Exploiting Peer Ontologies for Semantic Query Propagation *

(Extended Abstract)**

Silvana Castano, Alfio Ferrara, and Stefano Montanelli

Università degli Studi di Milano
DICo - Via Comelico, 39, 20135 Milano - Italy
{castano,ferrara,montanelli}@dico.unimi.it

Abstract. A challenging issue to advance the existing P2P query propagation protocols is related to the capability of developing a routing mechanism where a semantically rich description of the context of each peer is explicitly taken into account for selecting the query recipients. In this paper, we present the H-Link semantic routing approach designed to exploit peer ontologies and ontology matchmaking results for providing a semantic overlay network where peers having similar contexts are recognized and interlinked as semantic neighbors.

1 Introduction

In order to provide scalable infrastructures for peer communications, semantic-based P2P query propagation protocols are being proposed with the aim of identifying those peers that are most likely to provide relevant results according to the query content [1]. However, most of the existing approaches rely on a rather simplifying assumption of a centralized repository of knowledge where mappings among the distributed peer resource descriptions are maintained [2, 3]. In some other approaches, the knowledge model supported by the peers is kept quite poor (i.e., metadata rather than ontologies) in order to reduce the complexity of the matching process. In such cases, syntactic matching techniques (e.g., string- and keyword-based techniques) are generally employed to compute the similarity among the resource descriptions of the different peers, thus leading to a poor accuracy in query recipient selection [4, 5].

In this paper, we present the H-Link semantic routing mechanism designed to exploit the results of an ontology matchmaking process for providing a semantic overlay network where peers having similar contexts are recognized and interlinked as semantic neighbors. In particular, H-LINK aims at advancing the existing query propagation protocols by combining ontology-based peer context

* This paper has been partially funded by the ESTEEM PRIN project funded by the Italian Ministry of Education, University, and Research.

** An extended version of this paper is published on the Knowledge Engineering Review Journal, Special Issue on Contexts and Ontologies, 2007 (to appear).
descriptions and ontology matching techniques for providing query forwarding on a real semantic basis, in a completely decentralized way. Furthermore, H-Link aims at enforcing scalability in query forwarding by introducing a credit-based mechanism where the approximate number of desired replies is specified rather than the non-scalable number of hops to cross.

2 The H-Link approach to semantic query propagation

The key idea of H-Link is to exploit the results of knowledge discovery interactions to train the behavior of the query propagation mechanism. To this end, peers are connected through matching-based confidence measures that keep track of the semantic affinity among the knowledge of different peers. As a result, peers are organized in a semantic overlay network where nodes having similar knowledge are interlinked as semantic neighbors. The following main features characterize the H-Link mechanism.

*Use of a dynamic knowledge discovery approach.* In H-Link, peers interact by submitting discovery queries with the aim to identify relevant partners with respect to one or more target concepts of interest. Receiving a discovery query, a peer evaluates whether it is capable of providing concepts matching the target request. According to the results of the matching process, the list of concepts found to be relevant are replied to the requesting peer together with their associated semantic affinity values. Semantic affinity values provide a measure of the level of similarity between the target concepts of the query and the discovered matching concepts. In H-Link, the replying nodes are linked to the requesting peer as semantic neighbors and the returned affinity values are exploited to set the level of confidence of each semantic neighbor with respect to the discovered matching concepts. This way, as a peer learns about the network contents through discovery queries, also its network knowledge gradually evolves to reflect its newly acquired semantic neighbors.

*Use of peer ontologies.* In H-Link, both queries and peer resources are expressed in terms of ontological descriptions. In particular, each query contains a list of target concept(s) of interest with possible properties and semantic relations that further specify the request. Furthermore, each peer joining the system provides a peer ontology where knowledge is organized into a two-layer architecture, namely the content knowledge layer describing the knowledge the peer brings to the network, and the network knowledge layer describing the knowledge the peer has of the semantic neighbors it has interacted with. In particular, the network knowledge layer is seen as a set of network concepts $NC$ and location relations $LR$. A network concept $nc \in NC$ provides an abstract representation of a semantic neighbor (i.e., a peer) that has been identified during the knowledge discovery process. A location relation is defined to connect a network concept $nc$ with a concept $c$ in the content knowledge layer. A confidence value $cf$ is associated to a location relation to keep track of the discovered semantic affinity between $c$ and the peer ontology of the semantic neighbor represented with $nc$. 
Use of ontology matching techniques. In H-LINK, each peer is capable of providing ontology matching functionalities through the use of a semantic matchmaker. Ontology matching is employed by a peer during the dynamic knowledge discovery process in order to assess whether it can provide relevant knowledge in reply to an incoming discovery query. Furthermore, ontology matching is also exploited in H-LINK in order to select the recipients of a query according to the expected semantic affinity between the query contents and the peer ontology of the receiving peer.

3 Ontology matchmaking in H-Link

A key feature of H-LINK is the use of ontology matching techniques for query recipient selection that are currently performed by relying on the H-MATCH semantic matchmaker. 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. A threshold-based mechanism is enforced to set the minimum level of semantic affinity required to consider two concepts as matching concepts. Given two concepts $c$ and $c'$, H-MATCH calculates a semantic affinity value $SA(c, c') \in [0, 1]$ 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.
A detailed description of the H-MATCH models and related techniques is provided in [6]. 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 [6]. According to the obtained results, the choice to use H-MATCH for supporting H-LINK is motivated by the fact that H-MATCH can be dynamically configured to tune the tradeoff between performance and accuracy according to the requirements of the considered matching scenario. In this sense, other existing matching tools can however be used to enforce H-LINK in turn of H-MATCH provided that a dynamic and flexible configuration is supported.

4 Query propagation in H-Link

The H-LINK 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 distribution support according to semantic neighbor contents.

We consider a query $q$ with a target concept $tc$. 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 $O$. A list $MCL = \{\langle c_1, SA(tc, c_1)\rangle \ldots \langle c_n, SA(tc, c_n)\rangle\}$ of matching concepts $c_1 \ldots c_n \in O$ 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. Query replies are returned to the requesting peer by following the reverse query path in order to avoid a sudden burst of incoming messages as suggested in [5].

**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

\[ \text{For the sake of clarity, we consider the case of a single target concept in the query. The H-LINK mechanism can be easily extended to consider the case of multiple target concepts.} \]
**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 exploited 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.** 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 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:

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

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.

When \( MCL = \emptyset \), that is the peer ontology does not contain relevant concepts with respect to the target query, credits are proportionally distributed by ranking each peer according to a comprehensive expertise measure. For each peer \( nc \) in the network knowledge layer, expertise is computed as the average mean of the confidence values associated with all the location relations connected with \( nc \). A detailed description H-Link is provided in [7] where the H-Link proposals for addressing some typical query propagation issues are also discussed.

**Example.** As an example of H-Link query propagation, we consider a peer B with the associated peer ontology of Figure 1(b). The peer B intends to submit to the system the query \( Q \) described in Figure 1(a) with total number of credits to distribute \( N_{cr} = 5 \). The peer B uses H-Match to compare the query \( Q \) against its peer ontology. As a result, the following semantic affinity values are returned by H-Match: \( SA(Book, Volume) = 0.79 \) and \( SA(Book, Publication) = 0.49 \).

By invoking H-Link, we find that:

\( MCL = \{ \langle Volume, 0.79 \rangle, \langle Publication, 0.49 \rangle \} \)
\( SNL = \{ \langle peer A, \{ Volume, 0.74 \} \rangle, \langle peer E, \{ Publication, 0.81 \} \rangle, \langle peer F, \{ Volume, 0.875, Publication, 0.62 \} \rangle \} \)

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 follows:
The query \( Q \) is then submitted to the selected semantic neighbors together with the assigned number of credits. As shown in the query propagation schema of Figure 2, peer A receives the query, consumes one credit for replying to peer B, and forwards the query \( Q \) to peer D by assigning the last remaining credit. peer E consumes the unique credit received and 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.

![Fig. 2. The H-LINK propagation schema for the query Q](image)

5 Experimental results

The goals of the experiments is to evaluate H-LINK in terms of generated traffic and recall. As generated traffic, we mean the overall number of messages routed during a complete simulation run, while recall is measured as the ratio of the number of relevant concepts retrieved by a H-LINK query to the total number of relevant concepts that was available in the network. Several tests have been produced by varying the most important configuration parameters (i.e., \#nodes,
In this section, we report some snapshots that are appropriate for discussing the H-LINK performance. The complete experimental results are provided in [7].

(a) - H-Link scalability: traffic and recall

We observe that the generated traffic follows a sub-linear growing, while recall is not significantly affected by the #peer variation. The growing traffic is due to the random generation of queries and peer ontologies during the simulation. We believe that this promising result can be further improved by exploiting more realistic distribution models (e.g., the Zipf distribution model) that we plan to consider in future experiments. As another experiment we compare H-Link with a well-known query propagation mechanism, like Gnutella. In Figure 3(b), we show the results of the comparison when #nodes = 500 and the number of credits per query is a growing variable. In Gnutella, credit-based query propagation is not supported. For this reason, TTL is used to set the scope of Gnutella queries while credits are proportionally defined for H-Link. Furthermore, the results of H-LINK are also analyzed by varying the H-MATCH model from surface to intensive in the peer confidence computation. We note that excellent results are
obtained by H-LINK in terms of generated traffic. Moreover, we also note that
the variation in the adopted matching model does not significantly affect the
performance of H-LINK in terms of generated traffic. For what concern the recall
values, the optimal behavior of Gnutella is motivated by the fact that Gnutella
floods the network and succeeds in reaching a large part of either relevant and
irrelevant nodes, thus retrieving all the available concepts matching the target
query. On the other hand, H-LINK presents very interesting results in terms of
recall values especially with more accurate matching models.

6 Concluding remarks

In this paper, the main features of the H-LINK mechanism for query propagation
has been presented. The results obtained in the experiments show that H-LINK
succeeds in improving the effectiveness of traditional P2P query protocols by pro-
viding interesting results both in terms of scalability and accuracy. Future work
will regard the definition of further experiments in order to compare H-LINK with
recently proposed semantic routing protocols (e.g., REMINDIN’ [1]). Moreover,
H-LINK is the basic mechanism we plan to use for supporting the formation of
semantic communities of peers. On these topic, we are currently working in the
framework of the Esteem project and some initial results are presented in [8].

References

to-Peer Networks based on Social Metaphors. In: Proc. of the 13th Int. conference
2. Tsoumakos, D., Roussopoulos, N.: Adaptive Probabilistic Search (APS) for Peer-
to-Peer Networks. In: Proc. of the 3rd Int. IEEE Conference on P2P Computing,
Linköping, Sweden (2003)
with Semantic Topologies. In: Proc. of the Int. Conference on Semantics of a Net-
worked World (ICSNW), Paris, France (2004)
Proc. of the Int. Workshop on Peer-to-Peer Computing, Pisa, Italy (2002)
Conference on Pervasive Computing and Communications (PCC-04), Las Vegas,
Nevada, USA (2004)
7. Castano, S., Montanelli, S.: Semantically Routing Queries in Peer-based Systems:
Proc. of the ESWC Workshop on Ontologies in Peer-to-Peer Communities, Herak-
lion, Greece (2005)