we use recall and precision by adjusting variables. So we can determine the ability of proposed method to retrieved relevant document according user preference. The adjusted variables in recall and preci-sion are shown in Table 1. Recall is defined as aquation(8) and precision is defined as equation(9).

For experiment, we use several queries that shown in Table 2. Each query is a fragment taken from a document in group preference options. The query has some relevant documents in a variety of group preference options. Each query is combined with user preference options that shown in Table III. So, there are variation input such as Q1_P1, Q1_P2, ..., Q6_P3, Q6_P4.

In this experiment we compared proposed method with TF.IDF and TF.IDF.IBF term weighting method. For TF.IDF.IBF.IPF$_\alpha$ term weighting method, we use α=0.9. Based on experiment, α value is influence the level of user preference in document ranking. Higher value of α makes the documents included at the user preference are in the top rank. Figure 4 shows the average value of considered when providing a high recall and having the ability to retrieve documents. The highest average recall is achieved when α = 1, reaching 100%. the recall increases. In contrast, the average value of precision decreases. While the

TABLE 2
Query Experiment

#	*Query*
Q1	استعمال آنية من لا تحل ذبيحته *isti'malu aaniyatu man laa tahillu dzabiihatahu* Use the vessels of not *halal* the sacrifice
Q2	الماء الذي ينجس والذي لا ينجس *al-maa.ul-ladzi yanjisu wal-ladzi laa yanjisu* unclean water and not unclean
Q3	الماء المسخن *al-maa.ul-musakhkhon* heated water
Q4	الماء المشمس *al-maa.ul-musyammas* sunny water
Q5	الوضوء من النوم *al-wudhuu.u minan-naumi* ablution after sleep
Q6	غسل الجنابة *ghoslu al-janaabatu* wash janaabah

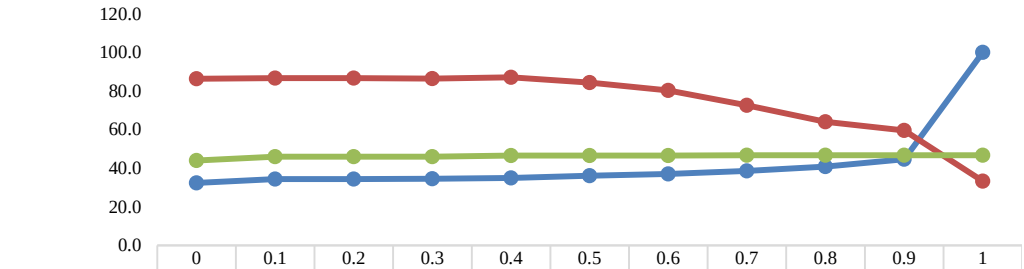
TABLE 3
User Prefrence Options

#	User Preference Options
P1	*Hanafiyah*
P2	*Malikiyah*
P3	*Syafi'iyah*
P4	*Hanabilah*

average F-mea-sure stable. In this research, the optimal value of α However, at α = 1 the average precision is only worth 33.3% because the system only retrie-ving documents that include user preference categorize and ignore other. Therefore, α is considered optimal with α is at 0.9 with an average recall of 44.7%, the average precision of 59.6%, and the average F-measure of 46.7%.

Result of experiment shown in Figure 5, Figure 6, and Figure 7. Figure 5 shows a comparison recall value between proposed method and other methods. From that figure we know, recall values of proposed method are higher than other method in 13 various input. For 5 other various input, recall TF.IDF.IBF.IPFα method is equal with TF.IDF and TF.IDF.IBF method. And 6 various input are no result founded. Among the queries that generate high recall is Q1_P4 with 9.5% value, Q2_P1 with 18.2% value, Q2_P2 with 15.9% value, Q2_P3 with 22.5% value, Q2_P4 with 40.9% value higher than others method, and so on. Highest recall TF.IDF.-IBF.IPFα method is 80% for input Q2_P3. This prove TF.IDF.IBF.IPFα method is capable to rank the document according to user preference rather than existing term weighting method.

Figure 6 shows a comparison between the precision value TF.IDF.IBF.IPFα method TF.IDF and TF.IDF.IBF method. From these image it is known that the TF.IDF and TF.IDF.IBF method has higher precision value compared with TF.-



	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Recall	32.4	34.4	34.4	34.5	35.0	36.1	37.0	38.6	40.8	44.7	100.0
Precision	86.3	86.7	86.7	86.4	87.0	84.4	80.3	72.6	64.0	59.6	33.3
F-measure	44.0	46.0	46.0	46.0	46.5	46.5	46.5	46.7	46.7	46.7	46.7

Figure 4. Average of recall, precision, and f-measure in various α value.

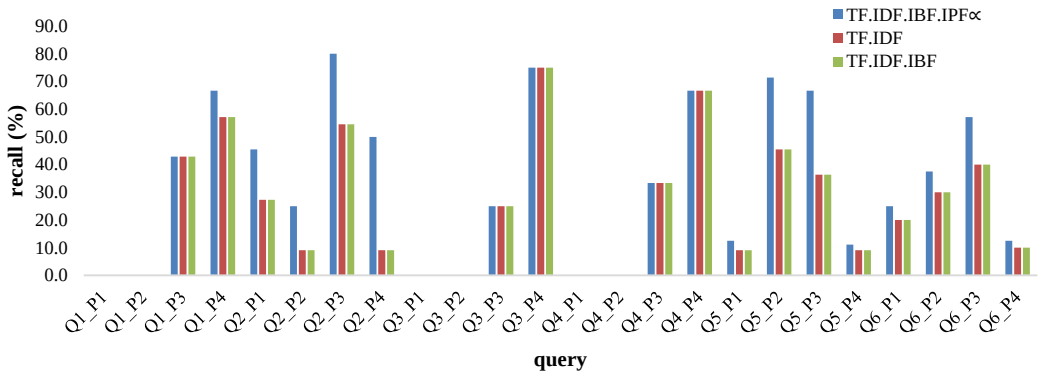


Figure 5. Comparison recall value.

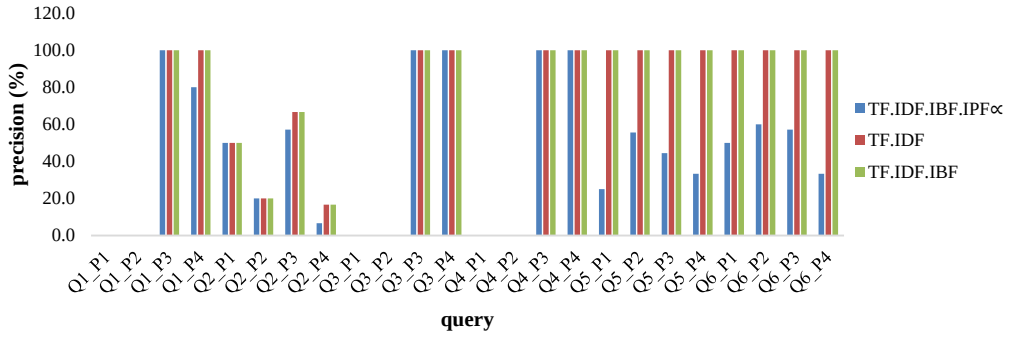


Figure 6. Comparison precision value.

IDF.IBF.IPFα method, i.e. at 11 queries used, while 7 other queries have the same precision value. This is because TF.IDF.IBF.IPFα method though relevant documents, but if it is not in according to user preference then the document will be placed on the lower order than the relevant documents according to user preference

Figure 7 shows a comparison between the F-measure value TF.IDF.IBF.IPFα method TF.IDF and TF.IDF.IBF method. Based on these images there are 3 inputs with TF.IDF.IBF.IPFα method which gives the higher value of F-measure than TF.IDF and TF.IDF.IBF method. For 15 other inputs, TF.IDF.IBF.IPFα method gives F-measure value equal to two other methods. The three inputs are Q2_P1 with 12.3% higher value, Q2_P2 with 9.7% higher value, and Q2_P3 with 6.7% higher value than the two other methods.

Overall, based on the comparison of the recall, precision, and F-measure between TF.IDF. IBF.IPFα method with TF.IDF and TF.IDF.-IBF method it can be concluded that TF.-IDF.IBF.IPFα method is affords to Arabic *fiqh* documsent ranking according to user preference rather than TF. IDF or TF.IDF.IBF. It is also evident from the average F-measure. TF.IDF.-IBF.IPFα method has average F-measure of 46.7% while the two other methods worth 45.1%. TF.IDF.IBF.IPFα has max-

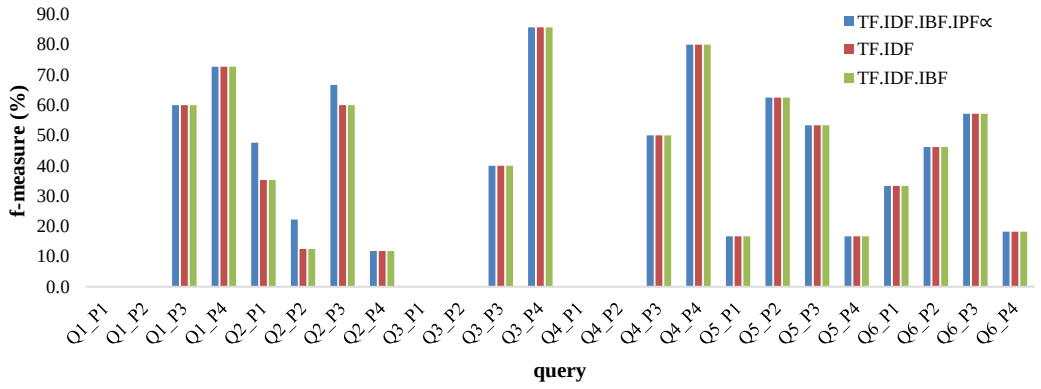


Figure 7. Comparison f-measure.

imum recall value 75%, precision 100%, and F-measure 85.7% for input Q3_P4.

4. Conclusion

From experiment of this research we can conclude that TF.IDF.IBF.IPFα method is capable to Arabic *fiqh* document ranking according to user prefernce with value of recall, precision and F-measure reached 75%, 100%, and 85.7% for Q3 query Q3 with chosen user preference options is P4. Proposed method may also can be implemented to rank other document that have a preference aspect.

However, preference in this study is based on a single preference. Therefore, further research can be done to fulfill the needs of multi preferences. Suitability of search results is also determined by the user query, further query expa-nsion is needed to improve search capabilities.

References

[1] C.D. Manning, P. Raghavan, & H. Schütze. *Introduction to Information Retrieval*, vol. 1. Cambridge: Cambridge university press, 2008.

[2] F. Harrag, A.H. Cherif, & E.E. Qawasmeh. "Vector Space Model for Arabic Information Retrieval—Application to “Hadith” Indexing." *Applications of Digital Information and Web Technologies, 2008. ICADIWT 2008. First International Conference on the*, IEEE, 2008.

[3] G. Salton & C. Buckley. "Term-Weighting Approaches in Automatic Text Retrieval." *Information Processing & Management* 24.5, pp. 513-523, 1988.

[4] S. Gerard. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of*. Addison-Wesley, 1989.

[5] F. Ren & M.G. Sohrab. "Class-indexing-based term weighting for automatic text classification." *Information Sciences* 236, pp. 109-125, 2013.

[6] M. A. Fauzi, “Term Weighting based on Book and Class Indices for Arabic Document Ranking,” B.S Thesis, Depart-ment of Informatic Engineering, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember, Indonesia, 2013.

[7] R. Chulyadyo & P. Leray. "A Personalized Recommender System from Probabilistic Relational Model and Users’ Preferences." *Procedia Computer Science*, vol. 35, pp. 1063-1072, 2014.

[8] D. Bogdanov, M. Haro, F. Fuhrmann, A. Xambo, E. Gommez, & P. Herrera, "Semantic audio content-based music recommenddation and visualization based on user preference examples", *Information Processing & Management* 49.1, pp. 13-33, 2013.

[9] T. Tegegne and Th P.V.D. Weide. "Enriching queries with user preferences in healthcare", *Information Processing & Management* 50.4, pp. 599-620, 2014.

[10] G.J. Hahm, M.Y. Yi, J.H. Lee, & H.W. Suh, "A personalized query expansion approach for engineering document retrieval", *Advanced Engineering Informatics* 28.4, pp. 344-359, 2014.

[11] A. Miele, E. Quintarelli, E. Rabosio, L. Tanca, "A data-mining approach to preference-based data ranking founded on contextual information", *Information Systems*, vol. 38, pp. 524-544, 2013.

[12] S.D. Amo, M.S. Dialo, C.T. Diop, A. Giacometti, D. Li, & A. Soulet, "Contextual preference mining for user profile construction" *Information Systems*, vol. 49, pp. 182-199, 2015.

[13] C. Boutilier, R.I. Brafman, C. Domshalk, H.H. Hoos, & D. Poole, "CP-nets: A Tool for Representing and Reasoning with Conditional Ceteris Paribus Preference Statements."

J. Artif. Intell. Res. (JAIR), vol. 21, pp. 135-191, 2004.

[14] S. Sulaiman, H. Mohamed, M.R.M. Arshad, N.A.A. Rashid, & U.K. Yusof, "Hajj-QAES: A Knowledge-Based Expert System to Support Hajj Pilgrims in Decision Making" *Computer Technology and Development, 2009. ICCTD'09. International Conference on*, vol. 1, 2009.

[15] A.M.A. Mesleh, "Chi Square Feature Extraction Based SVMs Arabic Language Text Categorization System" *Journal of Computer Science*, vol. 3, pp.430, 2007.

[16] L.S. Larkey, L. Ballesteros, & M.E. Connell, "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-Occurrence Analysis." *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2002.

[17] L.S. Larkey, L. Ballesteros, & M.E. Connell, "Light Stemming for Arabic Information Retrieval", *Arabic computational morphology*. Springer Netherlands, pp. 221-243, 2007.

[18] K.J. Cios, W. Pedrick, R.W. Swiniaraski, and L. Kurgan, "Data Mining: A Knowledge Discovery Approach", Springer, US, p.127, 2007.

[19] S. Tata, & J.M. Patel, "Estimating the selectivity of tf-idf based cosine similarity predicates." *ACM SIGMOD Record*, vol. 36, pp. 7-12, 2007.