Adaptive Model of Personalized Searches using Query Expansion and Ant Colony Optimization in the Digital Library

Wahyu Sulistiyo [1)]
Electrical Department, POLINES
Semarang, Indonesia
email:wahyu.sulistiyo@polines.ac.id

Bayu Surarso [2)] Aris Sugiharto [3)]
[2) 3)] MSI and Mathematics Department UNDIP,
Semarang, Indonesia
email: bayusurarso@yahoo.com [2)]
aris.sugiharto@undip.ac.id [3)]

***Abstract*--A system that can provide useful information for users will be able to increase user loyalty. Presentation of useful information can be done by providing the information needed by each user or personalization. Personalization is a form required when the system wants to interact with its users. A personalized system is built based on the needs of each user. In this research, a new personalization model was built for the interaction between digital library systems with users. The proposed model uses adaptive query expansion algorithm and ant colony optimization for searching documents in digital libraries.The model uses metadata to the search process. The metadata consists of titles, abstracts and keywords of the document. The User model is built to monitor user activities when selecting a document link (click) and visit page document (visit). Data of research uses 50 journal documents. The research results that the proposed model improves search scores by 60% and the order of the search documents up to 56%.**

Keywords*: *Adaptive Model* , *Query Expansion, Ant Colony Optimization

1. Introduction

A system that can provide useful information for users will be able to increase user loyalty system. Presentation of useful information can be done by providing the information needed by each user or personalization. Personalization is a form required when the system wants to interact with its users. A personalized system is built based on the needs of each user

Development of adaptive systems for web technology was done in [1], in which to serve the diverse web users, web site requires appropriate services for individual user desires. In [6] Model personalization with the knowledge base was created from the results of automatic machine learning. The model was built based on the model of personalized user content, navigation, information filtering and information retrieval. Other adaptive models were studied in [8] using knowledge base group with ACO method. The study developed the path adaptive learning system based on the log information from the data, in the form of attributes and activities of the learning object. In [7], ACO method was used to navigate the web. ACO algorithm imitated nature of ants in search of food. Research implementation was divided into three stages: build user profiles, classification of user profiles and personalize search results according to user profiles.

In the present study, ACO algorithm is used to assist in the search process of digital libraries with query expansion technique. An ACO algorithm that is utilized in this research is the Ant System algorithm to solve the Traveling Salesman Problem [2]. User model is built from the activity of the user Click and Visit

2. Review Of Literature

2.1 Ant Colony Optimization

The first Ant Colony Optimization method is Ant System (AS), made by Marco Dorigo in 1991. The characteristics of such algorithmare as follow [2]:
1. Versatile
2. Robust
3. Population Based

This algorithm can be compared to other optimization algorithms such as Simulated Annealing (SA) and Tabu Search (TS). Ant System algorithm mimics the characteristics of ants looking for food. At first (Fig. 1a), ants walk from the nest (A) to the food source (E) without passing through the barrier. Then the path is given the barrier (Fig. 1b). This condition causes the ants have to turn right (path ABCDE) or to the left (path ABHDE) to pass through the barrier. This choice will affect the intensity of the pheromone left by ants. At the time of the first cut, the ants (in position B and D) have the same possibility to choose the path because there is no pheromone trail left by previous ants. ABCDE path is shorter than the path ABHDE. Because pheromone changes according to the time, the intensity of pheromone left by ants to be higher in a shorter path. As a result, the next ant will tend to choose the path ABCDE that has a higher pheromone trail as Figure 1c.

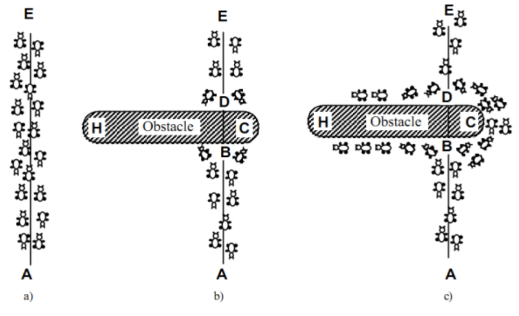

Figure 1 Experiments ants looking for food

Ant System (AS) can be used to solve the TSP with symmetric graph model. On this graph, the weight (cost) from city A to city B is equal to the weight of the city B to city A. Suppose Graph (N, E), where N is the set of cities and E is the set of edges (arcs) that connect each town with all other cities (fully connected graph). Let b_i (*t*) = (*i*=1,2,…,n) stating the number of ants in town *i* at time *t* and let $m = \sum_{i=1}^{n} b_i\,(t)$is total ants in the colony. Each ant is a simple agent that has the following characteristics [3]:

a. Ants choose a city based on a probability, which is a function of the distance between cities and the amount of pheromone contained in the connected arc.
b. In order to properly tour, the pathways that have been visited by ants no longer permitted to visit unless the journey is completed.
c. When traveling, each ant leaves pheromone on each path (*i, j*) itself.

Suppose τ_{ij} (*t*) is the intensity of pheromone on arc (*i,j*) at time t. Each ant at time *t* chooses the next city, so that the ants will be on time (*t+1*). Therefore if the Ant System algorithm iteration as *m* step by *m* ants in the interval (*t,t+1*) then every *n* iterations of the algorithm, means that each ant has completed a tour. At this point the intensity of pheromone is updated by the Equation (1).

$$\tau_{ij}(t+n) = \rho\tau_{ij}(t) + \Delta\tau_{ij} \qquad (1)$$

Where *ρ* is a coefficient such that (*1-ρ*) states evaporation pheromone trail between *t* and *t+n*, so the rest of the pheromone on arc (*i, j*) is expressed as Equation (2).

$$\Delta\tau_{ij} = \sum_{k=1}^{m} \Delta\tau_{ij}^{k} \qquad (2)$$

Where $\Delta\tau_{ij}^{k}$ is a quantity per unit of length of trail pheromone substances left on the arc(*i,j*) by ant *k*-th time interval between *t* and *t+n*. The quantity defined in Equation (3).

$$\Delta\tau_{ij} = \begin{cases} \frac{Q}{L_k}, & if\ ant\ k-th\ uses\ arc\ (i,j) \\ 0, & the\ others \end{cases} \qquad (3)$$

Where *Q* is a constant and L_k is the length of the tour generated by ant *k*-th. Coefficient *ρ* must be less than 1 to prevent the accumulation of infinite pheromone trail. Probability of ants move from one city to another following the taboo list as Equation (4).

$$P_{ij}^{k}(t) = \begin{cases} \frac{[\tau_{ij}(t)]^{\alpha}.[\eta_{ij}]^{\beta}}{\sum_{k \epsilon allowed_k}[\tau_{ik}(t)]^{\alpha}.[\eta_{ik}]^{\beta}} & ,\quad if\ j\ \epsilon\ allowed_k \\ 0 & ,\ the\ others \end{cases} \qquad (4)$$

Where *allowed$_k$* = (*N* – taboo), while *α* and *β* are two parameters that control the relative importance of the pheromone trail intensity and visibility.

2.2 *WordNet*

WordNet is a lexical data base developed by Princeton University. WordNet consists of nouns, verbs, adjectives and *adverbs* are grouped in the synsets. Each synset represents a different concept. Synsets are linked through conceptual-semantic and lexical relations. In 2006 the WordNet database has reached 155.287 words (unique strings), 117.659 synsets and 206.*941* word-sensepairs [10].

To determine the size of the WordNet synset similarity, can be used method of Wu-Palmer [9]. The method of Wu and Palmer measures synset similarity value in WordNet by calculating the length of the nearest point between synset as Equation (5).

$$Consim(C1,C2) = \frac{2*N3}{N1+N2+2*N3} \qquad (5)$$

3. Methodology

3.1 Material Research

The present study uses digital library journals document in PDF format. Data used in the study are described in Table 1.

Table 1 Data Research

No	Name	Total Amount
1	Library Catalog	50 documents
2	Metadata_Cat Table	50 records
3	Master_Term Table	823 records
4	Master_Cat Table	3254 records
5	Similarity_Cat Table	676506 records

3.2Research Methodology

The process of system developed using Waterfall method, consists of several steps: analysis and system design, system implementation and system testing.

4. Results and Discussions

4.1 The Results

In this study, an adaptive model of personalized searches is divided into three sub-systems, namely:

1. Text mining and classification
2. Search
3. Monitoring User Activity

4.2.1 *Text mining and classification*

This process is used for retrieval and classification of digital library documents. Block diagram of the process of text mining and classification can be seen in Figure 2.

Data retrieved from the data repository in the form of PDF files. Then, the data is processed to retrieve metadata. Metadata are taken from the documents consists of a title, abstract and keywords. Then, metadata is converted into tokens by the tokenizer in data normalization process. This process also eliminates unnecessary section of metadata (stop word). Stop word lists using general text stop word for English Language [4].Then, the text is converted to a form of stem metadata using K-Stemmer algorithm [5].

After that, the stem is classified using WordNet lexical database. Only stem nouns that are stored in database. Its similarity value is calculated as the pattern bipartite graph by using algorithms Wu-Palmer. Later, Wu-Palmer value is used to measure the length between the stem words as Equation (6).

$$Length = 1 - WUP \qquad (6)$$

At the end of this stage, the index file is created from the results of the previous process. This index file is useful to help the search process. The block diagram of this section is as Figure 2.

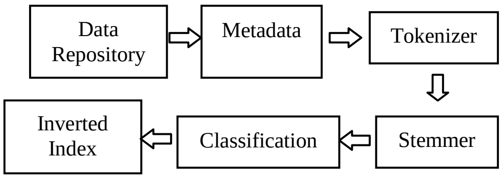

Figure 2 Text mining and classification

Data on initial processing are stored in database tables that have a relationship like Figure 3. Index file is generated as an inverted index. This file is created using Lucene.Net technology with a pattern similar to Table 2. Metadata fields in the index file to do the search process tokenized. Metadata field is a combination of Title Field, Abstract Field and Keyword Field (Table 2).

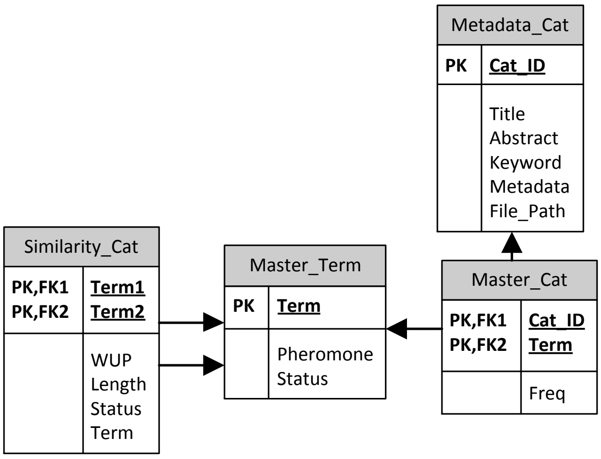

Figure 3 Relation Table on Database Systems

Table 2 Pattern in the index file

No	Field	Type
1	DocID	Field.Store.YES, Field.Index.NO
2	Title	Field.Store.YES, Field.Index.NO
3	Abstract	Field.Store.YES, Field.Index.NO
4	Keyword	Field.Store.YES, Field.Index.NO
5	Metadata	Field.Store.YES, Field.Index. TOKENIZED
6	Path_File	Field.Store.YES, Field.Index.NO

4.2.2 Search

This stage is used for the search process in a digital library system. Searches apply query expansion and Ant Colony Optimization. Query expansion is added from search keywords user using the ACO method. Maximum search keyword is restricted only 3 words. This process is adaptive according to the user system.

Block diagram of the process of query expansion as Figure 4. This process starts from a user input query. Then, continue to the process tokenizer and stemmer. This process produces a query in the form of stem. The next process, the query is developed by using the method of ant colony optimization. The results of this query are used to search.

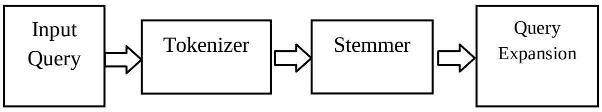

Figure 4 Personalized search process

Ant colony optimization is used in this system using Ant System (Equation 4). This equation is used to denote the probability of query expansion from one term to another term. Example of test data, algorithm term, there are 131 terms that can be used for query expansion. To select the term of keyword search algorithm, all connected terms are mapped according to the rules of probability Ant System in Equation 4. The all connected terms based on similarity distance between 2 terms. The total number of probability values is 1 such as Table 3. Then, Probability value is mapped in the position of the array element 0 to element 131 with the order of smallest to largest, as Figure 5. After that, random value between 0 and 1 is generated to select the term in the array element as the query expansion.

Table 3 Term Probability

No	Term	Length	Probability
1	Activity	0.2	0.021730163
2	Administration	0.4444444	0.004400359
3	Adoption	0.3684211	0.006403744
4	Analysis	0.3333333	0.00782286
…	…	…	…
131	Use	0.25	0.013907304

0	1	2	3		131
0	0.006	0.011	0.014	...	1

Figure 5 Probability array elements

4.2.3 Monitoring user activity

This section is used to change the pheromone value of each term data. Weight of pheromone used by the ACO algorithm to determine the probability of a query term expansion selected. This process can be seen in Figure 6.

In this process, the value of the term pheromone changed according to user activity. Activities are used to monitor the user model is the Click and Visit. Click the process occurs when the user selects the title of the journal. This process changes the pheromone terms contained in the document title. The Visit occurs when users view the contents pages of journals in a certain time period. This process changes the pheromone value of terms contained in the document metadata. To change the value of pheromone used Ant System Equation (1), (2) and (3). Pheromone value changes every search process carried out by the user when the user performs on the Click and Visit system. This process is a tour process of ant system algorithms.

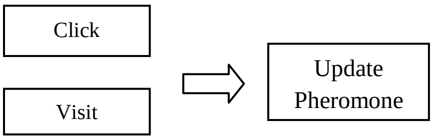

Figure 6 Process monitoring user activity

4.2 Discussion

Testing is done by changing the system model parameter values on the similarity threshold, compare the results between logical OR and logical AND Boolean Queryand compare the results before and after the method of the ACO tour. The testing process compared with Lucene analysis tool, Luke. Number of terms extraction of 50 documents journals are 1420 terms. Lucene score on the implementation of the system using Equation (7).

$$Score(q,d) = coord(q,d) * \sum_{t\ in\ q} \left(\left(idf(t) * querynorm(q) \right) * \left(tf(t\ in\ d) * idf(t) * norm(t,d) \right) \right) \quad (7)$$

Experiment uses 10 terms, each search process conducted Click and Visit by one each time. Testing parameters α = 1, β = 2, Q = 1.5, ρ = 0.5 and similarity threshold between 0.3 and 0.7. Query expansion uses Boolean Query OR method gives more document search results than the Boolean Query AND. Using fifty experiments performed query, before query expansion obtained 510 total documents, after query expansion obtained 627 total documents, after the tour with query expansion OR obtained 638 total documents and with query expansion AND got 87 total documents.

Table 4 Comparison of the number of documents found in the query results (0.3-0.7 similarity threshold)

Search Results	Query	Query Expansion without tour	Query Expansion OR	Query Expansion AND
Total	510	627	638	87

The next testing process using a reference document as the target search query using 10 terms as Table 5. Using a similarity threshold values 0.3, 0.4, 0.5, 0.6 and 0.7. Testing is done 50 times with each similarity threshold was tested 10 times.

Table 5 Document and query testing

Trees	DOC-0044
Schema	DOC-0030
Programming	DOC-0004
Benchmark	DOC-0050
Web	DOC-0006
Ontology	DOC-0025
Management	DOC-0019
System	DOC-0013
Information	DOC-0015
Data	DOC-0033

ACO Testing parameters α = 1, β = 2, Q = 1.5, ρ = 0.5. Testing conducted to observe the process before and after the query expansion. Query expansion process has not been doing the tour (Click and Visit) or pheromone update.The test results in Table 6 and the graph shown in Figure 7. In Table 6, the value of Lucene score before tour (Click and Visit) is still lower than the initial query before query expansion.

Table 6 Results of initial query score and query expansion before Click and Visit

	Initial Query	th=0.3	th=0.4	th=0.5	th=0.6	th=0.7
Trees	80.48	21.78	70.96	23.34	21.78	23.34
Schema	53.63	18.33	19.49	16.53	18.33	19.49
Programming	100	37.00	37.00	37.00	37.00	40.51
Benchmark	57.25	20.89	18.36	69.13	20.89	20.24
Web	44.26	12.34	13.20	14.87	13.20	14.87
Ontology	75.84	24.82	24.82	26.81	74.75	23.37
Management	26.56	7.92	7.40	8.64	8.64	7.60
System	31.12	6.43	6.43	6.43	8.44	9.64
Information	34.44	7.97	8.62	7.97	7.97	8.62
Data	40.22	6.16	7.64	7.64	6.16	5.35

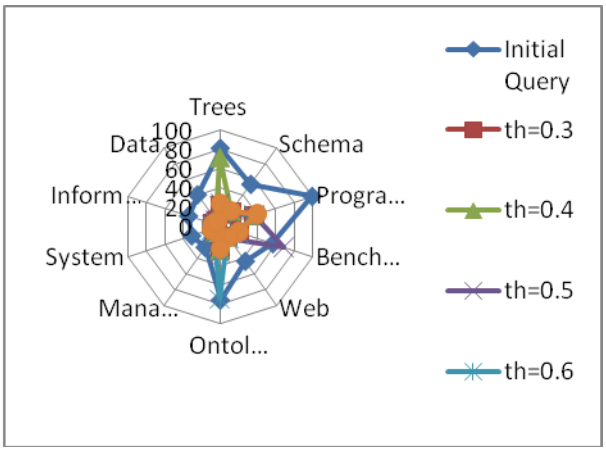

Figure 7 Graphs initial query and query expansion before Click and Visit

On the next testing, the search process is done, the tour (click and visit) or pheromone changes. The results are described in Table 7 and the graph of the testing results is shown in Figure 8. Lucene score of search's results are better than previous results (Table 6), with an increasing of Lucene score generally.

Table 7. Results of initial query score and query expansion after Click and Visit

	Initial Query	th=0.3	th=0.4	th=0.5	th=0.6	th=0.7
Trees	80.48	21.78	21.78	87.86	21.78	25.25
Schema	53.63	56.69	26.58	57.99	59.81	18.33
Programming	100	39.02	37.00	37.00	43.70	100.00
Benchmark	57.25	67.06	22.68	69.13	69.13	69.13
Web	44.26	60.21	60.21	12.34	47.33	13.20
Ontology	75.84	77.89	76.67	32.48	74.75	26.81
Management	26.56	7.40	44.53	47.64	44.53	44.53
System	31.12	6.43	6.43	53.49	7.43	9.64
Information	34.44	9.93	8.62	85.27	35.80	8.62
Data	40.22	5.35	7.64	4.35	7.64	6.16

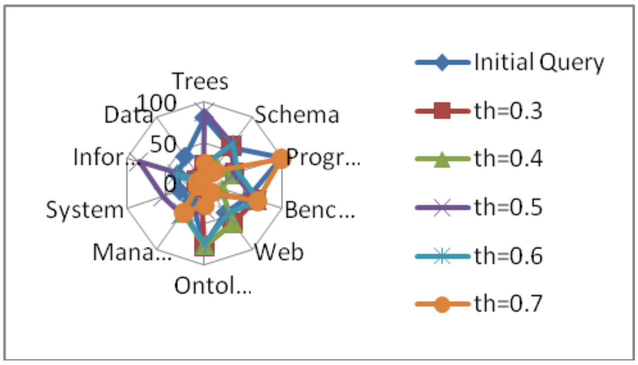

Figure 8 Graphs initial query and query expansion after Click and Visit

In the next test, to see the ranking order, parameter is used as the previous testing. On testing, the results of the search query expansion before the tour (Click and Visit) has a sequence that is worse than the initial search as Table 8. Graph rank order of search results can be seen in Figure 9.

Table 8 Ranking order initial query and query expansion before the Click and Visit

	Initial Query	th= 0.3	th= 0.4	th= 0.5	th= 0.6	Th =0.7
Trees	2	3	3	2	3	2
Schema	2	3	4	3	3	4
Programming	1	1	1	1	1	1
Benchmark	2	2	2	1	2	2
Web	1	2	2	3	2	3
Ontology	2	2	4	2	2	3
Management	5	6	6	9	9	6
System	3	4	4	4	8	10
Information	2	3	4	3	3	4
Data	1	2	5	5	2	1

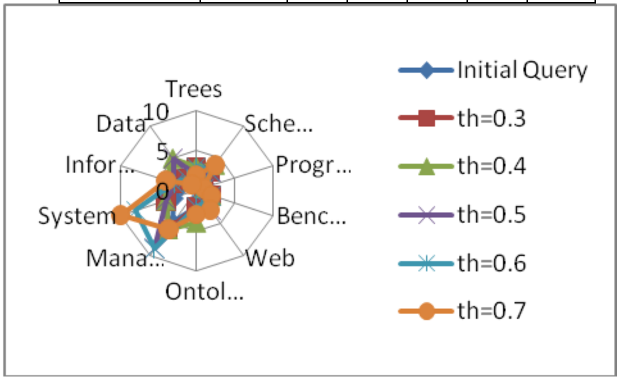

Figure 9 Graph ranking order initial query and query expansion before Click and Visit

From observation, improvement occurred after the tour sequences (Click and Visit) done, as in Table 9. The result graph of the document search process can be seen in Figure 10.

Table 9 Ranking order initial query and query expansion after the Click and Visit

	Initial Query	th=0.3	th=0.4	th=0.5	th=0.6	th=0.7
Trees	2	2	2	1	3	2
Schema	2	1	1	1	1	3
Programming	1	1	1	1	1	1
Benchmark	2	1	3	1	1	1
Web	1	1	1	2	1	2
Ontology	2	1	1	1	2	3
Management	5	6	1	1	1	1
System	3	4	4	2	6	10
Information	2	7	4	3	1	4
Data	1	1	5	9	5	2

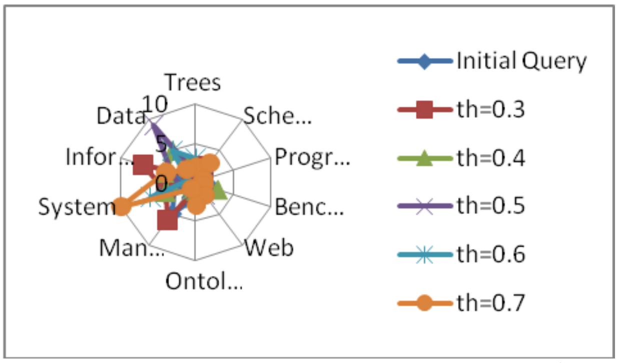

Figure 10 Graph ranking order initial query and query expansion after Click and Visit

By observing the score and rank order of the 50 tests, score can increased up to 60% and document ranking order becomes better up to 54% than earlier testing.Table results before and after Click and Visit can be seen in Table 10.

Table 10 Comparison of documents before and after the tour (Click and Visit)

	Score			Ranking Order		
	up	**equal**	**down**	**up**	**equal**	**Down**
th=0.3	6	2	2	6	3	1
th=0.4	5	4	1	5	4	1
th=0.5	6	2	2	6	3	1
th=0.6	7	2	1	6	3	1
th=0.7	6	2	2	4	5	1

5. Conclusion

This paper explains the development of the adaptive model using Ant Colony Algorithm to personalized search in digital library. Based on the results, the conclusions can be summarized as follows:

1. Boolean query OR generates recall document better than the Boolean query AND.
2. Comparison results of before and after process on clicking and visiting, tours of ACO improve the score up to 60%.
3. To rank order, ACO tours repair the sequence selected documents up to 54%.

Reference

[1] Brusilovsky, P., Maybury, T.M. 2002. From Adaptive Hypermedia to the Adaptive Web. *Journal of Communication of the ACM45*, 31-33

[2] Dorigo, M., Maniezzo, V., & Colorni, A. 1996. The Ant System: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics–Part B, Vol.26*, No.1, 1-13

[3] Dorigo, M., Stuzle, T. 2004. *Ant Colony Optimization*. MIT Press

[4] Fox, C. 1992. *Lexical Analysis and Stoplists, Information retrieval.* Prentice Hall,102-130

[5] Krovetz, R. 1993. *Viewing Morphology as an Inference Process*. Department of Computer Science, University of Massachusetts, Amherst, MA 01003

[6] Martinez, E. Frias, Magoulas G., Chen, S., & Macredie, R. 2006. Automated User Modeling for Personalized Digital Libraries. *International Journal of Information Management. Volume 26, Issue 3*, 234-248

[7] Phinitkar, P., Sophatsathit, P. 2010. Personalization of Search Profile Using Ant Foraging Approach. *ICCSA 2010, Part IV, LNCS 6019*, 209-224

[8] Wong, L.H., Looi, C.K. 2009. Adaptable Learning Pathway Generation with Ant Colony Optimization. *Educational Technology & Society*, 12 (3), 309–326.

[9] Wu, Z., Palmer, M. 1994. Verb semantics and lexical selection. *Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics*, 133-138.

[10] …, WordNet 3.0 database statistics, website:http://wordnet.princeton.edu/wordnet/ man/ wnstats .7WN.html diakses pada 14 Agustus 2012