

VOLUME 1

2011 INTERNATIONAL CONFERENCE ON

**UNCERTAINTY REASONING
AND
KNOWLEDGE ENGINEERING**

**4-7 AUGUST 2011
BALI, INDONESIA**



Proceedings of the International Conference on
Uncertainty Reasoning and
Knowledge Engineering

4-7 August 2011, Bali, Indonesia

URKE 2011

4
8
2
6
0
4
8

All right reserved.

Print Version

Copyright and Reprint Permission: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Operations Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved. Copyright 2011 by the Institute of Electrical and Electronics Engineers

URKE 2011
Table of Contents
Volume – 1

Preface – Volume 1.....	VI
Conference Organization – Volume 1	VII
Alternative implementation techniques of parallelized differential evolution for multi-core processors.....	1
<i>Kiyaharu Tagawa and Takashi Ishimizu</i>	
Designing of expert system for troubleshooting diagnosis on gas chromatography gc-2010 by means of inference method.....	5
<i>Suyanto M. Abduh Pancaputra, and Ruri Agung Wahyuono</i>	
Selection of districts for data entry sites in aceh household census.....	9
<i>Bonivasius Prasetya, Stefanie Intan Christienova, and Idaman</i>	
Collaborative internet content filtering on the internet infrastructure in Malaysia.....	12
<i>Fajri Achmad Maulana, Sahidan Abdulmana, and Fauzan Alfariti</i>	
Reinforcement mechanism and engineering applications of jointed rock mass.....	16
<i>Jingsui Zhang, Fei Wang, Shaojun Fu, and Tingting Liao</i>	
Uncertainty, protocol behavior prediction through chaos theory?.....	20
<i>Prabhu Shankar Kaliappan</i>	
Soaking with uncertainty for problem solving.....	24
<i>Wei Hui and Fu Xixu</i>	
A model for information systems evaluation.....	28
<i>Liang-Chuan Wu</i>	
Flexible decision making in uncertain R&D projects.....	32
<i>Juite Wang</i>	
Semantic technology classification - a defence and security case study.....	36
<i>Dirk Thorleuchter and Dirk Van den Poel</i>	
Companies website optimising concerning consumer's searching for new products.....	40
<i>Dirk Thorleuchter and Dirk Van den Poel</i>	
A framework for vision-based swimmer tracking.....	44
<i>Wen-Hui Chen, Po-Chuan Cho, Ping-Lin Fan, and Yi-Wen Yang</i>	
Initiating layers architecture design for software product line.....	48
<i>I Made Murwantara</i>	

Research on remote intelligent monitoring system.....	52
<i>Wang Xiao-li, Gao Xiao-hong, and Li Jian-wei</i>	
Empirical study on selection and evaluation of TPL based on CS.....	56
<i>Wei Xu, Songzheng Zhao, and Liangdong Lu</i>	
Research on the application of neural network PID to wipe film motor of molecular distillation.....	60
<i>Guan Changjun and Li Hui</i>	
Design of a pet drinking vessel based on TRIZ & growth design.....	64
<i>Hongwu Chen, Xuan Liu, and Can Wang</i>	
Knowledge sharing platform for project team based on web feeds.....	67
<i>I-Ching Hsu, Lee Jang Yang, and Der-Chen Huang</i>	
New switched filtering method for recurrent neural networks.....	71
<i>Choon Ki Ahn</i>	
Research on China's studying abroad situation and countermeasures.....	75
<i>Chen Aijuan and Zhang Qian</i>	
Classifier combination for telegraphese restoration.....	79
<i>Leo Willyanto Santoso</i>	
Design and implementation of the soft-sensing system for fermentation process.....	83
<i>Wu Zi-yue, Xu Zhe, and Jinfeng Geng</i>	
The study of mobile education development based on 3G technique and cloud computing.....	86
<i>Huang Bin</i>	
The analysis of human judgment accuracy using decision tree models.....	90
<i>Hsi-Peng Lu, Yi-Wen Yang, and Wen-Hui Chen</i>	
A weighted fuzzy time series model based on adoptive OWA operators.....	94
<i>Jing-Rong Chang, and Yu-Jie Huang</i>	
Research on randomized greedy algorithm for k-median problem.....	98
<i>Wang Shouqiang</i>	
An algorithm for splitting an orthogonal polyhedron with a polyplane.....	102
<i>Jefri Marzal, Hong Xie, and Chun Che Fung</i>	
Implementing user-oriented interfaces: from user analysis to framework's components.....	107
<i>Sergio Inzunza, Andrés Mejía, Reyes Juárez-Ramírez, and Manuel Gómez-Ruelas</i>	
Towards a formal model for knowledge transference: An initial proposal to support the adoption of process models.....	111
<i>Violeta ocegueda-miramontes, Reyes Juárez-Ramírez, Rafael Pimienta-Romo, and Manuel Gómez-Ruelas</i>	
A formal approach for measuring the lexical ambiguity degree in Natural language requirement specification.....	115
<i>Carlos Huertas, Reyes Juárez-Ramírez, Manuel Gómez-Ruelas, and Héctor Plata</i>	
Towards improving user interfaces: a proposal for integrating functionality and usability since early phases.....	119

<i>Reyes Juárez-Ramírez, Alan A. Gutiérrez, Manuel Gómez-Ruelas, and Pavel Negrete</i> Using iterative learning control methods for 2-D systems.....	124
<i>Zahra Maghsoodzade Sarvestani, Ahmadreza Argha, and Mehdi Roopaei</i> Direct torque control of brushless dc motor drives with reduced starting current using fuzzy logic controller.....	129
<i>N. Parhizkar, M. Shafiei, and M. Bahrami Kouhshahi</i> Application of topic based vector space model with wordnet.....	133 ←
<i>Adi Wibowo and Andreas Handojo</i> Processing comparisons and evaluations in business intelligence: a question answering system.....	137
<i>Kenston Choi, Rosalyn Margret Pacaña, Adrian Lester Tan, Jonathan Yiu, and Nathalie Rose Lim</i> A study of applying arima and svm model to software reliability prediction.....	141
<i>Jung-Hua Lo</i> Incorporating information quality criteria in blog search engine.....	145
<i>Fatemeh Azimzadeh, Abd. R. Ramli, Borhanoddin. M. Ali, and Hamidah Ibrahim</i> Study on the development trend of the digital printing.....	149
<i>Qian Hua</i> Using particle swarm optimization algorithm for image enhancement.....	154
<i>Hanxing Chen and Jun Tian</i> Chaos-based gray image watermarking algorithm.....	158
<i>Junmei Guo, Yong Yang, and Ning Wang</i> Special equipment safety analysis and countermeasures.....	161
<i>Jun Tian, and Hanxing Chen</i> A study on home information network.....	165
<i>Yong Yang and Junmei Guo</i> Non-subsampled contourlets and gray level co-occurrence matrix based images segmentation.....	168
<i>Zhang Jian and Chen Xiaowei</i> Preparation for structural path analysis from plain text input output matrix.....	171
<i>Nova Nurviana, Rizchi Wahyuni, and Idaman</i> A study on the design of half pipe.....	175
<i>Zhao Yang and Li zhaonan</i> Evaluation of building materials logistics provider based on FAHP—TOPSIS.....	180
<i>Sun Shusheng, Zhang Qin, and Zhang Min</i> An approach for multimodal biometric fusion under the missing data scenario.....	185
<i>Quang Duc Tran, Panos Liatsis, Bing Zhu, and Changzheng He</i> Hybrid-dimension association rules for diseases track record analysis at Dr. Soetomo general hospital.....	189
<i>Silvia Rostianingsih, Gregorius Satia Budhi, and Ni Wayan Yessy Dwijayanti</i> The forecast for the engine wear trend based on grey gm(1,1) and markov chain model.....	192
<i>Zhang Yong, Chen Liang, and Gu Xiaoxing</i>	

Data mining market basket analysis' using hybrid-dimension association rules, case study in minimarket X.....	196
<i>Djoni Haryadi Setiabudi, Gregorius Satia Budhi, I Wayan Jatu Purnama, and Agustinus Noertjahyana</i>	
Product cost estimation using supper efficiency data envelopment analysis.....	200
<i>Fantahun M. Defersha, Adil Salam, and Nadia Bhuiyan</i>	
Prioritization of information quality criteria in the blog context.....	205
<i>Mohammad Javad Kargar</i>	
Towards dedicated information quality frameworks in social network products.....	209
<i>Mohammad Javad Kargar</i>	
Study on the fast image registration based on genetic algorithm.....	213
<i>Shuai Feng, Yufeng Li, and Weiping Zhao</i>	
Compression and realization of power quality disturbance data based on wavelet analysis.....	217
<i>Zhao Hongtu and Xi Dongmei</i>	
Research on face detection algorithm in instant message robot.....	220
<i>Yufeng Li, Haijuan Zhang, and Yanan Zhang</i>	
Three-dimensional visual geographic information system.....	224
<i>Bo Hang</i>	
Runtime monitoring web services implemented in BPEL.....	228
<i>Zhang Haiteng, Shao Zhiqing, and Zheng Hong</i>	
Based on stm32f103 implement profibus-dp slave with high-speed transmission.....	232
<i>Xu Hongyan, Wu Guichu, and Chen Chong</i>	
LVMM: a lightweight virtual machine memory management Architecture for virtual computing environment.....	235
<i>Weizhe Zhang, Tao Cheng, Hui He, and Albert M.K. Cheng</i>	
Primary signal detection over rayleigh fading channel for cognitive radio.....	239
<i>Yufeng Li, Yanan Zhang, and Haijuan Zhang</i>	
A spt-routing key management scheme for heterogeneous wireless sensor networks based on ECC.....	243
<i>Gang Han, Rui Zhou, and Hua Yang</i>	
Design of video surveillance system based on 3G wireless network in underground coal mine.....	248
<i>Wei Zhou</i>	
A hybrid key management scheme for heterogeneous wireless sensor networks based on ecc and trivariate symmetric polynomial.....	251
<i>Rui Zhou and Hua Yang</i>	
Comprehensive evaluation and selection model of supplier under E-business.....	256
<i>Wang Mian</i>	

Voice conversion application (VOCAL).....259
Liliana, Resmana Lim, and Elizabeth Kwan
Clustering search engine at petra christian university library using suffix tree clustering.....263 ←
Andreas Handoyo, Adi Wibowo, and Jemmy Lay Santo
Author Index – Volume 1267

IEEE Conference Publications Management Group
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08854

IEEE Catalog Number: CFP1115N-PRT
ISBN: 978-1-4244-9983-0



Clustering Search Engine At Petra Christian University Library Using Suffix Tree Clustering

Andreas Handoyo
 Informatics Engineering
 Department
 Faculty of Industrial Technology
 Petra Christian University
 Email: handoyo@petra.ac.id

Adi Wibowo
 Informatics Engineering
 Department
 Faculty of Industrial Technology
 Petra Christian University
 Email: adiwb@petra.ac.id

Jemmy Lay Santo
 Informatics Engineering
 Department
 Faculty of Industrial Technology
 Petra Christian University
 Email:
 m26405118@john.petra.ac.id

Abstract—At Petra Christian University Library, book searching engine system is using a common keyword matching as search queries. So, the user must enter the correct keyword and this case often takes a long time for users to find book that needed. Therefore, required an application that's capable display the search results that classified into a group/cluster. Which is each cluster will contain the documents that have the same classification base on keyword that been input by user in order to assist users in perform book searching. In this study, have been build an application to classify the document search results into a group of documents that have the same classification/cluster. Input to the cluster obtained from regular search results then the classification/clustering done by using Suffix Tree Clustering (STC) uses the document phrase. Based on the testing, the resulting cluster has been able to cluster documents that correlate or have the same classification with the keyword that input by user. The average time that needed to process for each document is 0.1139 seconds.

Keywords- Suffix Tree Clustering, Suffix Tree, Clustering, Searching, Library

I. INTRODUCTION

Information retrieval is concerned with representing, searching, and manipulating large collections of electronic text and other human-language data [4]. In general, the result from search systems on the web showing a long list of documents, then the user should look for from start to finish finding the desired document. This long list of document that build by search results sometimes make user confused and more difficult to find the desired document and also consume a lot of time.

The library of Petra Christian University, right now still use the same kind of method. So its makes user more difficult to find the book that he/she wants among thousands of books that available on the library. Besides of that user also must input the correct keyword to give a correct input to the searching engine.

Therefore it is necessary to build a clustering searching engine, where the general searching results will be accepted as an input and the engine will create a group of documents with the same classification (as a cluster), so this will be facilitate users in finding the desired document or book.

In performing the clustering, this application will used Suffix Tree Clustering (STC) that uses phrase to identify a set of documents that share a common phrase and use this information to create clusters and classifying its contents to the user using web application.

II. SUFFIX TREE

Suffix Tree (also called PAT tree or, in an earlier form, position tree [5]) is a data structures covering problems in strings (a series of character) to be able to analyzed quickly. If $txt=t_1t_2...t_i...t_n$ was a string, so $T_i=t_it_{i+1}...t_n$ is suffix from txt, starting from position of i, example: [1]

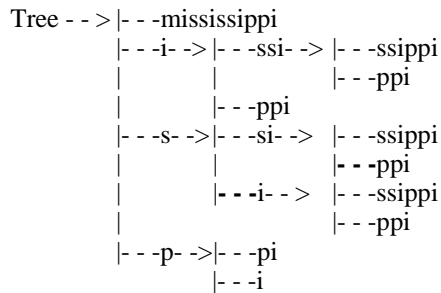
T ₁	=	mississippi = txt
T ₂	=	ississippi
T ₃	=	ssissippi
T ₄	=	sissippi
T ₅	=	issippi
T ₆	=	ssippi
T ₇	=	sippi
T ₈	=	ippi
T ₉	=	ppi
T ₁₀	=	pi
T ₁₁	=	i
T ₁₂	=	(empty)

Suffix tree could used to find a substring, pat [1 ... m] in O(m). There are $n(n+1)/2$ substrings in txt [1 ... n] and be more surprisingly that the suffix tree could formed in time O(n). The addition of single characters in the txt increase the number of substrings with n + 1. In the example above, if suffix is zero (empty) the process will be sorted as follows:

T ₁₁	=	i
T ₈	=	ippi
T ₅	=	issippi
T ₂	=	ississippi
T ₁	=	mississippi
T ₁₀	=	pi
T ₉	=	ppi
T ₇	=	sippi
T ₄	=	sissippi
T ₆	=	ssippi
T ₃	=	ssissippi

It is clear that some of them give the same prefix. There are substrings starting with 'i', 'm', 'p', and 's', but all

that starts with 'is' in fact start with 'issi'. Two or more of the same prefix gives the same path from the root of the suffix tree.



III. CLUSTERING

Clustering method is a method that has the ability to analyze and classify documents automatically. Clustering technique usually use words and documents as a collection of words without considering any order or so-called bag of words.

Cluster analysis is an effort to find a group of objects that represent a character of the same or nearly the same (similar) between one object with another object in a group and have a differences (not similar) with the objects in other groups. Of course the similarities and differences were obtained based on information provided by these objects and their relationships (relationship) between them.

IV. SUFFIX TREE CLUSTERING [3]

Suffix Tree Clustering (STC) is the first algorithm that uses phrase so the process is much simpler than other algorithms. STC is an incremental algorithm, the calculations is a linear time complexity $O(n)$ and meet the criteria for document clustering. [2]

Suffix Tree Clustering has three step of logic:

A. Document Cleaning

The sentence is marked and not remove, the original string documents is remaining in the keep.

B. Basic Cluster Identification

The identification of the basic cluster can be expressed as the making of an inverted index of the phrases from the document collections. This can be done efficiently with a data structure called a suffix tree. Suffix tree of a string 'S' is a compact tree that composed from the phrase that have suffix 'S'. Which is the document will be considered as a string of the words, not as a character. Therefore, the suffix will be consists of the one or a whole word.

Examples suffix tree of a set of strings 'cat ate cheese', 'mouse ate cheese too', and 'cat ate mouse too'. Node of the suffix tree is depicted in a circle (Figure 1). Each suffix node has one or more boxes that are connected to indicate a string of origin. The first number in each box indicates the string origin, while the second

number suffix indicating which of the label string, the suffix-node it.

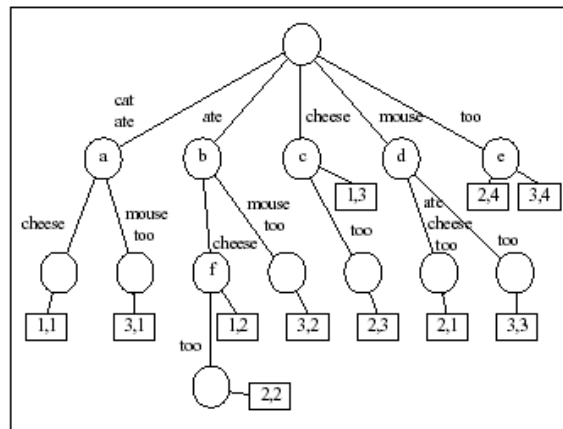


Figure 1. Suffix tree from string 'cat ate cheese', 'mouse ate cheese too', and 'cat ate mouse too'

Each node of the suffix tree represents a group of the documents and a phrase that is common to all. The label of the node represents a common phrase, a collection of the tagging documents its suffix node is a descendant of the node to form a group of documents (Table 1).

Therefore, basic cluster (consisting of the 2 or more documents) appear as a node in a suffix tree.

TABLE I. SIX NODES AT FIGURE 1 AND THE BASIC CLUSTER CORRESPONDENT

Node	Phrase	Documents
A	cat ate	1,3
B	ate	1,2,3
C	cheese	1,2
D	mouse	2,3
E	too	2,3
F	ate cheese	1,2

Each basic cluster is marked by a score that is a function of the the number of documents that are owned and words that make the phrase. Score from basic cluster B with the phrase P can be written:

$$s(B) = |B| \cdot f(|P|) \tag{1}$$

Where $|B|$ is the number of documents in basic cluster B
 s is the score than basic cluster B,
 $|B|$ is the number of documents in basic clusters B,
 $|P|$ is the number of words in P that has a score is not zero.

Words that include stop list or appears too few (3 or less) or too many (more from 40% from a collection of documents) were given score 0. The function f penalizes the 1 word phrase, linear to long-word phrases by 2 to 6 and became constant for longer phrases.

C. Basic Cluster Combination

Documents may share more than one phrase. As a result, a collection of documents from different basic clusters may overlap and may even be the same. A binary similarity measurement between the basic clusters is defined by the overlap from collection of documents. Given two basic clusters B_m and B_n , with size $|B_m|$ and $|B_n|$ each, and $|B_m \cap B_n|$ represents the number of documents to the two basic clusters, defined similarity from B_m and B_n to be 1 if:

$$\begin{aligned} |B_m \cap B_n|/|B_m| > 0,5 \text{ and} \\ |B_m \cap B_n|/|B_n| > 0,5 \end{aligned} \quad (2)$$

In contrast, the similarity is defined by 0.

V. IMPLEMENTATION AND TESTING

There are several main steps in from this system, which is:

- Documents Cleaning.
- Basic Cluster identification.
- Combination Basic Cluster Association.

Figure 3, will display the flowchart from the application

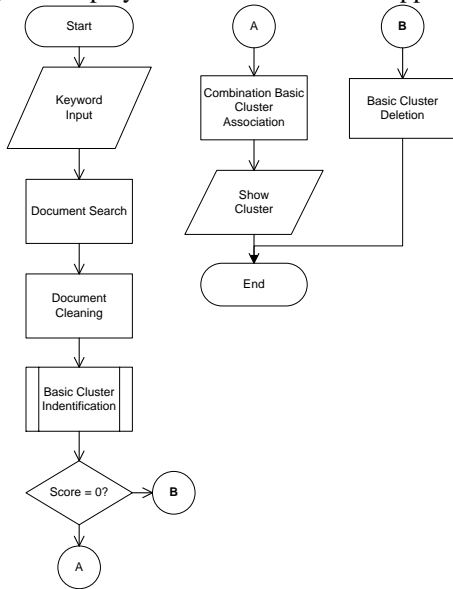


Figure 2. General System Flowchart from the Application

A. Document Cleaning

Document Cleaning is the first step to get the basic cluster from the documents that find from the common search results. In the process of cleaning document, punctuation in the title is removed. After completing the cleaning process then the result will be send to the basic cluster identification stage.

B. Basic Cluster Identification

The second step is to identify the basic cluster. At this stage there are a series

of basic cluster identification process will be carried out to obtain basic cluster. This process will include sorting process, tree making process and the process of calculating basic cluster score. From Nodes that formed, the number of documents owned by each node will be calculated, if there are nodes that have a number of documents as much as 2 or more then the node is stored as a cluster basic. Then, each basic cluster obtained score will be calculated

In the process of calculating a score, administrator could determine the lower and upper limit for the words that included in stop list or words that have a zero score. Every word in the sentence (title) will be taken and application will calculate the number of documents that containing the word. If the number is greater than the lower limit or smaller than the upper limit or included in the stop word the word is given a zero score. Then the words in basic cluster that doesn't have zero score will be counted and also the number of documents containing the basic cluster will be count. After getting the number of documents ($|B|$) and the words in basic cluster that has zero score then score than score from each basic cluster can be calculated

C. Basic Cluster Association

After getting the basic cluster, the next step is to combine these basic clusters. At this step, application will look for any relationship from existing basic cluster. Any basic clusters that have a relationship or similarity, the association value will be set by 1, if not then the value will be set by 0. To analyze the correlation from two basic cluster, the number of documents from each cluster will be count as $|B_m|$ and $|B_n|$, and $|B_m \cap B_n|$ will be indicates as the number of documents that owned together by that two basic clusters. Two basic clusters are said that have similarities or relationships if:

$$\begin{aligned} |B_m \cap B_n|/|B_m| > 0,5 \text{ and} \\ |B_m \cap B_n|/|B_n| > 0,5 \end{aligned}$$

Besides that, the value will be set as 0.

After getting the basic cluster relationship or similarity, then if the basic cluster included in stop list or has zero score, the basic cluster is removed. The examination will be given input query using keyword 'microcontroller'. The result of cluster base obtained as shown at tables 2.

TABLE II. RESULT OF BASIC CLUSTER WITH QUERY ' MICROCONTROLLER'.

No.	Base Cluster
1	68hc12 microcontroller
2	68hc12/11 microcontrollers
3	and the microcontroller
4	design of embedded systems using 68hc12/11 microcontrollers

5	digital
6	digital signal processing and the microcontroller
7	embedded systems using 68hc12/11 microcontrollers,
8	microcontroller
9	microcontrollers
10	of embedded systems using 68hc12/11 microcontrollers
11	processing and the microcontroller
12	signal processing and the microcontroller
13	systems using 68hc12/11 microcontrollers
14	the microcontroller
15	using 68hc12/11 microcontrollers

After getting the basic clusters, and based on the number of documents and words that form the phrase base clusters, the computed score than base cluster B with the phrase P using equation 1.

The score that calculate from table 2 base cluster number 11 'processing and the microcontroller' is as follows:

- Number of words = 4.
- The number of words that have a zero score = 3 ('and', 'the', and 'microcontroller'). The word 'and' and 'the' has zero score because included in stop word while the word 'microcontroller' has zero score because it has a too much appearance number (more than 40% from the number of documents with the query 'microcontroller'). So the number of words that do not have score zero is $(|P|) = 1$.
- The number of documents $(|B|) = 2$.

Thus, the obtained score for the base cluster 'processing and the microcontroller' as shown in table 2 Number 11 is 0. Furthermore, than base clusters obtained by the selected base clusters that has non-zero score to serve as a cluster. The result of cluster with query 'microcontroller' can be shown on table 3.

TABLE III.
RESULT OF CLUSTER WITH QUERY 'MICROCONTROLLER'.

No.	Cluster
1	68hc12/11 microcontrollers
2	design of embedded systems using 68hc12/11 microcontrollers
3	digital signal processing and the microcontroller
4	embedded systems using 68hc12/11 microcontrollers
5	signal processing and the microcontroller
6	systems using 68hc12/11 microcontrollers
7	using 68hc12/11 microcontrollers

The application from the clustering search engine using suffix tree can be shown on Figure 3.

The examination result from system performance that calculate the relation between number of documents that have been process and process time required it's presented by graphic (as shown on Figure 4).

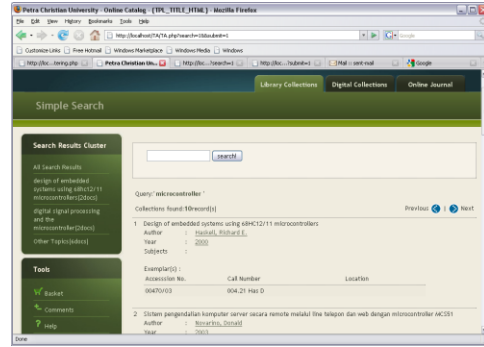


Figure 3. Library Clustering Search Engine

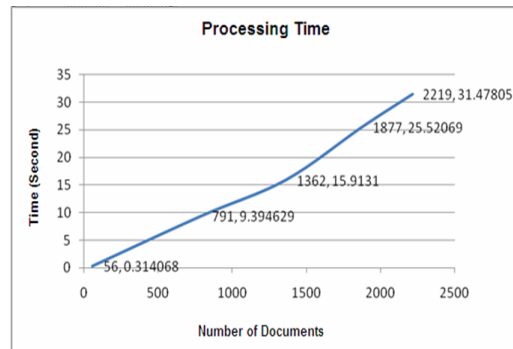


Figure 4. Correlation between Time and Number of Documents

VI. CONCLUSION

Based on the testing result on the system, we can conclude several things as follows:

- The cluster that generated by the system is good enough, in a sense to classify the documents that correlate to each other or have the same classification.
- The processing time is increase relatively equal to the number of documents.
- The process to combination the basic cluster is requires the longest time between any other processes.

REFERENCES

- [1] Allison, Lloyd. (n.d.). *Suffix trees*. Retrieved January 21, 2011, from <http://allisons.org/ll/AlgDS/Tree/Suffix/>
- [2] Zamir, Oren, Etzioni Oren, Omid Madani, Richard M Karp. (1997). *Fast and Intuitive clustering of web Documents*. KDD-97 Proceedings.
- [3] Zamir, Oren, and Etzioni Oren. (1998). *Web document clustering: a feasibility demonstration*. Washington: Department of Computer Science and Engineering University of Washington Seattle.
- [4] Stefan Büttcher, Charles L. A. Clarke, and Gordon V. Cormack. *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press, Cambridge, Mass., 2010.
- [5] *Suffix Tree*. Retrieved January 21, 2011. http://en.wikipedia.org/wiki/Suffix_tree