Jurnal Teknologi

Full Paper

DOMAIN BASED WEB RETRIEVAL DRIVEN BY SEMANTIC REPRESENTATION OF PERSONAL KNOWLEDGE ORGANIZER

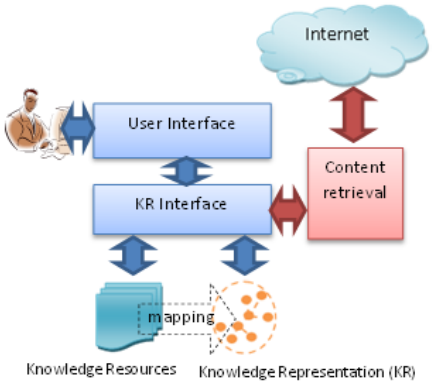
Istiadi[a,b*], Lukito Edi Nugroho[a], Paulus Insap Santosa[a]

[a]Dept.of Electrical Engineering and Information Technology, Universitas Gadjah Mada Yogyakarta, Indonesia
[b]Dept.of Electrical Engineering, Widyagama University of Malang, Indonesia

*Corresponding author
istiadi@widyagama.ac.id

Article history
Received
15 October 2015
Received in revised form
13 December 2015
Accepted
12 January 2016

Graphical abstract



Abstract

One important aspect in self-directed learning is to find learning resources, which can be referenced from the Internet. In the process of finding the resources, users have experience in accessing certain sites for specific information. But, when they want to refer back to related information, usersare supposed to recall the domain of sites and to determine the proportion of priority on a number of sites. The experience of accessing the Internet is mapped into semantic representation of Knowledge Organization (KO) like a memory thatit can be exploited to support in gaining new resources in the same domain of knowledge. This study aims to develop a web retrieval application to assist users based on their experience. The application will identify the domains of the sites that were previously accessed by tracing the representation. The domains are used as references to direct the search. Searching the web resources is undertaken by the composition of domain based search and search without domain limitation. The composition ofboth searchescan be determined by the users according to their expectations. The domain based search composition is determined by the resourcecontributions within the scope of a knowledge domain. The test results show that the system is able to identify the domain's location and the proportions as preferences for searching. The performance of the system in searching shows thatthe query returnsof relevant documents are dominant.

Keywords: Web search; agent; ontology; knowledge organization

1.0 INTRODUCTION

A part of self-directed learning process is to find resources[1]. Internet is one of the potential networks to obtain information that can be used as learning resources. The availability of huge and varied resources on the Internet will require efforts to find appropriate resources. Thus, the problem of the information-seeking has become one of the aspects that needs to be supported by cognitive tools [2].

The behaviors of internet users in obtaining information influence their experience or expertise [3]. In certain conditions, users with their knowledge tend to refer to sites on a URL base (domain name) for specific information. The sites contribute to provide information as needed. Therefore, to obtain other related information, the users are supposed to recall the domains and then browse through the pages or links by navigating or using search engines. This process would still require attention, for example the users need to remember the domains and to determine priority proportion of the domains that contribute to the development of their knowledge.

Meanwhile, another aspect of cognitive tools is Knowledge Organization [2]. In a large scale,

Knowledge Organization System (KOS) is used by an institutional or community level [4], but on a personal level KOS is needed to support the process of cognition [2,5]. Thus, KOS can be local and personal. KOS is needed to integrate chunks of information into a unified whole in a domain of knowledge. KOS is a medium to describe the resources and to provide relations to declare the association between resources [2]. Relationship with a particular meaning will form a semantic representation. The advantages of the semantic representation that use semantic technology such as ontology allow to be understood by a computer program. Thus, it can be used to trace and be utilized as reference to work autonomously [6,7].

Ontology representation as shareable knowledge can be seen from different purposes [8]. From a standpoint of KOS as a medium to express the classification of resources, in which there is a property of information sources that can be in the format of the URLs. The URLs are unique addresses so that the resources can be accessed on the Internet [9]. From another point of view, the model is seen as the representation of memory that stores users' access experience. This memory holds the addresses of access that are classified in a specific information context. The existence of this memory, it can be exploited by computer programs to gain access location preferences to direct internet access.

This study aims to develop a model of software to serve the obtainment of web content to complement information on personal KOS in the domain that has been known. The software is expected to explore the access location preferences of local representation and take action search access to the Internet. Furthermore, the results will be ordered to meet the criteria of similarity for users' choice.

2.0 RELATED WORK

Study on Web Information Retrieval (IR) emphasizes on centralized services system. On such systems, the organization of resources is placed in a repository that is accessible with a particular IR method. As in [10] which uses personalized browsing behavior and collaborative filtering methods. In [11], it uses the world knowledge base to support mapping the resources together. In [12], it uses an approach of users' model that utilizes relevant feedback services with evolutionary computation methods. Where as in [13], it uses adaptive neuro fuzzy method to evaluate the content and structure of retrieved web resources. Whereas in [14], it proposed query recommendation use Normalized Discounted Cumulative Gain (NDCG) and Coefficient of Variance (CV) methods.

The resources are organized locally and personally as in [5], which the optimization of content retrieval service has not been further exploited. When the user access address is recorded in organizer and is classified in the domain of knowledge as the context of information, then it is potential to support the discovery of a new resource in the same domain. This study utilizes a semantic representation of the organizing scheme to identify the domain of the access locations to direct the search.

3.0 SYSTEM DESIGN

The system design (Figure 1), illustrates the main parts of the expanded KOS with content retrieval features. Knowledge Representation (KR) is a scheme which maps descriptions and relationships of documents as knowledge resources. Organizing knowledge resources is basically to manage the KR through a User Interface. The KR uses semantic technologies such as OWL required interfaces (KR Interface) so that the applications can read or write data. Feature web retrieval service is an additional part that works based on data obtained from the KR. Based on these data, accessing the internet is done to obtain resources to complement existing resources within the same domain. Furthermore, the knowledge representation model and the software application model are presented below.

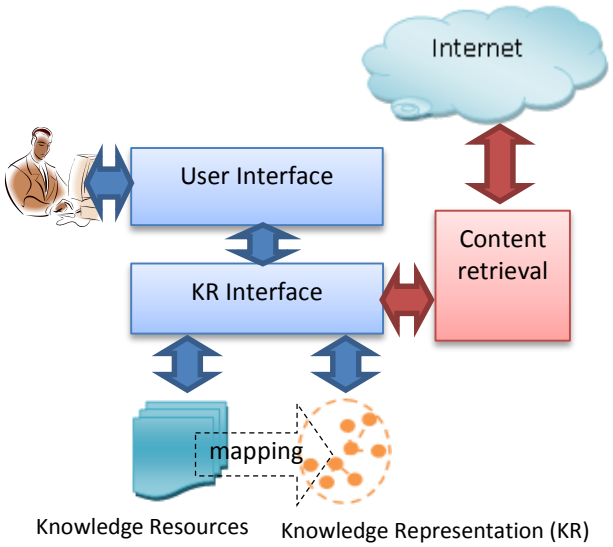


Figure 1 Architecture of KOS with web Retrieval feature

3.1 Knowledge Representation Model

The main part of the KOS is a knowledge representation (KR). The KR in the form of a scheme is to tag the resources and to provide relationship to be connected among others [2]. With this scheme users can construct resources integration. So, the scheme should accommodate description of resources elements and to represent knowledge structure. The design of the scheme's organizers uses the terms of a Knowledge Object (KO) and a Knowledge Domain (KD) refers in [15]. KO functions as a resource descriptor and KD represents of nodes or entities in the structure of knowledge. Furthermore, the KO and KD are used in the main concept of the Ontology representation (Figure

2).The advantages of the ontology are the capability to provide basic scheme as a body of knowledge that support semantic services [16],that it also possible to handle the heterogeneity of data [17].

KD concept contains objects as entities that use multiple relationships to represent the structure of knowledge. In the context of organizational knowledge, these relations should give a meaning that can be interpreted. In general, the relationships are categorized on hierarchical (BT/NT), associative (RT) and Equivalency (Use/Used For) as in [18]. But some organizers of resources use the standard relationships that are more specific such as DCMI[19], including has Part/is Part Of, References / is Referenced By, Requires/is Required By, is Based On/Is Basis For. In this study, the representation model refers to the specific relationships (DCMI) by adding Use/Used For to accommodate equivalency relations. Completing KD entities, properties of keywords and notes are added to complement the representation coverage of content.

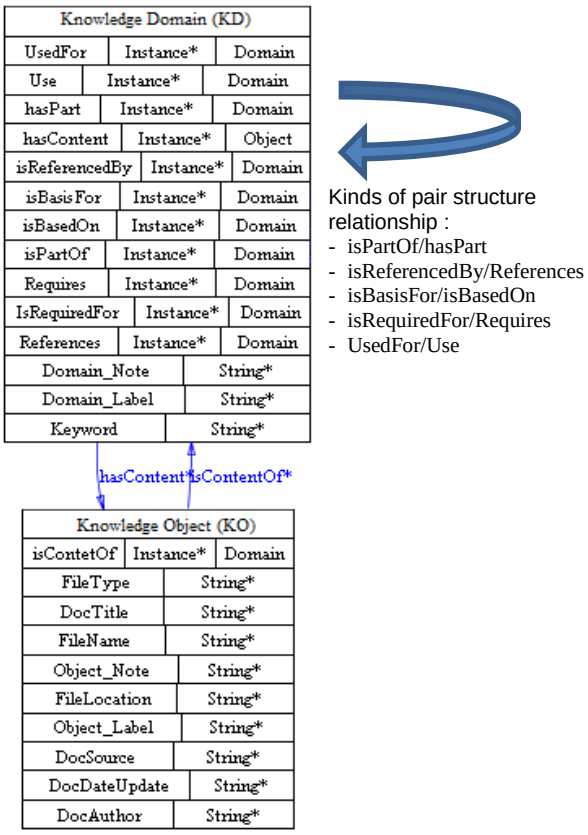


Figure 2 Model Ontology of KR

KO concept contains objects that consist of resource attributes as a digital document. Some attributes of the resources use have been derived from DCMI such as title (Doc Title), creator (Doc Author), description (Object_ Note), source (Doc Source), type (File Type), and several other attributes are added such as filename (File Name, file location (File Location) and date update(Date Update). KO is connected with KD as a content using is Content Of relation and uses has Content relation as the inverse. Property of Doc Source that contains the location of the resources obtained from networking using the URL format. The Doc Source properties will be explored further to support the content retrieval through Internet

Resources organized hierarchically (using is Part Of relation) will form aggregator [20],it means the composition a number of resources can be seen as a resources classifier. Classification of resources means grouping the resources based on information domain expressed in knowledge structure. A KD can cover narrower KD using has Part relationship, or another KD can be surrounded by wider KD using is Part Of relationship, so it will form the levels of information from the broad scope to the narrow scope or otherwise.

On the other side, with both types of relationships, KD may have or as a branch of another KD. Thus, it establishes a kind of areas within a scope. The levels and the areas will form like a map of the classification which allows exploration. If a KO as content of KD that has property of information source of its resources, it will obtain a map of resources classification. It will be some kinds of clue where to find resources in a knowledge domain. A scheme which provides information that can be used as the basis of a program operation can be seen as a kind of knowledge base (KB) [21].

According to the proposed concept, the retrieval information is based on URL base. This is analogy to the behavior of Internet users that are influenced by access experience. Specific domains will tend to be visited because it is expected to contribute [3]. This is in line with the advanced service of web search engines to limit a search to a domain host. Hence, the program should trace KB to obtain potential domains of sites that are expected to contribute information. Furthermore, in order the program can trace the KB, it requires a mechanism as in previous work in [22], as follows:

a. Starting from new defined KD (*Do*) as part of another KD, then finding KD (*Dx*) at upper level that covers it as expressed in (1).
   $Is\ Part\ Of(Do, ?Dx) \rightarrow coverage(?Dx)$ (1)
b. From the KD (*Dx*) which finds, query all KD (*Dy*) below and getting the document sources (*Sy*) in KO properties (*Oy*) as expressed in (2).

   $hasPart(Dx, ?Dy) \wedge hasContent(?Dy, ?Oy) \wedge DocSource(?Oy, ?Sy) \rightarrow contains(Dx, ?Sy)$ (2)

c. If the URLs are found then extract them to obtain the URL base that utilize as the preference for searching information. But if the location of the URL is not found, go to KD on a broader scope on the upper level and repeat step b.

3.2 Software Application Model

Based on the objectives to be achieved, there are tasks that should be fulfilled by the application of content retrieval services on KOS. First, the ability of system explores the organizing scheme to get data (URLs) and to extract domain (URL base) as an access preference.

Second, the ability to access the Internet based on the preferences and keywords that have been defined. Third, the ability of system to evaluate
the searching results and recommends it as a candidate of the knowledge content to be chosen by the users.

Appropriate to the requirements, a design of KOS software model with content retrieval features is shown in Figure 3. The design illustrates a behavioral model in

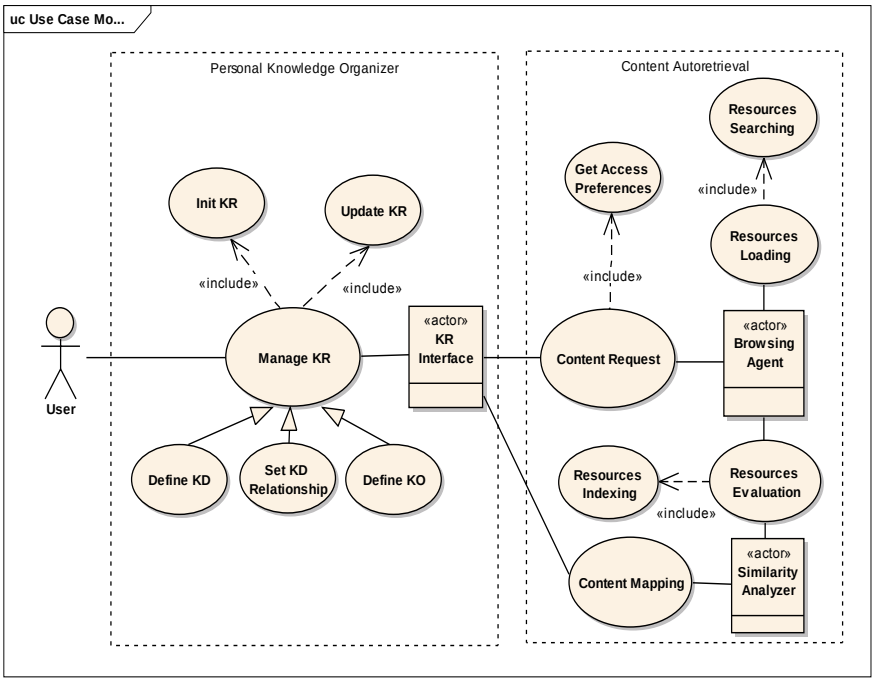


Figure 3 Use case diagram of application software

the form of a use case diagram, which refers to our previous work in [23].

The KOS design consists of two scopes, initially resources organizing services. Furthermore, it is expanded with web retrieval services. Organizing knowledge in particular provides an interface for users to manage KR. Managing the KR especially for tasks such as defining KD, setup relations and description of KO. KR Interface functions to mediate access to KR procedurally. Associated with the expansion of content retrieval services, there are elements that play a role internally as actor identified. Browsing agent roles to represent user to access resources on the Internet based on availability of the preferences. Similarity analyzer plays to evaluate the access results, and then it is sorted based on similarity scores for the content candidate which can be selected by users. In this study, similarity evaluation uses Vector Space Model (VSM) which based on TF-IDF method [24].

When a new KD is defined without including KO, users can submit a request to complete it. Subsequently, KR-Interface will explore KR to gain access preferences in the form of domains. The preferences are included to determine the portion of accessing the location in the search. This portion is determined based on its contribution in providing the information within a scope of KD.

The value of a domain contribution (C_d) is calculated based on number of the resource existences that come from that domain. Thus, the all resources contribution (C_R) is the sum of the domain contribution value (C_d) as expressed in (3).

$$C_R = C_{d1} + C_{d2} + \cdots + C_{dn} \quad (3)$$

The whole search (S) accommodates the domain based search (S_D) and search without domain restriction ($S_{\neg D}$) as expressed in(4).

$$S = S_D \cup S_{\neg D} \quad (4)$$

While the domain based search (S_D) is a combination of a proportion of each domain location search (Sd) as in(5)

$$S_D = S_{d1} \cup S_{d2} \cup \ldots \cup S_{dn} \quad (5)$$

If the S_D value comparable to the C_R value, then the proportion of the value of a domain based search (S_d) is defined as in (6)

$$S_d = S_D \left(\frac{C_d}{C_R}\right) \quad (6)$$

Once the domain location and the proportion can be determined, then the Browsing Agent will take action with direct internet access based on the preferences. In the searching action, users can configure the composition ratio of the domain based search (S_D) and the search without restrictions ($S_{\neg D}$). The combination of the both searching actions will complement each other. If the users are expecting resources from the domains that have contributed, they can define greater value of the S_D portion than the $S_{\neg D}$ portion, and vice versa.

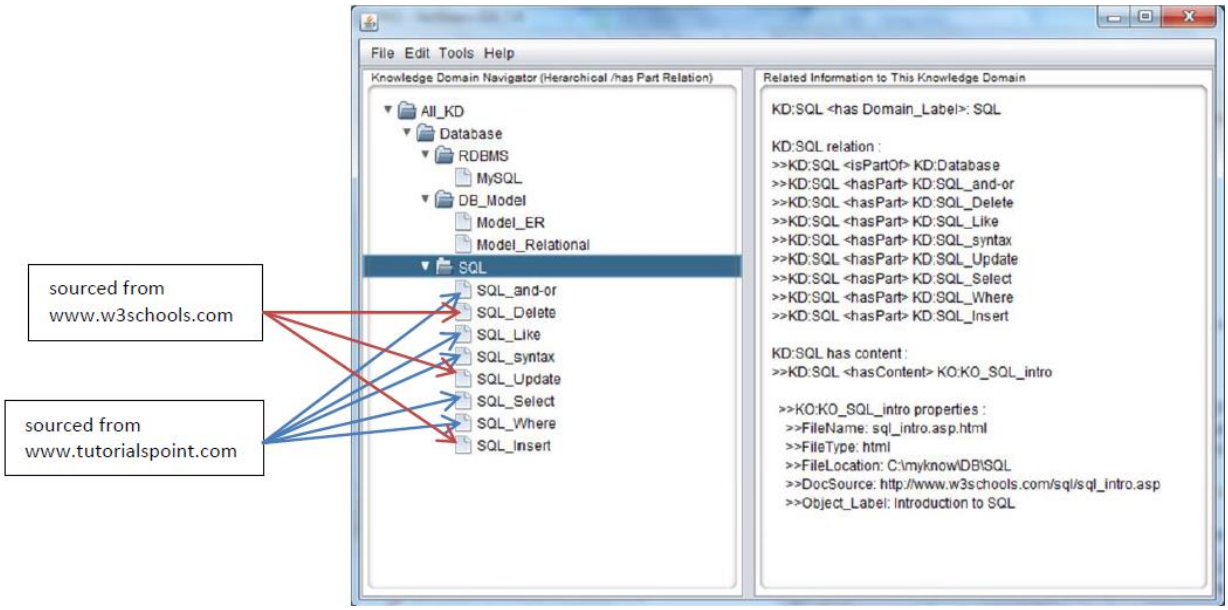


Figure 4 Case example of the resources organizing using the application

Similarity analyzer works further to index and to evaluate the searching results. VSM is used to rank the results based on a similarity score of a given keyword. The users can select the results after they are mapped into related KD as a content candidate.

4.0 EXPERIMENTAL

This system is implemented as a prototype that is developed using Java Platform. KR Interface is created using JENA library, which provides access to KR procedurally [26]. KR expressed uses OWL, which is generated with the aid of Protégé [27]. Java platform supports network programming library for accessing the Internet. The web searching capability utilizes Bing Search API. Similarity analyzer is based on Lucene library that supports the measurement of similarity with the VSM method.

System testing is carried out by using an example of a case in a scenario. For example, if a person is interested in studying Information Systems Development, and in the learning process there is learned knowledge of database. In the knowledge of database, previously the person needs to study SQL for data manipulation operations, such as sql syntax, sql select, sql insert, sql delete, sql update, sql and-or, sql like, and sql where clause with references to specific tutorial sites. Resources are organized as shown in the Figure4that illustrates the person who wants to increase knowledge of the operation of user grant, data sorting, join tables, table view and triggers. Based on the experience of access, the person expects to be assisted by the available tools to complete these resource needs. The person can set up the comparison of the domain based search more than the search without limitation, for example with a ratio of 60%:40%. The total amount of the desired results can also be specified in the configuration, such as the ten items of results.

Furthermore, the keyword as the content specification to process the searching are defined when a new KD is created. The users can define specific terms for keywords and can add common terms which internally available in the same domain. For example, the users select SQL term that is also used as a keyword of the resources that already exist.

In accordance with the scenario, the software application will run by each search resource. This search result is the composition of the domain based search and search without limitation domain. For each search query of resources, the results are sorted by similarity score assessment using VSM. Because the experiments were carried out directly on the Internet with a very large number of resources, the evaluation was done by calculating the Mean Average Precision (MAP) on a certain amount of top searching results[24,25].MAP calculates the mean of the average precision (AveP) of searching results for set queries (Q) as shown in (7).

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} AveP(q) \qquad (7)$$

5.0 RESULTS AND DISCUSSION

Table 1 presents the results of the access preferences identified. There are two domains detected, www.w3schools.com and www.tutorialspoint.com. The first domain contains three resources and the second domain contains five resources. This relates to the contribution value (C_d) are used to determine portion value (S_d) of the domain based searching as in (6).

Table 1 Identified Preferences

No	Identified URL base	Contribution Value (Cd)	Portion Value(Sd)
1	www.w3schools.com	3	2
2	www.tutorialspoint.com	5	4

Under the previous configuration, that the domain based search is 60% of the ten expected results. Thus, the first domain gets rounding two portions and the second domain gets rounding four portions. In accordance with the portion, other four parts remaining are the search without limitation. Furthermore, the search is carried out by tools with reference to the proportions.

Table 2 presents the search results evaluation of a given cases. There are five queries followed by keywords. Average Precision (AveP) is calculated for each query from ten items of search results. Relevant justification of the documents to the query that is intended is done by three lecturers who are experienced in teaching of database courses. Finally, the MAP calculated of all queries.

Table 2 Result of MAP calculation

Query	Keyword	AveP@10
q1	user grant sql	0.82
q2	table join sql	0.81
q3	data sorting sql	0.77
q4	database trigger sql	0.97
q5	table view sql	0.74
	MAP	0.822

According to the results in Table 2, q5 obtains the minimum score with a value of 0.74, while the maximum score is 0.97 in q4. MAP value of 0.822 indicates that the query returns of relevant documents are dominant for a given case.

6.0 CONCLUSION

Based on the design of systems that have been developed, the functionality of the system is able to meet the capabilities that are expected to add features in the form of web search services to support complement resources. This capability is indicated by the ability to provide preferences extracted from KR of KOS. The preferences are domains with a number of searching portion. The performance of the system in searching shows that the queries returns of relevant documents are dominant for a given case.

References

[1] Bouchard, P. 2009. Pedagogy Without A Teacher: What Are The Limits?. *International Journal of Self-Directed Learning*. 6(2): 13-22

[2] Iiyoshi, T., Hannafin, T.M., and Wang, F. 2005. Cognitive Tools And Student-Centred Learning: Rethinking Tools, Functions And Applications. *Educational Media International*. 42(4): 281–29

[3] White, R. W., Dumais, S. T., and Teevan, J.2009. Characterizing The Influence Of Domain Expertise On Web Search Behavior. *The Second ACM International Conference on Web Search and Data Mining*. ACM. 132-141

[4] Hjørland, B. 2008. What is Knowledge Organization (KO)?. Knowledge Organization. *International Journal devoted to Concept Theory, Classification, Indexing and Knowledge Representation*. 35(2/3): 86-10

[5] Vavoula, G. and Sharples, M. 2009. Lifelong Learning Organisers: Requirements For Tools For Supporting Episodic And Semantic Learning. *Educational Technology & Society*, 12(3): 82–97

[6] Clark, K.L., and McCabe, F.G.2007.Ontology schema for an agent belief store. *International Journal of Human-Computer Studies*. 65(7): 640-658. http://dx.doi.org/10.1016/j.ijhcs.2007.03.004.

[7] Su,CJ., and Peng.C. W. 2012. Multi-Agent Ontology-Based Web 2.0 Platform For Medical Rehabilitation. *Expert Systems with Applications*. 39(12):10311-10323http://dx.doi.org/10.1016/j.eswa.2011.09.089.

[8] Chungoora, N., Young, R. I., Gunendran, G., Palmer, C, Usman, Z., Anjum, N. A., Cutting-Decelle, AF.,Harding, J.A., and Case, K. 2013. A Model-Driven Ontology Approach For Manufacturing System Interoperability And Knowledge Sharing. *Computers in Industry*. 64(4): 392-401. http://dx.doi.org/10.1016/j.compind.2013.01.003.

[9] Ietf, R. F. C. 2616, 1999. Hypertext Transfer Protocol—HTTP/1.1. [Online] From http://www. rfc. net/rfc2616. html

[10] Mittal, N., Nayak, R., Govil, M. C., and Jain, K. C. 2011. Personalised Search–A Hybrid Approach For Web Information Retrieval And Its Evaluation. *International Journal of Knowledge and Web Intelligence*. 2(2-3): 119-137.

[11] Tao, X., Li, Y. andZhong. N. 2011. A Personalized Ontology Model for Web Information Gathering. *IEEE Transaction on Knowledge and Data Engineering.* 23(4): 496-511.

[12] Maleki-Dizaji, S., Siddiqi, J., Soltan-Zadeh, Y., and Rahman, F. 2014. Adaptive Information Retrieval System Via Modelling User Behavior. *Journal of Ambient Intelligence and Humanized Computing.* 5(1): 105-110.

[13] Iraji, M. S., Maghamnia, H., andIraji, M. 2015. Web Pages Retrieval with Adaptive Neuro Fuzzy System based on Content and Structure. *International Journal of Modern Education and Computer Science (IJMECS)*. 7(8): 69-84.

[14] Umagandhi, R., and Kumar, A. S. 2015. Evaluation of Reranked Recommended Queries in Web Information Retrieval using NDCG and CV. *International Journal of Information Technology and Computer Science (IJITCS)*, 7(8): 23-30.

[15] Koutsantonis, D. and Panayiotopoulos, J.C. 2011, Expert System Personalized Knowledge Retrieval. *Operational Research*. 11(2): 215-227.

[16] Mohd-Hamka, N., & Mohamad, R., 2014, OntoUji–Ontology to Evaluate Domain Ontology for Semantic Web Services Description, *JurnalTeknologi*, 69(6): 21-26.

[17] Yunianta, A., Yusof, N., Othman, M. S., Aziz, A., Dengen, N., Ugiarto, M., Haeruddin, Angelina, J. 2014. Semantic Data Mapping on E-Learning Usage Index Tool to Handle Heterogeneity of Data Representation. *Jurnal Teknologi*, 69(5): 1-6.

[18] Fayen, E. G. 2007. Guidelines For The Construction, Format, And Management Of Monolingual Controlled Vocabularies: A Revision Of ANSI/NISO Z39. 19 For The 21st Century. *Information Wissenschaft und Praxis*. 58(8): 1-184.

[19] DCMI-terms. 2012. DCMI Metadata Terms, [Online] from :http://dublincore.org/documents/2012/06/14/dcmi-terms/
[20] Guns, R. 2013. The Three Dimensions Of Informetrics: A Conceptual View. *Journal of Documentation*, 69(2): 295 – 308. http://dx.doi.org/10.1108/00220411311300084
[21] Russell, S. and Norvig, P. 2003. *Artificial Intelligence: A Modern Approach*, Prentice-Hall.
[22] Istiadi, Nugroho, L.E., and Santosa, P.I.2014. Providing Information Sources Domain for Information Seeking Agent From Organizing Knowledge. *The 1st International Conference on Information Technology, Computer and Electrical Engineering, (ICITACEE 2014)*. Semarang, Indonesia. 8 November 2014.
[23] Istiadi, Nugroho, L.E., and Santosa, P.I. 2014. An Agent Model of Domain based Information Seeking on Personal Knowledge Organizer. *International Conference on Information Technology Systems and Innovation*, Bandung-Bali , Indonesia. 24-27 November 2014.
[24] Manning, C.D., Raghavan, P. andSchütze, H. 2008. *Introduction To Information Retrieval*. 1. Cambridge: Cambridge University Press.
[25] Ali, R., and Beg, M. S.2011. An overview of Web search evaluation methods. *Computers & Electrical Engineering*. 37(6): 835-848.
[26] Ameen, A., Khan, K. U. R., and Rani, B. P., 2014, Reasoning in Semantic Web Using Jena. *Computer Engineering and Intelligent Systems*. 5(4): 39-47
[27] Horridge, M.2011.*A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools Edition 1.3*. The University Of Manchester