Development of Decision Related Engine Using Integration of Genetic Algorithm and Text Mining

EvianaTjatur Putri#1, Mardalena#2, Asmah#3

#Program StudiTeknikInformatika, STMIK PPKIA TarakanitaRahmawati
Jl. YosSudarso 8 Tarakan, Indonesia
1evianaputri@gmail.com
2mardalena.ahmad@gmail.com
3asmah_dp@yahoo.co.id

Abstract— **Text Mining as a popular method used to perform text-based information retrieval can not bedeniedits use. At the beginning of the textmining methodis based on the purpose of seeking to represent a collection of words from a document. Tokenizing, filtering, Stemming, Tagging andAnalyzing are the critical process sequencei n the methods of TextMining, which ultimately yield important information of the most dominantina document that analyzed.**

Applying Genetic Algorithms on TextMining method aims to create a Related Decision Engine, a machine capable of making decisions that relate among one another and the results of analysis carried out. Analysis carried out in this study is the analysis applied to digital libraries, the information starting from the synopsis, preface and the title ofthe book, which will be processed the information by the methods of TextMining. Users often only presented the information in accordance with what books you are looking for, without giving a few other books that likely her relationship with the books to be searched. Though likely will also be of interest to the user associated with the book that will besought. In this case, the users who will find the documents were also given another information that herrelationship with books that may also be an option in the user's search.

This application created a web-based, using the Apache Web Server and MySQL database to support the completeness of applications built on digital libraries.

Keywords— **Genetic Algorithms, Text Mining, Digital Library, Decision Related Engine, Apache Web Server.**

I. Introduction

The study took samples in the digital library. Where is the library that has a lot of literature the book has made the process of computerization of the identity of the books he owned. Data book title, author, up to a synopsis of the book have usually been recorded in the digital library.

Until now, the application is made to search for books at the library or the bookstore only similarity search method, namely by using the "like" in the SQL command. The results obtained are quite relevant to what is desired by the reader, but what happens if the book you are looking for does not exist by using the "like" them? Or the reader wants to find books that have the same topic with what you are looking for? The instruction can not perform actions like these.

This study tries to make a search step one book or several books that have the same subject matter with what is desired by the reader. A search form is capable of providing the decision to relate what is likely the same as the reader wishes.

As a simple example is if the reader wishes to find the book "PHP", then a standard search using the "like" will come up with a book with the same title or the synopsis that has elements of the word "PHP". The search process will end on the condition. In the related decision engine, namely engine perelasi decisions made in this study, will be raised also the result of several books on web programming. Programming language "PHP" as sought by the reader is about web programming. Therefore it would appear the results of several other books in the form of web programming, such as ASP (Active Server Pages) or JSP (Java Server Pages). The method used to obtain these results using genetic methods or commonly known as Genetic Algorithm (GA).

Recent Decision Engine is built in this study also provides an easy to search for books by giving a better approach by utilizing basic word search for the book you are looking for. The method used to search for books based on the basic word is Text Mining. Search on the word "programming" will lead to book deals with basic word "program". Books that have a title or synopsis using the base word, will be presented on a computer screen.

This application created a web-based, using the Apache web server, which used a lot on the web server on the Internet. With the selection of the many applications used by the web server, it is hoped this application would be easily attached to various server.

II. Theoritical Basis

2.1 Genetic Algorithm (GA)

Salvatore Mangano in the book Computer Design, in May 2005, stated that "Genetic Algorithms are good at taking large, Potentially huge search spaces and navigating them, looking for optimal combinations of things, solutions you Might not otherwise find in a lifetime.". GA developed by John Holland in 1970 at the University of Michigan, used to understand the adaptive processes of natural systems. GA is also used for design artificial systems software That retains the robustness of natural systems.

The process begins in the GA population initialization, subsequent to the evaluation process in each population. The next step is the iteration process to bring the condition of satisfaction on the results obtained. Iteration process is conducting the election "parent" or master of the population to make the process of reproduction. Further recombination and mutation activity, which ended in the evaluation process continued population.

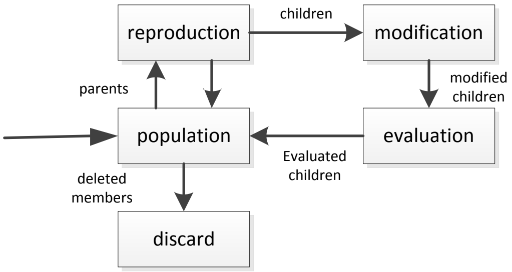


Fig.1 GA Cycles for Reproduction

Chromosome in the population there, which could be a number, a binary digit, rule or even a program element. Further selected parents (parent) are randomly air-relationship with chromosome to be evaluated, is called the reproduction process.

Next will be triggered by stochastic chromosome through the process of modification. Operators that affect chromosome modification is a mutation and recombination (crossover).

The next process, as shown in figure 1 is Evaluation. The evaluator will conduct process and decode the chromosome fitness value. The next issue can be resolved eventually.

The last process in the GA is Deletion. This process occurs on replacement of the entire population with the results that appear in each iteration.

2.2 Text Mining

Text Mining can be defined as a way to mine the data in the form of text, where the source data is usually in the form of documents. The purpose of this method is to find words that can represent the contents of the document, so as to analyze the relationship between the documents.

Stages in text mining are tokenizing, filtering, stemming, tagging and analyzing. Tokenizing the stage to cut the sentence into separate words are compiled. The second is the filtering process, which took the important words are generated from the Tokening. Get the important words can be a stop list (sign up words that will be discarded) or word list (sign up words that are important).

Stemming process to the next process, the process is quite complicated, because the search for basic words (root words) of each word. Proceed with the process of tagging, which is to seek forms of the words results stemming.

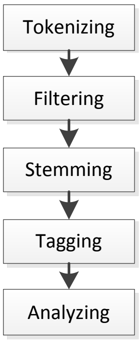


Fig.2Text Mining Cycles

Stages of relationships quest stages of determining how far is the connection between words and documents is the last stage, known for Analyzing methods.

2.3 Algoritma TF/IDF

This algorithm is part of the text mining. Analyzing the process used in the form of activity to perform the weighting (w) of each document on keywords.
The formula for this algorithm is as follows.

$$W_{d,t} = tf_{d,t} * IDF_t$$

Where as: d =dokumenke-d; t = kata ke-t dari kata kunci; W = bobotdokumenke-d terhadap kata ke-t

After the "w" is found, then do the sorting. A value of "w" means that the larger the closer the value of the document sought.

If there is a similarity value in the "w" or weight on the data book, it is not possible to do the sorting of search results, but should proceed with the application of the method of VSM (Vector Space Model). Formula for the application of VSM are as follows.

$$\text{Sqrt}(kk) = \text{Sqrt}(\textstyle\sum_{j=1}^{n} kk_j^2)$$

Where j is the value of words in the database. Further calculate the cosine angle between the keyword vector of each document with the following formula.

$$\text{Cosine}(D_i) = \text{sum}(kk \text{ dot } D_i) / [\ \text{sqrt}(kk) * \text{sqrt}(D_i)\ ]$$

III. DESIGN APPROACH

Data used in this study as the sample is 100 library book data, the data in the form of a book title, author and synopsis.

The next step is to perform the application of methods of Text Mining, beginning with the preparation of the stop list, then perform the tokenizing and filtering on the synopsis, which is then the result of both processes is stored in the table "TextMining". On the table there is a field "Filtering" and "Stemming" which contains a synopsis of the results of the filtering process and the process of finding the root of the words (stemming). Field is the key producer of information in the search process is performed.

Flowchart of the design of applications that can be observed in Figure 3.

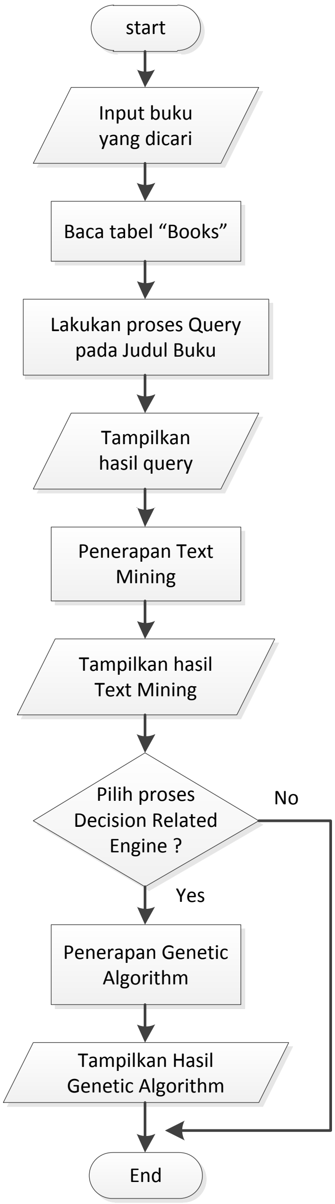
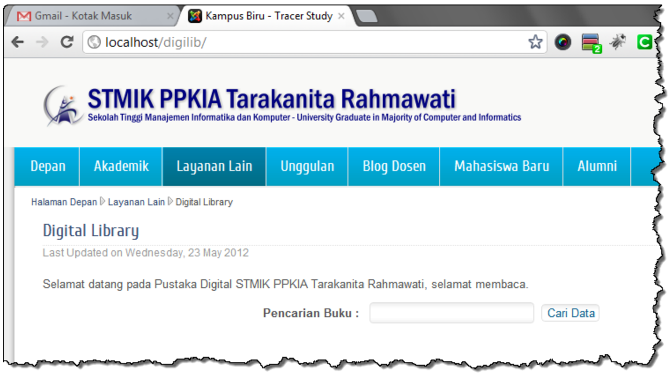

start
Input buku yang dicari
Baca tabel "Books"
Lakukan proses Query pada Judul Buku
Tampilkan hasil query
Penerapan Text Mining
Tampilkan hasil Text Mining
Pilih proses Decision Related Engine ?
No
Yes
Penerapan Genetic Algorithm
Tampilkan Hasil Genetic Algorithm
End

Fig. 3 Flowchart DesainAplikasi

Decision Process Related Engine, which become virtues in this study is an option for the user. If the user wants to search results, then the process will be implemented by applying a Genetic Algorithm, but if users are quite satisfied with the standard search results with a synopsis of the text mining, the process will not continue.

That is, if the user feels that the results have been simply using the results of the Text Mining, the search results with the method of Related Decision Engine can not be selected or omitted.


Gmail - Kotak Masuk
Kampus Biru - Tracer Study
localhost/digilib/
STMIK PPKIA Tarakanita Rahmawati
Sekolah Tinggi Manajemen Informatika dan Komputer - University Graduate in Majority of Computer and Informatics
Depan
Akademik
Layanan Lain
Unggulan
Blog Dosen
Mahasiswa Baru
Alumni
Halaman Depan ▷ Layanan Lain ▷ Digital Library

Digital Library

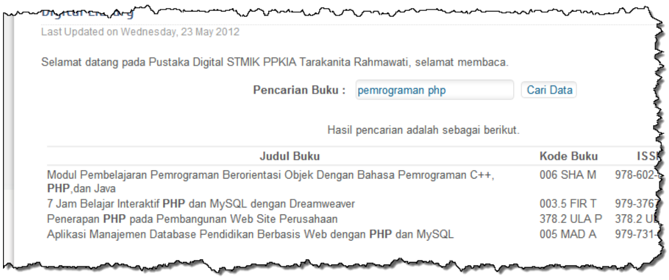
Last Updated on Wednesday, 23 May 2012

Selamat datang pada Pustaka Digital STMIK PPKIA Tarakanita Rahmawati, selamat membaca.

Pencarian Buku : Cari Data

Fig. 4 Web Design Interface of Digital Library

In Figure 4, presented the view to do a search on digital library data book. For example, the search is "Programming PHP", then the application will generate the information as shown in figure 5.



Last Updated on Wednesday, 23 May 2012

Selamat datang pada Pustaka Digital STMIK PPKIA Tarakanita Rahmawati, selamat membaca.

Pencarian Buku : pemrograman php Cari Data

Hasil pencarian adalah sebagai berikut.

Judul Buku	Kode Buku	ISS
Modul Pembelajaran Pemrograman Berorientasi Objek Dengan Bahasa Pemrograman C++, **PHP**,dan Java	006 SHA M	978-602-
7 Jam Belajar Interaktif **PHP** dan MySQL dengan Dreamweaver	003.5 FIR T	979-3767
Penerapan **PHP** pada Pembangunan Web Site Perusahaan	378.2 ULA P	378.2 U
Aplikasi Manajemen Database Pendidikan Berbasis Web dengan **PHP** dan MySQL	005 MAD A	979-731-

Fig. 5 Result for searching using Text Mining

Top results are the book "Learning Module Object Oriented Programming with the programming language C + +, PHP and Java", because the process of finding the weights (w) the search terms "php programming" will have a value of weight is greater, compared with titles another. Text Mining the calculation results obtained can be observed in Table 1.

TABLE I
RESULTS OF CALCULATIONS ON DATA SEARCH

token	tf					df	D/df	IDF
	kk	D1	D2	D3	D4			
Pemrograman	0	0	0	1	0	1	1.5	0.176
PHP	1	1	1	1	1	4	1.5	0.176

In the table, it can be observed, that the book D3 has two key words that match the desired search. Therefore it has a value of 2 for search results is.

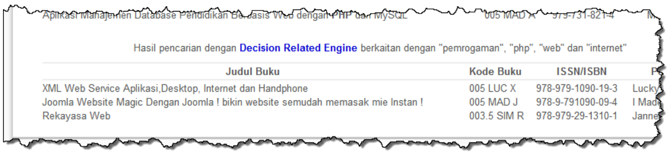
TABLE III
RESULTS OF WEIGHT CALCULATIONS ON DATA SEARCH

token	**IDF**	**w**					
		kk	**D1**	**D2**	**D2**	**D3**	**D4**
Pemro-gramam	0.176	0.176	0	0	0	0.176	0
PHP	0.176	0.176	0.176	0.176	0.176	0.176	0.176
Total	**0.352**	**0.352**	**0.176**	**0.176**	**0.176**	**0.352**	**0.176**

Book D1, D2 and D4 have the same value, then the process will be conducted Text Mining on the synopsis. The same process carried out during the process of Text Mining in the title (as is done in Table 1).

Thus the process is continued, if at the time Text Mining Synopsis still have the same value in the weighting, then count VSM, by finding the value of the word in question Cosine.

If the reader wishes to conduct the search process by the method of Related Decision Engine, it can press the button at the bottom of the program. The process will continue with the use of GA as the search for additional solution is sought in the book, as in figure 6.



Aplikasi Manajemen Database Pendidikan Berbasis Web dengan PHP dan MySQL 005 MAD A 979-731-821-4 M

Hasil pencarian dengan Decision Related Engine berkaitan dengan "pemrogaman", "php", "web" dan "internet"

Judul Buku	Kode Buku	ISSN/ISBN	P
XML Web Service Aplikasi,Desktop, Internet dan Handphone	005 LUC X	978-979-1090-19-3	Lucky
Joomla Website Magic Dengan Joomla ! bikin website semudah memasak mie Instan !	005 MAD J	978-9-791090-09-4	I Mad
Rekayasa Web	003.5 SIM R	978-979-29-1310-1	Janne

Fig. 6 Result of Decision Related Engine

How It Works Related Decision Engine is a search for other key words contained in the previous search results. Found two other keywords "web" and "internet", eventually combined searches for "programming", "php", "web" and "internet". If the search is performed using Text Mining it will take quite a long process, and therefore used Genetic Algorithm (GA).

IV. Conclusions

Get search results that her relationship with the information provided, is a form of new demand from users who are likely in the future will be done. In the digital library is a sample for this study. Search results displayed on the library books that relate to information to be searched, it bears no resemblance to the title, but the topics discussed read or have a relationship with a book that sought.

With the use of Text Mining and Genetic Algorithm (GA), the book search system was built. Not only in books, but in other conditions that likely can be applied to methods Related Decision Engine is done. For example, vehicle theft recidivist data search, data-relation is likely to air information about the convict fence stolen goods. Search vehicles corresponding to the types of vehicles in question, is the result of a form of applied Related Decision Engine.

References

[1] UnboShuai, Xiangguang Zhou. "A Genetic Algorithm Based on Combination Operators", Procedia Environmental Sciences, 2011, Vol 11, p.346

[2] GözdeBakırlı, DeryaBirant, Alp Kut. "An IncrementalGeneticAlgorithm for Classification and Sensitivity Analysis of Its Parameters", Expert Systems with Applications, 2011, Vol. 38, p.2609

[3] Huan-Yu Lin ; Jun-Ming Su ; Shian-Shyong Tseng. "An Adaptive Test Sheet Generation Mechanism Using Genetic Algorithm", Mathematical Problems in Engineering, 2012, Vol. 2012

[4] ZhanGangHao. "A New Text Clustering Method Based on KGA", Journal of Software. 2012. Vol. 7, P.1094

[5] Dr. Sanjay Tanwani ;NehaRahatekar ; ShrutiDubey ; DeepkaParmar. "Automated Personal Email Organizer with Information Management and Text Mining Application". International Journal of Computer Applications. 2012. Vol. NCRTC, p.9