ONTOLOGY MAINTENANCE BASED ON VOTING and SEMANTIC SIMILARITY APPROACH IN P2P ENVIRONMENT

Lintang Yuniar Banowosari[1], I Wayan Simri Wicaksana[2], Suryadi H.S[3]
University of Gunadarma, Jl. Margonda Raya no. 100, Depok, Indonesia
Email: {lintang, iwayan, suryadi_hs}@staff.gunadarma.ac.id

Abstract

Internet has contributed great value for data exchange, on other hand, Internet introduced some new issues. Currently, information sources are more massive, distributed, dynamic and open. Diversity is one of focus to overcome in Internet era. Some approaches have been delivered, such as semantic web and Peer-to-Peer (P2P). P2P allows community which common interest to be in a group or cluster (SON - Semantic Overlay Network). The similar interest in SON will reduce the problem of diversity in concept between peers. One of approach in semantic web is by implementation common ontology as reference for information sharing. However, P2P is very dynamic and autonomous, some adjustment of ontology is important to handle this situation. The common ontology in a period will be not satisfied anymore for the community members as reference of interoperability. An approach is needed to handle ontology maintenance in the P2P environment. Our approach is based on social approach in voting to choose the representative members. In other word, common ontology will be adjusted based on peers which represent 'appropriate' information among the cluster members.

Keywords: *maintenance, ontology, P2P, voting, similarity*

1. Introduction

Internet and Web as the information sources have advantages and problems. The main problems of the sources are more massive, distributed, dynamic, and open.

According to Sheth [21] there are heterogeneity of information and system. Information heterogeneity causes difference appearance of information system. Difference can be occurred at syntax, structure, and semantics level. To overcome the heterogeneity, some approaches have been developed. An approach based on semantic interoperability which coupled with P2P approach.

P2P make the possibility of forming the similar interest community or group. By developed the group, the semantics diversity can be reduced. This model is frequent referred with Semantic Overlay Network (SON). But this approach not yet adequate for information interoperability, so that it needs a bridge by utilizing semantic mediation approach which supported by ontology.

[1] Doctoral Student of University of Gunadarma Indonesia.

[2] Technical Advisor Researcher Partner, staff and researchers at University of Gunadarma Indonesia and University of Bourgogne France.

[3] Supervisor – University of Gunadarma Indonesia

Usage of an ontology and P2P has progressively expanded since last few years. Knowledge and content management in P2P architecture is easier then fully open system. In P2P model, ontology frequently assumed it has been already formed in the beginning. However, dynamic environment such as P2P, ontology which has been formed frequently has no longer fulfilled the concept of community member. Hence, it should be obtained a particular approach for the ontology maintenance in P2P environment.

The Semantic Web and Peer-to-Peer are two technologies that address a common need at different levels [20]:

- The *Semantic Web* addresses the requirement that one may model, manipulate and query knowledge and information at the conceptual level rather than at the level of some technical implementation. Moreover, it pursues this objective in a way that allows people from all over the world to relate their own view to this conceptual layer. Thus, the Semantic Web brings new degrees of freedom for changing and exchanging the conceptual layer of applications.
- *Peer-to-Peer* technologies aim at abandoning centralized control in favor of decentralized organization principles. In this objective they bring new degrees of freedom for changing information architectures and exchanging information between different nodes in a network.
- *Together*, Semantic Web and Peer-to-Peer allow for combined flexibility at the level of information structuring and distribution.

2. State of The Art and Related Works

2.1. Ontology Maintenance

In most applications, ontologies are not static. Instead, they have to be adapted to changing application domains, extensions of their scope, and evolving applications using them. Therefore, ontology evolution is one of the main aspects of ontology maintenance. Noy and Klein [14] argue that ontology evolution is closely related to schema evolution in databases, but that ontology evolution has certain peculiarities. Most notably, these are a different semantics and different usage paradigms. Klein et al. [11] distinguish conceptual changes (the way a domain is understood) from explication changes (the way how concepts are specified). In [12], changes to ontology are seen as sequences of individual update operations like a log file of a database system. They discuss minimal transformations between two given ontology states, i.e., how to go from one state to the other with the smallest set of individual updates and how to construct complex update operators from sequences of individual updates (represented as minimal transformations). These update operations can themselves be organized as an ontology and offered to the user in a menu. [4]

Ontologies are continuously confronted to evolution problem. Due to the complexity of the changes and maintenance process, an appropriate approach necessary to facilitate this task and to ensure its reliability. Gargouri propose a maintenance ontology model for a domain, whose originality is to be language independent and based on a sequence of text processing in order to extract highly related terms from corpus. According to [8], it deploy the document classification technique using GRAMEXCO to generate classes of texts segments having a similar information type and identify their shared lexicon, agreed as highly related to a unique topic. This technique allows a first general and robust exploration of the corpus. Further, it apply the Latent Semantic Indexing method to extract from this shared lexicon, the most associated terms that has to be seriously considered by an expert to eventually confirm their relevance and thus updating the current ontology. Finally, the result show how the complementarity between these two techniques, based on cognitive foundation, constitutes a powerful refinement process. However the method is difficult for dynamic and open environment in P2P.

The main purposes of ontology maintenance are:

- Fixing Bugs (inconsistent, inaccurate, inefficient)
- Enhancing (Tweaking{richness, correctness, organization, meta-level consistency, efficiency}, Extending {improving coverage, extending commitment, integration}, refactoring)
- Testing (regression tests, test suites, meta tag sets for test content, ablation tests) [18]

Maintenance of ontology can use some approaches. The approaches in general are:

- Mapping, where one ontology mapped to other ontology
- Merging, where two or more ontology joined becomes ontology
- Alignment, where ontology adjustment caused by change or adjustment of concept and knowledge. [1]

2.2. Ontology Maintenance with Semantic Similarity

How can we maintain a given explicit ontology in front of a dynamic world, characterized by continuously unstable textual data? How can we extract, from these texts, terms (or concepts) and their relations that are pertinent for an ontology and help maintain it? Because of the complexity of this problem, we will mainly deal in this paper, with only one dimension of this problem, which is the extraction of highly semantically related terms. Further dimensions, such the extraction of emergent terms in the texts that are related to certain ontologies, or the integration of new terms and relations with those of the current ontology, will be presented in our future work.[8]

The main issue in aligning consists of finding to what entity or expression in one ontology corresponds another one in the other ontology. Here are presented the basic methods which enable to measure this correspondence at a local level, i.e., only comparing one element with another and not working at the global scale of ontologies. Very often, this amounts to measuring a pair-wise similarity between entities (which can be as reduced as an equality predicate) and computing the best match between them, i.e., the one that minimizes the total dissimilarity (or maximizes the similarity measure). There are many different ways to compute such a dissimilarity with different methods designed in the context of data analysis, machine learning, language engineering, statistics or knowledge representation. Their condition of use depends of the objects to be compared, their context and sometimes the external semantics of these objects. Some of this context can be found in Figure 1.1 (From [19] enhanced in [11; 12] and [6]) which decomposes the set of methods along two perspectives: the kind of techniques (descending) and the kind of manipulated objects (ascending). [5]

2.3. Ontology Maintenance in P2P

In P2P settings assumptions that all parties agree on the same schema, or that all parties rely on one global schema (as in data integration) can not be made. Peers come and go in unpredictable period, import multiple schemas into the system, and have a need to interoperate with other nodes at runtime. In this activity we see schema alignment as the main process to enable nodes interoperation. Namely, when two peers “meet” on the network, they establish mappings between their schemas in a (semi) automatic alignment discovery process.

These attempts presume that ontologies have been constructed beforehand and what they are concerned about is how to use ontologies to exchange knowledge and to enable efficient and accurate semantic search in distrib-uted environments. In many application scenarios, such predefined ontologies cannot catch up with the ever-changing requirements of users. Instead, ontology should drift with the appearance of new application re-quirements. But just as [7] has stated, one cannot expect any maintenance to happen on the ontolo-gies in P2P environments (in fact, users will not often know what is in the ontologies on the machine, let alone that they perform maintenance on them) and as a result, we must design mechanisms that allow the ontologies to update themselves, in order to cope with ontological drift. [7] has proposed several informal mechanisms that use metaphors from social science (opinion-forming, rumour-speading, etc).

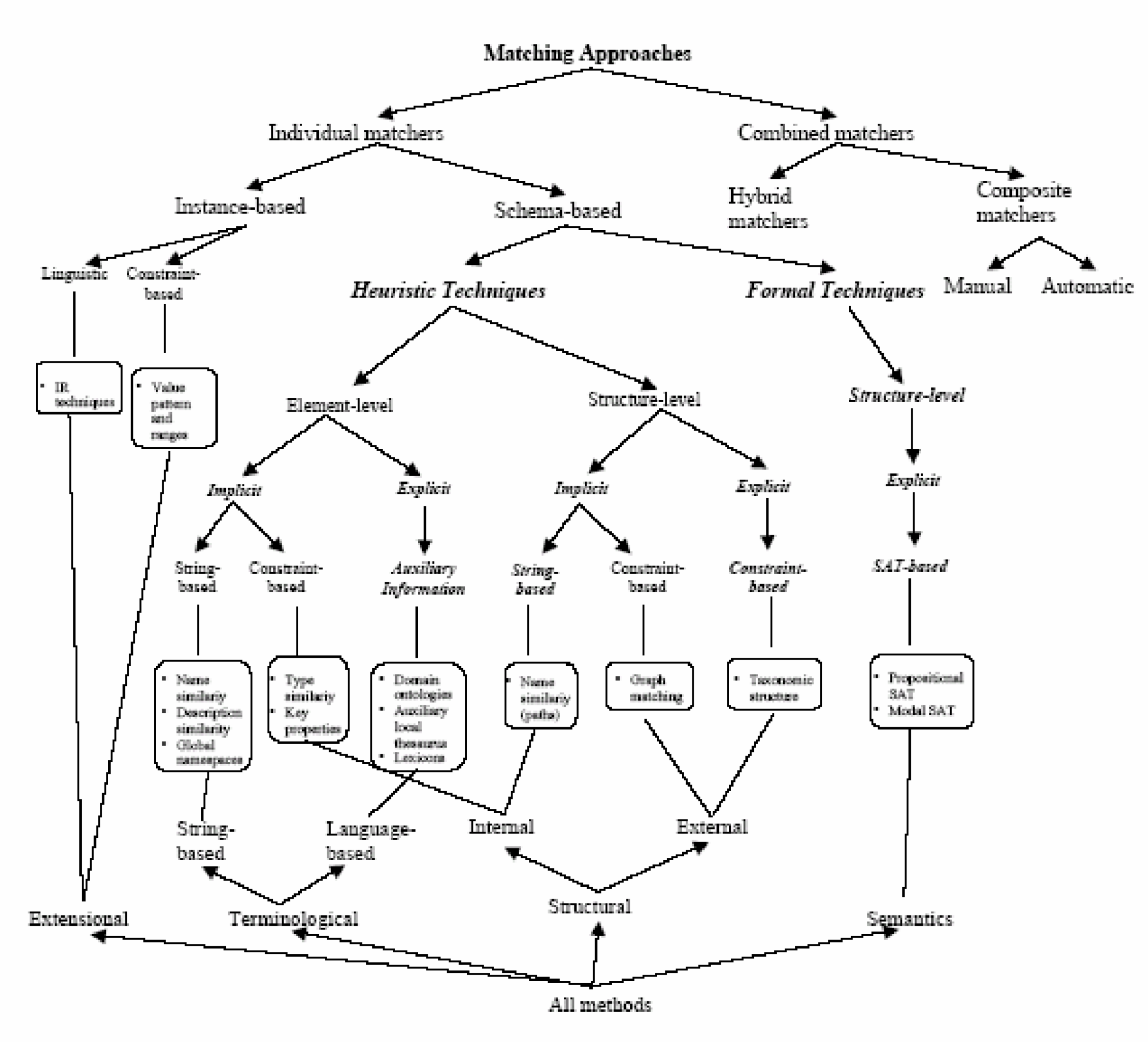

Figure 1.1: Classification of local methods [5]

In order to align concepts, to filter out noisy semantics, and to indicate the principal direction of the development of user requirements, *we propose these local ontologies be combined together to construct a common ontology*. With a common ontology, we can also improve the efficiency of semantic search by avoiding too many mappings between ontologies.

One possible way to combine the ontologies from all users is votes collecting: we collect the candidate of all users. The analysis to select of candidate considers respond in set of query. My approach of voting based on social model in the real world. Practically, a voting organizer (such as a chairman or a tally clerk) is needed to accomplish the voting task. This organizer can be considered as a server and serves for the common interests of a community by publishing messages to and receiving messages from all other voters. But in P2P environments, it may be hard to find any volunteer to serve the community for no evident good. Moreover, using a server to collect votes will bring about scalability and single node failure problems as discussed in many P2P researches. To get rid of such problems, we adopt Onto-Vote approach which is a scalable distributed votes collecting mechanism based on application-level broadcast trees, to collect votes on P2P platform. [10]

3. Research Background

3.1. Objective and Contribution of Research

Objective of the research is to find an appropriate approach to maintain common ontology based on community peer member in dynamic and open environment.

Contributions of the research are:

- Mechanism for selecting the candidate ontology by implements the voting method.
- Mechanism for maintaining common ontology by considers the similarity.

3.2. Research Questions

Refer to state of the art, the research based on some questions as initial step to conduct the research. The research questions can be seen at table 1.

Table 1 Research Questions and Proposed Solution

Step	Issues	Questions	Proposed Solution
Membership	Represent Content of Peers	How to represent content of peer?	RDF/S
		Using schema, what appropriate language?	OWL
		Is it peer in appropriate SP?	
	How and where to find an appropriate peers	What kind of P2P architecture?	Super Peer (SP)
		How to define SP?	Mechanism SP/P&SP/SP
		How to improve finding and reduce bandwidth?	SON
Ontology Voting - Election	election of ontology candidate and its source	How to handle dynamic peers? How to choose which provider peer to be used as input to maintain common ontology of super peer.	Hybrid Ontologies Voting, Representation, Similarity
		How to choose export scheme component of provider peer to utilize during alignment and merging.	Onto-Vote
Ontology	maintenance ontology	how to change the common ontology based on selected peers of the community	representation and voting-election

4. Approach

4.1. Overview of Approaches

4.1.1. Voting and Representation

Local Ontology can be represented in many models, like 'data dictionary', E-R Diagram, RDF up to logic mathematics expression. The approach refers to RDF and OWL graphic and expression. Problem of the election of ontology candidate and its source is how to choose appropriate peers as input to maintain the common ontology of super peer. The next problem is how to choose export scheme concepts of provider peer to be utilized during alignment and merging.

Approach of voting [8] is based on Onto-Vote approach and mix with general ontology integration approach. Idea of voting taken from common voting in social life. Selection of candidate PP as input for common ontology maintenance based on provider peer member which is most receive and respond appropriate query. Voting can be conducted based on a communication protocol. Representation is describing which provider peer give the satisfied query respond from request peer, and it is based on communication protocol in P2P.

The communications protocol of P2P has steps as follow:

- ***Delivery of quer***y, Request Peer (RP) writes a query based on view of CO and delivers the query to the community or cluster. Routing model of query can be in the form of ' broadcast', ' selected' or ' on-behalf-of'. 'Broadcast' is delivery of query to all community members, 'selected' is delivery of query to provider peer which have been selected by request peer based on selected criterion, and 'on-behalf'-of' is firstly by sending a query to super peer, then the super-peer determine with selected mechanism to resend the query to provider peers. Our approach will be more suitable with 'selected' model. Record query path which the interaction directly between provider and request is needed a mechanism. The mechanism is not being discussed in this paper because limited of space. Query information of RP will be recorded in SP in tuple Q_{RP} as following:

  $$Q_{RP}=<m_{ID,},Time,Q,RP_{ADDR},PP_{ADDR}> \qquad (1)$$

  Where: m_{ID} is unique ID created by SP, *Time* is the time of query delivery occurred, *Q* is content of query, RP_{ADDR} is address of peer query sender, PP_{ADDR} is destination address to provider peer.

- ***Query Negotiation***, deliver a query to provider peer, it frequently been occurred a perception differentiation although it has passed a common ontology. The common ontology is developed in general, so that it almost impossible to fulfill view of all community members (local ontology). Very often a query need query re-writing based on negotiation between the query and local ontology. To achieve better result of negotiation is by reduce semantic difference between common and local ontology. The reducing of the differences can be achieved by adjust local or common ontology. But in this case, the adjustment will be implemented in common ontology as community reference. Tracking mechanism to every negotiation is needed, although the tracking needs cost of computing process and communications. Negotiation will be noted in tuple as following:

  $$\mathrm{Q_{neg} = <m_{ID}, Time, Q, Neg, RP_{ADDR}, PP_{ADDR}>} \quad (2)$$

  Where: $\mathrm{m_{ID}}$ is unique ID created by SP for negotiation, Time is time of negotiation process occurred, Neg is result of conducted negotiation, $\mathrm{RP_{ADDR}}$ is address of peer query sender, $\mathrm{PP_{ADDR}}$ is destination address to provider peer.

- ***Query Respond*** is a respond to a query from an RP, RP will give a feed back to SP concerning respond given by RP whether it fulfill their requirement or not and it is expressed in the form of a tuple:

  $$RP_{resp} = <m_{ID}, Time, RP_{ADDR}, PP_{ADDR,}, Hsl> \quad (3)$$

  Where: m_{ID} is unique ID which value is same with equation 3, RP_{ADDR} is address of peer query sender, PP_{ADDR} is destination address to provider peer, *Hsl* is assessment result of RP headed for answer given by PP. In the early step, there are two values as satisfy and dissatisfy.

Calculation of voting and representation of common ontology will follow some steps. After some *T* time of duration (e.g. 3 months), SP will calculate mechanism by looking among Q_{RP} , Q_{NEG} and RP_{RESP}, and with same. Result of calculation give:

- The rank of PP based on number of query.
- The rank of PP based on number of negotiation.
- The ranks of PP based on number of satisfy answer.

From the above result, it can be done by ranking based on three criteria. Analysis of ranking can be done with some possibilities as follows:

- *A PP has high number of query but number of negotiation and responds satisfaction is low.* This result can be caused by usage of local ontology representation or export scheme inappropriate or the PP give less precise metadata. In this condition super peer need to inform to PP to enhance its local scheme/ontology. The goal is to reduce the network traffic caused by delivery of the query which always fails in respond.

- *A PP get high number of negotiation but the number of sufficient respond is low.* In this case it require analysis of its low quality of respond because of common ontology which need to be adjusted, or an appropriate wrapper to convert a query from concept level to data level.

- *A PP gives high number of related respond, but number of negotiation is low.* The PP has ‘high’ similarity concept to common ontology so that the PP is not ontology candidate for input in maintenance of common ontology.

From hit calculation result of amount of query, negotiation, and respond, then selection of local ontology of provider peer can be selected to fix it. Sequence step of the process calculation take into account at:

- Which PP is at most doing negotiations (voting), this show in the PP has high unrelated concept to common ontology.
- From PP above result, which is PP has most accepting query (voting), this show 'popularity' of provider peers.
- From second step, which is PP has most can give appropriate answer. In this case it will be selected from PP which give small number of satisfy answer. The final result of PP will utilize as input of common ontology maintenance.

Determination processes of PP candidate for the input of common ontology maintenance are:

- Sort the PP based on Q_{RP} , Q_{NEG} and RP_{RESP}

- Sequence result above will be selected again based on the cut-off minimum hit value criterion (Q_{RP}).
- Selection above result, if it is still too much, it can be selected again based on choosing a number of PP with biggest hit values (Q_{RP}).

Similarity

Ontology maintenance considers input from concepts of provider peers. A process will need mapping and merging process in reaching better common ontology. Before mapping and merging process, the similarity calculation is very important step. Every ontology can be represented in a label terminology hierarchy.

First step for similarity [23] is linguistic / label matching approach. There are two common processes in label matching. Started with linguistics analysis, like changing abbreviation, avoiding repeating, affixes-suffixes, then continued with referenced thesaurus like WordNet [25]. The calculation will calculate label by looked at its semantic relation by linguistically.

Result of this calculation can be expressed in tuple $< L^{I}_{CO}, L^{J-K}_{PP}, Sim_{label} >$, where L^{I}_{CO} is label of i'th at CO, L^{J-K}_{PP} is label to- at PP j'th, Sim_{label} is the similarity calculation based on Wordnet. Result from first step enriched with approach of internal and external structure comparison.

Internal structure comparison is comparing ' language' and ' real' attribute. Simply to calculate internally structure from two classes is looked at how many amount of the same attribute will be divided with amount of the biggest attribute from a class. $IS = similar\ attribute/[maxattributeataclass]$. This result is also expressed with tuple $< C^{I}_{CO}, C^{J-K}_{PP}, Sim_{IS} >$ where C^{I}_{CO} is i'th class at CO, C^{J-K}_{PP} is class to at PP j to k'th, Sim_{IS} is the calculation internal structure comparison.

External structure comparison is looked at the set from upper-class. Simply to calculate the external structure from two class is by looking at how many amount of the same upper-class will be divided with amount of the biggest upper-class from a class. $ES = upper - classsimilar/[maxupper - classataclass]$. This result is also expressed with tuple $< C^{I}_{CO}, C^{J-K}_{PP}, Sim_{IS} >$ where C^{I}_{CO} is i'th class at CO, $C^{J.\ K}_{PP}$ is class PP j to k'th, Sim_{ES} is the calculation of external structure comparison.

4.2. Experiment Design

Experiment planning will prepare as follow:

- Prepare simple example of common ontology and some local schema.
- Prepare some technology and tool to support the prototype, such as ontology editor (Protégé), similarity measurement (WordNet), language programmer (python), ontology language (OWL), operating system (Linux and MS Window) and network communication tools.
- Design evaluation model for the approach and prototype, some experts in related domain is needed.

4.3. Current Status and Planning

My research has started since April 2005, the first activity was to get better understanding in the area or research. Some surveys on related research have been conducted as state of the art and related works. The next step was developed the problem statement of the research. First approach has been proposed to overcome the research questions. Currently, some publications have been written to some conferences, the purpose is to get feedback from the research community.

The next planning is to develop a prototype of the approach. Evaluation of the approach can be executed by using the prototype. The developing of prototype is started at October 2006, and evaluation will be conducted at February 2007. Writing doctoral thesis can be started at January 2007, and thesis defence is scheduled around August 2007.

5. Summary

Ontology development is a difficult task, ontology maintenance activity is more difficult then ontology development. Our approach based on representation - voting of peers, and similarity calculation can demonstrate as basic methodology to maintenance common ontology in a dynamic P2P environment.

However, the available approach tool is in between manual and semi-automatic level. In big ontology, very dynamic peer and big member of peer, there is limitation on speed process of ontology maintenance. Big effort still needed to bring more automatic tool for ontology maintenance.

6. References

[1] Banowosari, LY, *Maintenance of Common Ontology in P2P environment with Voting and Representation* (Pemeliharaan Common Ontology di P2P dengan Voting dan Representation), Proc. Seminar National Information Technology (SNTI) 2005, University of Tarumanegara, Jakarta, 2005, (in Indonesia language).

[2] Banowosari, LY., *Ontology Maintenance Based on Voting and Similarity*, ICTS Proceeding, page 256-261, 2006

[3] Benjamins R., RB, 2000, *Knowledge System Technology: Ontologies And Problem Solving Methods*, 15th May 2004, http://www.swi.psy.uva.nl/usr/richard/pdf/kais.pdf.

[4] Cali, A., *State of the Art Surveys: TONES (Thinking Ontologies)*, Tones Consortium, Bolzano, 2005

[5] Euzenat, Jerome, *D2.2.3: State of the Art on Ontology Alignment*, KnowledgeWeb, 2004

[6] Euzenat, Jérôme and Petko Valtchev. An integrative proximity measure for ontology alignment. In *Proc. ISWC-2003 workshop on semantic information integration, Sanibel Island (FL US)*, pages 33–38, 2003.

[7] Fensel, Dieter, Steffen Staab, Rudi Studer and Frank van Harmelen. Peer-2-Peer Enabled Semantic Web for Knowledge Management. *Ontology-based Knowledge Management: Exploiting the Semantic Web*, Wiley, London, UK, 2002.

[8] Fernandez M.,Mf, ,Building Chemical Ontology a of Using MENTHONTOLOGY Ontology Design Environment , *IEEEE Expert (Intelligent Systems and Their Applications)*, 14(1), 1999, 37-46.

[9] Gargouri, Yassine, *Ontology Maintenance Usinf Text Analysis*, *Laboratory Cofnitive Information Analysis*, Montreal Canada, 2004

[10] Ge, Yanfeng, OntoVote*: Scalable Distributed Votes Collecting Mechanism for Ontology Drift on P2P Platform,* CEUR-WS Vol.71, Aachen Germany, 2004

[11] Giunchiglia, Fausto and Pavel Shvaiko. *Semantic matching*, In *Proc. IJCAI 2003 Workshop on ontologies and distributed systems, Acapulco (MX)*, pages 139–146, 2003.

[12] Giunchiglia,Fausto and Pavel Shvaiko. *Semantic matching*. The Knowledge Engineering Review, 18(3):265–280, 2004.

[13] Gruber T.R, TRG, *A Translation ontologies portable to approach*, Knowledge Acquisition,5(2),1993,199-220.

[14] Guarino N., NG, *Ad for Ontology Information Systems*,Proceedings FOIS98 of in, Trento, Italy, 6-8 June. Amsterdam, IOS Press, 1998,3-15.

[15] Klein, M. C. A., D. Fensel, A. Kiryakov, and D. Ognyanov. Ontology versioning and change detection on the web. In *Proc. of the 13th Int. Conf. on Knowledge Engineering and Knowledge Management – Ontologies and the Semantic Web (EKAW 2002)*, volume 2473 of *Lecture Notes in Computer Science*, pages 197–212. Springer, 2002.

[16] Klein, M. and N. F. Noy. A component-based framework for ontology evolution. In *Proc. of IJCAI 2003 Workshop on Ontologies and Distributed Systems*, 2003

[17] Milojick D.,Dm, etc., Peer-To-Peer Computing,2002.

[18] Noy, N. F. and M. C. A. Klein. Ontology evolution: Not the same as schema evolution. *Knowledge and Information Systems*, 6(4):428–440, 2004.

[19] Rahm, Erhard and Philip Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4):334–350, 2001.

[20] Staab, Steffen, Semantic *Web and Peer-to-Peer: Decentralized Management and Exchange of Knowledge and Information,* Springer, Berlin, 2006

[21] Sheth A.P, APS, *Changing Focus On Interoperability In Information Systems: From System, Syntax, Structure, To Semantics*, MITRE, Dec 3rd 1998.

[22] Welty, Chris, *Ontology Maintenance: Support, Text, Tools, Theory*, IBM Research, 2005

[23] Wicaksana, IWS,2005 *Important of Language in Information Interoperability to solve Semantic Diversity* (Pentingnya peranan Bahasa dalam Interoperabilitas Informasi untuk mengatasi Perbedaan Semantik), Proc. Seminar National (PESAT,2005), University of Gunadarma, Jakarta, 2005, (in Indonesia language).

[24] Wicaksana, IWS, *PhD Thesis: A Peer-to-Peer (P2P) Based Semantic Agreement Approach for Spatial Information Interoperability*, University of Gunadarma,Jakarta, 2006.

[25] WordNet homepage, access July 2006, http://WordNet.princeton.edu.