Enforcing a Semantic Routing Mechanism based on Peer Context Matching

Silvana Castano and **Stefano Montanelli** [1]

Abstract. In this paper we present the main features of the H-LINK semantic routing mechanism we are developing to combine ontology-based peer contexts and ontology matching techniques for providing P2P query forwarding on a real semantic basis. H-LINK defines a semantic overlay network where each edge represents a semantic link between two peers having similar contexts. Semantic links are exploited to address query propagation by identifying the semantic neighbors that can provide relevant knowledge with respect to a given target request.

1 INTRODUCTION

Recent schema-based P2P networks go beyond traditional file-sharing P2P networks, by providing infrastructures where peers can create and share knowledge [1]. In this scenario, peers join the system by providing their own context and need to cooperate by matching their respective context with the aim to discover similar partners and to enforce effective resource sharing. In order to provide scalable infrastructures for peer communications, P2P semantic routing protocols are being proposed with the aim to address query propagation on the basis of the local context of each peer [2, 7, 9, 12, 14]. At the current stage of development, a challenging issue regards the need of advancing the existing semantic routing protocols by combining ontology-based peer contexts and ontology matching techniques for providing query forwarding on a real semantic basis.

In this paper, we present the main features of the H-LINK semantic routing mechanism we are developing in the framework of our HELIOS peer-based system for knowledge sharing and evolution [5]. In HELIOS, the peer context is represented through a *peer ontology* describing the knowledge the peer brings to the network and the knowledge the peer perceives from the network. Peers act as independent agents with their own context (i.e., peer ontology) and interact each other by submitting discovery queries and by replying with relevant knowledge. In the HELIOS framework, the H-MATCH semantic matchmaker has been developed to evaluate the semantic affinity between an incoming discovery query and a peer ontology. On this basis, the H-LINK semantic routing mechanism is designed to exploit the matching knowledge acquired from the discovery process. The matching knowledge becomes *network knowledge* in the peer ontology, and it is exploited to provide a semantic overlay network where peers having similar contexts are interlinked as *semantic neighbors*. This way, as a peer learns about the network contents through discovery queries, also its network knowledge is gradually evolved to reflect its newly acquired semantic neighbors.

[1] Dipartimento di Informatica e Comunicazione, Università degli Studi di Milano. Via Comelico 39, 20135 Milano, Italy. Email:{castano,montanelli}@dico.unimi.it

Example of knowledge discovery in HELIOS. Considering the scenario of Figure 1, we suppose that peer A is interested in discovering peers capable of providing resources semantically related to the publishing domain. To this end, peer A composes and submits to the system a discovery query Q1 containing the target concepts of interest Publication and Book with the properties year and author, respectively. Moreover, Book is specified as a subclass of Publication. Receiving the query Q1, the peer (i.e., peer B, peer C, and peer D) uses the H-MATCH semantic matchmaker to compare the query target with its own peer ontology, with the aim to identify whether there are concepts matching the target request. According to their matching results, peer B and peer D send back to the requesting peer A a ranked list of concepts found to be semantically related to the target, and, for each entry, the corresponding semantic affinity value SA. In particular, peer B replies with the Volume matching concept as $SA(Book, Volume) = 0.82$, while peer D sends back two matching concepts, namely Newspaper and Magazine, with $SA(Publication, Newspaper) = 0.67$ and $SA(Book, Magazine) = 0.539$. On the other hand, peer C does not reply to peer A as no matching concepts are identified. The query replies represent the discovered knowledge of peer A that can be exploited to decide whether to further interact with the answering peers in order to access their relevant resources for data sharing. Before H-LINK, the discovery process relied on the conventional P2P infrastructure and associated routing protocols for addressing query propagation in the network. In H-LINK, we show how the discovered knowledge can be further exploited for semantic routing purposes by enforcing query forwarding according to peer context similarities.

2 ONTOLOGY MATCHING WITH H-MATCH

H-MATCH performs ontology matching at different levels of depth by deploying four different *matching models* spanning from surface to intensive matching, with the goal of providing a wide spectrum of metrics suited for dealing with many different matching scenarios that can be encountered in comparing concept descriptions of real ontologies. H-MATCH takes two ontologies as input and returns the mappings that identify corresponding concepts in the two ontologies, namely the concepts with the same or the closest intended meaning. H-MATCH mappings are established after an analysis of the similarity of the concepts in the compared ontologies. In H-MATCH we perform similarity analysis through affinity metrics to determine a measure of semantic affinity in the range $[0, 1]$. A threshold-based mechanism is enforced to set the minimum level of semantic affinity required to consider two concepts as matching concepts. Given two

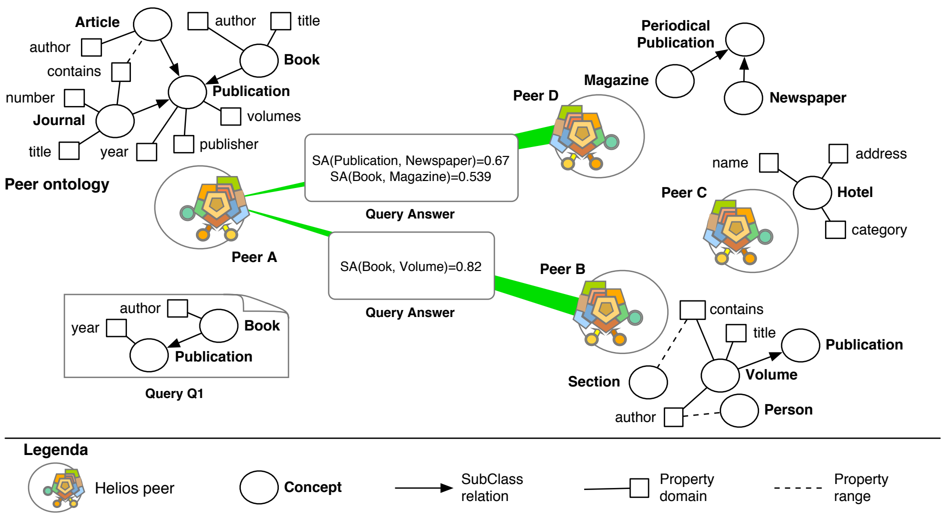

Figure 1. Example of knowledge discovery in HELIOS

concepts c and c', H-MATCH calculates a semantic affinity value $SA(c, c')$ as the linear combination of a linguistic affinity value $LA(c, c')$ and a contextual affinity value $CA(c, c')$. The linguistic affinity function of H-MATCH provides a measure of similarity between two ontology concepts c and c' computed on the basis of their linguistic features (i.e., concept names). For the linguistic affinity evaluation, H-MATCH relies on a thesaurus of terms and terminological relationships automatically extracted from the WordNet lexical system. The contextual affinity function of H-MATCH provides a measure of similarity by taking into account the contextual features of the ontology concepts c and c'. The context of a concept can include properties, semantic relations with other concepts, and property values. The context can be differently composed to consider different levels of semantic complexity, and four matching models, namely, *surface*, *shallow*, *deep*, and *intensive*, are defined to this end. In the surface matching, only the linguistic affinity between the concept names of c and c' is considered to determine concept similarity. In the shallow, deep, and intensive matching, also contextual affinity is taken into account to determine concept similarity. In particular, the shallow matching computes the contextual affinity by considering the context of c and c' as composed only by their properties. Deep and intensive matching extend the depth of concept context for the contextual affinity evaluation of c and c', by considering also semantic relations with other concepts (deep matching model) as well as property values (intensive matching model), respectively. The comprehensive semantic affinity $SA(c, c')$ is evaluated as the weighted sum of the Linguistic Affinity value and the Contextual Affinity value, that is:

$$SA(c, c') = W_{LA} \cdot LA(c, c') + (1 - W_{LA}) \cdot CA(c, c') \quad (1)$$

where W_{LA} is a weight expressing the relevance to be given for the linguistic affinity in the semantic affinity evaluation process.

H-MATCH has been extensively tested on several real ontology matching cases in order to evaluate the matching models with respect to performance and quality of results [4]. By analyzing the obtained results, we note that the most accurate and precise results are achieved with the deep and intensive matching models provided that the ontology descriptions are detailed enough. On the other side, we note that the best performance in terms of computation time are achieved with the surface and shallow matching models. For semantic routing purposes, the computation time of the semantic affinity evaluation is a crucial factor and needs to be performed as fastest as possible in order to avoid bottlenecks. To this end, possible lacks in matching precision and accuracy can be admitted in turn of rapid response time during the semantic affinity evaluation. For this reason, the shallow matching model is selected to work with H-LINK for identifying the semantic neighbors that have the highest chance to provide relevant knowledge with respect to a given query (see Section 4). A detailed description of H-MATCH and related matching models is provided in [4]. We note that H-MATCH can be suitably adopted to enforce semantic routing functionalities by relying on its flexible matching models that allow to dynamically configure the tradeoff between performance and accuracy according to the requirements of the considered matching scenario. Provided that a dynamic and flexible configuration is supported, other existing matching tools can however be used to enforce the H-LINK semantic routing mechanism in turn of H-MATCH [11]. In the remainder of the paper, we focus on the use of H-MATCH for semantic routing in H-LINK.

3 PEER ONTOLOGY ARCHITECTURE

The context of a HELIOS peer is described through a peer ontology that is organized in a two-layer architecture where the upper layer represents the *content knowledge* and the lower layer represents the *network knowledge* of the peer, respectively. The content knowledge layer describes the knowledge the peer brings to the network that is described as a graph of concepts, properties, and semantic relations [2]. The network knowledge layer describes the knowledge that the peer has of the semantic neighbors it has interacted with. With reference to the discovery example in Figure 1, when peer A receives a reply

[2] For the sake of internal representation of ontology specification languages, and in particular for Semantic Web languages like OWL, we rely on a reference model, called H-MODEL, that provides a graph-based representation of peer ontologies. For further details on H-MODEL, the reader can refer to [5].

from peer B and peer D as an answer to the discovery query Q1, it stores in the network knowledge layer a description of peer B and peer D. A peer description is given in the form of *network concept*, characterized by a set of properties describing the network features of the peer (e.g., IP address, bandwidth). A *location relation* is defined to connect a network concept nc with a concept c in the content knowledge layer. The location relation is labeled with a *confidence* annotation cf that keeps track of the discovered semantic affinity between c and the peer ontology of the peer represented with nc. The cf value corresponds to the semantic affinity value SA returned by the peer nc in its query answer. A new location relation is defined for each matching concept returned in the query answer. A comprehensive *expertise* measure is associated with a network concept nc and it is computed as the average mean of the confidence values associated with all the location relations connected with nc.

As an example, in Figure 2, we consider a portion of the peer ontology of the peer A after the knowledge discovery process described in Figure 1. In this example, peer B and peer D have answered to query Q1, then the corresponding network concepts are defined in the network knowledge layer. According to the query reply of peer B, a new location relation with a confidence value of 0.82 is defined to connect the peer B network concept with the Book concept in the content knowledge layer. As two matching concepts are returned in the query answer of peer D (i.e., Newspaper, Magazine), two location relations are defined by connecting the peer D network concept with the concepts Publication and Book in the content knowledge layer, and by setting a confidence value of 0.67 and 0.539, respectively. As a consequence, the expertise measures associated with peer B and peer D are 0.82 and 0.605, respectively.

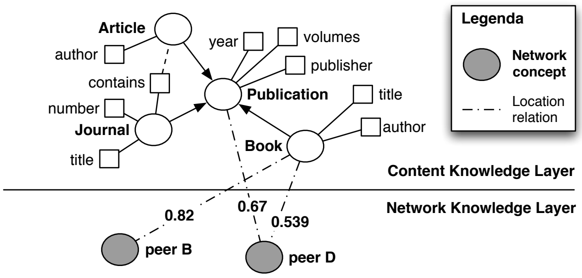

Figure 2. A portion of the peer ontology of peer A

Considerations. The confidence value associated with a location relation between c and nc is updated when a new semantic affinity value with c is returned by nc in reply to a discovery query. As proposed in [9], the confidence value cf associated with a given location relation between c and nc can be periodically updated by observing the ratio between the number of relevant replies provided by nc and the number of queries sent to nc with a target concept related to c. When the ratio has low values, cf can be decreased to denote that the original confidence (i.e., semantic affinity) is no more actual. In such way, only confirmed location relations are maintained in the peer ontology, while unreliable confidence values are gradually reduced and finally dropped. Furthermore, a number of information can be combined with the confidence measures for providing a more accurate evaluation of the network concept expertise and thus, of the associated semantic neighbor. For instance, a trust mechanism can be adopted to maintain reputation information about the semantic neighbors stored in the network knowledge layer [13]. Moreover, information regarding the network reliability of the semantic neighbors, such as connection stability and granted bandwidth, can be considered for expertise computation [2]. Confidence and expertise measures are exploited by H-LINK for addressing query routing on a semantic basis.

4 THE H-LINK SEMANTIC ROUTING MECHANISM

The H-LINK semantic routing mechanism is based on the idea of exploiting the network knowledge layer of a peer ontology by using the H-MATCH semantic matchmaker for providing query routing support according to semantic neighbor contents.

We consider a query q with a target concept tc [3]. Two different roles can be distinguished for a given peer p:

- *Requesting peer*. Peer p needs to submit to the network a query q in order to identify relevant partners for subsequent resource sharing. To this end, peer p invokes H-MATCH to compare the target concept tc against the content knowledge layer of its peer ontology $\mathcal{O}$. A list $MCL = \{\langle c_1, SA(tc, c_1)\rangle \ldots \langle c_n, SA(tc, c_n)\rangle\}$ of matching concepts $c_1 \ldots c_n \in \mathcal{O}$ and corresponding semantic affinity values $SA(tc, c_1) \ldots SA(tc, c_n)$ is returned as a result. Peer p sets the *number of credits* N_{cr} to distribute to the query recipients in order to define the number of replies that peer p wish to receive as answers to the query q. Therefore, H-LINK is invoked by passing the list MCL to select the semantic neighbors for query q submission.
- *Receiving peer*. When a peer p receives a query q together with the number of credits nc from a requesting peer r, it needs to evaluate whether matching concepts can be provided back to peer r. To this end, H-MATCH is invoked by peer p and the list MCL of matching concepts is still produced as a results. If $MCL \neq \emptyset$, the peer p sends MCL back to peer r by consuming one credit, otherwise no reply is sent back to peer r and all the received credits are still available for forwarding. If at least one credit is available, H-LINK is invoked by peer p to select the semantic neighbors for query q forwarding; otherwise the propagation mechanism stops.

H-LINK invocation. H-LINK is invoked for both query submission/forwarding provided that at least one credit is still available. Three main steps define H-LINK: *selection of semantic neighbors*; *ranking of semantic neighbors*; *distribution of credits*.

1- Selection of semantic neighbors. The network knowledge layer of the peer ontology is accessed to select the network concepts, together with the associated confidence values, that are connected to the concepts in MCL through a location relation. A list SNL of semantic neighbors is returned as a result. A semantic neighbor $sn \in SNL$ is described in the form $sn = \langle nc, \{c_1, cf_1 \ldots c_m, cf_m\}\rangle$, where nc is the network concept featuring sn, while $c_1 \ldots c_m \in MCL$ are the concepts of MCL connected to nc through a location relation, and $\{cf_1 \ldots cf_m\}$ the corresponding confidence values.

2- Ranking of semantic neighbors. The semantic neighbors in SNL are ranked with respect to their relevance for the query target tc. To this end, the harmonic mean is used to combine the

[3] For the sake of clarity, we consider the case of a single target concept in the query. The H-LINK semantic routing mechanism can be easily extended to consider the case of multiple target concepts.

confidence values associated with the semantic neighbors in SNL and the semantic affinity values in MCL. Given a semantic neighbor $sn \in SNL$, the ranking value r_{sn} corresponds to the following formula:

$$r_{sn} = \frac{\sum_{i=1}^{m} \frac{2 \cdot cf_i \cdot SA(tc, c_i)}{cf_i + SA(tc, c_i)}}{m} \qquad (2)$$

Finally, a ranked list $RSNL$ of semantic neighbors with the corresponding ranking value is returned as a result. A threshold mechanism can be used to rule out the semantic neighbors with a ranking value lower than a predefined threshold t.

3- Distribution of credits. The semantic neighbors in $RSNL$ determine the recipients of the query q. Available credits A_{cr} are proportionally distributed to the semantic neighbors in $RSNL$ according to their ranking value. Then, the number of credits nc_{sn} assigned to the semantic neighbor $sn \in RSNL$ is computed as follows:

$$nc_{sn} = \lfloor \frac{A_{cr}}{\sum_{\forall sn_i \in RSNL} r_{sn_i}} \cdot r_{sn} \rfloor \qquad (3)$$

We note that if H-LINK is invoked with $MCL = \emptyset$, selection and ranking of semantic neighbors are not performed and credits are proportionally distributed according to the expertise measure of the network concepts in the network knowledge layer.

Example. As an example of H-LINK semantic routing, we consider the peer B of Figure 1. Peer B intends to submit to the system the query Q2 described in Figure 3(a) with total number of credits to distribute $N_{cr} = 5$. The peer B uses H-MATCH to compare the query

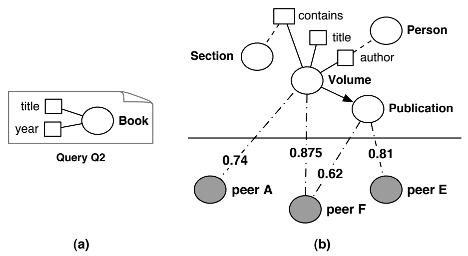

Figure 3. (a) The Query Q2 example and (b) a portion of the peer B ontology

Q2 against its peer ontology (see Figure 3(b)). As a result, the following semantic affinity values are returned by H-MATCH:

SA(Book,Volume)=0.79
SA(Book,Publication)=0.49

By invoking H-LINK, we find that:

MCL={⟨ Volume,0.79 ⟩, ⟨ Publication,0.49 ⟩}
SNL={⟨ peer A,{Volume,0.74}⟩, ⟨ peer E,{Publication,0.81}⟩,
⟨ peer F,{Volume,0.875,Publication,0.62}⟩}

On the basis of such results, H-LINK computes the ranking of the semantic neighbors in SNL and assigns the corresponding number of credits, as shown in Table 1. The query Q2 is then submitted to the selected semantic neighbors together with the assigned number of credits. As shown in the routing schema of Figure 4, peer A receives the query, consumes one credit for replying to peer B, and forwards the query Q2 to peer D by assigning the last remaining credit. Peer E

Table 1. Example of semantic neighbor ranking and credit distribution

Semantic neighbor	Ranking value	Assigned credits
peer A	0.764	2
peer E	0.611	1
peer F	0.689	2

consumes the unique credit received and soon stops the forwarding process, while the peer F forwards all the received credits to peer G as no reply is sent back to peer B.

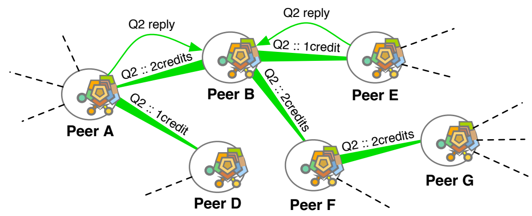

Figure 4. The H-LINK routing schema for query Q2

Considerations. A possible side effect of the H-LINK mechanism is due to the fact that credits are distributed on the basis of the knowledge discovered during past interactions. This means that the knowledge of new peers joining the system is hardly discovered and it is not considered for semantic neighbor selection. H-LINK deals with this by introducing a perturbation during the credit distribution phase. As proposed in [14], a small set of random peers is picked and it receives a percentage of the credits available for distribution. As a result, a larger part of the network is explored with the aim to discover additional knowledge and to include new peers in the semantic routing process.

5 RELATED WORK

Semantic query routing techniques are required to improve effectiveness and scalability of current discovery and search processes for resource sharing in P2P systems. In this direction, the notion of P2P Semantic Link Network is introduced in [15] to emphasize the need of typed semantic links specifying semantic relationships between peers in order to maintain information about nodes with similar contents. Each peer defines its own XML Schema (source schema) describing the contents to share and adopts SOAP-based messages to communicate with the other members of the network. As a difference with location relations in H-LINK, semantic links are exploited with cycle analysis and functional dependency analysis in order to select the query recipients according to the types of the semantic links as well as to the similarity between elements and structures of peer schemas. We note that semantic links need to be actively updated, while location relations are automatically maintained in H-LINK by relying on conventional discovery processes. In [12], the REMINDIN' multi-step query propagation mechanism is described to enforce selected propagation of queries by observing which queries are successfully answered by other peers, by storing these observations, and by subsequently using this information for peer selection. A similar approach is presented in [14] where the Intelligent Search Mechanism (ISM) is introduced to provide an efficient and scalable solution for improving the information retrieval problem in P2P systems. Each ISM peer is composed of four basic elements: i) the *profiling structure* that is used to store the most recent replies of each known peer, ii) the *query similarity function* that is used to identify the similarity between different search queries, iii) the *RelevanceRank algorithm* which exploits the profiling structure to select the peers that can provide relevant answers with respect to a given query, and iv) the *search mechanism* that is used to send the query to the selected peers. As another example of P2P semantic routing approach, the NeuroGrid adaptive decentralized search system is proposed in [9]. In such work, semantic routing is intended as content-based query forwarding, and a learning mechanism is defined to dynamically adjust the relevance of known peers for each query. In NeuroGrid, each node maintains a knowledge base that contains associations between keywords and other nodes. Queries are then forwarded to the nodes that may store matching documents according to the actual knowledge base. In [2], the *Seers* search infrastructure is presented. In Seers, each shared resource is described through a XML *meta-document* and a *matching policy* is used to define how to evaluate the similarity between resources and queries and to assign scores. Scores are then exploited to select the most relevant documents and to rank neighbors for query forwarding. In recent work, ontology-based frameworks are also being proposed to address the lack of semantics in actual P2P routing algorithms. A RDF-based semantic routing architecture is presented in [10]. Nodes are clustered in structured trees according to their interests and intra-/inter-cluster routing algorithms are defined for providing a scalable query forwarding mechanism. In [7], peers advertise their experience in the P2P network according to a shared common ontology. Based on the semantic similarity between a query and the expertise of other nodes, a peer can select appropriate peers for query forwarding.

Original contribution of H-LINK. With respect to the above approaches, we observe that current content-based P2P query propagation algorithms are essentially based on statistical observations and exploit, in some cases, a shared ontology, often mainly a taxonomy. In order to evaluate the similarity between a target query and resources, keyword-based strategies and basic matching techniques (e.g., string matching) are actually supported. The main contribution of H-LINK is related to the use of independent ontologies, rather than a single shared ontology, and to the use of ontology matching techniques to build a network knowledge layer reflecting the gradual learning of semantic neighbors. A further contribution of our approach regards the fact that H-LINK is capable of addressing emergent semantics requirements, by extending current techniques to work in multi-ontology contexts and thus releasing the constraint of having an initial common shared knowledge.

6 CONCLUDING REMARKS AND FUTURE WORK

In this paper, we have presented the H-LINK mechanism we are developing for matching-based semantic routing in P2P systems. Preliminary experimentations show that the H-LINK approach is effective. Our future work will be focused on the extensive experimentation of H-LINK by means of simulation techniques with the aim to assess the real scalability of the proposed approach. Furthermore, we plan to i) investigate the opportunity to refine the credit distribution procedure by considering the recommendation adjustment techniques developed in the field of document retrieval in distributed environments [8], and ii) compare H-LINK with other existing P2P routing approaches in order to evaluate the performance for what concern generated traffic and single peer workload. We will also investigate the opportunity to use flexible ontology evolution techniques for extending the peer ontology with the new concepts that are mostly queried in the network [3], thus improving also peer routing capabilities. Finally, we note that the network concepts keep track of peer context similarities. In this respect, the network knowledge can be exploited for the formation of emergent communities of peers on the basis of their common perspective and context. Some initial results on this topic are presented in [6].

ACKNOWLEDGEMENTS

This paper has been partially funded by the ESTEEM (Emergent Semantics and cooperaTion in multi-knowledgE EnvironMents) PRIN project and by the WEB-MINDS (Wide-scalE, Broadband, MIddleware for Network Distributed Services) FIRB Project funded by the Italian Ministry of Education, University, and Research.

REFERENCES

[1] S. Androutsellis-Theotokis and D. Spinellis, 'A survey of peer-to-peer content distribution technologies', *ACM Computing Surveys*, **36**(4), 335–371, (2004).

[2] N.T. Borch and N.K. Vognild, 'Searching Variably Connected Networks', in *Proc. of the Int. Conference on Pervasive Computing and Communications (PCC-04)*, Las Vegas, Nevada, USA, (2004).

[3] S. Castano, A. Ferrara, and S. Montanelli, 'Evolving Open and Independent Ontologies', *International Journal of Metadata, Semantics and Ontologies*, (2006). To Appear.

[4] S. Castano, A. Ferrara, and S. Montanelli, 'Matching Ontologies in Open Networked Systems: Techniques and Applications', *Journal on Data Semantics (JoDS)*, **V**, (2006).

[5] S. Castano, A. Ferrara, and S. Montanelli, *Web Semantics and Ontology*, chapter Dynamic Knowledge Discovery in Open, Distributed and Multi-Ontology Systems: Techniques and Applications, Idea Group, 2006.

[6] S. Castano and S. Montanelli, 'Semantic Self-Formation of Communities of Peers', in *Proc. of the ESWC Workshop on Ontologies in Peer-to-Peer Communities*, Heraklion, Greece, (2005).

[7] P. Haase, R. Siebes, and F. van Harmelen, 'Peer Selection in Peer-to-Peer Networks with Semantic Topologies', in *Proc. of the Int. Conference on Semantics of a Networked World*, Paris, France, (2004).

[8] M.N. Huhns, U. Mukhopadhyay, L.M. Stephens, and R.D. Bonnell, 'An Intelligent System for Document Retrieval in Distributed Office Environments', *Journal of the American Society for Information Science*, **37**(3), 123–135, (1986).

[9] S. Joseph, 'NeuroGrid: Semantically Routing Queries in Peer-to-Peer Networks', in *Proc. of the Int. Workshop on Peer-to-Peer Computing (co-located with Networking 2002)*, Pisa, Italy, (2002).

[10] J. Li and S. Vuong, 'A Scalable Semantic Routing Architecture for Grid Resource Discovery', in *Proc. of the 11th Int. Conference on Parallel and Distributed Systems*, pp. 29–35, Fukuoka, Japan, (2005).

[11] P. Shvaiko. Ontology Matching Projects, 2006. http://www.ontologymatching.org/projects.html.

[12] S. Staab, C. Tempich, and A. Wranik, 'REMINDIN': Semantic Query Routing in Peer-to-Peer Networks based on Social Metaphors', in *Proc. of the 13th Int. conference on World Wide Web (WWW 2004)*, New York, NY, USA, (2004).

[13] L. Xiong and L. Liu, 'PeerTrust: Supporting Reputation-based Trust for Peer-to-Peer Electronic Communities', *IEEE Transactions on Knowledge and Data Engineering*, **16**(7), 843–857, (2004).

[14] D. Zeinalipour-Yazti, V. Kalogeraki, and D. Gunopulos, 'Exploiting Locality for Scalable Information Retrieval in Peer-to-Peer Networks', *Information Systems*, **30**(4), 277–298, (2005).

[15] H. Zhuge, J. Liu, L. Feng, and C. He, 'Semantic-Based Query Routing and Heterogeneous Data Integration in Peer-to-Peer Semantic Link Networks', in *Proc. of the Int. Conference on Semantics of a Networked World*, pp. 91–107, Paris, France, (2004).