Journal of
Information Systems Engineering
and Business Intelligence

Vol.5, No.2, October 2019
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

# Relevance Feedback using Genetic Algorithm on Information Retrieval for Indonesian Language Documents

**Salman Dziyaul Azmi [1], Retno Kusumaningrum[2] ***

*[1][2]Department of Informatics Universitas Diponegoro, Indonesia*
*Jl. Prof. H. Soedarto, SH, Semarang*
[1] s.azmi29@gmail.com, [2]retno@live.undip.ac.id

*Article history:*

Received 3 June 2019
Revised 21 September 2019
Accepted 2 October 2019
Available online 28 October 2019

*Keywords:*

Genetic Algorithm, Information Retrieval, Indonesian language document, Mean Average Precision, Relevance Feedback

***Abstract***

**Background:** The Rapid growth of technological developments in Indonesia had resulted in a growing amount of information. Therefore, a new information retrieval environment is necessary for finding documents that are in accordance with the user's information needs.
**Objective:** The purpose of this study is to uncover the differences between using Relevance Feedback (RF) with genetic algorithm and standard information retrieval systems without relevance feedback for the Indonesian language documents.
**Methods:** The standard Information Retrieval (IR) System uses Sastrawi stemmer and Vector Space Model, while Genetic Algorithm-based (GA-based) relevance feedback uses Roulette-wheel selection and crossover recombination. The evaluation metrics are Mean Average Precision (MAP) and average recall based on user judgments.
**Results:** By using two Indonesian language document datasets, namely abstract thesis and news dataset, the results show 15.2% and 28.6% increase in the corresponding MAP values for both datasets as opposed to the standard Information Retrieval System. A respective 7.1% and 10.5% improvement on the recall value at 10th position was also observed for both datasets. The best obtained genetic algorithm parameters for abstract thesis datasets were a population size of 20 with 0.7 crossover probability and 0.2 mutation probability, while for news dataset, the best obtained genetic algorithm parameters were a population size of 10 with 0.5 crossover probability and 0.2 mutation probability.
**Conclusion:** Genetic Algorithm-based relevance feedback increases both values of MAP and average recall at 10th position of retrieved document. Generally, the best genetic algorithm parameters are as follows, mutation probability is 0.2, whereas the size of population size and crossover probability depends on the size of dataset and length of the query.

## I. INTRODUCTION

Nowadays, the massive growth of technology has reshaped and transformed the way tasks are carried out in the digital area [1], where the ease of receiving and conveying information had now led to a larger pool of information resources and the increased use of databases. Since information retrieval has now become a necessary tool for users to locate and retrieve important information that is relevant for their use and purpose, a more comprehensive information retrieval system is therefore needed for supplying the right information that meets the user's specific needs.

The initial process of an information retrieval process is an input query from users into the system. However, even with the query given by the user, the search process often provides a partial description of the actual information desired [2]. As a result, users would have to conduct a few more search queries in locating required documents as the search results had only matched a small portion of information to their needs [3]. One of the ways of addressing this problem is using relevance feedback.

Relevance feedback or also known as RF is a query optimization process, which increases the number of relevant documents that appear on the search results. Some of relevance feedback techniques that had been used in the information retrieval systems for Indonesian language documents were the Rocchio algorithm [4] and the Ide Dec-Hi and Ide Regular methods [5]. The Rocchio algorithm allows weights adjustment for the relative input of relevant

[*] Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print)
doi: http://dx.doi.org/10.20473/jisebi.5.2.171-182

documents and non-relevant documents. The weighting scheme is based on a normalized version of the document weights rather than the actual document weights themselves. However, the problem with these Rocchio variants is that the weighting schemes for Rocchio formulation must be a static one, it means the weights is appropriate only for the respected corpus. However, it is often requiring fine-tuning parameters for the specific query and for its respected corpus. It is assumed that the tuned parameters are the optimal ones for all users, which may not necessarily be the case [6]. The Ide Dec-Hi and Ide Regular methods are more effective than the Rocchio method, as both had based on a modified Rocchio formula in forming a new query. While Ide Dec-Hi method uses all the identified relevant items and one retrieved non-relevant item for query modification, the Ide Regular method, on the other hand, utilizes previously retrieved non-relevant documents and does not provide weights like the Rocchio method [5]. These standard optimization techniques, however, are not used as the newly generated query can be assumed to be true in one context but not optimal in another and must be verified by further optimization algorithms. For this reason, genetic algorithm is the best-fitted optimization technique since it had been previously applied in several expansion contexts with great success [7].

There had been several studies conducted on the application of genetic algorithms for relevance feedback in the information retrieval system [8][7][9][10]. Since the method had been found to improve the performance of the information retrieval system by providing more relevant search results to the user, this implies that genetic algorithms can also be used to support the information retrieval for Indonesian language documents. However, as the Indonesian language must be subjected to the crucial stemming process, this study had used the Sastrawi stemmer to transform the word variants to their common root word [11]. Since some of the Indonesian text documents also contain many important English terms that describe the contents of the documents, the English words will not be subjected to stemming under the use of an Indonesian stemmer as a way of avoiding a different search result.

Therefore, this study had attempted to compare the performance of an information retrieval system with relevance feedback using genetic algorithms and those systems without relevance feedback on the Indonesian language documents.

## II. Methods

### A. Information Retrieval System

Information retrieval system is a system that can be used to obtain the information resources that are relevant to an information need from a collection of information resources [8] with the application of a search engine. In search engines, users evaluate the relevance of the search results displayed and will stop the search if the results had suited their needs, otherwise, the user will continue to alter the existing keywords until the displayed information had matched their desired requirements [3]. Fig. 1 shows a portion of the information retrieval system.

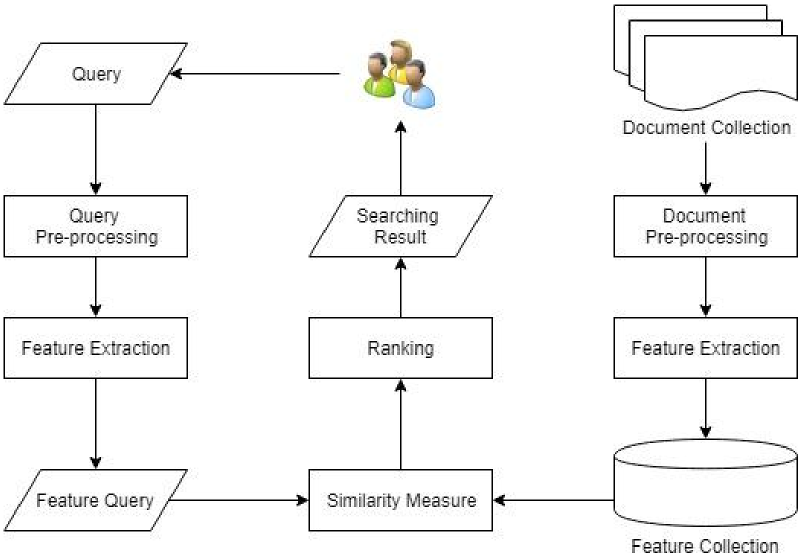


Fig. 1 Architecture of General Information Retrieval System

### *B. Pre-processing*

Pre-processing is a process that is used to transform text documents into a simple word format. In this study, the three steps of pre-processing, namely tokenization, stop word removal and stemming, are used in the analysis of the Indonesian language documents.

Tokenization is a step to split longer strings of text into smaller pieces or tokens. Given a character sequence and a defined document unit, tokenization not only involves the cutting of text documents into words, but also the removal of certain characters, such as punctuation marks and symbols. Tokenization also converts the texts into lower cases or case-folding as a way of homogenizing the array of words in each document so that the words that have the same meaning but in different case forms are not considered as two different words. As an example, this will allow instances of "Technology" at the beginning of a sentence to match with a query of "technology" after case folding. Punctuation marks, symbols and so on are also omitted in this step to produce results that contain exclusively of words.

Stop word removal is an important pre-processing technique involve the removal of a commonly used word. In the pre-processing stage, the elimination of stop words not only result in a reduction of the term space dimensions, but the presence of a stop word can also make the text weight appear larger than it should be. Since stop words had mostly consisted of articles, prepositions, and pronouns that do not give a specific meaning to the document [12], this study had included a list of the most used Indonesian stop words such as "dan" (and), "atau" (or), "di" (in, on, at), etc.

Stemming is commonly applied at index time to reduce morphological variants to a common root form. The purpose of stemming is to create a match between the user query and document terms that are related to the same meaning, but which appear in different morphological forms [12]. For example, the word "penyaringan" will be changed to its basic word of "saring". In this study, the stemming process had used the Sastrawi stemmer, which is a high-quality special stemmer library developed for the Indonesian language. The algorithm for stemming in Sastrawi library had been based on the Nazief-Adriani algorithm and was further developed by the Confix Stripping (CS) Enhanced Confix Stripping (ECS) and the Modified ECS algorithms [11][13][14].

### *C. Vector Space Model and Feature Extraction*

The basic idea of the Vector Space Model is that a document can be represented as a vector of keywords, where the comparisons of documents can be made by cosine similarity and dimensionality reduction [15]. The Vector Space Model was first introduced by Salton [16] and since then, had become one of the most widely applied retrieval models for evaluating the relevance of a web page.

In information retrieval, the term frequency-inverse document frequency (also called TF-IDF) is a well-known method for evaluating how important a word is in a document. The TF-IDF value generally increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some of the words may appear more frequently than the others [17].

The TF-IDF formula is shown as follows [2]:

$$w_{i,j} = tf_{i,j} \times idf_j \tag{1}$$

$$idf_j = log \frac{N}{df_j} \tag{2}$$

where:

$w_{i,j}$ = weight of the $j$-th word in the $i$-th document

$tf_{i,j}$ = term frequency of the $j$-th word in the $i$-th document

$idf_j$ = inverse document frequency of the $j$-th word

$df_j$ = number of documents containing the $j$-th word

$N$ = number of documents in the collection

$i$ = document index

$j$ = word index

After the weight of the word has been determined, a ranking function is subsequently required to measure the similarity between the queries and the document vectors. In the Vector Space Model, the cosine similarity is used by calculating the angle between the query vector and the document vector when represented in a V-Euclidean space, where V is the number of vocabularies.

The cosine similarity formula is shown below [2]:

$$sim\left(q,d_i\right)=\frac{d_i \square q}{\left|d_i\right|\times\left|q\right|} \tag{3}$$

Where:

$sim\left(q,d_i\right)$ = similarity of query $q$ with $i$-th document

$d_i$ = $i$-th document vector

$q$ = q query vector

$|\,d_i\,|$ = length of vector $i$-th document

$|q|$ = length of the $q$ vector query

$i$ = document index

### D. Relevance Feedback using Genetic Algorithm

The purpose of including relevance feedback in the information retrieval system is to involve the user in the retrieval process as a way of improving the result set. Based on the feedback given by the user on the relevance of the documents from the initial set of results [6], the system will then compute a better representation of the information need and display a revised set of retrieval results. Fig. 2 show the architecture of GA-based Relevance Feedback.

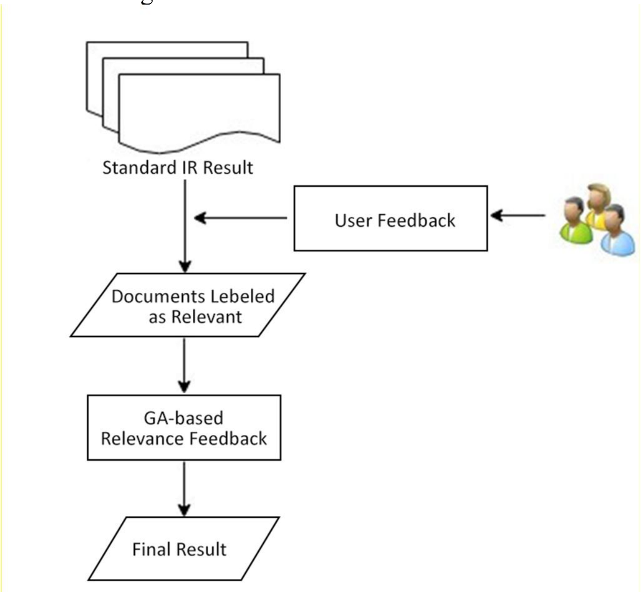


Fig. 2 Architecture of GA-based Relevance Feedback [9]

Genetic algorithm is a probabilistic algorithm that produces an approximate solution to a problem through genetic selection and is often used as an optimization method in solving problems where not much is known about the objective function within a large search space. While many of a genetic algorithm's processes are random, this optimization technique, however, allows one to set the level of randomization and the level of control. Therefore, these processes are far more powerful and efficient than other random search and exhaustive search algorithms. Since genetic algorithm had displayed a more powerful search mechanism in terms of its robustness and quick search capabilities in a large and complex search space, it was therefore incorporated as part of the information retrieval process. As such, by employing the evolutionary principles of natural and genetic selection as well as the manipulation of a population of queries rather than a single query, genetic algorithms typically provide the best solution to the problems, which in this case is producing the information that matches the user's needs.

The genetic algorithm was first proposed by John Holland in 1975 and since they are evolutionary algorithms that had replicated the occurrence of biological evolutionary processes in living things [9], they had been successfully implemented in various fields of research such as in the science and engineering industries [18].

Basically, genetic algorithm is the process of manipulating a chromosome population, which is a representation of the solution needed so that the best solution for the existing problem is obtained. In the manipulation stage, the genes of the parents are mixed and recombined via crossovers or mutations to produce offspring in the next generation.

*1) Chromosome:* In genetic algorithm, this term refers to a number of values representing candidate solution for problem to solve using genetic algorithm. Firstly, a population of chromosomes is randomly generated. Then, a genetic algorithm carries out fitness-based selection and recombination in order to yield a successor population. During recombination phase, some genetic material of chosen parent chromosomes is recombined for producing child chromosomes. This process continuous to repeat for a sequence of evolved successive generations. During the process, the average fitness of the chromosomes tends to increase before the repetition stop for a given criterion. These chromosomes can be bit strings, real numbers, permutation of elements, lists of rules, program elements or any various data structures [9]. On a chromosome, there are several elements called genes with values in them called alleles [18].

*2) Fitness:* The fitness function that is used in the application of genetic algorithms in the information retrieval system is shown below [19]:

$$F_{q,\,C} = \frac{\sum_{\forall d_k \in A(q)} \frac{1}{d_k \text{ position on } Ranking(q,C)}}{\sum_{k=1}^{N_{A(q)}} \frac{1}{k}} \tag{4}$$

Where:

$F_{q,C}$ = the $C$ chromosome fitness value in q query

$A(q)$ = relevant documents according to the user

$d_k$ = $k$-th relevant document

$N_{A(q)}$ = number of relevant documents in $A(q)$

$Ranking(q,C)$ = document ranking that appears to query with C chromosome

$k$ = index of relevant document in $A(q)$

*3) Selection:* Selection is the process of randomly selecting chromosomes that are based on their fitness values, where chromosomes with high fitness values are most likely to survive in populations than those with a low fitness value [20]. The most common method used in the selection process is the Roulette-wheel or the fitness proportional selection (see Fig. 3). This method involves allocating each chromosome with a probability of being selected as proportional to its relative fitness, where the relative fitness is the respective chromosome's fitness as a proportion of the summed chromosomes' fitness values in the population [18]. This technique is analogous to a roulette wheel with each slice being proportionally sized to their fitness values. Implementation of the roulette-wheel technique is to calculate the probability of selection for each chromosome based on its fitness value. Then the cumulative probability is calculated and used as a percentage of chromosome selection in the wheel in the selection process [21].

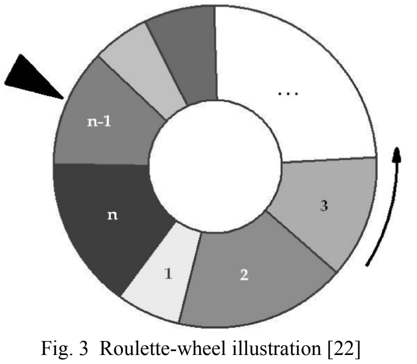
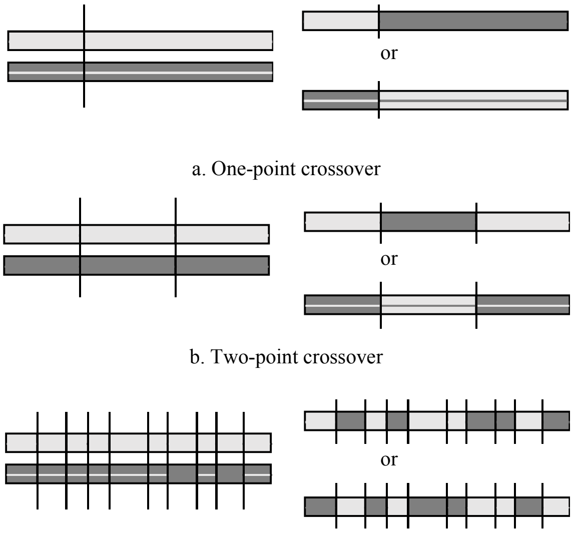


Fig. 3 Roulette-wheel illustration [22]

*4) Recombination:* Recombination is the selected chromosomes from a source population are recombined to produce members of a successive population. This is done by mixing the genetic material in the reproduction phase. Recombination consists of two main components, being crossover and mutation [18].

Crossover denotes combination process of the bits from some different chromosomes in order to create an offspring for successive generation. The child must inherit combined characteristics of its parents [18]. Crossover can be categorized into three types, namely one-point crossover, two-point crossover and uniform crossover [23] as illustrated in Fig. 4.



a. One-point crossover

b. Two-point crossover

c. Uniform crossover

Fig. 4 Types of crossovers [9]

Mutation is a genetic operator that is used to maintain the genetic diversity of a population of chromosomes between generations and works by altering one or more gene values in a chromosome from its initial state [9]. In the case of chromosome representations, the most common mutation operator used is the uniform mutation. The uniform mutation operator works by changing the value of a chosen gene with a uniform random value that was selected from the user's specified upper and lower bound for that gene [24]. The uniform mutation forms new values on the chromosome as a way of increasing its genetic diversity.

### *E. Evaluation*

The evaluation methods that are commonly used in the information retrieval systems are precision and recall. The formulas for both precision and recall are shown as such [8]:

$$precision = \frac{number\ of\ retrieved\ relevant\ documents}{number\ of\ retrieved\ documents} \tag{5}$$

$$recall = \frac{number\ of\ retrieved\ relevant\ documents}{number\ of\ all\ relevant\ documents} \tag{6}$$

Another evaluation method that is used in this study is the Mean Average Precision (MAP), which shows the mean of the average precision scores for each query. The formula of the Mean Average Precision Formula is shown below [6]:

$$MAP(Q) = \frac{1}{|Q|}\sum_{n=1}^{|Q|}\frac{1}{relDoc_n}\sum_{p=1}^{relDoc_n} precision(R_{np}) \qquad 7)$$

Where:

$MAP(Q)$ = mean average precision for $Q$ query set

$|Q|$ = the number of queries in a $Q$ query set

$relDoc_n$ = the number of relevant documents retrieved for the $n$ query

$n$ = *query* index

$p$ = retrieved relevant document index

$precision(R_{np})$ = precision for the $p$-th relevant document on the $R$ search results ranking for $n$ query

### F. Dataset and Experimental Scenarios

In this study, the datasets that had been used in this study were the abstract thesis datasets that consisted of 513 thesis documents from the Department of Computer Science / Informatics at Universitas Diponegoro and the news documents dataset, which comprised 1,000 news documents that were obtained from online news sites such as Detik, Kompas, Oso Securities and HaloMoney. There are two scenarios of experiment in this study.

*1) First Experiment:* The first test had involved the performance comparison of information retrieval without relevance feedback and relevance feedback with genetic algorithms. Apart from the top-10 mean average precision and the 10th recall position, this comparison test had also used the 8 parameter combinations as shown in Table 1.

TABLE 1
PARAMETER COMBINATIONS

| Combination-ID | Population Size | Crossover Probability | Mutation Probability |
|---|---|---|---|
| K-1 | 10 | 0.5 | 0.02 |
| K-2 | 10 | 0.5 | 0.2 |
| K-3 | 10 | 0.7 | 0.02 |
| K-4 | 10 | 0.7 | 0.2 |
| K-5 | 20 | 0.5 | 0.02 |
| K-6 | 20 | 0.5 | 0.2 |
| K-7 | 20 | 0.7 | 0.02 |
| K-8 | 20 | 0.7 | 0.2 |

In each of the dataset, this study had used the 5 queries as shown in Table 2 and Table 3.

TABLE 2
QUERY FOR THESIS ABSTRACT DATASET

| No. | Indonesian Query |
|---|---|
| 1 | Aplikasi Berorientasi Objek Unified Process<br>(Object Oriented Application Unified Process) |
| 2 | Jaringan Saraf Tiruan Learning Vector Quantization<br>(Artificial Neural Network Learning Vector Quantization) |
| 3 | Data Mining Menggunakan Naive Bayes<br>(Data Mining Using Naïve Bayes) |
| 4 | Steganografi Pada Citra Digital<br>(Steganography of Digital Image) |
| 5 | Prediksi dengan Backpropagation<br>(Prediction Using Backpropagation) |

TABLE 3
QUERY FOR NEWS DATASET

| No. | Indonesian Query |
|---|---|
| 1 | Nilai Tukar Rupiah<br>(Rupiah Exchange Rate) |
| 2 | Indeks Harga Saham Gabungan<br>(Composite Stock Price Index) |
| 3 | Gubernur Jawa Tengah<br>(Governor of Central Java) |
| 4 | e-Commerce<br>(e-Commerce) |
| 5 | Transfer Pemain Bola<br>(Soccer Player Transfer) |

*2) Second Experiment:* The second test was conducted to determine the best genetic algorithm parameters that can be used in the relevance feedback of the Indonesian document information retrieval system. The parameters that were tested are shown in Table 1.

## III. RESULTS

### *A. Result of First Experiment*

Fig. 5 and Fig. 6 show the MAP value for Information Retrieval System employed GA-based Relevance Feedback for two different datasets, i.e. dataset of Thesis dataset and News dataset, respectively.

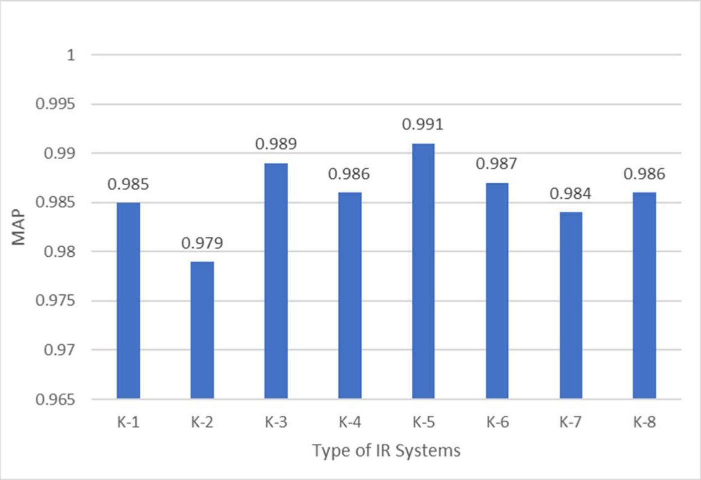


Fig. 5 MAP Graph for the Dataset of Thesis Abstract

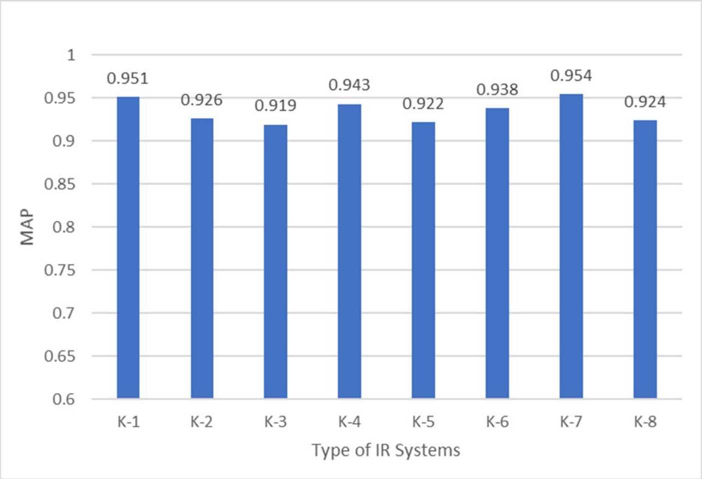


Fig. 6 MAP Graph for the News Dataset

Fig. 7 and Fig. 8 depict the average recall value for both datasets at the 10th position prior to using the relevance feedback method.

### *B. Result of Second Experiment*

The best genetic algorithm parameters were determined based on the average runtime or the time needed for the genetic algorithm to obtain chromosomes with the fitness value 1 and to display the relevance feedback results. Fig. 9 and Fig. 10 show the running time of GA-based Relevance Feedback for both datasets.

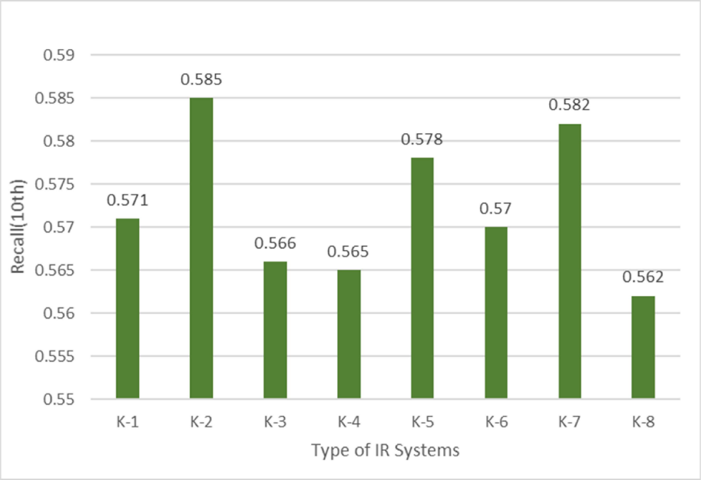


Fig. 7 Average recall (10th position) graph for abstract thesis dataset

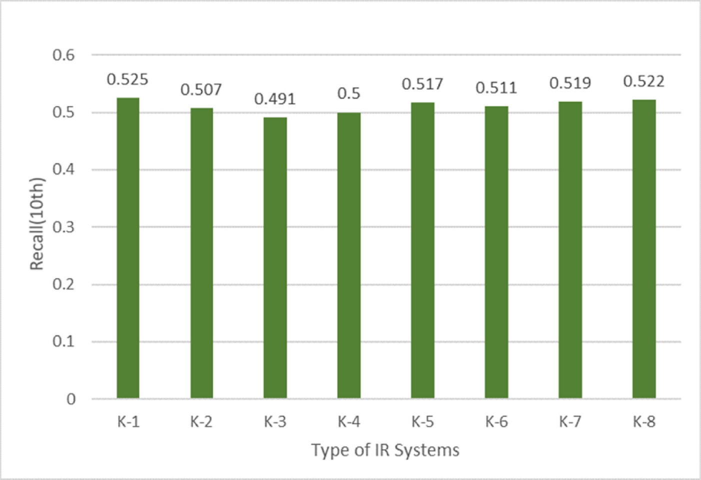


Fig. 8 Average recall (10th position) graph for news dataset

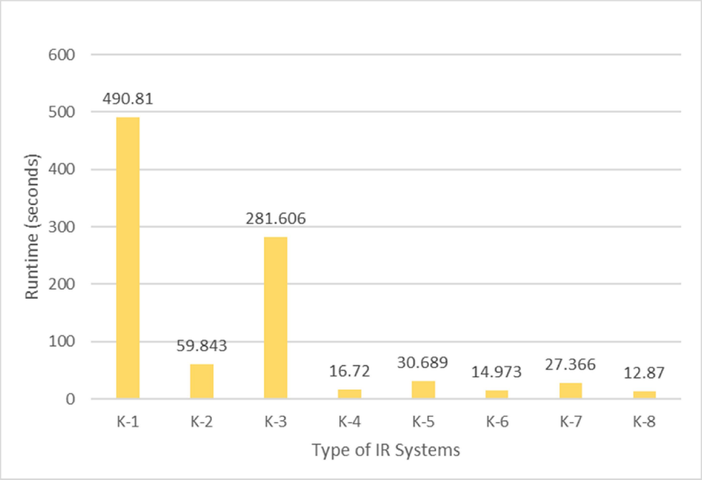


Fig. 9 Average runtime graph for abstract thesis dataset

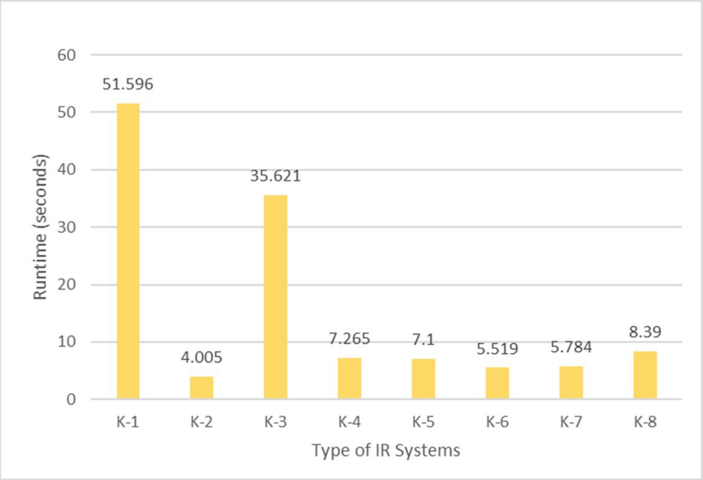


Fig. 10 Average runtime graph for news dataset

## IV. Discussion

### *A. Discussion of First Experiment*

As the comparison study, all the results from first experiment will be compared to the standard IR System or IR System without Relevance Feedback mechanism. The MAP and recall values of standard IR System for abstract thesis dataset are 0.834 and 0.501, respectively. In addition, the MAP and recall values of standard IR System for news dataset are 0.649 and 0.407, respectively.

Details of the difference between MAP values and recall values between the implementation of GA-based Relevance Feedback (RF) and Standard IR System can be seen in Table 4.

Table 4
Difference of MAP and Recall Values

| Dataset | Difference to Types of GA-based RF | MAP | Recall | Dataset | Difference to Types of GA-based RF | MAP | Recall |
|---|---|---|---|---|---|---|---|
| Abstract Thesis Dataset | K-1 | 0.151 | 0.070 | News Dataset | K-1 | 0.302 | 0.118 |
| | K-2 | 0.145 | 0.084 | | K-2 | 0.277 | 0.100 |
| | K-3 | 0.155 | 0.065 | | K-3 | 0.270 | 0.084 |
| | K-4 | 0.152 | 0.064 | | K-4 | 0.294 | 0.093 |
| | K-5 | 0.157 | 0.077 | | K-5 | 0.273 | 0.110 |
| | K-6 | 0.153 | 0.069 | | K-6 | 0.289 | 0.104 |
| | K-7 | 0.150 | 0.081 | | K-7 | 0.305 | 0.112 |
| | K-8 | 0.152 | 0.061 | | K-8 | 0.275 | 0.115 |
| | Average | 0.152 | 0.071 | | Average | 0.286 | 0.105 |

Note: K-1 through K-8 are parameter combinations as shown in Table 1.

As shown in Table 4, the usage of relevance feedback with genetic algorithms had resulted in an average increase of 15.2% MAP value for the abstract thesis datasets, whereas an average increase of 28.6% MAP value for the news dataset. Furthermore, the employed GA-based Relevance Feedback also increases the value of average recall values for both datasets, i.e. abstract thesis dataset and news dataset, are 7.1% and 10.5% respectively. Generally, the increase in the MAP value is observed from the usage of GA-based relevance feedback as the fitness function had ensured all the relevant documents from the initial search results to appear at the top. Hence, the obtained average MAP values had demonstrated higher or at least the same values as those from the initial search results or standard IR system. Moreover, this condition also allows for obtaining relevant documents that have not previously been retrieved by the standard IR system. Hence, it also increases the recall value in general.

### *B. Discussion of Second Experiment*

Based on the experiment, 12.87 seconds had been the lowest runtime obtained at K-8 (population size is 20, crossover probability is 0.7, and mutation probability is 0.2) for the abstract thesis dataset, while in Fig. 10, 4.005 seconds of lowest runtime was observed at K-2 (population size is 10, crossover probability is 0.2, and mutation probability is 0.2) for the news dataset. Since the action of natural selection on mutation rate is related to population size, the low mutation probability value that was shown by the above two images had indicated a lower mutation rate in low population size. As such, it can be said that the optimal mutation probability value for both datasets had been 0.2. Both datasets had also experienced different optimal population sizes and crossover probabilities. In the abstract thesis dataset, the optimal parameters obtained were a population size of 20 with a 0.7 probability of crossover, while for the news dataset, the optimal parameters had been from a population size of 10 with a 0.5 crossover probability.

However, one of the drawbacks of determining the optimal parameters based on runtime is that although the chromosomes that were obtained through a longer evolution process were more likely to provide more relevant search results, the evolution process would take a long time to complete if they are fed into a relevance feedback process that contains large quantities of words. In this study, the evolution process had taken a total of 6 hours to complete. Therefore, while runtime criteria are used in selecting the best genetic algorithm parameters, the relevance feedback process, however, would become less relevant in the information search process as a result of the long evolution process.

## V. Conclusions

In this paper, we have implemented the usage of relevance feedback with genetic algorithms on the information retrieval system for Indonesian language documents. The successful application of the same technique in related research with different subjects is concluded to provide similar results in Indonesian language documents. The implementation using Sastrawi stemmer and genetic algorithms with described parameter combination has been able to conclude an improvement in the information retrieval process.

Based on results of the experiment, (i) the GA-based relevance feedback increases both values of MAP and average recall at 10th position of retrieved document as shown in Table 4, (ii) the best genetic algorithm parameters for the abstract thesis datasets were obtained at a population size of 20 with a 0.7 crossover probability and a 0.2 mutation probability, while for the news dataset, the best genetic algorithm parameters obtained were with a population size parameter of 10, a 0.5 crossover probability and a 0.2 mutation probability.

From this study conducted on the relevance feedback with genetic algorithms in the information retrieval system for Indonesian documents, it is suggested that further research can be conducted by way of combining relevant documents with other techniques prior to the population formation process so that the reduced chromosome length had only required a shorter runtime process. Another research on stemming techniques can be conducted because there are many foreign terms in Indonesian language documents and how they affect the effectiveness of genetic algorithms.

## References

[1] Setiawan, W., 2017. *Era Digital dan Tantangannya.* Universitas Pendidikan Indonesia.

[2] Lee, D. L., Chuang, H. & Kent, S., 1997. Document Ranking and the Vector Space Model. *IEEE Software,* 14(2), pp. 67-75.

[3] Agbele, K., Adesina, A., Ekong, D. & Ayangbekun, O., 2012. State-of-the-Art Review on Relevance of Genetic Algorithm to Internet Web Search. *Applied Computational Intelligence and Soft Computing,* Volume 2012.

[4] Pamungkas, Z. Y., Indrianti & Ridok, A., 2015. Query Ekspansion pada Sistem Temu Kembali Informasi Dokumen Berbahasa Indonesia menggunakan Pseudo Relevance Feedback (Studi kasus: Perpustakaan Universitas Brawijaya). *Jurnal Mahasiswa PTIIK UB,* 6(3)

[5] Agusetyawan, A. W., Ridha Ahmad & Adisantoso, J., 2006. Relevance Feedback pada Temu Kembali Teks Berbahasa Indonesia dengan Metode Ide-Dec-Hi dan Ide-Regular. *Jurnal Ilmiah Ilmu Komputer,* 4(2).

[6] Manning, C. D., Raghavan, P. & Schutze, H., 2008. *An Introduction to Information Retrieval.* New York: Cambridge University Press.

[7] Perez-Aguera, J. R. & Santesmases, J. G., 2007. *Using Genetic Algorithms for Query Reformulation.* Glasglow, BCS Learning & Development Ltd.

[8] Erwin, M. & Mandala, R., 2004. Relevance Feedback pada Temu Kembali Informasi Menggunakan Algoritma Genetika. Yogyakarta, SNATI 2004.

[9] Kusumaningrum, R., 2012. *Reducing Semantic Gap Using GA-Based Relevance Feedback In Remote Sensing - Image Retrieval System.* Depok, Universitas Indonesia.

[10] Ligade, A. N. & Patil, M. R., 2013. Optimized Content Based Image Retrieval Using Genetic Algorithm with Relevance Feedback Technique. *International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR),* 3(4), pp. 49-54.

[11] Librian, A., 2014. *Sastrawi.* Github.

[12] Vijayarani, S., Ilamathi, J. & Nithya, 2014. Preprocessing Techniques for Text Mining - An Overview. *International Journal of Computer Science & Communication Networks,* 5(1), pp. 7-16.

[13] Tahitoe, A. D. & Purwitasari, D., 2010. *Implementasi Modifikasi Enchanced Confix Stripping Stemmer untuk Bahasa Indonesia dengan Metode Corpus Based Stemming,* Surabaya.

[14] Arifin, A., Mahendra, I. & Ciptaningtyas, H., 2009. *Enchanced Confix Stripping Stemmer and Ants Algorithm for Classifying News Document in Indonesian Language.* s.l., Proceeding of International Conference on Information & Communication Technology and Systems (ICTS).

[15] Chen, Y.-L. & Chiu, Y.-T., 2012. Vector Space Model for Patent Document with Hierarchical Class Labels. *Journal of Information Science,* 38(3), pp. 222-223.

[16] Salton, G., Wong, A. & Yang, C. S., 1975. A Vector Space Model for Automatic Indexing. *Communications of the ACM,* November, 18(11), pp. 613-620.

[17] Kumari, M., Jain, A. & Bhatia, A., 2016. Synonyms Based Term Weighting Scheme: An Extension to TF.IDF. *Procedia Computer Science,* Volume 89, pp. 555-561.

[18] McCall, J., 2005. Genetic Algorithms for Modelling and Optimisation. *Journal of Computational and Applied Mathematics,* Volume 184, pp. 205-222.

[19] da Silva, S. F., Batista, M. A. & Barcelos, C. A. Z., 2007. *Adaptive Image Retrieval through the use of a Genetic Algoritm.* 19th IEEE International Conference on Tools with Artificial Intelligence.

[20] Srinivas, M. & Patnaik, L. M., 1994. Genetics Algorithm: A Survey. *Computer,* 27(4), pp. 17-26.

[21] Mahmudy, W. F., 2013. *Algortima Evolusi.* Malang: Program Teknologi Informasi dan Ilmu Komputer Universitas Brawijaya.

[22] Sharapov, R. R., 2007. Genetic Algorithms: Basic Ideas, Variants and Analysis. *Vision systems: segmentation and pattern recognition*, pp. 407-422.

[23] Umbarkar, A. & Sheth, P., 2015. Crossover Operators In Genetic Algorithms: A Review. *ICTACT Journal On Soft Computing,* 6(1), pp. 1083-1092.

[24] Soni, N. & Kumar, T., 2014. Study of Various Mutation Operators in Genetic Algorithms. *International Journal of Computer Science and Information Technologies,* 5(3), pp. 4519-4521.