Proc. of The Second Intl. Conf. On Advances In Computing, Control And Networking - ACCN 2015
Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-073-6 doi: 10.15224/ 978-1-63248-073-6-16

Ontology Model Development Combined with Bayesian Network

Ionia Veritawati, Ito Wasito, T. Basaruddin

Abstract— **Recently, development of methods in extracting knowledge from a text collection is still explored. In this work, the proposed approach utilize important words or key words that represent a domain of text. The key words may have relations among them and the relational keywords in the text domain can be organized become an ontology model as a domain knowledge. The proposed method for forming knowledge represented the text consists of three stages process. First, Vector Space Model (VSM) of key words from text is clustered using bottom-up approach and each clustered data is categorized to be an input of structure learning in a Bayesian network concept. The next stage, structure development of each clustered data using Markov Chain Monte Carlo (MCMC) method such that key words as nodes are related each other as in DAG (directed acyclic graph) form. The result of structure learning process of each cluster produces a clustered DAG. The same learning process is also applied to the original data and it produces a general DAG. The third stage is an analysis process using some rules applied to clustered DAGs and the general DAG to determine connector nodes. A connector node is located in a clustered DAG and it has a relation (edge) to other node in another clustered DAG. It causes cluster of DAGs to be a union graph called an Ontology Model which represent knowledge of the text domain. Data in this works consist of simulation data using a small number of key words from natural science. The ontology model resulted is evaluated manually and it shows that the knowledge of text can be represented visually. The experiment of ontology development still has some challenges to be improved.**

Keywords— bottom-up clustering, MCMC, Connector Node, Ontology

I. Introduction

Ontology development has been studied in many application domain areas, such as in automotive industries to develop knowledge of services [1], in bioinformatics to represent protein-protein interaction [2], in medical fields to analyze diagnosis requirements [3], in semantic web to make links between web pages [4], etc. Approach for ontology development are various including formal method approach [5] and machine learning approach [6]. The ways of the approach are developed manually, semiautomatically or automatically. It means a domain expert is sometimes needed to develop relevant ontology as a knowledge of the domain [5]. The proposed method of ontology development is an automatic process without a domain expert. It can be applied to text data which can be derived from any domain.

Ionia Veritawati
Department of Informatics
Pancasila University
Indonesia

Ito Wasito; T. Basaruddin
Faculty of Computer Science
University of Indonesia
Indonesia

I. Ontology and Text

Ontology is broadly defined as "*a formal, explicit specification of a shared conceptualization*" [7]. Generally, domain ontology representation has spectrum covered ranging from lightweight ontology which the structure is represented by a taxonomy (tree or graph) to formal ontology represented by a relational data base [5].

Text as a data collection consists of meaningful words as key words or key phrases, and stop words which are meaningless words and are usually removed. The used of text data in machine learning approach is initated by extracting only frequencies of meaningful words from the data and by arranging the frequencies in a vector space (table of key words versus documents) [8].

Collection of meaningful words from a domain represents knowledge of the domain itself. It can be arranged more specifically by determining relations among the meaningful words. The meaningful words related to each other is called as an ontology [9]. In this work, text data models are created and used to develop an ontology model by using the proposed methodology (Fig. 1).

II. Methodology

Fig. 1 is a methodology for ontology development proposed. Text as data are numbers of collection of key words from documents in a Vector Space Model (VSM). Three types of DAG are defined including modeled DAG, clustered DAG and general DAG. Modeled DAG is determined manually as a model, clustered DAG and general DAG are resulted from a structure learning. In data modeling step, the data are modeled by creating manually two or more modeled Directed Acyclic Graphs (DAG) including labels as key words for each node and relations (edges) between them. Each modeled DAG will represent a cluster of key words collection and its relations (cluster of knowledge). Further, the modeled DAGs are sampled by a bayesian network approach. The samplings from all modeled DAGs are combined as a table of categorical data. The process is continued by converting the categorical data to real numbers as a vector data. This vector data model is an inputted data. Preprocessing is applied to the data by using tf-idf and normalization.

A hierarchical clustering is applied to the vector data which functions to separate their data elements. The clustered data are categorized and they are as an input data for structure learning process in bayesian network. A scoring function is applied to each clustered data to predict a graph structure

SEEK DIGITAL LIBRARY

Proc. of The Second Intl. Conf. On Advances In Computing, Control And Networking - ACCN 2015
Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-073-6 doi: 10.15224/ 978-1-63248-073-6-16

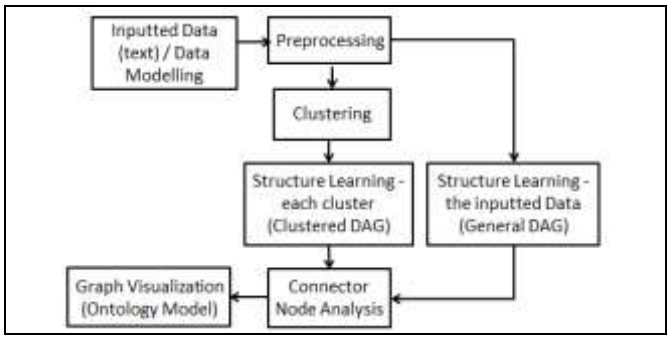

Figure 1. Methodology of experiment

(Clustered DAG). The structure learning is also applied to the data before clustered which has been preprocessed and categorized and the result is a graph structure called a general DAG.

Clustered DAGs and a general DAG become an input in Connector node analysis process. The process is to find nodes that have relations between the separately clustered DAGs to be a combined graph called an ontology model. The detail process will be described in the next section.

III. Clustering for Text

Clustering for text is a process to separate key words of the text data in a vector space. The process can be used to categorize or classify infomation [10], to extract information [11] and also to retrieve information [12]. Clustering can be implemented as a flat clustering technique such as k-means [13] and as a hierarchical clustering technique [14] such as top-down or bottom-up approach.

Clustering in this experiment uses bottom-up approach (hierarchical clustering) that the data is clustered by calculating distances between elements of the data (Fig. 1). Two data elements with the smallest distance are combined and the process is done repeatedly until all data elements become one cluster. The number of clusters to be separated are determined manually. Bottom-up approach as a hierarchical clustering has more accurate performance to separate a sparse data in a vector space compared to flat clustering such as k-means.

IV. Structure Learning in Bayesian Network

Bayesian network specifies a joint probability distribution (JPD) from a Directed Acyclic Graph (DAG) structure. Each node in the graph represents a random variable and the edge which connect two nodes representing probabilistic dependencies between the nodes [15]. The joint distribution represented in the structures of graphs is equivalent with conditional independence (CI) relations between nodes, and it is formulated in (1).

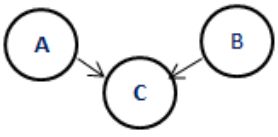

Figure 2. A Simple Bayesian Network.

$$P(x_1, x_2, \ldots x_n) = \prod_i P(x_i \mid parents\ (x_i)) \quad (1)$$

where $x_1, x_2, .., x_i, .., x_n$ are random variables of a domain. If $X = \{x_1, x_2, .., x_i\}$ then Parents $(x_i) = \{x_1, x_2, .., x_{i-1}\}$.(a,b).

Fig. 2 is a simple bayesian network which joint probability distribution is P(A,B,C) = P(C|A,B) P(A) P(B).

Structure learning is an important part in bayesian network to find dependencies between nodes. Data from a collection of features in a number of records represent a multinomial distribution. A DAG can be predicted from the data by using a scoring approach with bayesian analysis for the score prediction [16]. Bayesian rule (2) is adopted by bayesian network to calculate a posterior of a graph $P(G|D)$, given data (D).

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)} \quad (2)$$

where $P(D|G)$ is a marginal likelihood, $P(G)$ is a prior graph and $P(D)$ is a probability of data.

In structure learning, the best structure of graph (G) can be determined by maximizing the marginal likelihood (3).

$$P(D|G) = \int P(D|G,\theta)\, P(\theta|G)\, d\theta \quad (3)$$

where $P(D|G,\theta)$ is a likelihood and $P(\theta|G)$ is a prior graph over parameters.

Derived from (2), a bayesian score (4) is a sum of family score of $\boldsymbol{X_i}$.

$$score(G:D) = \sum_i FamScore(x_i, Parents(x_i) : D) \quad (4)$$

where family score is only from x_i and its parents given data (D).

A. Scoring and MCMC Method

Markov Chain Monte Carlo (MCMC) method is a scoring method using bayesian approach which calculates a posterior distribution from a prior distribution given a data (2). It uses monte carlo integration for a complex distribution (5) .

```
Read a tabular data
Data Categorization
Determine number of repeat : jRepeat
For i = 1 to jRepeat
   Score data using bayesian score in MCMC method
      Determine structure (id from combination graph) with
            the maximum score
   Collect the id in a List of idGraph
End

For i = 1: length of List of idGraph
     Collect a unique id Graph from List of idGraph
End
For each unique id Graph
    Calculate number of parents of the unique graph
End
Determine the unique idGraphs with the maximum parents as a
predicted DAG from the data
```
Figure 3. Algorithm to Predict a DAG Structure from Data

SEEK DIGITAL LIBRARY

Proc. of The Second Intl. Conf. On Advances In Computing, Control And Networking - ACCN 2015
Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-073-6 doi: 10.15224/ 978-1-63248-073-6-16

$$\int_a^b h(x)dx = E_{p(x)}\,[f(x)] \cong \frac{1}{n}\sum_{i=1}^{n} f(x_i) \quad (5)$$

where *h(x)* is the production of a function *f(x)* and a probability density function *p(x)* defined over the interval (a,b).

A Markov Chain is a sequences of density (θ) which make a stationary distribution. Each density is determined by

sampling the data that initiated with a random value θ_0 and evaluated by using a ratio of density, α, (6) in each step of scoring according to Metropolis Algorithm) [17].

$$\alpha = min\left(\frac{f(\theta^*)}{f(\theta_{t-1})}, 1\right) \quad (6)$$

where $f(\theta^*)$ is a candidate of current point and $f(\theta_{t-1})$ is a point before the current. If the candidate point value is lower than a point before it, the candidate is accepted, and becomes an element θ_i of sequences of markov chain.

Structure learning in this experiment uses MCMC method from Murphy [18]. In learning the structure, it does not always result in a true structure, but an approximate structure, called markov equivalence. The structure predicted sometimes has different direction of arcs (edges), or has more or less arcs. It is because the search space ranging of a DAG is very wide includes all combinations of DAG of N nodes. The number of combinations increases super exponentially when the number of nodes increases. To maximize the search of prediction structure, an algorithm (Fig. 3) is used.

V. Ontology Development

Ontology in this work is developed by a method with machine learning approach using a combination of techniques of clustering and bayesian network (Fig. 1). Clustered data are processed by the method for predicting a structure (DAG) and the process uses the prediction algorithm (Fig. 3). All nodes from the clustered DAGs are analyzed to find connector nodes between the clustered DAGs.

A. *Connector Node Analysis*

A connector node is a node from a clustered DAG (Fig. 8a) which has relations (edges) to other nodes from different clustered DAG. These relations facilitate the development of ontology. The algorithm for connector node analysis consists of three steps as follows:

Step I : determine candidates of connector nodes

- define status of nodes which can be as parents in clustered DAGs (Fig. 8a) and the general DAG (Fig. 8b). The status will be used to determine candidates of connector nodes. The candidates are nodes which have relations (edges) within nodes in the same cluster and also between nodes in different clusters.
- select the similar candidate nodes which are parents in clustered DAGs and the general DAG

Step 2: find relations of candidate nodes by determining a terminal node as a couple of a connector node.

- Find nodes from different clusters which have relations with a candidate of connector node.
- Select a terminal node that has the most children, if the candidate of a connector node has more than one relation to a cluster.

Step 3:Eliminate the same relation and do a mapping

- Determine a couple of nodes (connector node – terminal node) as a coordinate (x, y) in an adjacency matrix (graph representation) and put them in a list.
- Find the cluster of each node as a couple of cluster.
- Remove a couple of node from the list if its couple of cluster has existed.
- Map the final couple of nodes (connector nodes and terminal nodes) as coordinates in a new adjacency matrix of DAG. The matrix combines clustered DAGs and the new coordinates as a graph structure of an ontology model.
- Visualize the ontology structure.

Fig. 10 is an example of the ontology model. It is a result of the node connectors analysis process as a part of methodology of this work (Fig. 1).

VI. Experimental Results and Analysis

The experiments consist of two parts. Experiment I is a comparison of scoring methods in structure learning by calculating averages and its standard deviations of maximum score from a collection of modeled DAGs with 4 nodes and 4 edges (Fig. 4). The visualization of maximum score versus size of data using some different bayesian scoring methods [18] (BIC, BDeu and MCMC) is presented in Fig. 5. Fig. 6 shows deviations of averages of maximum score (Fig. 5). Each maximum score of a specific size of data is related to a number of DAGs (Fig. 7).

In this experiment, the curves of score using several methods show the same trend (Fig. 5), following the change of size of data. The most maximum scores are reached in size of data about 200, for case of this modeled DAGs (4 nodes). According to the values of curves, MCMC method shows the highest score along with the size of data. Fig. 6 shows the deviation scores from the averages of maximum scores using MCMC and K2 methods with BIC score. The two methods present the deviation of the average scores, which become larger after size of data exceed 200, for this case. Following the average scoring, the trend curves of the numbers of DAGs which matches with the maximum score in every size of data (Fig. 7) become smaler. The numbers is minimum in size of data is 200, which related to the maximum score. It means, in the maximum score, number of candidate of prediction DAG after the structure learning process is small, and the range area to predict the DAG is more specific.

SEEK DIGITAL LIBRARY

Proc. of The Second Intl. Conf. On Advances In Computing, Control And Networking - ACCN 2015
Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-073-6 doi: 10.15224/ 978-1-63248-073-6-16

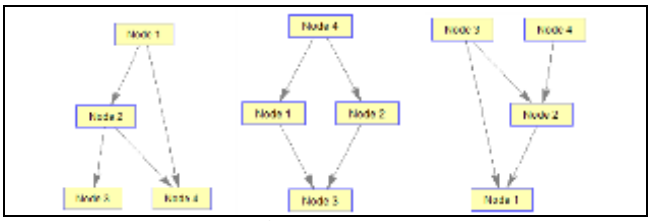

Figure 4. Examples of random Modeled DAGs with N=4

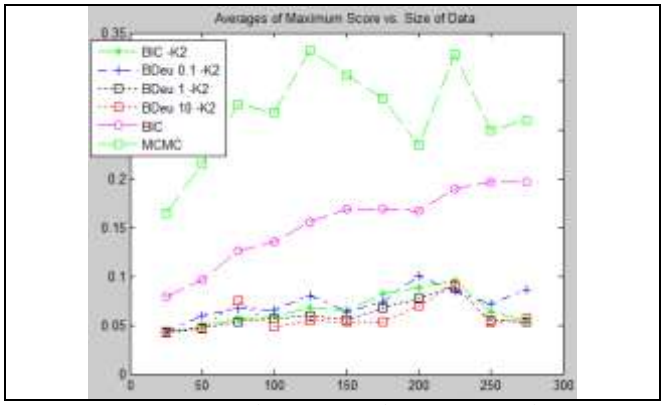

Figure 5. Average of Max. Score of DAGs vs Size of Data

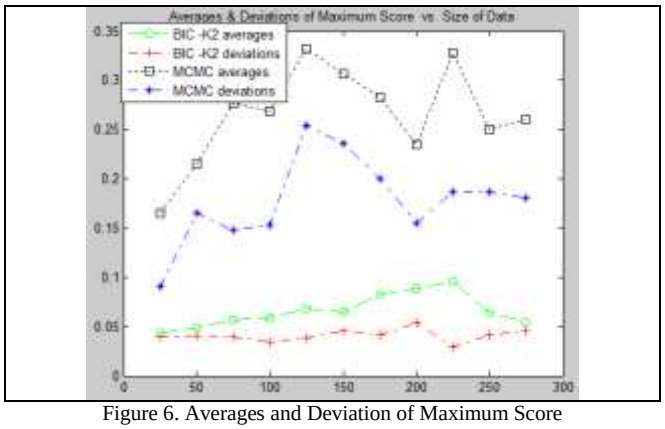

Figure 6. Averages and Deviation of Maximum Score

The second experiment uses modeled data to implement the methodology proposed (Fig. 1). This experiment needs three types of DAG described in section III (modeled DAG, clustered DAG and general DAG). The modeled DAGs is determined (Fig. 8a) and sampled and combined to be a vector data as an inputted data. Clustering and structure learning is applied to the inputted data and the results are clustered DAGs (Fig. 10a). Structure learning without clustering is also applied to the inputted data and the result is a general DAG (Fig. 8b).

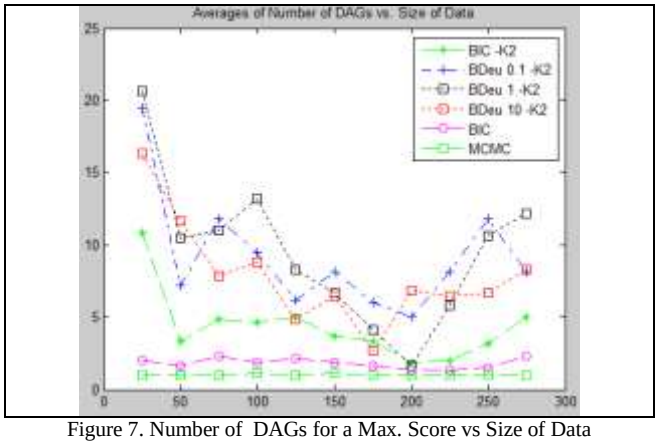

Figure 7. Number of DAGs for a Max. Score vs Size of Data

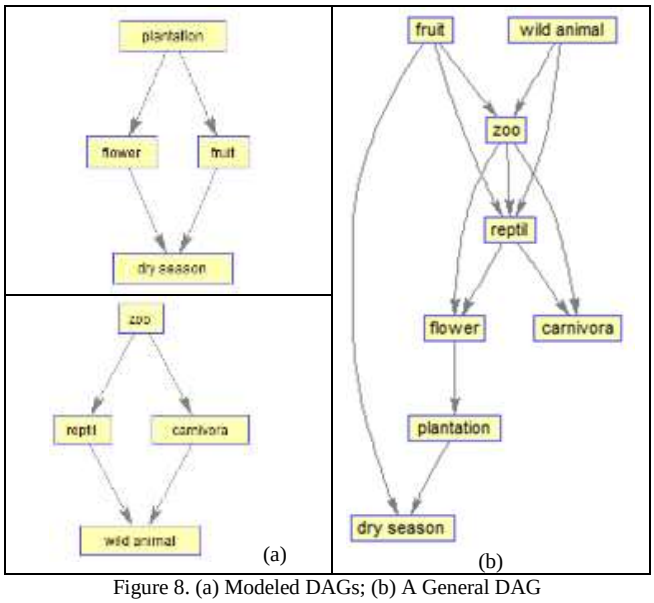

Figure 8. (a) Modeled DAGs; (b) A General DAG

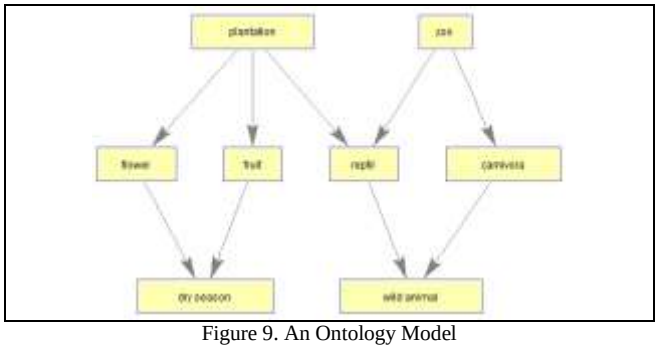

Figure 9. An Ontology Model

From the experiment I, it shows that by calculating and visualizing averages of maximum scoring for a DAG collection and a specific number of nodes and edges, a certain number of data size needed can be determined as an

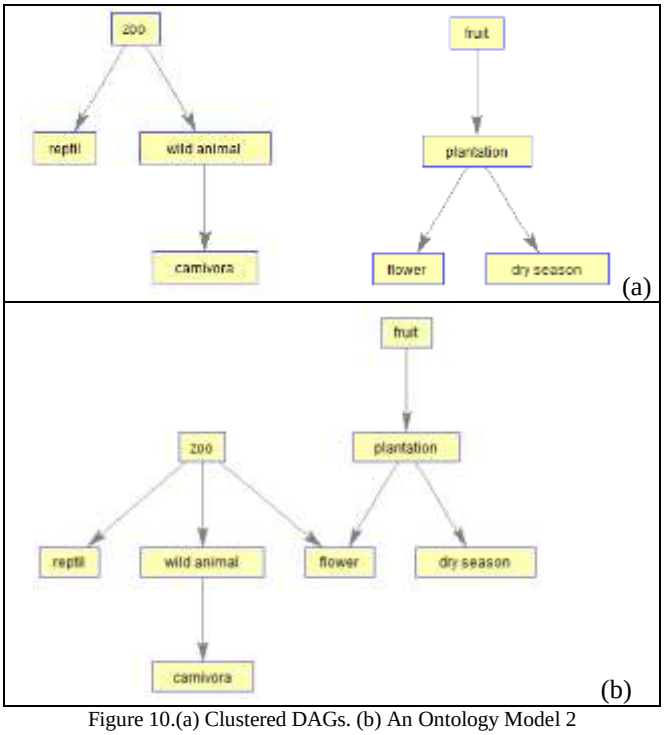

Figure 10.(a) Clustered DAGs. (b) An Ontology Model 2

SEEK DIGITAL LIBRARY

Proc. of The Second Intl. Conf. On Advances In Computing, Control And Networking - ACCN 2015
Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-073-6 doi: 10.15224/ 978-1-63248-073-6-16

information about size of data for structure learning in experiment II. Experiment II is an ontology development process. The learning results show the clustered DAGs (Fig. 10a) which are different from the modeled DAGs (Fig. 8a). By analyzing the general DAG and clustered DAGs, the connector nodes between clustered DAGs are determined. A combined clustered DAGs (Fig. 10b) is formed and called an ontology model from a vector space data. An ideal ontology model is shown in Fig. 9, in which clustered DAGs is the same as modeled DAGs.

VII. Conclusion

The automatic ontology development proposed offers a different approach which combines clustering and a probabilistic approach. The result depends on size of data, clustering process, and structure learning process. With a probabilistic approach, a DAG prediction (clustered DAGs) can approximate the modeled DAG. A connector node which is a node having the strongest relation to its cluster and also to different cluster can be determined. It has an important role to relate to all clustered DAGS as an ontology.

For future work, the experiments will process the structure learning with optimization to get a better DAG prediction, develop more rules to improve the process of connector node analysis and apply the methodology to a real data.

Acknowledgment

We wish to acknowledge Prof. Boris Mirkin for his contributions to this research. This research was supported partially by Grant from Directorate of Higher Education of Indonesia.

References

[1] D. G. Rajpathak, "Computers in Industry an Ontology Based Text Mining System for Knowledge Discovery from the Diagnosis Data in The Automotive Domain," in Computers in Industry, no. 64, pp.565-580, April 2013. doi:10.1016/j.compind.2013.03.001

[2] A. Mitrofanova, V. Pavlovic, and B. M. Fellow, "Prediction of Protein Functions with Gene Ontology and Inter-Species Protein Homology Data," in Journal of IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 6, no. 1, pp. 1–10, October 2009.

[3] P. Waraporn, P. Meesad, and G. Clayton, "Ontology-supported Processing of Clinical Text using Medical Knowledge Integration for Multi-label Classification of Diagnosis Coding," in International Journal of Computer Science and Information Security, vol. *7*, no. 3, pp. 30–35, 2010.

[4] M. M. Taye, M. M., "Understanding Semantic Web and Ontologies : Theory and Applications," in Journal of Computing, vol. 2, no. 6, pp. 182–192, June 2010.

[5] W. Wong, W. E. I. Liu, and M. Bennamoun, "Ontology Learning from Text : A Look Back and into the Future," in *Computing*, vol. 44, no. 4, pp. 1-36, 2012.

[6] M. Ramezani, H. F. Witschel, S. Braun, and V. Zacharias, "Using Machine Learning to Support Continuous Ontology Development, " in SAP Research and FZI Forschungszentrum Informatik, Karlsruhe, Germany, 2010.

[7] R. Studer, V. R. Benjamins, and D. Fensel, "Data & Knowledge Engineering : Principles and Methods," in Data & Knowledge Engineering, vol. 25, pp. 161-197, 1998.

[8] I. Veritawati, I. Wasito, and T. Basaruddin, "Text Preprocessing using Annotated Suffix Tree with Matching Keyphrase," in International Journal of Electrical and Computer Engineering, vol. 5, no. 3, pp. 409~420, June 2015. http://iaesjournal.com/online/index.php/IJECE.

[9] C. S. Lee, "Automated Ontology Construction for Unstructured Text Documents," in Data & Knowledge Engineering, vol.60, pp. 547-566, 2007.

[10] Y. Li, S. M. Chung, and J. D. Holt, "Text Document Clustering Based on Frequent Word Meaning Sequences," in Data & Knowledge Engineering, vol. 64, no. 1, pp. 381–404, January 2008.

[11] G. Zhee, L. Dong, L. Qi, Z. Jianyi, X. Yang, and N. Xinxin, "An Online Hot Topics Detection Approach using the Improved Ant Colony Text Clustering Algorithm," in Journal of Convergence Information Technology, vol. 7, no. 2, pp. 243–252, February 2012. doi:10.4156/jcit.vol7.issue2.29

[12] R.Jensi and Dr.G.Wiselin Jiji, "A Survey on Optimization Approaches to Text Document Clustering**,** " in International Journal on Computational Sciences & Applications (IJCSA), vol.3, no.6, pp. 31 – 44, December 2013. DOI:10.5121/ijcsa.2013.3604 31.

[13] L. Jing, M. K. Ng, X.Yang, and J. Z. Huang, "A Text Clustering System based on K -Means Type Subspace Clustering and Ontology," in World Academy of Science, Engineering and Technology, vol. , no. 4, pp. 1070-1082, 2008.

[14] V. Schickel-zuber, V. "Using Hierarchical Clustering for Learning the Ontologies used in Recommendation Systems," in. Artificial Intelligence, vol 1, pp. 599-608, 2007.

[15] B. Gal I,"Bayesian Networks, in Ruggeri F., Faltin F. & Kenett R.,Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons, 2007.

[16] R. E. Neopolitan, "Learning Bayesian Network", Prentice Hall Series in AI, 2004.

[17] B.Walsh, "Markov chain Monte Carlo and Gibbs sampling" ----- Lecture notes for EEB 581, version April 2004. http://web.mit.edu/~wingated/www/introductions/mcmc-gibbs-intro.pdf

[18] K. Murphy, A Brief Introduction to Graphical Models and Bayesian Networks, 1998. http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html.

SEEK DIGITAL LIBRARY