

The ESTEEM platform: enabling P2P semantic collaboration through emerging collective knowledge

Stefano Montanelli · Devis Bianchini · Carola Aiello · Roberto Baldoni ·
Cristiana Bolchini · Silvia Bonomi · Silvana Castano · Tiziana Catarci ·
Valeria De Antonellis · Alfio Ferrara · Michele Melchiori · Elisa Quintarelli ·
Monica Scannapieco · Fabio A. Schreiber · Letizia Tanca

Received: 25 July 2009 / Revised: 26 May 2010 / Accepted: 26 May 2010 /
Published online: 24 June 2010
© Springer Science+Business Media, LLC 2010

Abstract In this paper, we present Esteem (Emergent Semantics and cooperation in multi-knowledge Environments), a community-based P2P platform for supporting semantic collaboration among a set of independent peers, without prior reciprocal knowledge and no predefined relationships. Goal of Esteem is to go beyond the existing state-of-the-art solutions for P2P knowledge sharing and to provide an integrated platform for both data and service discovery. A distinguishing feature of Esteem is the use of *semantic communities* to explicitly give shape to the collective knowledge and expertise of peer groups with similar interests. Key techniques of Esteem will be presented in the paper and concern: *shuffling-based communication, ontology and service matchmaking, context management, and quality-aware data*

S. Montanelli (✉) · S. Castano · A. Ferrara
Università degli Studi di Milano - DICO, Via Comelico, 39, 20135 Milano, Italy
e-mail: montanelli@dico.unimi.it

S. Castano
e-mail: castano@dico.unimi.it

A. Ferrara
e-mail: ferrara@dico.unimi.it

C. Aiello · R. Baldoni · S. Bonomi · T. Catarci · M. Scannapieco
Università di Roma “La Sapienza” - DIS, Via Ariosto, 25, 00185 Roma, Italy

C. Aiello
e-mail: caiello@dis.uniroma1.it

R. Baldoni
e-mail: baldoni@dis.uniroma1.it

S. Bonomi
e-mail: bonomi@dis.uniroma1.it

T. Catarci
e-mail: catarci@dis.uniroma1.it

M. Scannapieco
e-mail: monscan@dis.uniroma1.it

integration. An application example of data and service discovery in the health-care domain will be presented, by also discussing results of system and user evaluation.

Keywords Ontologies · Semantic collaboration · Emergent collective knowledge · Data and service discovery

1 Introduction

Traditional information integration architectures, characterized by moderately static scenarios, shared understanding of the domain of interest, and closed, or at least access-controlled, set of participating sources (Lenzerini 2002) are leaving the floor to modern knowledge sharing infrastructures, characterized by dynamic collaboration, emerging knowledge of the domain, and peer-to-peer participation. Systems are actually shifting from hierarchical, supervised networks, where a set of possibly predefined nodes are in charge of coordinating the system organization, to open networked infrastructures, where local agreements and peer interactions autonomously emerge and automatically disappear when they are no longer required (Aberer et al. 2004). In such a scenario, autonomy and independence of network peers become crucial requirements which need to be addressed by proper techniques, for allowing nodes to act as completely decentralized agents available for collaboration. In this respect, ontologies, along with Semantic Web technologies like ontology matching (Shvaiko and Euzenat 2005) and semantic search (Fagin et al. 2005), come into play as key solutions for effectively addressing the various building blocks of these innovative collaborative platforms. Examples of these building blocks are semantics- and context-driven resource discovery (Haase et al. 2008), semantic routing (Löser et al. 2007), and P2P data management (Halevy et al. 2004). Open issues still concern the ability to go beyond simple peer-to-peer collaboration, moving towards more

D. Bianchini · M. Melchiori · V. De Antonellis
Università degli Studi di Brescia - DEA, Via Branze, 38, 25123 Brescia, Italy

D. Bianchini
e-mail: bianchin@ing.unibs.it

M. Melchiori
e-mail: melchior@ing.unibs.it

V. De Antonellis
e-mail: deantone@ing.unibs.it

C. Bolchini · E. Quintarelli · F. A. Schreiber · L. Tanca
Politecnico di Milano - DEI, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy

C. Bolchini
e-mail: bolchini@elet.polimi.it

E. Quintarelli
e-mail: quintare@elet.polimi.it

F. A. Schreiber
e-mail: schreiber@elet.polimi.it

L. Tanca
e-mail: tanca@elet.polimi.it

sophisticated interaction models where peers with similar interests are interlinked in a sort of *collective knowledge space*.

In this paper, we present Esteem (Emergent Semantics and cooperation in multi-knowledge Environments), a comprehensive platform for data and service discovery/sharing in a community-based P2P environment.¹ Peculiar feature of Esteem is the integration in a unique and comprehensive platform of a number of advanced Semantic Web techniques, concerning i) *shuffling-based communication*, for supporting P2P interactions and the autonomous formation of semantic communities of peers, ii) *semantic matchmaking*, for enforcing data and service discovery at different levels of flexibility and granularity, iii) *context management*, for profiling the peer behavior and for filtering the available resources according to the peer current context and preferences, and iv) *quality-aware data integration*, for specifying different levels of peer/data reliability. Focus of the paper is to present architectural and functional aspects of the Esteem approach. A semantic community is the core notion of Esteem around which the *collective knowledge* belonging to a group of peers with similar interests has the opportunity to be transparently recognized and to become explicitly aware, without sacrificing the autonomy of each single node. To clearly motivate the proposed approach, an application example of data and service discovery in the health-care domain will be presented. The platform demonstrator, that has been accessed by real users (i.e., doctors) during the evaluation phase of Esteem, will be described by means of examples concerning the main Esteem functionalities. Examples will show, on one side, how a doctor can join the Esteem network to get in contact with focused communities of highly specialized experts, with the aim of submitting queries and invoking services, and, on the other side, how users can bring their personal data and knowledge into the network for collaboration purposes.

We want to remark that the value of this paper primarily consists in i) the capability of addressing both data and service discovery and sharing in a unique picture, rather than separately considering each of them through ad hoc techniques as in most of the existing approaches, and ii) the proposal of a comprehensive infrastructure where the gain is on the integration aspects of the different techniques involved, rather than in the specific contribution of each of them. The Esteem platform is the result of a joint research project combining techniques and tools that the participant groups have separately developed either for overlay-based network management and for data or service semantic interpretability. In this paper, to highlight the project achievements, the various Esteem techniques are described by focusing on their roles and functionalities in the integrated platform and by characterizing their original contribution with respect to other similar solutions proposed in the literature. Appropriate references to more technical publications on the Esteem techniques are provided throughout the paper for the interested reader.

The paper is organized as follows. Motivations and original contributions of Esteem are presented in Section 2. In Section 3, we discuss the role of semantic communities for Esteem and we present the architectural organization of the proposed platform. In Section 4, we provide details about semantic community formation and advertisement. Esteem techniques to enforce community-based data

¹<http://www.dis.uniroma1.it/~esteem/>

and service discovery are presented in Section 5 and 6, respectively. In Section 7, we discuss selected experimental results we performed to evaluate the effectiveness of the Esteem platform. Finally, concluding remarks are provided in Section 8.

2 Motivations and goals of Esteem

The goal of Esteem is to enhance the existing solutions for P2P knowledge sharing and to provide an integrated platform for both data and service discovery. A distinguishing feature of Esteem is the use of *semantic communities* to explicitly *give shape* to the collective knowledge and expertise of peer groups with similar interests. In this sense, the Esteem platform can be considered as a concrete attempt to combine the benefits of recently emerging social-based approaches with more consolidated Semantic Web technologies, as argued in Gruber (2008). In particular, the Esteem techniques have been developed to provide an integrated and comprehensive approach capable of addressing the following open issues in the current state-of-the-art of P2P systems.

To overcome the lack of techniques for supporting semantic interpretability across heterogeneous depastures Semantic interpretability is a crucial problem in open distributed systems like P2P, due to the need of enabling peers to a seamless access of the right information resource, both in case of data retrieval and service invocation. Currently, most of the existing solutions in the field rely on the creation of a *semantic overlay network* where the links among peers, called *semantic links*, are established according to the content similarity of nodes rather than to their topological proximity (Zeinalipour-Yazti et al. 2005; Löser et al. 2007). In this respect, a number of matching techniques have been proposed for supporting the P2P discovery of semantic links and to enable the overlay management. However, the choice of the most suitable family of matching functions is still a matter of research due to the inherent trade-off between the need of scalability and accuracy that are both crucial requirements for enforcing effective semantic collaboration.

Ontology and service matching techniques have been developed in Esteem to enforce both data and service discovery within a single comprehensive framework. Specific modules for ontology and service matching are included in the Esteem architecture, based on the HMatch 2.0 ontology matching engine (Castano et al. 2006b) and on the FC-MATCH (FunctionalCompatibility-Match) service matchmaking approach (Bianchini et al. 2008), respectively. Focused versions of both HMatch 2.0 and FC-MATCH have been developed for Esteem to specifically work in the P2P environment, by strengthening the *dynamic configurability* of the matching process instead of proposing a novel set of matching techniques. This means that the execution of the matching process can be customized at runtime according to the features of the specific matching scenario to consider, by flexibly combining the invocation of the matching techniques that are more appropriate. An extensive experimentation of the ontology matching techniques adopted in Esteem have been performed over the 2006, 2007, and 2009 benchmarks of the Ontology Alignment Evaluation Initiative (OAEI)² (Castano et al. 2006a).

²<http://oaei.ontologymatching.org/>

To overcome the lack of techniques for enforcing efficient and unsupervised recognition/aggregation of peers with similar content In the recent years, the idea to move from a network of random peer interconnections towards a grid of semantic links, each one denoting the existence of a relationship between the knowledge of the involved nodes, is getting more and more importance. The idea is that grouping nodes according to similarity-based criteria can have a positive impact on both traffic load and effectiveness of the search processes. In most cases, the notion of peer group is not explicitly supported and each node only maintains a *semantic neighborhood*, namely a set of peer-to-peer connections with other nodes storing similar contents (Hidayanto and Bressan 2007; Haase et al. 2008). In other approaches, the notion of peer community is explicitly supported, but the maintenance of a community structure requires some forms of supervising authority, thus loosing the inherent autonomy and independence of peers (Wang and Vassileva 2004).

On this topic, the Esteem contribution is twofold. First, community formation in Esteem is unsupervised and it is enforced through *lightweight handshake techniques*. In contrast with most of the existing solutions, community membership is open and approval/rejection of a peer is not determined by the decision of a supervisor. Moreover, community maintenance is efficient due to the fact that peers can autonomously join/leave communities at any moment, according to their collaboration needs, without requiring community re-organization or structural adjustment. Second, query propagation in Esteem is enforced through a *routing-by-community* mechanism. The idea of routing-by-community is to use communities as query recipients, thus allowing to evaluate the relevance of a query against the topic of a community, and to reach all the members of a relevant community with a single request. Routing-by-community has been specifically developed for Esteem and it is an extension of H-Link (Castano and Montanelli 2007), a content-based routing approach based on semantic similarities among peer contents and on the use of single-peer recipients.

To overcome the lack of techniques for assessing the reliability of peers In some P2P scenarios, the effectiveness of a semantic collaboration depends on the capability of delimiting the peer-to-peer interaction among a restricted set of peers that are considered as more *reliable* than others. In Esteem, we deal with two different kinds of reliability, namely *trust-based reliability* and *context-based reliability*.

Trust-based reliability allows a peer to select the collaboration partners according to their level of trust and to the quality of the data they provide. Trust and data quality issues in P2P systems are only marginally considered in the current literature, though some interesting proposals are available (Xiong and Liu 2004; Rana and Hinze 2004). When considering quality-aware data integration systems, both traditional and P2P, two principal types of conflicts may arise: key-level conflicts (i.e., conflicts on key attributes) and attribute-level conflicts (i.e., conflicts on attributes that are not keys) (Fan et al. 2001; Sattler et al. 2003). Most of the existing systems are focused on attribute-level conflicts, and they do not explicitly deal with key-level conflicts. The DaQuinCIS system (Scannapieco et al. 2004) provides key conflict resolution in the framework of traditional data integration systems. The Esteem platform adopts the DaQuinCIS system to support trust and data quality, by solving key-level conflicts at query processing time. Moreover, Esteem extends the DaQuinCIS system by introducing specific solutions targeted to work in a P2P

environment, such as the full automation of the discovery procedure of the key attributes to match.

Context-based reliability allows a peer to select the collaboration partners according to the level of match of their corresponding contexts. This way, it is possible to assign a higher priority to those resources coming from peers with a matching context and to filter-out/discard information coming from other peers. The notion of context in Esteem is introduced to enable a peer to reduce the load of irrelevant (or loosely relevant) resources collected for a potentially high number of replying peers during a search interaction. The notion of context, formerly emerged in various fields of research like psychology and philosophy (Chalmers 2004), is acquiring a great importance also in the computer science field. In the last few years, sophisticated and general context models have been proposed to support context-aware applications, and a rather comprehensive survey can be found in Bolchini et al. (2007a). Community-based approaches to context definition are also being proposed for the P2P environment, with the aim at enabling peers to incrementally build a shared knowledge base (Chen et al. 2003; Ouksel 2003). However, in these kinds of approach, the notion of context is embedded in the considered P2P system, and it is hard to adapt to a different application scenario. The notion of context proposed in Esteem is the P2P application of a more general context model based on the *Context Dimension Tree (CDT)* (Bolchini et al. 2009, 2007b). The CDT allows the flexible representation of all the possible profiles of a peer and it can be combined with other techniques (i.e., routing-by-community, ontology/service matching) to enforce semantic collaboration in a more effective way.

Running example Throughout the paper, we consider an application scenario in the health-care domain with the goal of presenting the main Esteem contributions described above and the associated demonstrator. In the example, we consider a doctor working in a hospital of the Central Africa who is aimed at finding data and services for healing a patient with malaria and adrenal insufficiency. The Unified Medical Language System³ is exploited as a reference knowledge source for defining both the community interests and the peer ontologies. From the system point of view, the device used by the doctor for interacting with the network (e.g., a laptop, a PDA, a smartphone) represents a network peer equipped with the Esteem demonstrator.

3 Emerging collective knowledge in Esteem

Esteem is defined as a community-based P2P platform for supporting semantic collaboration among a set of independent peers, without prior reciprocal knowledge and no predefined relationships. Such a collaboration scenario is *multi-knowledge*, in that no centralized authority is defined to manage a comprehensive view of the resources shared by all the nodes in the system, while preserving the autonomy of each participating peer.

³<http://umlsinfo.nlm.nih.gov/>

3.1 Knowledge equipment

An Esteem community autonomously emerges as a peer group built around a declared interest expressed in the form of an ontology-based *manifesto*. The community manifesto can be seen as a conceptual means to enable the system peers to become aware of their common interests and, thus, to recognize the existence of a collective knowledge that can be handled and organized in a concrete structure (i.e., the community). In other words, an Esteem community and the associated manifesto enables a peer to move from a *peer knowledge space*, where it is considered as a single agent with its personal knowledge, towards *collective knowledge space*, where it is considered as a member of multiple groups storing a portion of the group knowledge (see Fig. 1). In particular, in Esteem, the knowledge that a peer can bring in the system is characterized as follows.

The **peer ontology** is the core knowledge of a peer and it provides a semantically rich description of the peer data that are available for sharing. The peer ontology is exploited to evaluate whether matching knowledge can be returned to a requesting peer in reply to an incoming query. Furthermore, the peer ontology is also exploited for deriving the current peer interests and for determining the semantic communities to join. Besides the classical methodologies and editing tools for manual ontology engineering, tool-supported approaches can be adopted for creating a peer ontology (Gomez-Perez et al. 2003). A viable approach is based on (semi-)automated derivation of OWL axioms from ER/UML schemas and from relational database schemas of the peer resources (e.g., Baader et al. 2003; Curino et al. 2009). This way, domain knowledge already encoded in data schemas can be reused in form of peer ontologies, thus sensibly reducing the manual effort required. In more recent work, approaches suitable for non-specialist users are being proposed to generate the peer ontology by relying on the results of semantic annotation of the peer resources (e.g., Specia and Motta 2007; Mukherjee and Ramakrishnan 2008). As an additional feature available

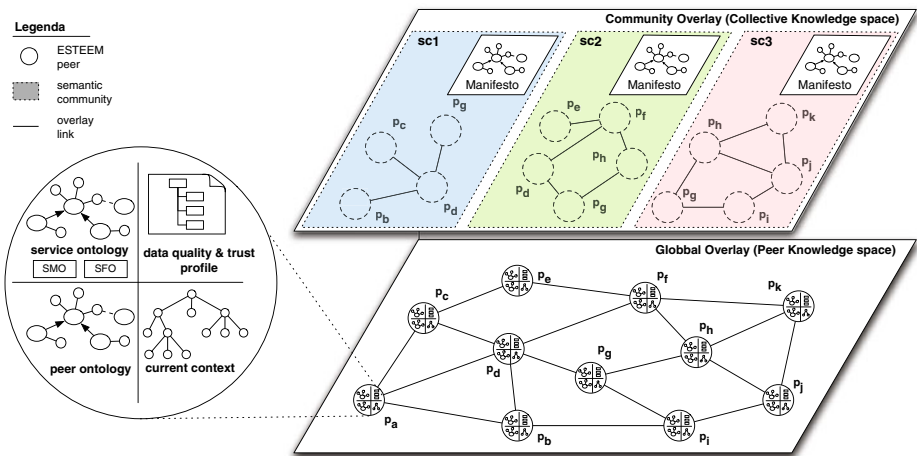


Fig. 1 The knowledge equipment of Esteem

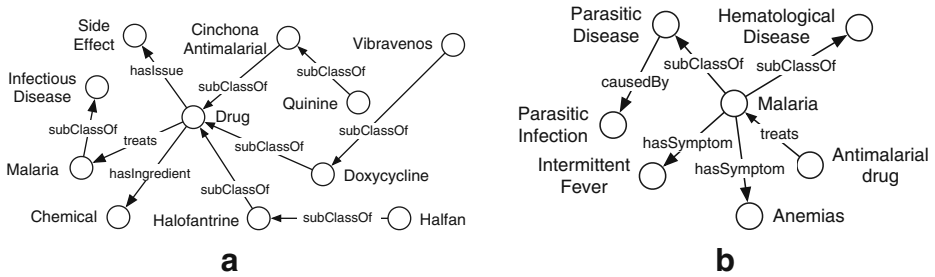


Fig. 2 Example of peer ontology for **a** the peer p_d , and **b** the peer p_f

in Esteem, the peer ontology can be built starting from an initial list of keywords denoting the interests of the final user (i.e., the doctor in the considered application example). A reference peer ontology can be obtained by acquiring one or more ontology fragments from the Semantic Web and/or from the network nodes with similar interests according to the specified keywords. This is possible, for example, in the health-care domain considered in the Esteem scenario, where a number of taxonomies/ontologies are available and can be exploited/downloaded by a peer (e.g., MeSH - Medical Subject Headings,⁴ SNOMED - Systematized Nomenclature of Medicine⁵). As an example, in Fig. 2, we show a portion of the ontologies of the peer p_d and p_f , respectively.

The **service ontology** provides a semantically rich description of the peer services that are available for sharing. Service description represents functional aspects of a service, based on the WSDL standard for service representation, in terms of service category, service functionalities (operations) and input/output messages (parameters). According to Bianchini et al. (2009), semantic service descriptions are obtained by means of a *Service Message Ontology* (SMO), whose concepts are used to add semantics to service I/O parameters, and a *Service Functionality Ontology* (SFO), whose concepts are used to add semantics to service functionalities (operations). In particular, in Esteem, where semantic interpretability of both data and services is supported, the SMO coincides with the peer ontology. The SFO is used to conceptualize the functionalities that a service provides. In Fig. 3, we show an example of SFO derived from the HL7 health-care standard.⁶ The SFO is formalized as a taxonomy, where generic terms represent service categories, that are specialized into concepts representing single service operations.

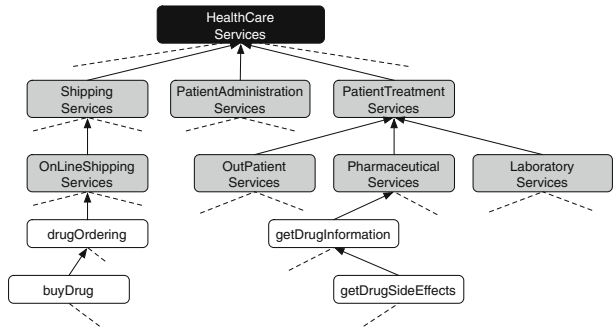
The **current context** of a peer is a subtree of the Context Dimension Tree (CDT) and it describes the peer profile, its interests, situation, and spatial/temporal coordinates at the time of the interaction with the Esteem network. The CDT expresses the several perspectives (*dimensions*) determining what portion of data is interesting in the various situations. The user category, *actor*, the *situation*, and

⁴<http://www.nlm.nih.gov/mesh/meshhome.html/>

⁵http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html/

⁶HL7 (Health Level 7): <http://www.hl7.org/>.

Fig. 3 Example of Service Functionality Ontology from the HL7 health-care standard



the *interest topic* are some of the most commonly significant *dimensions*, driving the selection of relevant information/services. A dimension value can be further analyzed w.r.t. different viewpoints, generating further (sub)dimensions organized in a tree-like structure. The current context of a peer is obtained by appropriately choosing a set of dimension values of the CDT. Each context determines a portion of the entire data set (a *data chunk*), specied as a view, representing the data that are relevant when the corresponding context becomes the current one. In Fig. 4, we show an example of CDT modeling all the possible contexts in a health-care scenario, like the one considered by Esteem. The grey area identifies the specieic context of a doctor in the field, interested in textual information on drugs useful for pathologies present in the specieic region she is in.

The **data quality and trust profile** involves the computation of data quality metrics on the peer data that are available for sharing with the other peers. Each peer has the opportunity of semi-automatically associating *quality metadata* with the exported data. Such metadata represent data quality measures corresponding to specieic quality dimensions. Currently, the implemented metrics refer to the dimensions most commonly defined for data quality, namely: *column completeness*, *format consistency*, *accuracy* and *internal consistency* (Batini and Scannapieco 2006, chap. 2).

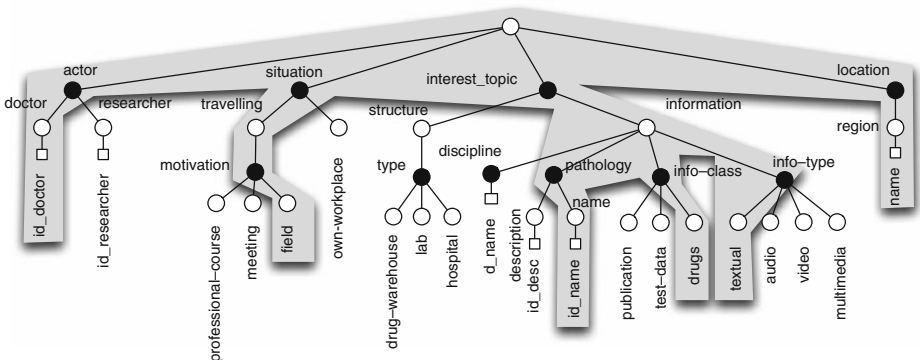


Fig. 4 An example of Context Dimension Tree

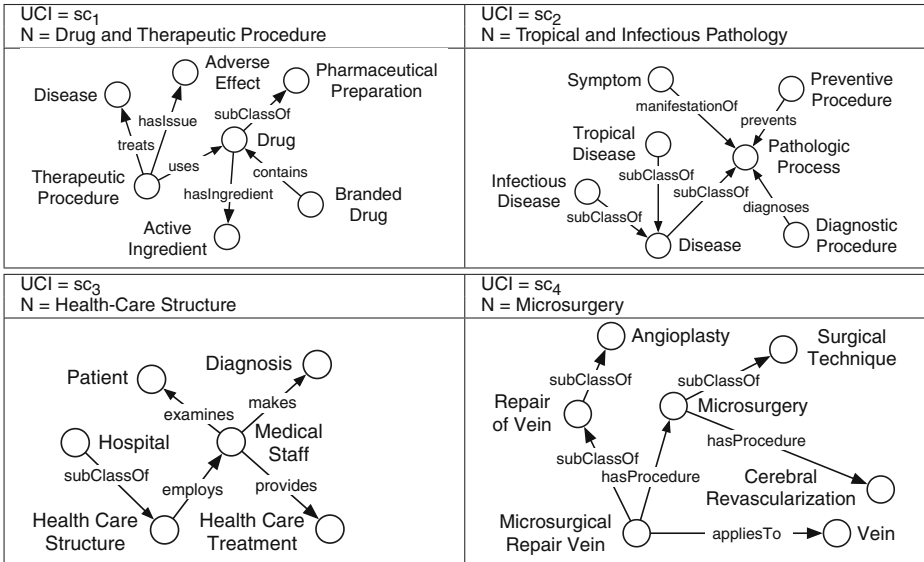


Fig. 5 Example of community manifestos

The **community manifestos** of the joined semantic communities are locally stored by each peer. A community manifesto is defined according to the preferences of the community founder (i.e., a peer) which proposes the community formation. In general, the community manifesto is extracted from the peer ontology of the founder and consists of a focused ontology. The level of detail used for specifying the community manifesto depends on the community goal. Portions of the SFO, the CDT, and the data quality and trust profile can be also included in a community manifesto to further specify the community target. For example, by associating a CDT with the focused ontology of the manifesto, the founder specifies which context dimensions are suitable for the community. In this respect, the founder specifies the correspondence between each given context of the CDT and the portion of the manifesto ontology (i.e., data chunk) that is relevant to it.

As an example of community manifesto extracted from a peer ontology, in Fig. 5, we show the manifestos of four different communities, namely sc_1 , sc_2 , sc_3 , and sc_4 , featuring their specific focus of interest in the health-care domain.

3.2 System architecture

The Esteem platform relies upon an unstructured P2P network where a semantic community is defined as an *overlay*, namely a logical layer built on top of a basic P2P infrastructure called *global overlay* (see Fig. 1). In Esteem, a semantic community sc is defined as a 4-tuple of the form $sc = \langle UCI, N, L, M \rangle$ where UCI is the Universal Community Identifier that univocally characterizes the community sc , N and L are a symbolic name and a natural-language description of the community interests, respectively, and M is the community manifesto. In particular, UCI and M are used in Esteem to enforce identification and characterization of a semantic community

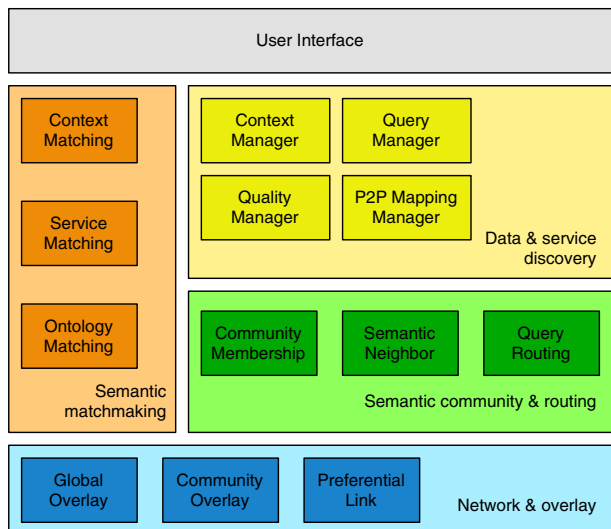
at the system level, while N and L provide a community description at the user-interface level. According to its interests, an Esteem peer can be included in zero or more communities by joining the corresponding community overlay. For instance, in the example of Fig. 1, the peer p_g joins all the existing communities sc_1 , sc_2 , and sc_3 , while the peer p_a does not participate in any semantic community and it is included only in the global overlay.

A *probe/search mechanism* is defined in Esteem to distinguish:

- the **discovery phase**, based on ontology matching, where *probe queries* are defined to identify the communities/peers that are capable of providing relevant knowledge with respect to a given topic of interest;
- the **sharing phase**, based on P2P mapping definition, where standard *search queries* are defined to point-to-point interact with a previously discovered peer for actual data acquisition and/or service invocation.

To this end, an Esteem peer is organized in a component-based architecture as described in Fig. 6. The user exploits Esteem for satisfying its collaboration needs (**Data & service discovery**). Techniques for context and quality/trust management can be invoked during peer interactions to improve the effectiveness of the sharing phase according to the specific requirements of the collaboration that is considered in a given moment. At the system level, semantic communities are used during the discovery phase as the enabling infrastructure for allowing the request of a single user to be propagated towards the groups of potential replier (**Semantic community & routing**). As a peculiarity of Esteem, ontology and service matching techniques are used to guide the system choices in handling both community formation and message routing (**Semantic matchmaking**). Appropriate techniques are provided for managing the peer connectivity and for maintaining the community overlays (**Network & overlay**).

Fig. 6 The architecture of an Esteem peer



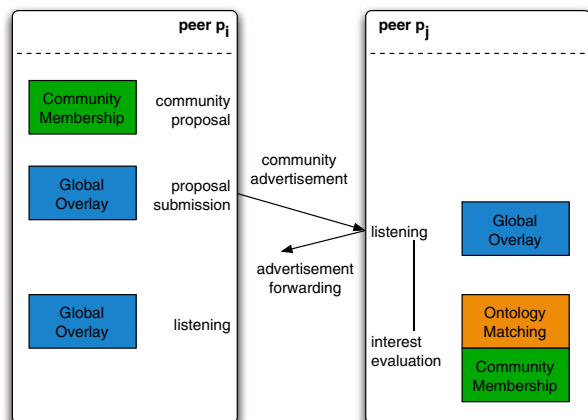
4 Esteem communities for P2P semantic collaboration

Esteem peers are initially inserted in a global overlay where interactions are enforced among all the system nodes and peer connectivity is maintained by means of a *shuffling-based mechanism* (Voulgaris et al. 2005). A community overlay, namely an Esteem community, emerges in the system when a peer p_i , called community founder, triggers the formation of the community sc by manually specifying a manifesto and by invoking the lightweight handshake techniques as follows (see Fig. 7).

Community advertisement It is performed by the community founder p_i and it consists in the dissemination of an advertisement message containing the identifier (UCI_i) and the manifesto (M_i) of the emerging community sc_i . Advertisement propagation throughout the network is performed in an automatic way through the Global Overlay module by appending these messages to ordinary shuffling communications.

Membership identification Receiving the incoming advertisement, each candidate peer p_j in the global overlay has to decide whether to join the emerging community sc_i . The peer p_j invokes the Community Membership module for assessing its level of interest in the community by relying on the Ontology Matching module. The underlying idea is that the level of interest of a peer in a community can be measured as the level of similarity between the ontology of the peer p_j and the incoming manifesto M_i . The user of the peer p_j has to specify a set of join constraints for configuring the ontology matchmaker (i.e., HMatch 2.0) and for defining the minimal matching conditions that are required to join a community (e.g., at least one concept in the peer ontology that matches the manifesto M_i with a given semantic affinity value). In case that the join constraints are satisfied, the peer p_j automatically joins the community by storing the associated UCI_i and M_i , otherwise the advertisement message is discarded. At the same time, the advertisement message is also exploited by the Global Overlay module for forwarding.

Fig. 7 Lightweight handshake of a semantic community



We note that the Global Overlay module maintains a listening socket as long as the peer is active in the system in order to receive community advertisements and to update the list of available communities. Through lightweight handshake, each peer maintains in the Community Membership module a list JSC of the joined semantic communities with $JSC = \{\langle UCI_1, M_1 \rangle \dots \langle UCI_k, M_k \rangle\}$ where $\langle UCI_i, M_i \rangle$ are the identifier and the manifesto of a joined community, respectively. Such a list is managed with a LRU policy in order to drop the less interesting communities. As a result, no explicit procedure for community disbandment is required. An Esteem community is considered as disbanded when the associated identifier is dropped by all the peers in the network and the associated overlay is no longer exploited. Furthermore, we note that the Esteem peers are not responsible for member deletion and failure events. A leaving member is not required to notify the system about its change of status, since the network organization is unstructured and since no supervising authority is defined to coordinate the peer behavior.

Creation of a new semantic community In Esteem, the choice of a peer to become community founder and to initiate the formation of a new community is autonomous. Community formation can be manually triggered by the user of a peer when the query results collected from the members of the currently joined communities are poor and the semantic collaboration is not satisfactory. Moreover, the proposal of a new community is “tool-assisted”, in the sense that Esteem automatically suggests to create a community by sniffing the sharing traffic among the peers and by collecting statistical data to detect the concepts that are mostly queried in the network. Statistical observations on traffic are also useful to support a sort of semi-automatic *community splitting* mechanism. The formation of a new community with a more focused manifesto can be promoted by the user of a peer when the traffic of an existing community is becoming intensive. The idea is that if a community is highly exploited for collaboration, more fine-grained overlays can be generated to satisfy more focused collaboration needs.

4.1 Esteem contribution

The following contributions characterize formation and advertisement of the Esteem communities.

Shuffling-based communication Shuffling techniques are used in the Network & Overlay component of the Esteem architecture to maintain the peer connectivity within the basic global overlay and within the various community overlays. The Global Overlay module (GO) and the Community Overlay module (CO) are defined to this end. The GO module relies upon the use of an Overlay Management Protocol (OMP) to arrange the peer connections despite the continuous and interleaved process of arrival/departure of nodes. In Esteem, a shuffling-based OMP is adopted. Shuffling is a robust gossip-based mechanism which exploits randomness to disseminate information across the P2P network (Allavena et al. 2005; Voulgaris et al. 2005). The basic idea of this OMP is to keep a node connected to a small set of other nodes that are continuously changed through random exchange of neighborhood (i.e., shuffling operation). This way, message propagation in the global overlay is epidemic and inexpensive in terms of traffic overhead since piggy-backed on the basic

shuffling operations. A shuffling-based OMP has been chosen for Esteem due to its scaling capability without affecting the peer load and the communication latency. Moreover, experiments show that shuffling-based networks like Cyclon (Voulgaris et al. 2005) are robust and self-healing to multiple peer failures under the assumption that the dynamism of the network is not too high (as occurs in the Esteem scenario). In the GO module, each peer maintains an Access Point (AP) table, where it stores information about the discovered communities. Besides community advertisement, a peer can speed up the discovery of existing communities by querying a random set of peers in the global overlay through the GO module. The querying peer defines a *random walk* in the global overlay and it acquires from the peers on the walk their AP tables, thus discovering the semantic communities contained in them. The manifesto of the discovered communities is then evaluated for possible join as described in the membership identification step. Within a community sc , the shuffling-based OMP is executed by the CO module to maintain the peer connectivity within the community overlay of sc .

Ontology matchmaking In Esteem, ontology matchmaking is performed by relying on the HMatch 2.0 engine and it has the role of measuring the level of match between concept descriptions of different peers through a process of semantic affinity evaluation. In particular, HMatch 2.0 is invoked during the handshake of an Esteem community to measure the level of semantic affinity between the proposed community manifesto and the peer ontology of a receiving peer. Moreover, HMatch 2.0 is invoked during the discovery phase of Esteem i) to select the more adequate query recipients in the routing-by-community mechanism, and ii) to evaluate whether a peer can provide matching knowledge in reply to an incoming probe query (see Section 5). Matching in HMatch 2.0 is defined as a process which takes as input two ontologies and that returns as output the mappings between those pairs of concepts in the two ontologies that are considered as similar with respect to their name and their structure/context. Given two concepts c' and c'' , the function $SA(c', c'') \rightarrow [0, 1]$ calculates a semantic affinity value 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 HMatch 2.0 provides a measure of similarity between two concepts computed on the basis of their linguistic features, namely their names. For the linguistic affinity evaluation, HMatch 2.0 relies on a thesaurus of terms and terminological relationships automatically extracted from the WordNet lexical system. The contextual affinity function of HMatch 2.0 provides a measure of similarity between two concepts by taking into account their contextual features, namely their semantic relations and corresponding range values. In this respect, four matching models are available in HMatch 2.0 to customize the behavior of the contextual affinity function by allowing to choose the different kinds of ontology axioms to consider in the context of c' and c'' (e.g., `equivalentClass` and `subClassOf` axioms). The comprehensive semantic affinity value computed by $SA(c', c'')$ is defined as follows:

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

where W_{LA} is a weight expressing the relevance assigned to the linguistic affinity in the semantic affinity evaluation process. A threshold-based mechanism is enforced

to set the minimum level of semantic affinity required to consider two concepts as matching concepts.

4.2 Example

To join the Esteem network, the doctor needs to specify a list of keywords representing her interests, and/or other possible knowledge equipments (i.e., peer ontology, service ontology, current context, and data quality & trust profile) when available. The following list of keywords are specified to describe the doctor interests in the considered running example: medicines, contagious diseases, and medical structures. The doctor enters the global overlay of the Esteem network where it is denoted as peer P . By receiving the community advertisement and by performing random walks, the peer P becomes aware of a number of existing communities. In particular, the community manifestos in the example of Fig. 5 are discovered by the peer P . The doctor interests are exploited by the Community Membership module to evaluate the possible participation of the peer P in these communities and the Ontology Matching module is invoked to this end. By relying on HMatch 2.0, the community manifestos of Fig. 5 are matched against the doctor interests and a list of semantic affinity results SA is produced as follows:

$SA(\text{Contagious Disease, Disease}) = 0.94$ $SA(\text{Medicine, Drug}) = 0.74$ $SA(\text{Medical Structure, Hospital}) = 0.74$ $SA(\text{Medical Structure, Health-Care Structure}) = 0.60$ $SA(\text{Medicine, Surgical Technique}) = 0.13$
--

In the standard configuration, the Esteem demonstrator is set to automatically join a community sc when at least one matching concept is found between the doctor interests and the manifesto of sc (a threshold $t = 0.5$ is set in HMatch 2.0). For this reason, the community sc_4 is not joined and its manifesto is discarded. On the opposite, the communities sc_1 , sc_2 , and sc_3 match the doctor interests and they are joined. The list of joined communities can be visualized in the demonstrator by the doctor, in order to eventually change the choice of the communities to join.

5 Community-based data discovery

Data discovery in Esteem allows a node to detect which communities/peers are capable of providing relevant matching knowledge with respect to a given target request. Two steps are defined as follows.

Query formulation and routing Discovery in Esteem starts when a requesting peer p_i formulates a probe query q to identify potential partners for data/service sharing. As shown in Fig. 8, the user formulates the query q by relying on the Query Manager with (optional) involvement of the Context Manager in case of context-oriented queries. The Query Routing module is then invoked to execute the routing-by-community mechanism of Esteem and to select the query recipients by considering the joined semantic communities stored in the JSC list. Semantic communities are used as “virtual query recipients” on the basis of their associated community manifesto. This way, queries are propagated throughout the network

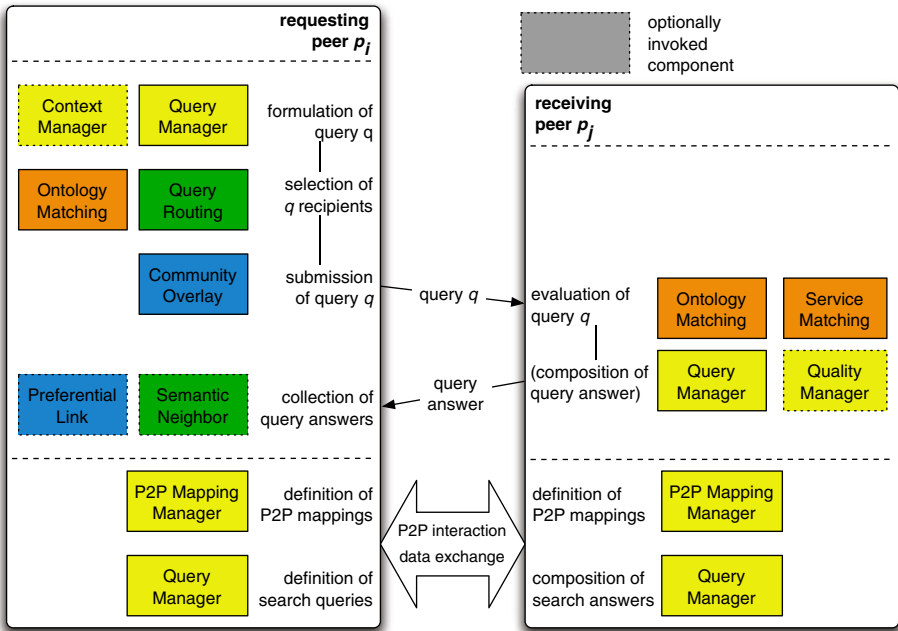


Fig. 8 Community-based data discovery

according to shuffling and peer members of a community sc can filter the incoming queries by considering (i.e., processing) only requests having sc (and the other joined communities) as recipient. The following steps constitute the routing-by-community mechanism.

1. *Selection of candidate communities.* It is performed by matching q against the community manifestos in the JSC list. A community sc_i is considered as a candidate recipient for the query q when at least one matching concept $c \in M_i$ is found for q by HMatch 2.0.
2. *Selection of recipient communities.* Candidate communities are ranked according to the semantic affinity values produced by HMatch 2.0. Candidate communities with the highest ranking are then selected as recipients of the query q . In this respect, different strategies can be adopted by the peer to govern the scale of recipients. For example, top- k communities in the ranking can be selected. This way, by setting the parameter k , the peer can specify the exact number of query recipients to contact. Alternatively, a threshold-based mechanism can be enforced. In this case, through the threshold, the peer specifies the minimum level of semantic affinity of a recipient community, thus obtaining a (potentially) larger set of recipients than in the previous case.

The Community Overlay module is then invoked to propagate the query q throughout the overlays of the semantic communities selected as recipients by the routing-by-community mechanism.

Answer composition and management Receiving a query q , a peer p_i invokes its Ontology Matching module to compare q against its peer ontology and to identify possible semantic affinities. A (possibly empty) ranked list of matching concepts (i.e., concepts semantically related to the target request q) is then returned. If a non-empty result is produced, a query answer is composed by the Query Manager and returned back to the requesting peer p_i . In answer composition, the Quality Manager can be optionally invoked to attach the list of discovered matching concepts with associated trust and quality metadata. Collecting request answers, the peer p_i exploits the obtained results for deciding the subsequent actions. On one side, the matching values provided by an answering peer p_j are used by peer p_i to set a confidence value featuring p_j as a *semantic neighbor* on the q topics. To this end, those answering peers that provided high matching results for the request q (i.e., matching results over a predefined threshold) are passed to the Semantic Neighbor module for being stored. Furthermore, the Preferential Link module is used to establish a direct connection to the semantic neighbors with the highest confidence values in order to foster the interactions with the (potentially) most interesting peers. On the other side, the peer p_i has to decide whether to further “point-to-point interact” with one or more of the answering peers (i.e., semantic neighbors) for sharing. The choice of the peer p_j highly depends on the collaboration goal that is pursued and on the quality of the obtained answers. The received matching results are exploited for assessing the relevance of the discovered data sources. In this respect, interaction with the user is supported for semi-automatic selection of the most appropriate semantic neighbor to interact with. When the collaboration partners are chosen, the sharing phase can take place. According to Fig. 8, the P2P Mapping Manager and the Query Manager are invoked for defining appropriate P2P mappings and for subsequently specifying search queries of interest, respectively. The solution adopted in Esteem is similar to the approach proposed in Halevy et al. (2004), and it is characterized by the use of P2P mappings for allowing the requesting peer p_i to access the resources (i.e., data and services) provided by another peer (p_j) discovered during probing.

5.1 Esteem contribution

The following contributions characterize community-based data discovery in Esteem.

Context use and management In Esteem, the use of context supports data discovery by enabling a peer to specify more focused probe queries, thus allowing to collect a more precise set of matching results. In particular, the Context Manager module supports the user in defining the mappings between each context represented in the CDT and portions of the peer ontology of the user, in turn corresponding to chunks of relevant data. This phase is semi-automatic and the Ontology Matching module is invoked to suggest to the user the possible correspondences between elements of the CDT and sets of concepts and roles of the peer ontology. Two possible uses of the context are available in Esteem.

- *Context-based query formulation.* The Context Manager module is invoked during query formulation to attach the current context of a peer to the probe

query. A receiving peer invokes the Context Matching module to evaluate the level of match between its context and the context of the requesting peer. For a peer receiving the probe query, context matching is an additional requirement to satisfy besides the semantic affinity evaluation of the probe query previously described, which allows to obtain a more precise and reliable set of results.

- *Context-based preferential link establishment.* The Context Manager module can be invoked by a peer to discover other peers using a similar context and to establish a preferential link with them for subsequent uses in the sharing phase. To this end, the user can trigger the submission to the network of a specific kind of probe query only containing the current context of the peer. A receiving peer invokes the Context Matching module to evaluate the level of match between the context of the requesting peer and its own current context. A reply is sent to the requesting peer when a matching context is discovered. A preferential link with the peers storing the most similar contexts is established by the requesting peer in a semi-automatic way according to the user's preferences.

Quality-aware data integration The metadata provided in the data quality and trust profile of an Esteem peer are used to support community-based data discovery as described in the following.

- *Trust-aware query answer.* A peer replying to a probe query can attach the quality metadata to its answer. This implies that an object identification step is performed by a replying peer to choose those attributes called *matching keys* that have to be used to solve potential key-level conflicts. In Esteem, an additional quality metadata, called *identification power*, is defined to specify how much a given attribute is discriminating when trying to match objects. The identification power of an attribute is calculated in Esteem in a fully automatic way through the Quality Manager module (Bertolazzi et al. 2003). The attributes with higher identification power are then chosen as matching keys and they are used to solve key-level conflicts during query processing.
- *Peer trust evaluation.* The quality metadata of an Esteem peer are used to calculate a comprehensive trust measure of the node. In Esteem, we adopt the generally-accepted approach to trust a peer as a whole, with respect to the totality of the exchanged data, and/or to the number of transactions performed with the other peers (e.g., Aberer and Despotovic 2001). However, the Esteem approach is characterized by two main extensions. First, we consider only a specific kind of transactions, namely, the data exchange transactions. Second, we measure the peer trust with respect to a specific kind of data. The atomic unit of trust considered in Esteem is the couple $\langle p_i, \mathcal{D} \rangle$, where \mathcal{D} is an element of the peer ontology of p_i . The trust of a peer p_i is computed at the level of semantic communities in a fully automatic way on the basis of the number of complaints fired by other community members, for which p_i worked as a data provider (De Santis et al. 2003). In the data discovery phase, a peer can exploit the trust value of another peer p_i in order to assess the reliability of a query answer provided by p_i in reply to a certain probe query.

5.2 Example

We consider the scenario where the doctor (i.e., the peer P) is interested in discovering network nodes that are capable of providing relevant knowledge about malaria and possible drugs for treating this disease. The Esteem demonstrator is used by the doctor to input the search keywords that will be used for probing. By invoking the Query Manager module, a probe query q_1 is formulated containing the user-defined keywords Medicine, Vaccine, Disease, Treats, and Malaria. The routing-by-community mechanism is exploited to select the semantic communities to choose as q_1 recipients and the query q_1 is matched against the manifesto of the joined communities sc_1 , sc_2 , and sc_3 . Through HMatch 2.0, the following semantic affinity results are calculated.

HMatch(q_1, M_{sc_1})	HMatch(q_1, M_{sc_2})
SA(Disease, Disease) = 0.76	SA(Disease, Tropical disease) = 0.65
SA(Medicine, Drug) = 0.64	SA(Disease, Disease) = 0.6
SA(treats, treats) = 0.50	SA(Disease, Infectious Disease) = 0.54


According to these results of HMatch 2.0, the communities sc_1 and sc_2 are selected as q_1 recipients since their manifestos contain similarities with the query topics, while the community sc_3 is not selected since no semantic affinities are detected. The selected communities are then passed to the Community Overlay module for subsequent submission to the network on the corresponding overlays (i.e., the overlays of sc_1 and sc_2). Each peer in the selected communities receives the request through shuffling and it invokes the Ontology Matching module for comparing the query q_1 against its peer ontology. In this example, we consider the peer ontologies of peer $D \in sc_1$ and peer $F \in sc_2$ that are shown in Fig. 2a and b, respectively. For what concern the peer D and peer F, by using HMatch 2.0, the matching results obtained by comparing the query q_1 with their respective peer ontologies are the following.

HMatch(q_1, O_{peerD})	HMatch(q_1, O_{peerF})
SA(Malaria, Malaria) = 1.0	SA(Malaria, Malaria) = 0.8
SA(Disease, Infectious Disease) = 0.94	SA(Medicine, Antimalarial Drug) = 0.74
SA(Medicine, Drug) = 0.74	
SA(Medicine, Doxycycline) = 0.54	
SA(Medicine, Quinine) = 0.54	

As a consequence, both peer D and peer F reply to the requesting peer P and the results are visualized in Fig. 9. Collecting the replies, the peer P invokes its Semantic Neighbor module to store both the replying peers as semantic neighbors and to set a preferential link with them by relying on the Preferential Link module. Moreover, the peer P can decide to further interact with one of the replying peers for data acquisition. By exploiting the received replies, the doctor recognizes that the peer ontology of the peer D provides knowledge about drugs to treat malaria, while the peer ontology of the peer F is concerned with a symptomatic description of malaria. Since the doctor is interested in retrieving data about drugs for malaria, the peer D is selected as a partner for the subsequent sharing phase and for data acquisition.

ESTEEM

Home Registration Community search Data search Service search Context search Help



Emergent Semantics and
cooperation in multi-knowledge
Environments

Advanced methods and tools for
semantic cooperation in Web virtual
communities

Search results

The following results have been retrieved:

Malaria Infectious Disease Drug Doxycycline Quinine	from peer D	Click here to acquire data
Malaria Antimalarial Drug	from peer F	Click here to acquire data

Please, specify the keywords for your search:

Filter by quality
 Filter by context

Fig. 9 Esteem demonstrator: probe query results

6 Community-based service discovery

The members of a community can share functional facilities by means of services that are made available and invocable via their interfaces. The two steps of query formulation and routing and of answer composition and management previously introduced support the discovery of both data and services. However, due to the peculiarities of service conceptualization, specific techniques for service discovery are enforced in Esteem. In particular, service discovery is characterized by the following aspects.

- *Service descriptions represent functionalities provided on the network.* A service request and a service advertisement are described in terms of: (i) a concept representing the required service functionality (e.g., drug ordering, product delivery, remote diagnosis, laboratory testing), (ii) a set of concepts representing the desired results (outputs), (iii) a set of concepts representing the data that must be provided for service execution (inputs).
- *Different matching information are considered.* Matching between two service descriptions can be differentiated by taking into account: (a) the *kind* of match (total, partial, mismatch), to establish if a service advertisement fully or partially satisfies the requested functionalities or does not satisfy the request at all, (b) the *degree* of match, to establish how much the service advertisements satisfies (even partially) the request.

The Service Matching module is in charge of performing service comparison for discovery purposes. The adopted service matchmaking approach, namely FC-MATCH (FunctionalCompatibility-Match), is presented in Bianchini et al. (2008), where also a comparative analysis with other existing approaches in literature is given. In FC-MATCH, service matchmaking is performed on the basis of concept definitions in the Service Functionality and Message Ontologies, taking into account the description of services in terms of operations, inputs and outputs. In particular, two different matching models are combined. First, a deductive model is used to qualify the kind of match $\text{MatchType}(\mathcal{R}, \mathcal{S})$ between the descriptions of a service request \mathcal{R} and a service advertisement \mathcal{S} . The deductive model checks if: (i) for each operation in the request, there is an equivalent or more generic operation in the advertisement, according to the SFO; (ii) for each output in the request, there is an equivalent or more generic output in the advertisement, according to the SMO; (iii) for each operation in the request, there is an equivalent or more generic operation in the advertisement, according to the SFO. Inputs are used in a dual way w.r.t. outputs, looking for an equivalent or more generic input the requester is able to provide for each input found in the advertisement. This distinction between inputs and outputs is derived from the definition of plug-in between software components (Zaremski and Wing 1995). According to the deductive model, it is possible to state if \mathcal{S} exactly provides the required functionalities (\mathcal{S} EXACT \mathcal{R}), if \mathcal{S} provides additional functionalities with respect to the required ones (\mathcal{S} EXTENDS \mathcal{R}), if there is a non empty intersection between functionalities provided by \mathcal{R} and \mathcal{S} , but not all the required functionalities are provided (\mathcal{S} INTERSECTS \mathcal{R}), or if \mathcal{R} and \mathcal{S} have nothing in common (\mathcal{S} MISMATCH \mathcal{R}). In case of partial match (\mathcal{S} INTERSECTS \mathcal{R}), a similarity-based matchmaking model is used to quantify the degree of match $G\text{Sim}(\mathcal{R}, \mathcal{S})$ between service descriptions through coefficients properly defined to compare input and output names (Entity-based service similarity $E\text{Sim}(\mathcal{R}, \mathcal{S})$) and between functionality names (Functionality-based service similarity $F\text{Sim}(\mathcal{R}, \mathcal{S})$). These coefficients are based on SFO and SMO and also rely on a thesaurus extracted from WordNet likewise the computation of semantic affinity value between ontological concepts explained in Section 4. The $E\text{Sim}$ and $F\text{Sim}$ coefficients are linearly combined to obtain a comprehensive measure of the degree of match. EXACT and EXTENDS match correspond to the case $G\text{Sim}(\mathcal{R}, \mathcal{S}) = 1.0$ (total match), while MISMATCH corresponds to the case $G\text{Sim}(\mathcal{R}, \mathcal{S}) = 0.0$. In the literature, techniques for service matchmaking have been proposed by separately considering deductive approaches (Horrocks and Li 2004; Kawamura et al. 2002) and similarity-based ones (Dong et al. 2004). In Esteem the FC-MATCH approach has been used in combination with the routing-by-community mechanism to implement different forwarding policies, as explained in the following.

6.1 Esteem contribution

When a service request is formulated on a peer p_i by relying on the Query Manager, a Service Semantic Affinity value SSA is evaluated to select the communities to be considered as candidate request recipients. This value is defined as $SSA = \alpha \cdot SA_1 + (1 - \alpha) \cdot SA_2$, where SA_1 represents the semantic affinity value obtained by comparing the requested service functionality against the concepts of the Service Functionality Ontology of the community manifesto; SA_2 represents the

semantic affinity value obtained by comparing the requested inputs/outputs against the concepts in the Service Message Ontology of the community manifesto; α is a weight expressing the relevance assigned to the previous semantic affinity values. SA_1 and SA_2 are computed by applying the ontology matching techniques that will be described in Section 4. According to the SSA value, query formulation and propagation are performed towards other peers p_j . The FC-MATCH algorithm is applied on peer p_j to obtain a list $MS(\mathcal{R}) = \{\langle S1_X, GSim_1, mt_1 \rangle, \dots, \langle Sn_X, GSim_n, mt_n \rangle\}$ of matching service descriptions such that $GSim_i = GSim(\mathcal{R}, Si_X) \geq \delta$ and $mt_i = MatchType_e(\mathcal{R}, Si_X) \in \{EXACT, EXTENDS, INTERSECTS\}$, where δ is the similarity threshold set up experimentally to filter out non relevant results. Peer p_i can designate an answering peer p_j as its *semantic neighbor* and can decide whether to interact with it according to trust and quality of the obtained results. The degree of match is used as the confidence value of the semantic neighbor and the kind of match is used to state if the semantic neighbor is able to provide additional functionalities with respect to those already provided by local services on p_i . This information will be used to serve future service requests, according to different forwarding policies. We distinguish between two policies: an *exhaustive* and a *minimal* policy.

According to the minimal policy, service discovery stops when matching services which fully satisfy the service request have been found. Formally, if $\exists Si_X \in MS(\mathcal{R})$ such that $Si_X \text{ EXACT} \mid \text{EXTENDS } \mathcal{R}$, it is not necessary to forward the service request to the semantic neighbors of the current peer, since Si_X already satisfies the request. Otherwise, the list of semantic neighbors is investigated to find remote services that potentially offer additional functionalities with respect to local services. A semantic neighbor sn is considered for service request forwarding if the kind of match labeling the semantic neighbor is EXTENDS or INTERSECTION. In fact, according to the meaning of these kinds of match, this means that sn provides services with additional functionalities with respect to those locally found.

According to the exhaustive policy, service discovery does not stop when matching services that fully satisfy the request have been found, in order to find other services that could present, for example, better non functional features, such as QoS, contextual information, service peer reliability. A Time-To-Live mechanism (TTL) is implemented to avoid cycles and unbounded request propagation on the network (Bianchini et al. 2009).

According to this service discovery model, a peer raising a service request: (i) identifies the communities acting as request recipients, through the application of routing-by-community mechanism; (ii) collects answers and sets or updates information about its semantic neighbors. On the other hand, a peer receiving a request: (i) applies FC-MATCH between the request and locally available services; (ii) applies forwarding policies to send the request towards semantic neighbors. If no semantic neighbors can be identified according to the forwarding policies (e.g., no local matching services have been found or no semantic neighbors have been set yet), a peer randomly selects a subset of peers in the communities acting as request recipients. In Esteem, the choice of FC-MATCH is motivated by the fact that the combined use of the kind and the degree of match between services is useful both to reduce the complexity of matching and to implement the forwarding policies that aim at keeping low the network overload (e.g., a peer forwards only to semantic neighbors featured by high degree of match and whose services provide

additional operations with respect to the local services, according to the kind of match).

6.2 Example

We consider that the peer P needs to find nodes on the network providing services for on-line drug ordering and purchasing. By invoking the Query Manager, a service query S_q^1 is formulated as follows:

OPERATION: buyDrug
 INPUTS: Kinin, Quantity
 OUTPUTS: ShippingTime, Price

Firstly, peer P is interested in identifying which communities are able to satisfy this service query. The peer P applies the routing-by-community mechanism and compare the service query S_q^1 against the manifesto of each joined semantic community sc_1 , sc_2 , and sc_3 by invoking the Ontology Matching module. According to the obtained matching results, only the community sc_1 is selected as S_q^1 recipient. Peer P sends the query S_q^1 to peers B and peer G that are known to belong to the semantic community sc_1 . On peers B and peer G the service query is matched against the local service descriptions as shown in Fig. 10. Similarly, peer G forwards the query to peer C since it knows that peer C is in the list of peers belonging to sc_1 and obtained results are collected and sent back to peer P, where they are displayed to the doctor through the Esteem demonstrator (see Fig. 11). Furthermore, as shown

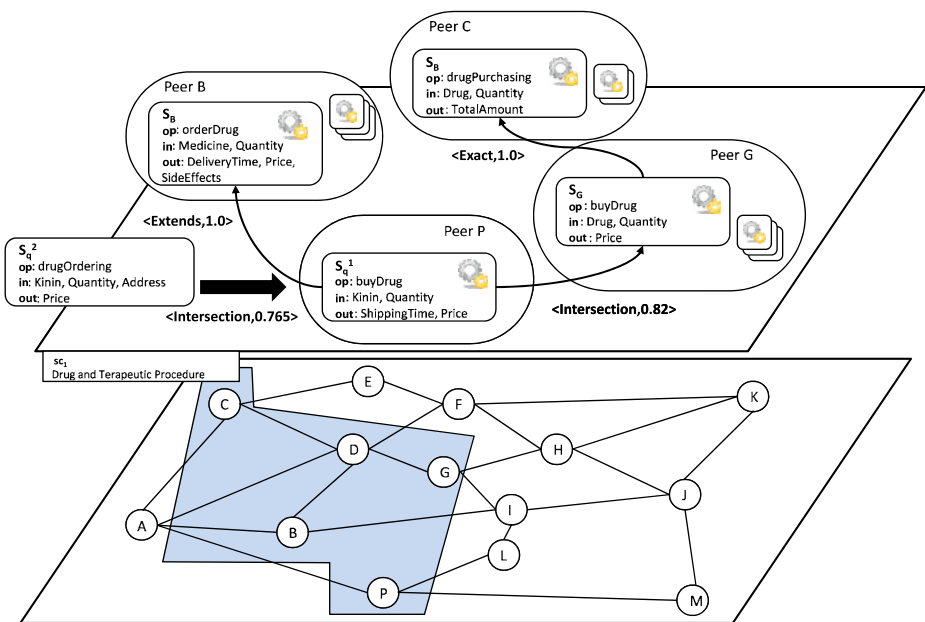


Fig. 10 Community-based service discovery

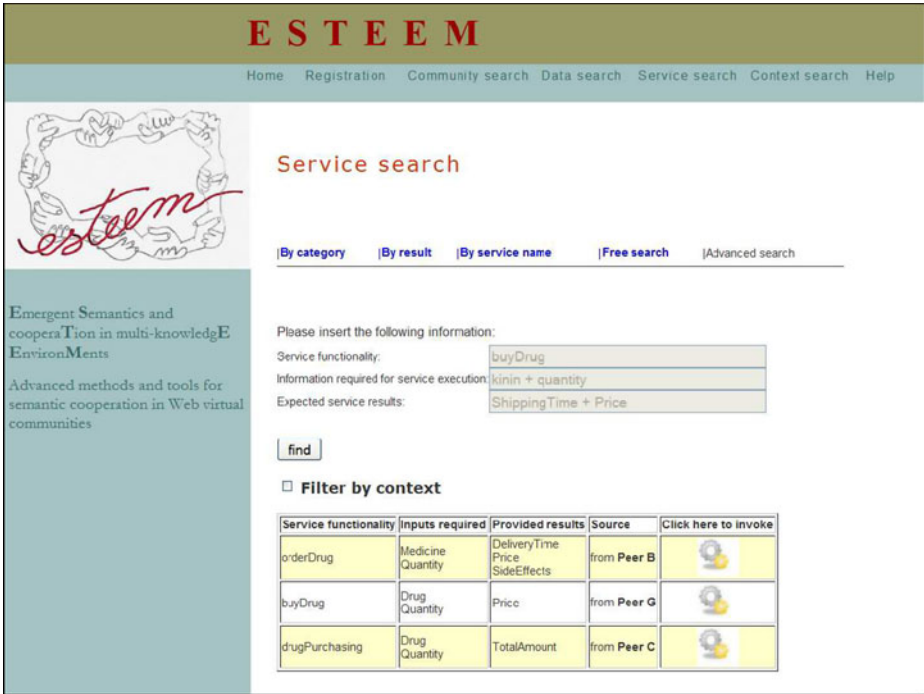


Fig. 11 Esteem demonstrator: service request results

in Fig. 10, peer P sets preferential links towards peer B and peer G on the basis of received results, labeling them with matching information. In the same way, peer G sets a preferential link towards the peer C.

Let suppose now that peer P formulates another service request S_q^2 to find a service for drug ordering and shipping:

OPERATION: drugOrdering
 INPUTS: Kinin, Quantity, Address
 OUTPUTS: Price

The Peer P matches S_q^2 against the previous query S_q^1 and finds that $match(S_q^2, S_q^1) = INTERSECTS$ and $GSim(S_q^2, S_q^1) = 0.85$. At this point peer P can exploits the preferential links previously set to speed up service discovery over the network. Peer P forwards S_q^2 to semantic neighbors (B and G) with respect to S_q^1 . Service query S_q^2 is matched against the service descriptions provided by the peer B and the peer G and matching results are sent back to the peer P. According to the minimal forwarding policy, since on peer G the query S_q^2 is only partially satisfied by the service description provided locally (S_G), but semantic neighbors of peer G (in this example, peer C) do not add further capabilities with respect to those already provided on peer G (EXACT match), then the query S_q^2 is not further propagated. The results are visualized as in Fig. 11.

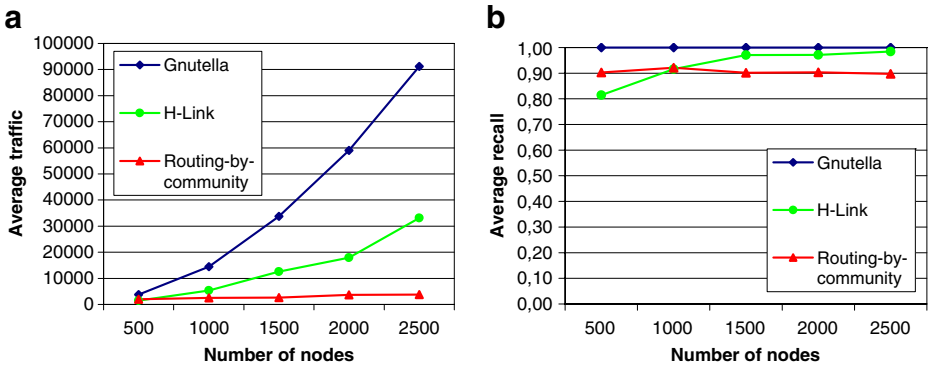


Fig. 12 Simulation results with variation of the network size: **a** Traffic **b** Recall

7 Experimental results

The evaluation of the Esteem platform has been performed at both system and user-interface level.

System evaluation and comparison with existing approaches A dedicated evaluation session has been specifically executed for each component of the Esteem architecture. These tests were devoted to measure the performance of Esteem for what concerns community formation and routing-by-community. In particular, the routing-by-community mechanism has been compared with the Gnutella protocol (The Gnutella Protocol 2001) and with the H-Link routing mechanism (Castano and Montanelli 2007). The choice of Gnutella is due to the fact that this routing protocol is well-known and it is frequently considered as a reference example. The choice of H-Link is motivated by the need of a comparison with a content-based routing approach enforcing the use of single-peer recipients. Moreover, both Gnutella and H-Link are based on network overlays posed on top of an unstructured P2P infrastructure, as for Esteem. In the evaluation, the Neurogrid P2P simulator⁷ has been employed. The evaluation is expressed in terms of *generated traffic* and *recall*. By generated traffic, we mean the overall number of messages routed during a complete simulation run. According to the classical definition of Information Retrieval, recall is defined as the ratio of the number of relevant concepts retrieved by an Esteem query to the total number of relevant concepts available in the network. Detailed experimental results are provided in Catarci et al. (2007b). In the following, we report a selection of the most interesting results.

The first test is devoted to measure scalability of Esteem and results are shown in Fig. 12. In this test, the number of network nodes *#nodes* gradually grows in the range [*#nodes* = 500, *#nodes* = 2,500]. The number of connections per node are proportionally increased during the simulation, while the number of queries per simulation is fixed and it corresponds to *#queries* = 1,000. This test highlights the excellent behavior of the routing-by-community mechanism in terms of scalability.

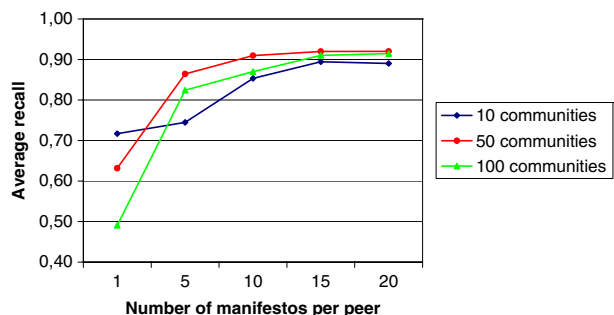
⁷<http://www.neurogrid.net/>

The test shows that the traffic performance of both Gnutella and H-Link are affected by the growing size of the network, while the traffic generated by routing-by-community is stable due to the efficiency of shuffling-based communications within semantic communities. Moreover, we want to stress that the better performance of routing-by-community when compared with H-Link are also motivated by the simulation of peer volatility during the experiment. To simulate the event of peer disconnection that frequently occurs in real P2P sharing networks, a random subset of the network peers was hidden at a certain step of the test. The disconnection event of a peer negatively impacts on the performance of single-recipient routing mechanisms like H-Link since new paths have to be discovered to replace those previously learned and no more active. On the opposite, routing-by-community is not affected by peer volatility since an Esteem community can be viewed as a *virtual peer* that is persistently active in the system apart from the availability of the single community members. Concerning recall, we observe that although both Gnutella and H-Link outperform routing-by-community, the results of the Esteem approach are very interesting on the whole. Indeed, if we consider both generated traffic and recall, routing-by-community enforces a better trade-off than Gnutella and H-Link, since it provides a stable recall value around 90% while requiring a low network traffic, at the same time.

The second test is devoted to assess how the number of communities joined by a peer impacts on the recall measures of routing-by-community. Results are shown in Fig. 13. The test has been performed on a network with $\#nodes = 500$ and by varying i) the number of communities that each peer can join (i.e., $\#manifestos$), and ii) the total number of communities available (i.e., created) in the network (i.e., $\#communities$). The experiment shows that by increasing the number of communities that can be joined by a peer, the recall value also increases as a result. This is due to the fact that the higher is the number of possible communities to join, the higher is the likelihood of a peer to get reached by a query. However, the experimental results also indicate that with $\#manifestos \geq 15$, the impact on the recall measure becomes stable. Experimental results indicate that the $\#manifestos$ parameter should be set to 20% of the $\#communities$ value for enforcing high recall measures.

User-interface evaluation We believe that the relevance/success of the Esteem platform resides both in its functionalities and in its usability. In fact, since Esteem is conceived for a wide range of users, most of which are non ontology-expert users, the usability aspects must be considered as an essential requirement of the

Fig. 13 Recall simulation results with variation in the number of existing communities



system. The design phase of the demonstrator has been based on the results of an initial interview posed to a sample personnel employed in different Italian medical structures. Subsequently, the Esteem demonstrator has been validated by a selected group of doctors belonging to different medical fields and with different expertise. These users have been recorded during their usage of the demonstrator and data have been gathered with a think aloud method (Dix et al. 2003). After the experience, a satisfaction questionnaire concerning both the research topics of Esteem and the usability issues of the demonstrator has been submitted to the doctors. The questionnaire results show that knowledge exchange between medical operators with similar expertise (i.e., community-based knowledge exchange) is very positively considered. The potentially interesting role of the Esteem project in improving the communication based on Semantic Web technologies has been also stressed. Details about the user-interface evaluation are reported in Catarci et al. (2007a).

8 Concluding remarks

In this paper, we have presented the Esteem platform for P2P semantic collaboration based on collective knowledge emerging from peer semantic communities. Esteem is the result of a national research project funded by the Italian Ministry of Education, University, and Research. A key feature of Esteem is the preservation of the autonomous and spontaneous nature of peer community formation while offering an integrated approach to data and service discovery/sharing at the same time. Original contributions of Esteem can be summarized as follows: i) the combination of a shuffling-based communication mechanism with ontology matchmaking techniques to enforce the construction of a semantic overlay network, ii) the definition of a probe query approach for both data and service discovery in P2P systems, and iii) the capability of incorporating information about the user's context and the trust & quality of data into the query answer and thus into the knowledge sharing process.

The positive feedback of the first Esteem evaluation, both in terms of satisfaction and usability of the designed functionalities, encourages to continue working on community-based P2P semantic collaboration. By relying on the research achievements of Esteem, new research directions are raising. In particular, we plan to work on *on-the-fly* data/service integration techniques specifically conceived for highly dynamic scenarios where quality and context are essential system requirements.

Acknowledgements The Esteem platform has been developed within a PRIN Project funded by the Italian Ministry of Education, University and Research. Authors wish to thank anonymous referees for their insightful comments that led us to an improved presentation of the paper. A special acknowledgement is due to Carlo A. Curino, Diego Milano, Giorgio Orsi, Antonella Poggi, Leonardo Querzoni, Denise Salvi, Sara Tucci for their work in the Esteem project.

References

- Aberer, K., & Despotovic, Z. (2001). Managing trust in a peer-2-peer information system. In *Proc. of the 10th int. conference on information and knowledge management* (pp. 310–317). Atlanta, Georgia, USA.
- Aberer, K., et al. (2004). Emergent semantics principles and issues. In *Proc. of the 9th int. DASFAA conference* (pp. 25–38). Jeju Island, Korea.

- Allavena, A., Demers, A., & Hopcroft, J. E. (2005). Correctness of a Gossip based membership protocol. In *Proc. of the 24th ACM PODC* (pp. 292–301). Las Vegas, USA.
- Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P. (Eds.) (2003). *The description logic handbook: Theory, implementation, and applications*. Cambridge University Press Cambridge, Cambridge, UK
- Batini, C., & Scannapieco, M. (Eds.) (2006). *Data quality: Concepts, methodologies, and techniques*. Springer Verlag, Berlin
- Bertolazzi, P., De Santis, L., & Scannapieco, M. (2003). Automatic record matching in cooperative information systems. In *Proc. of the ICDT Int. DQCIS workshop* (pp. 13–20). Siena, Italy.
- Bianchini, D., De Antonellis, V., & Melchiori, M. (2008). Flexible semantic-based service matchmaking and discovery. *World Wide Web Journal*, 11(2), 227–251.
- Bianchini, D., De Antonellis, V., & Melchiori, M. (2009). P2P-SDSD: On-the-fly service-based collaboration in distributed systems. *Int Journal of Metadata, Semantics and Ontologies*, 5(3), 222–237.
- Bolchini, C., Curino, C. A., Quintarelli, E., Schreiber, F. A., & Tanca, L. (2007a). A data-oriented survey of context models. *SIGMOD Record*, 36(4), 19–26.
- Bolchini, C., Schreiber, F. A., & Tanca, L. (2007b). A methodology for very small database design. *Information Systems*, 32(1), 61–82.
- Bolchini, C., Curino, C., Quintarelli, E., Schreiber, F. A., & Tanca, L. (2009) Context information for knowledge reshaping. *International Journal on Web Engineering and Technology*, 5(1), 88–103.
- Castano, S., & Montanelli, S. (2007). Semantically routing queries in peer-based systems: The H-Link approach. *The Knowledge Engineering Review*, 23(1), 1–22.
- Castano, S., Ferrara, A., & Messa, G. (2006a). ISLab HMatch results for OAEI 2006. In *Proc. of the ISWC int. OM workshop*. Athens, Georgia, USA.
- Castano, S., Ferrara, A., & Montanelli, S. (2006b). Matching ontologies in open networked systems: Techniques and applications. *Journal on Data Semantics (JoDS)*, 5, 25–63.
- Catarci, T., et al. (2007a). *Integrated mock-up prototype and experimental results on the ESTEEM application scenario*. Deliverable DALL4, MIUR Esteem Project.
- Catarci, T., et al. (2007b). *Mock-up prototypes and data collected during the testing phase*. Deliverable DALL3, MIUR Esteem Project.
- Chalmers, M. (2004). A historical view of context. *Computer Supported Cooperative Work*, 13(3), 223–247.
- Chen, H., Finin, T., & Joshi, A. (2003). An intelligent broker for context-aware systems. In *Proc. of the 5th int. Ubicomp conference* (pp. 183–184). Seattle, Washington, USA.
- Curino, C., Orsi, G., & Tanca, L. (2009). Accessing and documenting relational databases through OWL ontologies. In *Proc. of the 8th int. conference on flexible query answering systems (FQAS)* (pp. 431–442). Roskilde, Denmark.
- De Santis, L., Scannapieco, M., & Catarci, T. (2003). Trusting data quality in cooperative information systems. In *Proc. of the 11th int. CoopIS conference* (pp. 354–369). Catania, Italy.
- Dix, A., Finlay, J. E., Abowd, G. D., & Beale, R. (Eds.) (2003). *Human-computer interaction* (3rd ed.). Prentice-Hall, Englewood Cliffs, NJ
- Dong, X., Halevy, A. Y., Madhavan, J., Nemes, E., & Zhang, J. (2004). Similarity search for web services. In *Proc. of the 30th int. VLDB conference* (pp. 372–383). Toronto, Canada.
- Fagin, R., Kolaitis, P. G., Miller, R. J., & Popa, L. (2005). Data exchange: Semantics and query answering. *Theoretical Computer Science*, 336(1), 89–124.
- Fan, W., Lu, H., Madnick, S., & Cheung, D. (2001). Discovering and reconciling value conflicts for numerical data integration. *Information Systems*, 26(8), 635–656.
- Gomez-Perez, A., Fernandez-Lopez, M., & Corcho, O. (2003). *Ontological engineering*. Springer Verlag, Berlin
- Gruber, T. (2008). Collective knowledge systems: Where the social web meets the semantic web. *Journal of Web Semantics*, 6(1), 4–13.
- Haase, P., Siebes, R., & van Harmelen, F. (2008). Expertise-based peer selection in peer-to-peer networks. *Knowledge and Information Systems*, 15(1), 75–107.
- Halevy, A., Ives, Z., Madhavan, J., Mork, P., Suciu, D., & Tatarinov, I. (2004) The Piazza peer data management system. *IEEE Transactions on Knowledge and Data Engineering*, 16(7), 787–798.
- Hidayanto, A. N., & Bressan, S. (2007). Towards a society of peers: Expert and interest groups in peer-to-peer systems. In *Proc. of the OTM international IFIP workshop on semantic web & web semantics (SWWS 2007)* (pp. 487–496). Vilamoura, Portugal.
- Horrocks, I., & Li, L. (2004). A software framework for matchmaking based on semantic web technology. *International Journal of Electronic Commerce*, 8(4), 39–60.

- Kawamura, T., Paolucci, M., Payne, T., & Sycara, K. (2002). Semantic matching of web services capabilities. In *Proc. of the 1st int. ISWC conference* (pp. 333–347).
- Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proc. of the 21st ACM SIGMOD-SIGACT-SIGART PODS* (pp. 233–246). Madison, Wisconsin, USA (invited tutorial).
- Löser, A., Staab, S., & Tempich, C. (2007). Semantic social overlay networks. *IEEE Journal on Selected Areas in Communication*, 25(1), 5–14.
- Mukherjee, C., & Ramakrishnan, I. (2008). Automated semantic analysis of schematic data. *World Wide Web Journal*, 11(4), 427–464.
- Ouksel, A. M. (2003). In-context peer-to-peer information filtering on the web. *SIGMOD Record* 32(3), 65–70.
- Rana, O. F., & Hinze, A. (2004). Trust and reputation in dynamic scientific communities. *IEEE Distributed Systems Online*, 5(1), 8.
- Sattler, K., Conrad, S., & Saake, G. (2003). Interactive example-driven integration and reconciliation for accessing database integration. *Information Systems*, 28(5), 393–414.
- Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., & Baldoni, R. (2004). The DaQuinCIS architecture: A platform for exchanging and improving data quality in cooperative information systems. *Information Systems*, 29(7), 551–582.
- Shvaiko, P., & Euzenat, J. (2005). A survey of schema-based matching approaches. *Journal on Data Semantics (JoDS)*, IV, 146–171.
- Specia, L., & Motta, E. (2007). Integrating folksonomies with the semantic web. In *Proc. of the 4th European semantic web conference (ESWC 2007)* (pp. 624–639). Innsbruck, Austria.
- The Gnutella Protocol (2001). *The Gnutella Protocol specification v0.4*. http://www9.limewire.com/developer/gnutella_protocol_0.4.pdf.
- Voulgaris, S., Gavidia, D., & van Steen, M. (2005) CYCLON: Inexpensive membership management for unstructured P2P overlays. *Journal of Network and Systems Management*, 13(2), 197–217.
- Wang, Y., & Vassileva, J. (2004). Trust-based community formation in peer-to-peer file sharing networks. In *Proc. of the IEEE/WIC/ACM int. conference on web intelligence (WI'04)* (pp. 341–348). Beijing, China.
- Xiong, L., & Liu, L. (2004). PeerTrust: Supporting reputation-based trust for peer-to-peer electronic communities. *IEEE Transactions on Knowledge and Data Engineering*, 16(7), 843–857.
- Zaremski, A., & Wing, J. (1995). Specification matching of software components. In *Proc. the 3rd int. ACM symposium on foundations of software engineering (SIGSOFT)* (pp. 6–17). Washington DC, USA.
- Zeinalipour-Yazti, D., Kalogeraki, V., & Gunopulos, D. (2005). Exploiting locality for scalable information retrieval in peer-to-peer networks. *Information Systems* 30(4), 277–298.