# Semantic Routing for Effective Search in Heterogeneous and Distributed Digital Libraries

Federica Mandreoli*, Riccardo Martoglia*, Wilma Penzo**, and Simona Sassatelli*
*DII – University of Modena and Reggio Emilia, Italy
{fmandreoli,rmartoglia,sassatelli}@unimo.it
**DEIS – University of Bologna, Italy
{wpenzo}@deis.unibo.it

***Abstract*— Next generation Digital Libraries (DLs) will offer an entire ensemble of systems and services designed to help users to easily find and access the information they are looking for. However, much work is still required in order to achieve this vision. In this paper, we concentrate our attention on devising techniques allowing an effective routing of queries, which we think can be of the utmost importance in providing effective and efficient querying in heterogeneous and distributed DLs, identifying the best ways to navigate the available nodes and, thus, the documents (or their parts) which are most suitable to best answer the user needs. We describe a routing mechanism, which we call *routing by mapping*, in which the query is sent to the DL peers whose subnetworks best approximate the concepts required. To this end a distributed index mechanism is adopted, which we call *Semantic Routing Index (SRI)*. We also present some exploratory experiments showing the effectiveness of the proposed approach.**

## I. Introduction

In recent years, the constant integration and enhancements in computational resources and telecommunications, along with the considerable drop in digitizing costs, have fostered development of systems which are able to electronically store, access and diffuse via the Web a large number of digital documents and multimedia data. In such a sea of electronic information, the user can easily get lost in her/his struggle to find the information (s)he requires. For these reasons, the concept of *Digital Library* (*DL*) has become a pivotal one: exactly as a physical library, a DL contains a collection of documents that are at the users' disposal. The most advanced DLs becoming available today have the following features, among others: (i) documents (textual documents or even metadata on multimedia items) are not limited to free text, but are most likely also expressed in semistructured formats, such as XML associated to XML Schemas; (ii) they come from different sources, usually available on the web, and are *heterogeneous* for what concerns the structures adopted for their representations but related for the contents they deal with; (iii) the underlying architecture is more and more often *distributed* over a number of nodes (peers), each one, for instance, managing specific document collections.

Along with the documents themselves, a good next generation DL should offer an entire ensemble of systems and services designed to help users to easily find and access the information they are looking for. Indeed, querying and accessing distributed and heterogeneous DL information in an effective and efficient way requires to devise a whole series of techniques in several synergic areas. Consider for instance Figure 1 as a sample scenario of a portion of a distributed DL containing data about publications. Each peer composing the DL network ("DL Peer" in the picture) is enriched with a schema that represents the peer's domain of interests, and semantic mappings, represented as grey bold lines, are locally established between peers' schemas [1], [2], [3]. In order to query a peer in the DL, its own schema is used for query formulation and mappings are used to reformulate the query over its immediate neighbors, then over their immediate neighbors, and so on. Thus, query answers can come from any peer in the DL that is connected through a semantic path of mappings [4]. In such a setting, effectively answering a query means propagating it towards the peers which are semantically best suited for answering the user needs. However, it is not always convenient for a peer to propagate a query towards all other peers. In particular, a query posed over a given DL peer should be forwarded to the most relevant peers that offer semantically related results among its immediate neighbors first, then among their immediate neighbors, and so on. As an example, let us consider the following query, posed on the schema of peer A: "Retrieve the titles of the scientific publications of author XY". The peer A's neighbors peer B and peer C are very similar as to the portion of the schemas involved in the query above; as to the second step of query reformulation, peer E is more relevant than peer D and peer F, since it deals with scientific publications, instead of magazines and newspapers. For these reasons, the answers obtained from path peer C - peer E fit better the query conditions than those from paths peer B - peer D - peer F and peer B - peer F.

In this paper, we concentrate our attention on devising techniques allowing an effective routing of queries in a distributed environment, which we think can be of the utmost importance in providing effective and efficient querying in next generation DLs, identifying the most relevant documents (and documents' portions) in their network. We describe a routing mechanism, which we call *routing by mapping* [5], in which the query is sent to the peers whose subnetworks best approximate the concepts required. To this end a distributed index mechanism is adopted: each peer in the DL owns a *Semantic Routing Index (SRI)* which summarizes the ability of its subnetworks to semantically approximate the concepts

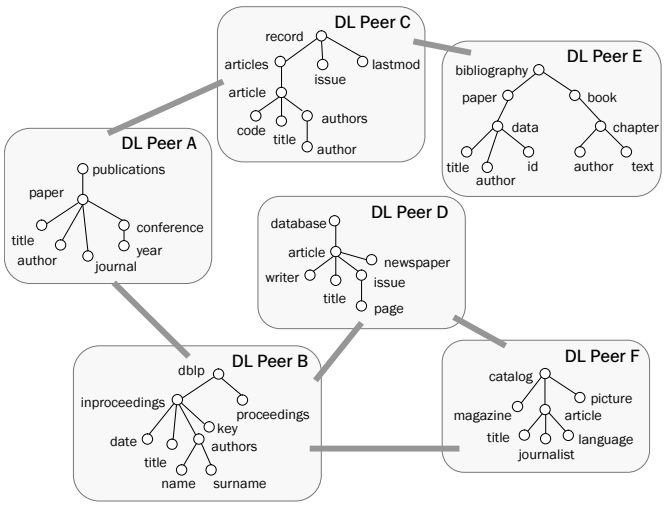
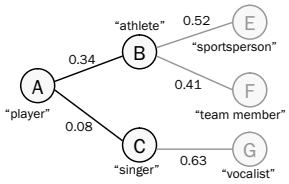



Fig. 1. A small example of a distributed bibliography DL




(a) Semantic mappings.

| | sport | team | coach | player |
|---|---|---|---|---|
| **B** | 0.19 | 0.44 | 0.32 | 0.28 |
| **C** | 0.02 | 0.08 | 0 | 0.05 |

(b) SRI of peer A.

Fig. 2. Example of the semantic routing by mapping mechanism.

of its schema. Such data structures are dynamically computed exploiting the available semantic mappings and evolve with the network topology following specifically devised algorithms. The paper is organized as follows: in Section II we introduce our approach, present the SRI structure and its use for routing, Section III shows the results we obtained in our preliminary experimental tests, while Section IV presents a short discussion of related works and concludes the paper.

## II. Semantic Routing by Mapping

In our work we rely on the notion of summarized subnetworks as in [6], and we propose the *routing by mapping* mechanism, where the selection of the best answering peers in the DL is based on the semantic information about the peers' contents. It relies on the *semantic mappings* (originally described in [7] for a heterogeneous centralized environment) that each peer establishes between its schema and the ones of its neighbors by performing apposite schema matching operations. By means of these mappings each concept of the peer schema is associated to the most similar concepts of the neighbors schemas and each of these associations is characterized by a numerical score, belonging to the interval [0,1] and quantifying the level of semantic approximation in moving from the first to the second concept. In the scenario we consider, a query originating from a given peer is always expressed in terms of its reference schema. If routing was limited to the semantic knowledge each peer has on its neighbors, every query reaching a peer would be forwarded to the neighbors having the highest scores for the required concepts, since these peers have the highest probability to produce correct results.

*Example 1:* Let us consider a portion of a multi-topic distributed DL (Figure 2-a). Peers A, B, E and F are nodes containing documents about sports, while Peers C and G's topic is music. Peer A has established appropriate semantic mappings with its two neighbors, peers B and C. We now suppose that, according to these mappings, concept "player" of peer A schema is associated to concept "athlete" of peer B and to concept "singer" of peer C; the scores for these mappings, indicated on the connecting arcs, reflect the similarity between the two couples of concepts and are 0.34 and 0.08, respectively. This means that, according to our routing mechanism, a query posed to peer A and asking for concept "player" would be preferably forwarded to peer B, because it has an higher approximation score for concept "player" w.r.t. peer C, signifying an higher probability to obtain useful documents parts. □

### A. From Mappings Scores to Summarized Information

A good routing mechanism should not be limited to the exploitation of the information about the neighbors alone. Indeed, in the neighbors selection, each peer should also consider the approximation capability of the peers belonging to the subnetworks routed by its neighbors (i.e. peer E, F and G in the Figure 2-a), as the query would likely be propagated to these subnetworks too. Ideally, it would be desirable for each peer to calculate a semantic mapping with each other peer of the DL, so that this information could be exploited in the routing process. However, an approach of this kind is clearly not applicable in a real-life distributed DL context, due to the excessive amount of data to be stored because of the potentially very large number of peers.

Instead, in our approach, each peer creates and maintains cumulative information summarizing the approximation capabilities of the whole subnetworks routed by each of its neighbors. This summarized information is calculated by each peer by appropriately combining the semantic mappings scores towards its neighbors with the summarized information each neighbor has about its own subnetwork. Being such information computed in the same manner, we obtain that the knowledge about mappings is propagated throughout the whole DL and each peer can learn about all other peers without being directly connected or interacting with them. Further, in order to avoid the presence of cyclic paths in the updates propagation, when a peer connects to the network a cycle detection mechanism based on global unique identifiers, as in [6], may be adopted. To obtain the cumulative information we apply two different types of operations, named *aggregation* and *composition*, to the original mapping scores. Before introducing in detail the data structures we devised for conveniently maintaining this summarized information, let us show by means of an example of use of these operations.

*Example 2:* Consider again the DL scenario of the previous example (Figure 2-a). Peer B is connected to peer E and F other than peer A, and so the score for the mapping that associates concepts "player" and "athlete" must be revised considering the subnetwork of B. To this end, peer A computes its semantic score towards B by *composing* the similarity score between "player" and "athlete" (i.e. 0.34) with a score obtained from peer B indicating how well concept "athlete" can be approximated in the subnetwork including peer E and F. This last score is computed by peer B by *aggregating* the scores characterizing its mappings for concept "athlete" towards neighbors E and F. Specifically, these mappings involve concepts "sportsperson" (peer E) and "team member" (peer F) with two scores of 0.52 and 0.41, respectively. Peer B sends the aggregated result to A, which composes it with its score of the mapping "player"-"athlete" and obtain a final score expressing how well the concept "player" can be semantically approximated by the subnetwork routed at peer B. Similarly, the score for "player" (peer A) toward peer C must be computed considering the subnetwork of peer C, i.e. the only peer G. Thus, the score for "player" toward peer C is calculated, by peer A, by *composing* the similarity score between "player" and "singer" (0.08) with the aggregation of the scores characterizing the mappings of peer C toward its neighbors (i.e. the only peer G), that corresponds to the only score 0.63. □

As to the actual execution of the aggregation and composition operations, they are performed by applying appropriate mathematical functions, adequately expressing the meaning of aggregation and composition.

### B. Semantic Routing Indices

To maintain the information about mapping scores, each DL peer owns a specially devised data structure called *Semantic Routing Index (SRI)*. The index is represented by a matrix and, for each peer of the system, the rows are associated to the peer neighbors, while the columns refer to the concepts of its schema. An example of such data structures is represented in Figure 2-b for peer A, whose schema is supposed to include only four concepts: "sport", "team", "coach" and "player". As can be seen, the columns of the matrix are associated to the concepts of peer A schema, while its rows are associated to peer A neighbors (i.e. peer B and peer C).

Our idea is that each cell of the matrix stores a score representing how the concept associated to that column is semantically approximated by the subnetwork routed by the neighbor associated to a given row. For example, the number 0.28 in the cell corresponding to the last column and the first row, means that concept "player" of peer A can be approximated through the subnetwork routed by peer B with a similarity score of 0.28.

The scores stored into the indices are computed by the involved peers in an incremental way, on the basis of the peer connections to the system, and following its evolution. Specifically, when entering the system, each peer queries its neighbors about their previous mappings and can consequently compute its scores performing the appropriate aggregation and composition operations. The sequence of operations to be performed is specified by a protocol whose detail can be found in [5].

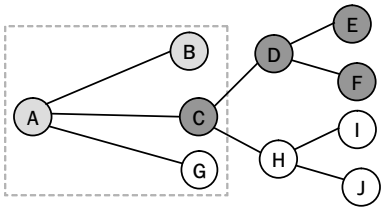



Fig. 3. The experimental scenario

## III. Experimental Evaluation

In this section we present a summary of a set of experiments we performed to test the effectiveness of the techniques presented in the previous sections. In this initial phase of our research, instead of relying on particular distributed DL architectures or implementations, we chose a simulation framework, able to reproduce the main conditions characterizing such an environment. In particular, for our experiments we used SimJava 2.0, a discrete, event-based, general purpose simulator, which allows us to verify the behaviour of our algorithms without using real systems. In this way we were able to abstract from additional DL management, network and communication issues, while maintaining full control on the internal dynamics of routing indices management, which is essential to our effectiveness evaluation ends. In our experiments we chose not to perform efficiency evaluation, since this would involve many other aspects related to networks issues, requiring an exhaustive analyses of them, and is therefore beyond our current goals.

We performed two types of experiments: the first type to verify the comparability of mapping scores and the second to show the usefulness of semantic routing indices. Due to the lack of space, we present only a small selection of results for each type of these tests. The common scenario we modelled through the simulator corresponds to a small distributed and multi-topic DL. Each node (peer) is supposed to contain documents different from the others and describing a particular reality. Further, in order to deepen the tests at different levels of semantic heterogeneity, we considered peers belonging to a small set of topics, where the schemas of the peers about the same topic describe the same reality from different points of view. In Figure 3 a portion of this network is depicted, where peers about the same topic are identified by the same shade of grey. In particular, peers in the figure contain documents about sport (peers A and B), music (C, D, E and F) and scientific publications (G, H, I and J). Notice that, since we currently are only in the initial phase of our testing, the considered network scenarios are not particularly complex. In the future, we will enrich them with more complicated network topologies and consider a larger number of peers, stressing our approach on real-life DL scenarios.

| PeerA | PeerB | | PeerC | | PeerG | |
|---|---|---|---|---|---|---|
| sport | sport | 0.1965 | storage | 0.0193 | | |
| team | club | 0.1202 | track | 0.0356 | article | 0.0585 |
| coach | trainer | 0.3858 | signboard | 0.0765 | journal | 0.0606 |
| player | athlete | 0.1721 | singer | 0.0834 | author | 0.0962 |

| | PeerC >D | PeerC >H | Gr Ratio |
|---|---|---|---|
| tracklist | 0.0093 | 0.0026 | 3.51 |
| track | 0.0025 | 0.0002 | 14.77 |
| singer | 0.0253 | 0.0161 | 1.58 |
| albumTitle | 0.0269 | 0.0002 | 174.44 |

Fig. 4. Results of first (top) and second (bottom) experiment

For the first type of experiments, we considered the part of the network in Figure 3 surrounded by the broken line, including peers A, B, C and G. The top part of Figure 4 shows the mapping scores of peer A, and the concepts these scores refer to. As can be seen, the matching algorithm correctly maps each peer A concept to the corresponding peer B concept. Also for peer C and G, whose schemas belong to different categories, associations are built between concepts considered the most similar for their semantics and positions, however in this case the mapping scores are very low. Nevertheless, mapping scores comparability is demonstrated because, for each peer A term, the mapping with the highest score is towards peer B; this reflects the fact that peer B, which is about the same topics of peer A, can semantically approximate peer A concepts in a better way than peer C and G do.

For the second type of experiments we focused our attention on peer C. In the bottom part of Figure 4 we show how the scores in peer C routing index are different between the subnetwork including three peers about the same subject (peers D, E, F, first column of the table), and the subnetwork of peers containing documents about different topics (peers H, I, J, second column). The scores are computed by applying the product as composition function and, as aggregation, the revision of a function commonly used in travel demand applications when modeling the aggregation of several alternatives [8]. In this type of tests, the key parameter for effectiveness evaluation is the growth ratio, i.e. the measure of how bigger are the scores towards the same topic subnetwork w.r.t. the other one. We can see, as we expected, that the scores are significantly higher in the first case (growth ratio greater than 1), reflecting that the SRI correctly captures the fact that the subnetwork of peers on the same topic of peer C contain documents whose concepts are semantically more similar to the peer C ones. Thus, these and other tests show that it can be possible to rely on SRIs in order to identify, in a distributed DL, the subnetworks containing the documents (or their parts) which are most likely to satisfy a given user query.

## IV. Related Work and Concluding Remarks

In order to design an efficient routing mechanism for a distributed DL environment, we analyzed and tried to exploit and significantly enhance the best ideas available in the distributed information management and search field, specifically from the P2P and PDMS areas. P2P systems provide very basic data management capabilities and rarely offer mechanisms to represent and exploit their semantic, with negative consequences for localization and retrieval operations. Therefore, basic P2P architectures alone are not flexible enough in order to provide the required search features of next generation DLs. On the other hand, recent PDMSs [9], [10] offer a decentralized and easily extensible architecture for advanced data management, in which anytime every node can act freely on her/his data, while in the meantime accessing data stored by other nodes. However, even in the most advanced systems, such as [11], [12], the routing mechanism is limited to the only local information provided by the neighboring peers or, as in [6], it is only based quantitative information.

In our work we tried to combine the routing capabilities of P2P systems and the semantic richness of PDMSs, enabling effective searches in a distributed, totally dynamic and flexible environment, not depending from a centralized server and without losing the semantic richness of queries. From the tests we performed, we can see that exploiting SRIs in a distributed DL environment could indeed be beneficial for a more effective querying process, since in this way it is possible to identify the best way to navigate the available nodes and, thus, the documents (or their parts) which are most suitable to best answer the user needs. In the future, we plan to deepen the test activity, stressing our approach on large real-life DL scenarios.

## References


[1] A. Halevy, Z. Ives, J. Madhavan, P. Mork, D. Suciu, and I. Tatarinov, "The Piazza Peer Data Management System," *IEEE TKDE*, vol. 16, no. 7, pp. 787–798, July 2004.

[2] W. Nejdl, B. Wolf, S. Staab, and J. Tane, "EDUTELLA: Searching and Annotating Resources within an RDF-based P2P Network." in *Proc. of WWW Intl. Workshop on the Semantic Web*, 2002.

[3] M. Arenas, V. Kantere, A. Kementsietsidis, I. Kiringa, R. Miller, and J. Mylopoulos, "The hyperion project: from data integration to data coordination," *SIGMOD Record*, vol. 32, no. 3, pp. 53–58, 2003.

[4] I. Tatarinov and A. Halevy, "Efficient Query Reformulation in Peer Data Management Systems," in *Proc. of SIGMOD*, 2004.

[5] F. Mandreoli, R. Martoglia, W. Penzo, and S. Sassatelli, "SRI: Exploiting Semantic Information for Effective Query Routing in a PDMS," in *Proc. of WIDM*, 2006.

[6] A. Crespo and H. Garcia-Molina, "Routing indices for peer-to-peer systems," in *Proc. of ICDCS*, 2002.

[7] F. Mandreoli, R. Martoglia, and P. Tiberio, "Approximate Query Answering for a Heterogeneous XML Document Base," in *Proc. of WISE*, 2004.

[8] M. Ben-Akiva and S. R. Lerman, *Discrete Choice Analyses: Theory and Application to Travel Demand*. The MIT Press, 1985.

[9] S. Gribble, A. Halevy, Z. Ives, M. Rodrig, and D. Suciu, "What Can Databases do for Peer-to-Peer?" in *Proc. of WebDB*, 2001.

[10] A. Y. Halevy, Z. G. Ives, D. Suciu, and I. Tatarinov, "Schema mediation in peer data management systems," in *Proc. of ICDE*, 2003.

[11] A. Castano, S. Ferrara, S. Montanelli, E. Pagani, and G. Rossi, "Ontology-addressable contents in P2P networks," in *Proc. of SemPGRID*, 2003.

[12] P. Haase, R. Siebes, and F. van Harmelen, "Peer Selection in Peer-to-Peer Networks with Semantic Topologies," in *Proc. of ICSNW*, 2004.