World Wide Web (2010) 13:3–31
DOI 10.1007/s11280-009-0080-6

Emergent Semantics and Cooperation in Multi-knowledge Communities: the ESTEEM Approach

Devis Bianchini · Stefano Montanelli · 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: 30 April 2009 / Revised: 4 August 2009 /
Accepted: 7 December 2009 / Published online: 24 December 2009

Abstract In the present global society, information has to be exchangeable in open and dynamic environments, where interacting users do not necessarily share a common understanding of the world at hand. This is particularly true in P2P scenarios, where millions of autonomous users (peers) need to cooperate by sharing their resources (such as data and services). We propose the Esteem approach (Emergent Semantics and cooperaTion in multi-knowledgE EnvironMents), where a comprehensive framework and a platform for data and service discovery in P2P systems are proposed, with advanced solutions for trust and quality-based data management, P2P infrastructure definition, query processing and dynamic service discovery in a context-aware scenario. In Esteem, semantic communities are built

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

D. Bianchini (✉) · V. De Antonellis · M. Melchiori
Dipartimento di Elettronica per l'Automazione, Università degli Studi di Brescia,
via Branze 38, 25123 Brescia, Italy
e-mail: bianchin@ing.unibs.it

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

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

S. Montanelli · S. Castano · A. Ferrara
Dipartimento di Informatica e Comunicazione, Università degli Studi di Milano,
via Comelico 39/41, 20135 Milan, Italy

S. Montanelli
e-mail: montanelli@dico.unimi.it

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

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

around declared interests in the form of manifesto ontologies and their autonomous nature is preserved by allowing a shared semantics to naturally emerge from the peer interactions. Inside the borders of semantic communities data and services are discovered, queried and invoked in a resource sharing scenario, where the context in which users interoperate and the trust of exchanged information are also relevant aspects to take into account.

Keywords emergent semantics · P2P data and service discovery · context-aware data and service discovery · trustworthy data and service discovery

1 Introduction

In the present global society, users and organizations communicate and share data and services on the network through applications that rely on decentralized structures, that handle a variety of heterogeneous information sources. Actually, the problem of providing transparent access to heterogeneous sources, while maintaining their autonomy, is not new and has been almost solved by information integration techniques, where interaction between clients and sources is performed through a centralized access point and uniform query interfaces give users the illusion of

C. Aiello · R. Baldoni · S. Bonomi · T. Catarci · M. Scannapieco
Dipartimento di Informatica e Sistemistica "A.Ruberti",
Università di Roma "La Sapienza", via Ariosto 25, 00185 Rome, 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

C. Bolchini · E. Quintarelli · F. A. Schreiber · L. Tanca
Dipartimento di Elettronica e Informazione, Politecnico di Milano,
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

querying a homogeneous system [22, 27]. However, these techniques work under certain hypotheses, including moderately static scenarios, shared understanding of the domain of interest (as a global schema or ontology), a closed, or at least access-controlled, set of participating sources. These hypotheses do not hold in the current evolving P2P paradigm, where millions of autonomous users (peers) need to cooperate by sharing their resources (such as data and services) without having a common understanding of the world. In this scenario, they dynamically build new information or knowledge, create new semantic communities and establish a new form of context-aware semantic interoperability, that we refer to as "emergent semantics", essentially based on dynamic trustful agreements on common interpretations within a given task [2].

Only few research efforts have faced the new requirements of emergent semantics, due to the high dynamism of the network, the lack of any agreed-upon global ontology, as well as the need of distributing the computation on the nodes when processing queries and composing services in a P2P environment. Semantic Web technologies allow for identifying relevant results in spite of terminological discrepancies (e.g., due to synonymies or homonymies) through the use of ontologies, that give a shared conceptualization of the resources at hand. The use of ontologies increases the efficacy of the discovery process with respect to traditional search engines, where users specify desired data and services through a set of keywords; search results are constituted by a set of documents, Web pages, Web services or portals which "contain" specified terms, but the user must refine his/her search by selecting only the relevant results. However, ontologies perform well when dealing with domains which are clearly identified, while P2P environments feature high semantic heterogeneity, due to the adoption of different (often deep-rooted) standards or terminologies.

Forcing interoperating peers to change their reference standards by adopting a common ontology is not always feasible. A significant example is constituted by healthcare resources, stored in a large amount of data sources that are accessed by medical personnel on the basis of their own reference terminologies. Existing systems allow for querying locally stored data and services, but the user should be able to interoperate in a P2P environment with other users with similar interests by using his/her own vocabulary, despite the presence of different underlying terminologies on the network. Furthermore, data and services in P2P systems can be accessed through different devices (with different hardware and software equipments) and in different contexts. For example, healthcare resources could be accessed via smart phones, workstations, palmtops or laptops by: (i) a doctor operating in a nursing home, who uses them to cure his/her patients and needs drugs or active principles that are available in commercial products; (ii) a laboratory technician, who uses this information to test new research approaches and needs pharmacological substances to perform his/her experiments; (iii) a student looking for up-to-date pharmacological information and services to order on-line books tailored to those specific topics. Not all search results are relevant for a given user, since each user operates in his/her own context and used devices often are not suitable for displaying all kinds of multimedia contents. In such circumstances, and in general in order to reduce the users' confusion in the presence of large, scarcely manageable amounts of knowledge (information and services), the peer's context should be taken into account. Finally, in a huge and dynamic environment, reputation and quality of data and services retrieved on the network are also relevant. This is particularly true for the healthcare domain, where quality and trustworthiness of privacy-sensitive information are very important.

The Esteem approach (Emergent Semantics and cooperaTion in multi-knowledgE EnvironMents) [33] proposes a comprehensive framework and platform for data and service discovery in P2P systems, with advanced solutions for trust and quality-based data management, P2P infrastructure definition, query processing and dynamic data and service discovery in a context-aware scenario. In particular, the system has been validated in the healthcare domain, which presents the features mentioned above. The system supports a doctor/specialist that is looking for healthcare resources (data and services) and joins the P2P network as a peer with his/her own interests, quality requirements and contextual characteristics. Common peers' interests identify *semantic communities*, which emerge in an autonomous way by collecting information sources whose contents present high similarity, enabling peer aggregation despite terminological differences (semantic communities do not constrain participants to adhere to a global ontology). Data and service discovery is performed inside the borders of such communities. Moreover, the Esteem system also supports context-aware data and service selection—excluding from the search results the resources that are not accessible in the current user's context—and is in charge of preventing the users from retrieving data and services in untrustworthy information sources.

In this paper we present the main capabilities of the Esteem system through a prototype that implements the proposed approach: in Section 2 we discuss about requirements that the Esteem system meets in a healthcare application scenario, where all the above-mentioned issues are relevant. We will see what is the knowledge equipment of a peer joining the Esteem network (Section 3), how to find, join or create semantic communities, where users share their interests (Section 4), how data and service discovery is performed within the semantic communities (Section 5), taking into account context-aware (Section 6) and trust aspects (Section 7). Section 8 presents system validation results, while Section 9 discusses its contributions with respect to the state of the art. Finally, in Section 10 we will give some concluding remarks.

2 System requirements

To illustrate the healthcare scenario we have in mind, we take the perspective of a doctor/specialist who uses the Esteem system to effectively and efficiently find desired data and services in a P2P network. The doctor is working in a small hospital in Central Africa and has patients with severe clinical conditions. They suffer from malaria, but they also have a strong adrenal insufficiency and are weakened by a chronic disease due to inadequate nutrition. Such a clinical condition could cause side effects to the standard malaria cure. Firstly, the doctor has to perform a Web search to find information about the best drugs to use and their last known side effects in the presence of concurring diseases. Given the particular emergency situation in which the doctor is operating, besides of looking for information on already known sites, he/she would prefer to formulate a generic request, that is spread over a network of healthcare institutes, laboratories or other specialists who agree on sharing their own data. After retrieving the right drug information, the doctor searches for a delivery service to order the required quantity of pharmacological substance. Laboratories, hospitals, nursing houses and drugstores on the network could provide services, intended as capabilities (e.g., drug ordering, product delivery, diagnosis services, lab test reading) to obtain a benefit or a satisfaction of the users' needs. The doctor has also the problem of understanding how trustworthy the information is, and he/she has found, to check if the provided results are up-to-date. Moreover, information and services are accessed and invoked through devices,[1] whose hardware and software resources might be scarcer than those of a laptop or a powerful workstation. According to the considered application scenario, the Esteem system supporting the doctor in his/her daily activity should satisfy the following requirements.

Community-based data and service discovery The size and dynamics of the P2P networks make data and service retrieval a difficult task; collaborative peers are aggregated according to their interests into semantic communities to efficiently guide users' requests through the network; community management is automatically and transparently performed due to the great number of peers leaving and joining the network at any moment.

User-oriented data and service discovery The system enables data and service discovery in a user-friendly way, hiding technical details for non-expert users and supporting them in all steps of the discovery process. In particular, the system supports the users with an intuitive Web interface that assists them in joining the semantic communities that share their interests and in identifying data and services they are looking for.

Semantic-driven data and service discovery Shared data and services are semantically described to enhance searching facilities on the network, yet allowing the peers to keep their own terminologies or standards. Ontologies provide semantics for both data and services as a means to bridge the gap between the different terminologies, but collaborating parties are not forced to adopt a common global ontology.

Context-aware data and service discovery Discovery facilities take into account in a transparent way the context from which a user accesses data or invokes services. Context identifies the user's current situation, actual values for spatial and temporal coordinates, his/her interests among those specified for the available communities and the preferential kinds of multimedia contents.

Trust-aware data and service discovery Not all the retrieved results are trustworthy or present the required quality level. A way to attach quality and trust metadata to data and with services has been studied and trust-based mechanism to filter out non-reliable data and services is applied transparently to the user.

[1] The medical personnel in the described situation could be endowed with personal computation means, like smartphones or palm computers, but also with obsolete devices which have low computational and storage capabilities.

3 The ESTEEM peer knowledge equipment

The ESTEEM system is characterized by the presence of a set of independent peers, that dynamically need to cooperate by sharing data and services without prior reciprocal knowledge or relationship. Such a collaboration scenario is *multi-knowledge*, in that no centralized authorities are defined to manage a comprehensive view of the resources shared by all the nodes in the system, due to the high dynamism and variability of collaboration and sharing requirements. For example, while there exist several well-known vocabularies, such as MeSH[2] (Medical Subject Headings) or SNOMED[3] (Systematized Nomenclature of Medicine), to describe healthcare information, that have been semantically represented by means of ontologies, healthcare parties are accustomed to adopt their own standards and terminologies and it is not realistic to constrain them to use different ones.

An ESTEEM peer provides its own ontology-based representation of the resources it intends to share with the other nodes of the system. In particular, an ESTEEM peer is equipped with: (i) a semantic description of shared data and services, expressed through ontologies, to properly identify its interests; (ii) the representation of (possible) context(s) from which the peer envisages to access data and invoke services (*context model*); (iii) the representation of quality and trust metadata attached to its data and services (*quality profile*). When joining semantic communities that share its interests, the peer also maintains information about joined communities, but this aspect will be better discussed in the next section. The peer ontology is also exploited for deriving the current peer interests and for determining the semantic communities to join. In Figure 1 a graphical representation of a portion of a peer ontology extracted from the Unified Medical Language System (UMLS[4]) is shown.

The peer ontology is characterized by a semantic network of concepts (e.g., `Sign or Symptom`) and semantic relationships between them (e.g., a sign or symptom is a `manifestationOf` a biological function, that in turn can be a pathologic or a physiologic function). Moreover, instances of medical concepts (e.g., `Abdominal Pain` and `Paresis` are instances of `Sign or Symptom`) are also extracted from available vocabularies and included in the peer ontology through a metathesaurus and are related to the concept names by means of terminological relationships (e.g., synonymy, hyperonymy), to bridge the gap between different ontologies on different peers and allowing users to use their own vocabularies in the open P2P environment.

Besides existing methodologies and editing tools for manual ontology engineering, tool-supported approaches can be adopted for creating a peer ontology. 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., see [16, 23, 28]). Domain knowledge already encoded in data schemas can thus be reused in the form of peer ontologies, sensibly reducing the required manual effort. 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., see [25, 31]). Furthermore, a reference peer ontology can

[2] http://www.nlm.nih.gov/mesh/meshhome.html

[3] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

[4] http://www.nlm.nih.gov/research/umls/

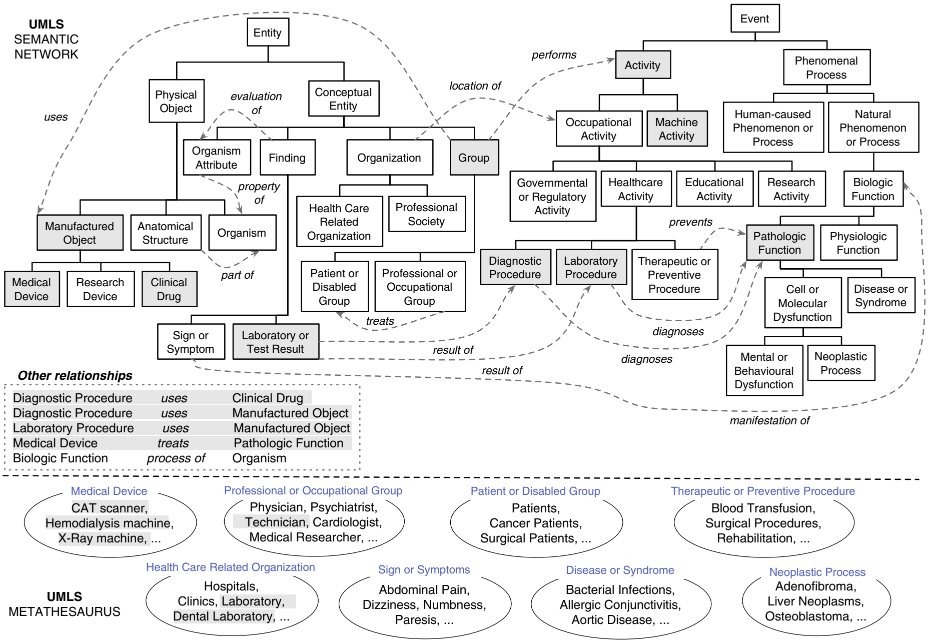

UMLS SEMANTIC NETWORK
Entity
Physical Object
Conceptual Entity
uses
evaluation of
Organism Attribute
Finding
Organization
Group
property of
Manufactured Object
Anatomical Structure
Organism
part of
Medical Device
Research Device
Clinical Drug
Health Care Related Organization
Professional Society
Patient or Disabled Group
Professional or Occupational Group
treats
Sign or Symptom
Laboratory or Test Result
result of
result of
Event
Activity
Phenomenal Process
performs
location of
Occupational Activity
Machine Activity
Human-caused Phenomenon or Process
Natural Phenomenon or Process
Governmental or Regulatory Activity
Healthcare Activity
Educational Activity
Research Activity
Biologic Function
prevents
Pathologic Function
Physiologic Function
Diagnostic Procedure
Laboratory Procedure
Therapeutic or Preventive Procedure
Cell or Molecular Dysfunction
Disease or Syndrome
diagnoses
diagnoses
Mental or Behavioural Dysfunction
Neoplastic Process
manifestation of
Other relationships
Diagnostic Procedure uses Clinical Drug
Diagnostic Procedure uses Manufactured Object
Laboratory Procedure uses Manufactured Object
Medical Device treats Pathologic Function
Biologic Function process of Organism
Medical Device
CAT scanner, Hemodialysis machine, X-Ray machine, ...
Professional or Occupational Group
Physician, Psychiatrist, Technician, Cardiologist, Medical Researcher, ...
Patient or Disabled Group
Patients, Cancer Patients, Surgical Patients, ...
Therapeutic or Preventive Procedure
Blood Transfusion, Surgical Procedures, Rehabilitation, ...
UMLS METATHESAURUS
Health Care Related Organization
Hospitals, Clinics, Laboratory, Dental Laboratory, ...
Sign or Symptoms
Abdominal Pain, Dizziness, Numbness, Paresis, ...
Disease or Syndrome
Bacterial Infections, Allergic Conjunctivitis, Aortic Disease, ...
Neoplastic Process
Adenofibroma, Liver Neoplasms, Osteoblastoma, ...

Figure 1 An example of peer ontology for the healthcare application domain.

be obtained by combining fragments of ontologies downloaded from the Semantic Web with other ontology specifications acquired from the network nodes with similar interests. This is especially possible in the healthcare domain, where a number of taxonomies/ontologies are available and can be exploited by a peer to classify its own resources.

Besides the peer ontology, a *service ontology* provides a semantically rich description of the peer services that are available for sharing. Service descriptions represent functional aspects of a service, based on the WSDL standard for service representation [15], in terms of service categories, service functionalities (WSDL operations), data required from the user for service invocation (inputs) and results produced by the service execution (outputs). Service categories classify services according to standard taxonomies (e.g., UNSPSC or NAiCS): for example, patient administrative services include patient registration services, laboratory reporting services provide sample analysis, that in turn can be blood sample tests or urine tests. Services are registered in a service registry (implementing the UDDI standard [34]) together with their WSDL documents. A WSDL document constitutes a syntactic, low-level representation of the service functional interface and cannot be used to express the semantics of the services, neither to support the user in specifying service requests. 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 ESTEEM, context is described through the *Context Dimension Tree (CDT)* [9], a context model that has been conceived to support the tailoring of the peer data and services according to the current context. For example, the doctor considered in our application scenario is interested in acquiring information on the diseases and symptoms common in Central Africa and on the available care facilities; in another scenario, a laboratory technician needs/offers information and services related to the devices, procedures and analysis to be performed within the lab structure. In Figure 2 an example of CDT, modeling the possible contexts of our medical application, is shown. In this example, context is analyzed with respect to the dimensions shown in Figure 2, which are common to most applications: the *actor*, representing the user's role (e.g., doctor or researcher), the *situation* he/she may be in, the *location*, the *interest topics*. A dimension value can be further analyzed with respect to different viewpoints, generating further (sub-)dimensions in the tree-like structure. A *context* is a subtree of the CDT, obtained by appropriately choosing a set of (sub-)dimension values. The CDT designer is in charge of establishing which dimensions are appropriate for the current application domain and of specifying the correspondence between each context and the portion of the peer and service ontologies that is relevant to it (called *data chunk*). As an example, the gray parts of Figure 1 represent the data chunk associated with the context represented by the gray part of Figure 2 (namely a *doctor*, who is *on field*, is in *Africa* and is interested in the malaria *pathology* and its related *drugs*). Context information can be collected during the registration to the system, as discussed in Section 4. Since building the CDT is a very hard task for non-expert users, the Esteem system supports them providing the Web interface shown in Figure 3. The user has just to answer simple questions about his/her current situation and the system will define the user's personal subtree of the CDT. Note that appropriate fields allow the user to specify parameters like the pathology, the discipline and the user's country (location). A complete and formal definition of the CDT and its usage for data tailoring can be found in [9].

The data quality and trust profile involves the computation of peer data quality metrics, making them available to other peers. More specifically, each peer has the possibility of associating the exported data with *quality metadata*, that represent data quality measures for some specific quality dimensions. We have currently implemented metrics for those quality dimensions that are considered the most

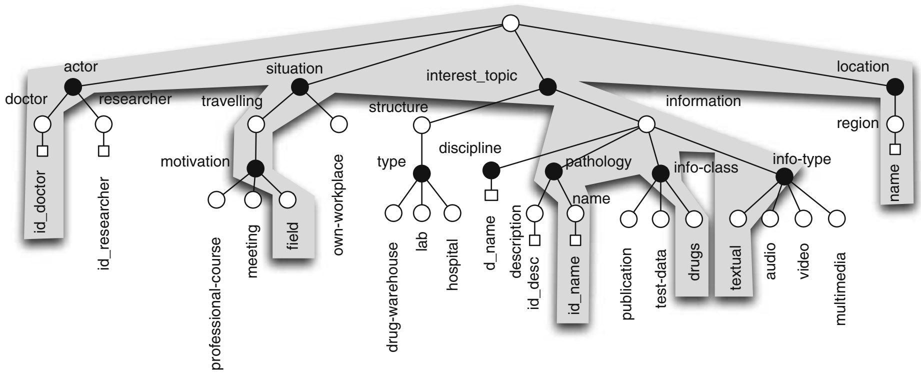
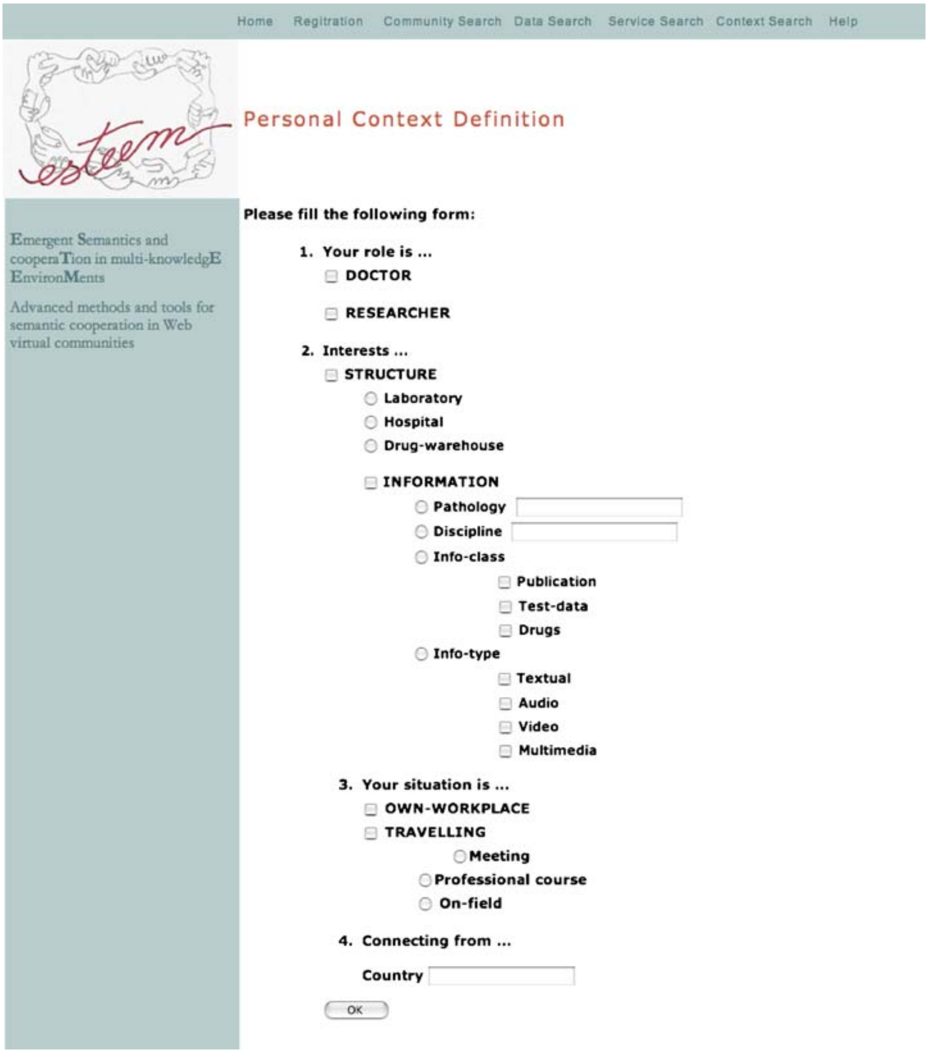


Figure 2 An example of Context Dimension Tree.


Home
Regitration
Community Search
Data Search
Service Search
Context Search
Help
esteem
Emergent Semantics and cooperaTion in multi-knowledgE EnvironMents
Advanced methods and tools for semantic cooperation in Web virtual communities
OK

Personal Context Definition

Please fill the following form:

1. Your role is ...
   - □ DOCTOR
   - □ RESEARCHER
2. Interests ...
   - □ STRUCTURE
     - ○ Laboratory
     - ○ Hospital
     - ○ Drug-warehouse
   - □ INFORMATION
     - ○ Pathology
     - ○ Discipline
     - ○ Info-class
       - □ Publication
       - □ Test-data
       - □ Drugs
     - ○ Info-type
       - □ Textual
       - □ Audio
       - □ Video
       - □ Multimedia
3. Your situation is ...
   - □ OWN-WORKPLACE
   - □ TRAVELLING
     - ○ Meeting
     - ○ Professional course
     - ○ On-field
4. Connecting from ...

   Country

Figure 3 Esteem Context Description page.

common among the ones that define data quality, namely: *completeness*, *format consistency*, *accuracy* and *internal consistency* (see [4] for the metrics definition).

4 Semantic community discovery

The goal of the Esteem system is to support semantic cooperation among a set of autonomous and independent peers. To this end, Esteem relies on an overlay P2P network, where *semantic communities* are defined to aggregate peers with similar interests, and a search mechanism within the communities is adopted to enforce data and service discovery. Community interests are described through a *manifesto*, that is composed of a community ontology, providing a formal representation of the common interpretation (i.e., perspective) of the community interests and, optionally, a Context Dimension Tree representing all the possible contexts supported by the community members. Accordingly, an ESTEEM 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 manifesto. In particular, UCI and M are used by the ESTEEM system to enforce identification and characterization of a semantic community at the system level, while N and L are exploited for providing a community description at the user-interface level.

The emergence of an ESTEEM semantic community is autonomous, in that it originates from a proposal of a *community founder* (i.e., a peer) which initiates the community formation through dissemination of an advertisement message containing the UCI and M of the emerging community. The community manifesto M is defined according to the founder's wishes. In general, the community manifesto is extracted from the peer ontology of the founder and consists of a focused ontology. Moreover, portions of the service ontology, the CDT, the data quality and trust profile can be also included in the community manifesto to further specify the community objectives. For instance, the founder could establish which context dimensions are appropriate for the current scenario, design the CDT and specify 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. We stress that the level of detail used for specifying the community manifesto depends on the community goal. For instance, by using only the first level dimension nodes of the CDT, the founder selects the high-level concepts to specify the interests of the semantic community.

Each receiving peer P_i autonomously decides whether to join the community on the basis of its level of interest in the received manifesto M. Such a level of interest is computed by invoking an *ontology-based semantic matchmaker* (see Section 5) and by evaluating the semantic affinity between M and the peer ontology of P_i. An ESTEEM peer can join zero or more semantic communities according to the results of the semantic matching process, as shown in Figure 4. In this figure it is evident how communities are exploited as a *semantic overlay* on top of the basic P2P overlay (i.e., the *global overlay*) in order to enforce effective data and service sharing according to a probe/search mechanism that will be introduced in the next section.

The peer user, which joins the ESTEEM network, tries to identify the communities that are capable of providing relevant knowledge with respect to his/her interests. The first task the user has to perform is registration. The interface for this step allows one to insert name, surname, job and specialization. Even though the issue of identity checking in a P2P system is not a goal of the ESTEEM project, we decided to envisage a registration step in the GUI because, as emerged from the user requirements, users declared to care about the identity of the community members. After registration, the user can act in two ways: (1) he/she can visualize all the communities currently in the system, without providing any information about himself/herself, (2) he/she can visualize only the communities that have a certain affinity with his/her interests. In the last case, the user must specify his/her own interests. As already stated, the concept of *"ontology"* can be unknown to the users in the considered application scenario, so he/she would not understand what an *"ontology-based representation"*

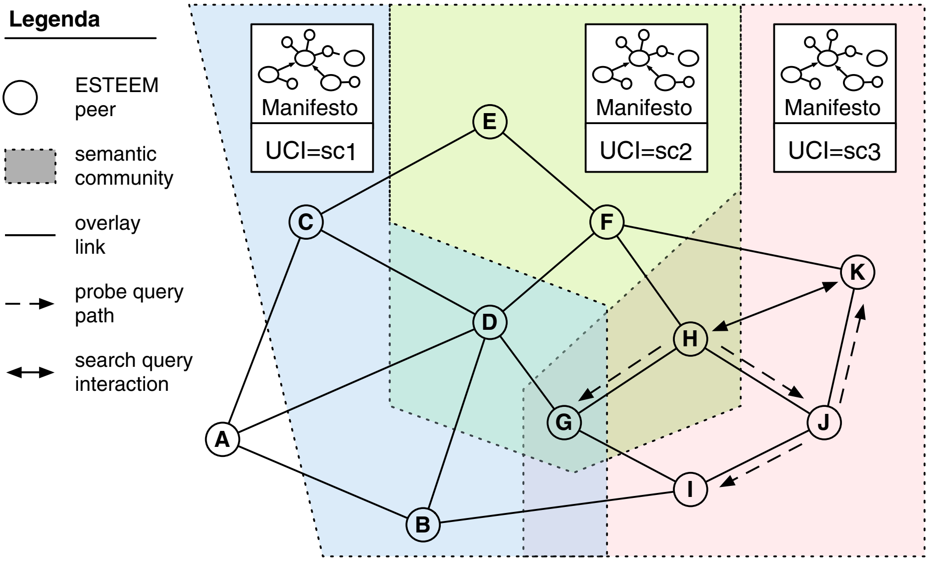


Figure 4 An ESTEEM P2P network.

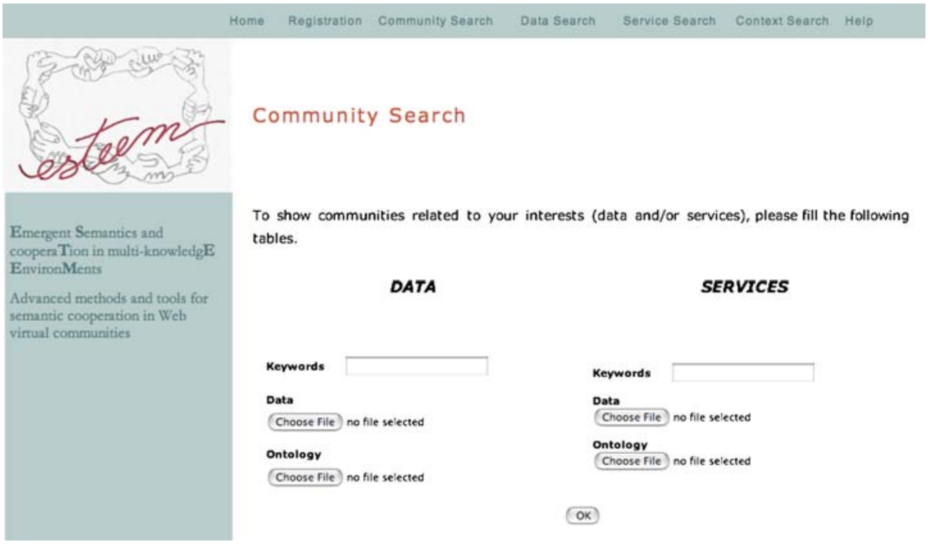
is, neither he/she could provide an ontology-based representation of his/her interests. To overcome this problem, in the interface the user can provide the representation of his/her resources in three different modalities (mutually exclusive):

- by means of a set of keywords;
- by uploading his/her data;
- in case of an expert user, by defining or uploading an ontology as a set of concepts and relationships among them meant to express his/her interests.

The interface for this task is shown in Figure 5. The user's interests are matched against the manifestos of available communities by applying the ontology-based semantic matchmaker to obtain an affinity value quantifying the degree of overlapping between the user's interests and the manifestos of available communities in the ESTEEM network. If the user's interests are provided in one of the first two forms, then the provided representation is automatically converted into an ontology.

The user can visualize the available ESTEEM semantic communities, their name, a brief description, the affinity level and the community manifesto as shown in Figure 6. To join one or more communities he/she has just to click on the ones he/she is interested in. When the user doesn't provide a description of his/her interests, an analogous page will show all the available communities, without providing the affinity rate. Once one or more communities have been joined, each page of the system also displays the list of the joined (available) communities in order to help the user to remember where he/she is, i.e., when performing a query (data and/or service), but also in case of a further access.

Furthermore, when joining a semantic community, the *current context* of a peer is derived by choosing the user's profile, its current situation, actual values for spatial and temporal coordinates and the user's interests among those specified in the CDT



Emergent Semantics and cooperaTion in multi-knowledgE EnvironMents

Advanced methods and tools for semantic cooperation in Web virtual communities

Community Search

To show communities related to your interests (data and/or services), please fill the following tables.

DATA

Keywords

Data

Ontology

SERVICES

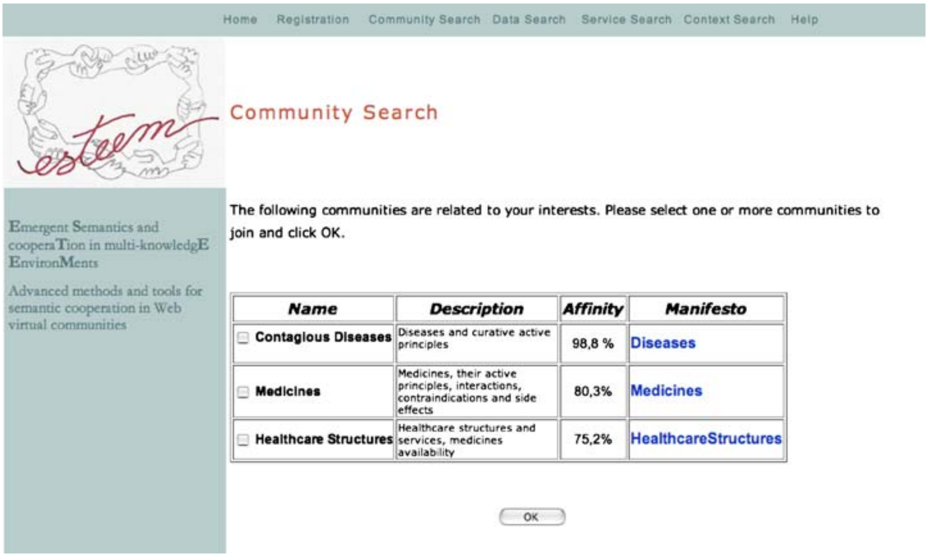
Keywords

Data

Ontology

Figure 5 Esteem community discovery page.

in the manifesto. These values identify a subtree of the CDT, composed by a unique value for the *actor*, *situation*, *time* and *space* dimensions, whereas the various user's preferences on the *interest-topic* values determine a more or less rich context. The user's context can be updated whenever the user desires to change his/her current context information, in particular, it can be refreshed each time the user comes in contact with the community. The definition of the current context is supported



Emergent Semantics and cooperaTion in multi-knowledgE EnvironMents

Advanced methods and tools for semantic cooperation in Web virtual communities

Community Search

The following communities are related to your interests. Please select one or more communities to join and click OK.

Name	Description	Affinity	Manifesto
□ Contagious Diseases	Diseases and curative active principles	98,8 %	Diseases
□ Medicines	Medicines, their active principles, interactions, contraindications and side effects	80,3%	Medicines
□ Healthcare Structures	Healthcare structures and services, medicines availability	75,2%	HealthcareStructures

Figure 6 Esteem community join page.

through the interface shown in Figure 3 for non expert users. Once a peer becomes part of a semantic community, it starts advertising the new knowledge acquired by the community to its neighbors. The advertisement mechanism is related to the community manifesto propagation and it is performed over the global overlay.

A new semantic community may be founded by the user when no communities of interest, namely no matching communities, are discovered in the network. In order to create a new community, a community manifesto must be defined, as described above, considering together the peer ontology, the service ontology and the Context Dimension Tree. Once the community manifesto is defined, a UCI is determined and the new community is created, comprising the community founder as unique member. From now on, the new community manifesto will be spread over the global overlay to advertise it to the other network peers.

We stress that semantic overlay management is performed in a way transparent to the user. A new peer joining the network is inserted in the global overlay through a *bootstrap node* chosen among the already connected peers. A *join message* is sent by the joining node to this bootstrap node, that forwards the request to a subset of its neighbors (maybe all). A set of independent random walks is thus initiated to visit a random set of nodes and the last visited nodes will become neighbors of the joining node. After joining, the new peer starts a thread named *shuffling* thread [36], that periodically renews the peer's neighborhood in order to remove disjoined peers and to insert newly joining ones.

Each peer maintains a table, called Access Point Table (APT), where it stores, for each discovered community, a tuple $\langle UCI, M, N_{ap}\rangle$, where UCI and M are the community identifier and the manifesto, respectively, and N_{ap} is the peer which acts as access point for that community. To join a community, a peer has first to retrieve information about existing communities by querying the peers in the global overlay through a random walk. Each node involved in the random walk will return the complete content of its own APT. For each entry of retrieved $APTs$, the joining peer applies the ontology-based semantic matchmaker in order to establish if the community matches the peer interests or not. If the result is affirmative, the peer will join the community by means of the access point node (N_{ap}) stored in the APT.

The joining peer might not be able to discover any matching community even if it exists in some region of the network, due to the impossibility for a peer to have a complete list of existing communities and to the probabilistic nature of the random walks technique. In order to avoid the existence of several semantically equivalent communities, a community merging mechanism has been implemented. Each time a peer receives an advertisement carrying an UCI not present in the local APT, the peer tries to detect if the received manifesto is similar to the others, stored in the $APTs$, through the ontology-based semantic matchmaker. If some matchings are detected, the community merging procedure is executed by forcing a shuffle among two members of the two communities (exchange of neighbors) and by choosing one triple (UCI, manifesto, access point) as representative for the merged community.

5 Community-based data and service discovery

Once the peer has joined the community, it can share its resources (data and services) within the selected communities. The idea behind the community-based data and service discovery is to select request recipients by considering the joined semantic communities and peers providing similar contents (*semantic neighbors*) discovered during previous interactions. Given a request Q (for data or services) spread over the network by a peer p, the system firstly identifies joined semantic communities within which the request Q could be satisfied by comparing the concepts contained in the request (describing data or I/O parameters and functionalities in case of service request) against each community manifesto. The request Q is then spread over selected communities, where each peer r receiving the request Q compares it against concepts in its peer ontology or its own service descriptions to identify possible semantic affinities. As a result, a (possibly empty) ranked list of matching concepts or of matching services (i.e., services that provide similar functionalities with respect to required functionalities contained in Q) is compiled and returned back to the requesting peer p. After collecting request answers, the peer p exploits the obtained results for deciding the subsequent actions. On the one side, answering peers that provided high matching results for the request Q (i.e., matching results over a predefined threshold) are stored as *semantic neighbors* of p. On the other side, the user on peer p has to decide whether to further "point-to-point interact" with one or more of the semantic neighbors for effective data acquisition and exchange or service invocation. The list of semantic neighbors can be updated each time a new request is answered. Semantic neighbors can be exploited for efficient propagation of future data and service requests according to *semantic routing policies*, in order to foster the interactions with (potentially) most interesting peers.

Since our user is not an expert, he/she is not necessarily able to write a query or a service request compliant with the service interface description introduced in Section 3. To this end, a very simple interface that assists the user in data and service request formulation has been designed. Furthermore, to enable effective data sharing, the user can also benefit of a quality checking mechanism (*quality filter*) and a *context filter* for restricting the results depending on the current user's context.

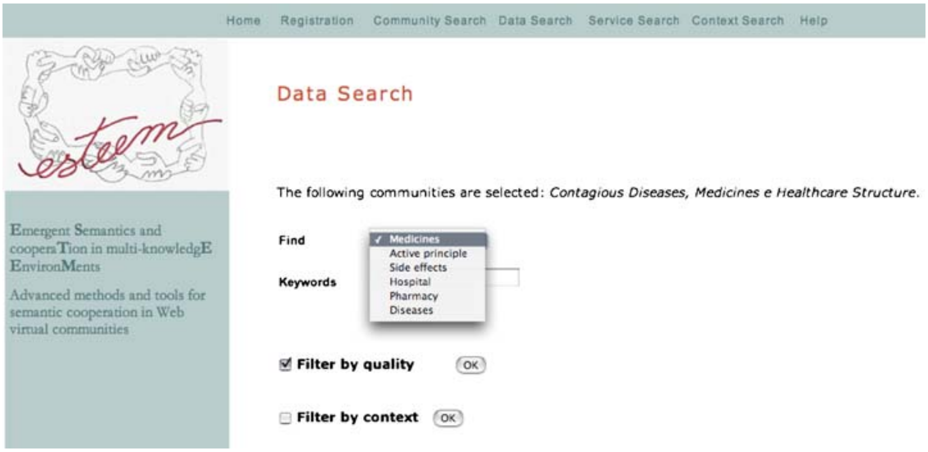
5.1 Data discovery in semantic communities

Query formulation A data-oriented request (i.e., a probe query) provides an ontological description of target concept(s) of interest for the requesting peer. In particular, a request is composed of the following clauses:

- `find`: the list of target concept(s) names;
- `with`: the (optional) list of properties of the target concept(s);
- `where`: the (optional) list of conditions to be verified by the property values and/or the (optional) list of concepts related to the target by a semantic relationship.

The list of target concept(s) of interest (`find` clause) is a mandatory requirement to formulate a data-oriented request in Esteem, while the specification of additional clauses (i.e., properties, property values and semantic relationships) is an optional requirement. As a consequence, different levels of richness in query formulation are enforced in the Esteem system according to the user expertise. For instance, the `find` clause is adequate for non-expert users with search-engine-like requests (i.e., keyword-based queries), while `with` and `where` clauses are required for expert users with Semantic-Web-like requests (i.e., ontology-based queries). To support non-expert users in query formulation, a very simple interface has been designed, shown in Figure 7. In the first field the user must provide 'what' he/she wants to find, by choosing from a list of concepts extracted from the community manifesto. This list is uploaded for each joined community. In the second field the user must provide additional keywords to better specify the target concepts. The data-oriented request is then automatically converted into an ontological description of target concept(s) of interest for the requesting user (containing only the `find` clause). Expert users could refine the request to build queries with the support of an ontology-based editor.

Data matching Ontology matching has the role of measuring the level of match between concept descriptions of different peers through a process of semantic affinity evaluation with the goal of enabling effective comparison of independent peer ontologies with heterogeneous vocabularies. Ontology matching is invoked in different moments of the Esteem approach. Firstly, it is invoked at community formation to measure the level of semantic affinity between the proposed manifesto of a new semantic community and the peer ontology of a receiving peer. Moreover, during data discovery, ontology matching is invoked upon reception of a probe query to evaluate whether a peer can provide matching knowledge in reply to it. To this end, in Esteem, we rely on the HMatch [11] ontology matching system, which has been specifically conceived to work in open distributed systems, like P2P systems. The HMatch system provides flexibility and dynamic configurability features, that are essential requirements to work in open environments. HMatch takes two ontologies as input and returns a semantic affinity value $SA(c, c') \in [0, 1]$ between corresponding concepts c and c' in the two ontologies, along with the mappings between them. Since concepts can be seen as portions of the respective ontologies where they are defined, HMatch is able to compare two single concepts to evaluate their semantic affinity. Semantic affinity is calculated as the linear combination of a linguistic affinity value $LA(c, c')$ and a structural affinity value $TA(c, c')$. The linguistic affinity provides a measure of similarity between two ontology concepts c and c' computed on the basis of their linguistic features (i.e., concept names). For the linguistic affinity



Data Search

The following communities are selected: *Contagious Diseases, Medicines e Healthcare Structure.*

Find: Medicines

Keywords

☑ Filter by quality

□ Filter by context

Figure 7 Esteem Data Search page.

evaluation, HMatch relies on a thesaurus of terms and terminological relationships automatically extracted from the WordNet lexical system. The structural affinity provides a measure of similarity by taking into account the structural features of the ontology concepts c and c', including properties, semantic relationships with other concepts and property values. A comprehensive semantic affinity value $SA(c, c')$ is evaluated as the weighted sum of the linguistic affinity value and the structural affinity value, that is, $SA(c, c') = W_{LA} \cdot LA(c, c') + (1 - W_{LA}) \cdot TA(c, c')$, where $W_{LA} \in [0, 1]$ 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.

Definition of data semantic neighborhood Semantic affinity $SA(c, c')$ is used to set the list of semantic neighbors of peer p within a community. When a peer r receives a concept c in the peer ontology of the sending peer p through a probe query, it replies with a set of concepts $\{c'\}$, where $SA(c, c')$ exceeds a given threshold. If this set is not empty, peer p sets an *inter-peer semantic link* towards peer r for each matching concept c', labeled with $SA(c, c')$. Note that there can be more than one inter-peer semantic link between p and r. An overall measure of the semantic neighborness between p and r is evaluated as the arithmetic mean of the semantic affinity values associated to the inter-peer semantic links between p and r. Semantic neighborhood is updated each time peer r replies to a query Q sent by p during discovery phase. Definition and maintenance of the semantic neighborhood is totally transparent to the user.

Semantic routing mechanism When the peer p receives a query Q containing a target concept tc, data matching with HMatch is applied between tc and each concept c in the peer ontology of p, obtaining the list of the matching concepts MCL = $\{\langle c_1, SA(tc, c_1)\rangle \ldots \langle c_n, SA(tc, c_n)\rangle\}$, where $c_1 \ldots c_n$ are matching concepts in the peer ontology of p with semantic affinity values $SA(tc, c_1) \ldots SA(tc, c_n)$. The list MCL is used to select the best recipients for the query Q among the semantic neighbors of p. Query forwarding is performed as follows:

1) **selection of semantic neighbors -** a semantic neighbor list SNL is defined, where each element is the identifier of a semantic neighbor (i.e., a peer) that is associated with at least one of the matching concepts in MCL through an inter-peer semantic link; each element $sn \in$ SNL is defined as $sn = \{\langle n_{sn}, \{c_1, cf_1, \ldots c_m, cf_m\}\}$, where n_{sn} is the identifier of the semantic neighbor sn, $c_1 \ldots c_m$ are the concepts in MCL connected with sn through an inter-peer semantic link and $cf_1 \ldots cf_m$ are the semantic affinity values associated with such inter-peer semantic links;
2) **ranking of semantic neighbors -** the semantic neighbor list SNL is ranked with respect to the target concept tc by combining the semantic affinity values of inter-peer semantic links and the semantic affinity values of matching concepts in MCL; the harmonic mean is adopted to this end; given a semantic neighbor $sn \in$ SNL, the ranking value r_{sn} is defined as

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

a threshold mechanism is used to filter out the semantic neighbors with a lower ranking since they are less relevant for *tc*; finally, a ranked list RSNL of semantic neighbors is returned;

3) **distribution of credits (optional) -** the semantic routing mechanism can be extended with an optional step for enabling a peer *p* that receives *Q* to in turn forward the query to its semantic neighbors, thus enlarging the scope of *Q*; in particular, a credit-based strategy can be adopted to associate with a probe query a certain amount of credits; credits are progressively consumed by the answers of the receiving peers; the credit-based strategy has been implemented with positive results in the HLink mechanism for semantic query routing in P2P systems; the interested reader can refer to [10] for technical descriptions and simulation results; if MCL = ∅ (no matching concepts found locally), the query is forwarded to all the semantic neighbors of the peer *p* according to the value of their semantic neighborness; if no semantic neighbors have been set yet, the query *Q* is forwarded to a randomly chosen subset of known peers in the joined semantic communities.

5.2 Service discovery in semantic communities

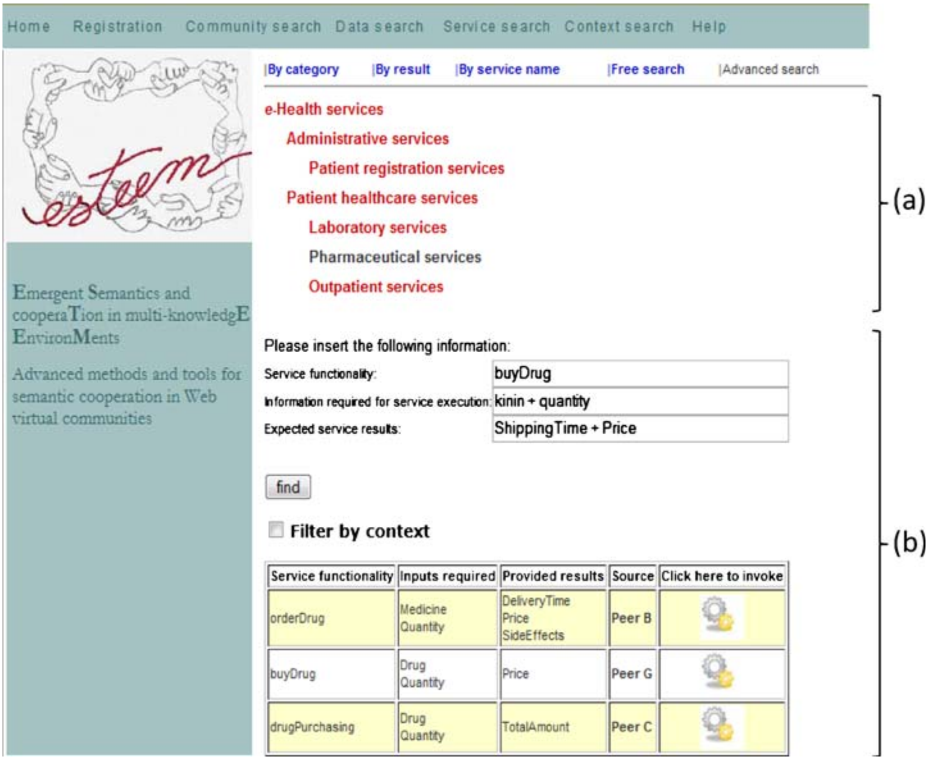
Service request formulation A service request contains the following kinds of information to identify suitable matching services:

- the desired service category;
- a concept representing the required service functionality (e.g., drug ordering, product delivery, remote diagnosis, laboratory testing);
- given the desired functionality, a set of concepts representing the desired results (outputs) and a set of concepts representing data that the requester is able to provide for service execution (inputs).

Also in this case, the non-expert user is supported in service request formulation through a Web interface, where he/she can look for a service through selection of four different options:

- he/she can browse the tree of service categories from standard taxonomies already included in the underlying UDDI Registry (e.g., administrative services, laboratory services, outpatient services), as shown in Figure 8a;
- he/she can specify the results that he/she expects from the service (e.g., documentation, diagnosis, product, drug);
- he/she can specify the kind of service he/she is looking for, denoted by the service functionality (e.g., drug ordering, product delivery, remote diagnosis, laboratory testing);
- in case of more expert user, he/she can perform an advanced search by specifying in the same request the results expected from the service, the kind of service he/she is looking for and data he/she is able to provide for service execution (for example, to buy on-line from a drugstore, the user must provide the active principle and the address where the product must be shipped), as shown in Figure 8b.

The Esteem system automatically builds the service request by filling the category field with service categories selected by the user, the functionality field, outputs



Please insert the following information:

Service functionality:	buyDrug
Information required for service execution:	kinin + quantity
Expected service results:	ShippingTime + Price

Service functionality	Inputs required	Provided results	Source	Click here to invoke
orderDrug	Medicine Quantity	DeliveryTime Price SideEffects	Peer B	
buyDrug	Drug Quantity	Price	Peer G	
drugPurchasing	Drug Quantity	TotalAmount	Peer C	

Figure 8 ESTEEM Service Discovery page: *a* by category, *b* advanced.

and inputs fields with expected results and information the user is able to provide for service execution, respectively. Unspecified fields (depending on the adopted search option) are filled with the `Any` concept to mean that “any matching concept is accepted”.

Non expert users can also perform a traditional keyword-based service search: the specified keywords are matched against the categories, the functionalities or the results of available services on the network. In this case, service search is performed like data search, where service descriptions and requests are simply viewed as sets of concepts not distinguishing among functionalities, input or output parameters.

Search results are presented to the user by specifying all service information (categories, service functionality, provided results and data required for service execution), together with the URL where the service is effectively provided and can be invoked.

Service matching As for data matching, service matching is invoked upon the reception of a service request to evaluate whether a peer can provide services that match the request. To this purpose, we rely on FC-MATCH (FunctionalCompatibility-Match) [7], a service matching approach that performs a comparison between the service request R and each service advertisement S_i available on the peers. In this paper, the FC-MATCH approach has been applied in a P2P environment, to enable the service comparison in presence of different peer ontologies. In FC-MATCH service matching is performed on the basis of concepts contained in the service request and service advertisement and defined in the Service Functionality and Message Ontologies of the peer, combining together two different matching models. Firstly, a deductive model is used to qualify the kind of match MatchType($\mathcal{R}$,$\mathcal{S}_i$). According to this matching model, it is possible to state if $\mathcal{S}_i$ provides the same functionalities required in $\mathcal{R}$ ($\mathcal{S}_i$ EXACT $\mathcal{R}$), if $\mathcal{S}_i$ provides additional functionalities with respect to the required ones ($\mathcal{S}_i$ EXTENDS $\mathcal{R}$) or viceversa ($\mathcal{S}_i$ EXTENDED-BY $\mathcal{R}$), if there is a non empty intersection between provided and required functionalities ($\mathcal{S}_i$ INTERSECTS $\mathcal{R}$) or if $\mathcal{S}_i$ and $\mathcal{R}$ have nothing in common ($\mathcal{S}_i$ MISMATCH $\mathcal{R}$). In case of partial match ($\mathcal{S}_i$ EXTENDED-BY|INTERSECTS $\mathcal{R}$), a similarity-based matching model is used to quantify the degree of match $GSim(\mathcal{R}, \mathcal{S}_i)$ between service descriptions through coefficients properly defined to compare input and output names (Entity-based service similarity $ESim(\mathcal{R}, \mathcal{S}_i)$) and between functionality names (Functionality-based service similarity $FSim(\mathcal{R}, \mathcal{S}_i)$). These coefficients are finally linearly combined to obtain the degree of match. EXACT and EXTENDS match correspond to the case $GSim(\mathcal{R}, \mathcal{S}_i) = 1.0$, while MISMATCH corresponds to the case $GSim(\mathcal{R}, \mathcal{S}_i) = 0.0$.

Also in case of service discovery, a threshold-based mechanism is enforced to set the minimum level of global similarity to consider two services as matching services. The service request R and each service advertisement S_i match if the kind of match is not MISMATCH and $GSim$ exceeds the threshold. The rationale behind the use of an hybrid matching model is that services represent software components that provide their functionalities over the network and the user is interested not only in establishing how much an available service satisfies his/her requests, but also if the available service meets user requirements fully or only partially. Combination of two matching models enhances the flexibility of the ESTEEM system for service discovery.

Definition of service semantic neighborhood The definition of semantic neighborhood for services follows the same principles used for data semantic neighborhood, where the $GSim()$ value is used instead of the semantic affinity value $SA()$. When a peer r receives a service S_i in the service ontology of the sending peer p through a probe request, it replies with a set of services $\{S_j\}$, where $GSim(S_i, S_j)$ exceeds a given threshold. If this set is not empty, peer p sets an *inter-peer semantic link* towards peer r for each matching service S_j, labeled with $GSim(S_i, S_j)$ and the kind of match among EXACT, EXTENDS, EXTENDED-BY or INTERSECTS. An overall measure of the semantic neighborness between p and r is evaluated as the arithmetic mean of the global similarity values, but in this case information about the kind of match between S_i and S_j must be considered. In particular, two distinct situations are considered for peer r:

- peer r provides services that add functionalities with respect to those already provided by p, that is, there is at least one inter-peer semantic link between p and r labeled with a EXTENDS or an INTERSECTS match;
- peer r does not provide services that add functionalities with respect to those already provided by p, that is, all inter-peer semantic links set between p and r are labeled with an EXACT or a EXTENDED-BY match.

Semantic routing mechanism The semantic routing mechanism applied during service discovery is similar to the one applied for data, where $GSim()$ values are used instead of the $SA()$ ones. However, the kinds of match are used to further refine the selection of semantic neighbors according to various strategies:

- only semantic neighbors that add functionalities with respect to those already provided by matching services found locally must be selected (*minimal strategy*); in this case, neighbors that do not provide additional functionalities are not selected; moreover, also matching services in the MCL list that present an Extends or an Exact match with the request R are not considered in the semantic neighbor selection, since they already satisfy the required functionalities;
- also semantic neighbors that do not add functionalities with respect to those already provided by matching services found locally must be selected (*exhaustive strategy*); in this case, the same procedure exposed for data is applied.

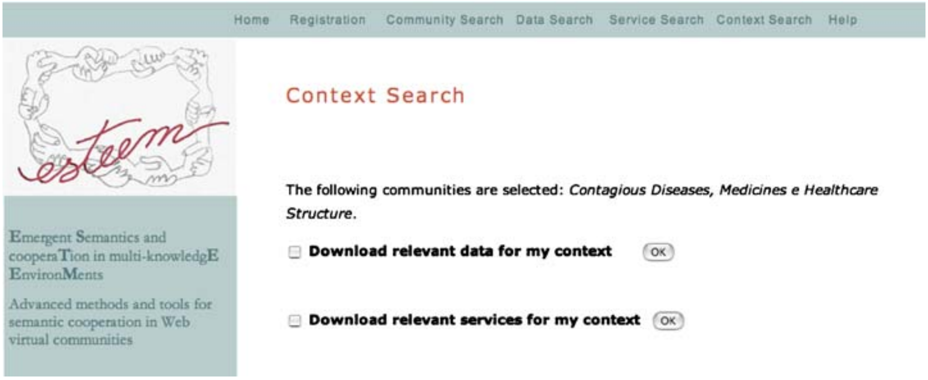
It is important to note that the second strategy is applied when also services that are equivalent from the functional point of view are presented to the user as search answers. These services will be filtered out in a second moment according to quality-based and context-aware aspects. The first strategy can be used in emergency situations, when the response time is often more relevant than some quality features (e.g., product costs) of the answers.

6 Context-aware data and service discovery

In the Esteem system, context is used, in the initial phase, when the user joins or creates a semantic community, as well as later on, during the community life-cycle. The peer context is exploited to support the user in formulating queries according to different approaches.

On the one side, the peer context can be used to directly formulate a query when a peer is interested in discovering nodes using a similar context. In this case, the query contains a context the peer is interested in (i.e. a subtree of the CDT, containing one or more dimension values, built with the support of the Web interface shown in Figure 3). The query containing the context is spread over the semantic community and each peer receiving it applies context matching to establish if it shares the same context. Context matching is another capability provided by the Esteem system. The goal is to compare the concepts (i.e., nodes) belonging to different CDTs and to identify possible correspondences. To this end, traditional string-based matching techniques are adopted. Moreover, some peculiar aspects of context matching are considered. In particular, concepts need to be compared with respect to the dimensions represented by the possibly matching nodes. Two contexts might be equal, incomparable or one strictly contained into the other. The first case is obviously the easy one, since there is full correspondence of contexts. Also the containment case is easily dealt with, because in this case the more general context is chosen as the common one and an affinity value 1.0 is returned. In the case of incomparable contexts, an affinity value is computed, based on the global affinity of the data chunks associated with the two contexts under analysis. If the computed affinity between two contexts exceeds a given threshold, the receiving peer notifies to the sending one that it shares the same or very similar contexts.

On the other side, the peer context can be used to indirectly formulate a query by supporting the user in specifying the concepts of interest to be inserted. In this case,



Emergent Semantics and cooperaTion in multi-knowledgE EnvironMents

Advanced methods and tools for semantic cooperation in Web virtual communities

Context Search

The following communities are selected: *Contagious Diseases, Medicines e Healthcare Structure.*

□ **Download relevant data for my context** OK

□ **Download relevant services for my context** OK

Figure 9 Esteem Search by Context page.

the user selects the context of interest from the peer CDT. Mappings between CDT and peer ontologies are then exploited to define the query contents according to the ontological concepts associated to the context of interest. This approach is exploited during data and service discovery in a transparent manner by checking the “Filter by context” button (see, for example, Figures 7 and 8).

Moreover, the Esteem system allows the user to perform a novel type of query that we have called *Search by context*. In fact, a user may need to query his/her semantic neighbours not only for a precise query, but also to collect a set of information that could interest him/her later on. This could be the case when the network connection to the joined semantic communities is not always available, for instance, when a small device (e.g., palm computer, smartphone) is considered, or for caching purposes. As a consequence, the user downloads the portion of relevant data for an upcoming context when the connection is available. Downloaded information can be used later on. The interface for this task is shown in Figure 9.

7 Obtaining trustworthy data and services through the Esteem network

Once quality metadata are available in the system, they can be used for: (1) implementing a quality filter to be used in the discovery process and (2) evaluating the trust of a peer providing data and services to other peers of the community.

The quality filter is invoked during query processing and exploits quality metadata introduced in Section 3 in order to tackle data inconsistencies. More specifically, in Esteem, we assume that data can exhibit key-level conflicts [4]. This implies that an object identification step must be performed in order to provide answers to user’s queries. Due to the specific requirements of the Esteem system, this step should be performed in a fully automatic way. Object identification involves the choice of one or more attributes, referred to as matching keys. Such a choice is normally performed by humans, while, in our case, we need it to be performed automatically. Therefore, we have added to the metadata calculated for quality profiling a further metadata, named identification power, that specifies how much a given attribute is discriminating when trying to match objects. For instance, a `Sex` attribute is quite surely less discriminating than a `Surname` attribute, when matching records are referred to persons. The automatic method for matching key computation, which is based on the identification power and on quality metadata, is fully described in [6]. Once this phase has been made automatic, we are able to run an object identification process with the objective of solving key-level conflicts, thus allowing query processing to be carried out.

The trust support of the Esteem system is also based on the quality metadata that are available at each peer. Indeed, a peer is more trustworthy than another one if it declares quality values for the data and services it exports that are verified as "reliable" by other peers. When deciding the atomic unit to trust in an emergent semantics system, a first hypothesis could be to trust the peer as a whole with respect to the totality of exchanged data, or more generally to the transactions performed with the other peers. The method proposed in [1] is an example of this case. We follow the approach of associating trust to a peer as a whole, but we propose two major modifications: first, we consider a specific type of transaction, i.e., data exchanges; second, we evaluate trust of a peer with respect to a specific type of provided data. The key idea can be summarized as follows: (i) the atomic unit of trust is the couple $\langle P_i, \mathcal{D}\rangle$, where $\mathcal{D}$ is an element of the peer ontology of the peer P_i; (ii) the trust level of a peer P is computed on the basis of the number of complaints fired by other peers of the community, for which P had been a data provider. The details of the model that we use for trust computation are provided in [17]. The major adaptation to the Esteem architecture is to consider each semantic community as a newly constituted cooperative information system, thus requiring a community specific trust computation service. Trust evaluation prevents the user considered in our application scenario from receiving non-reliable data and services and it is performed in a completely transparent way.

8 System validation

We believe that the success of the Esteem platform resides both in its functionalities and in its usability. Consistently with the objectives of the project, we have tested the Esteem system with the kind of users for whom the system is intended. To this end the design phase was preceded by the collection of information about medical personnel activities both in terms of functionalities and in terms of their acquaintance with Web-based technologies. In this phase 18 persons that operate in the medical sector have been interviewed and the results of this activity have been exploited in the design of all the Esteem functionalities. To validate these functionalities and their usability, the Graphic User Interface (GUI) has been tested with doctors that have been asked to use it for executing some specific tasks. In this section we describe the validation of the Esteem system. Firstly, we provide an overview of the evaluation techniques used to test the GUI, then we present the results of the validation.

8.1 Experimental design

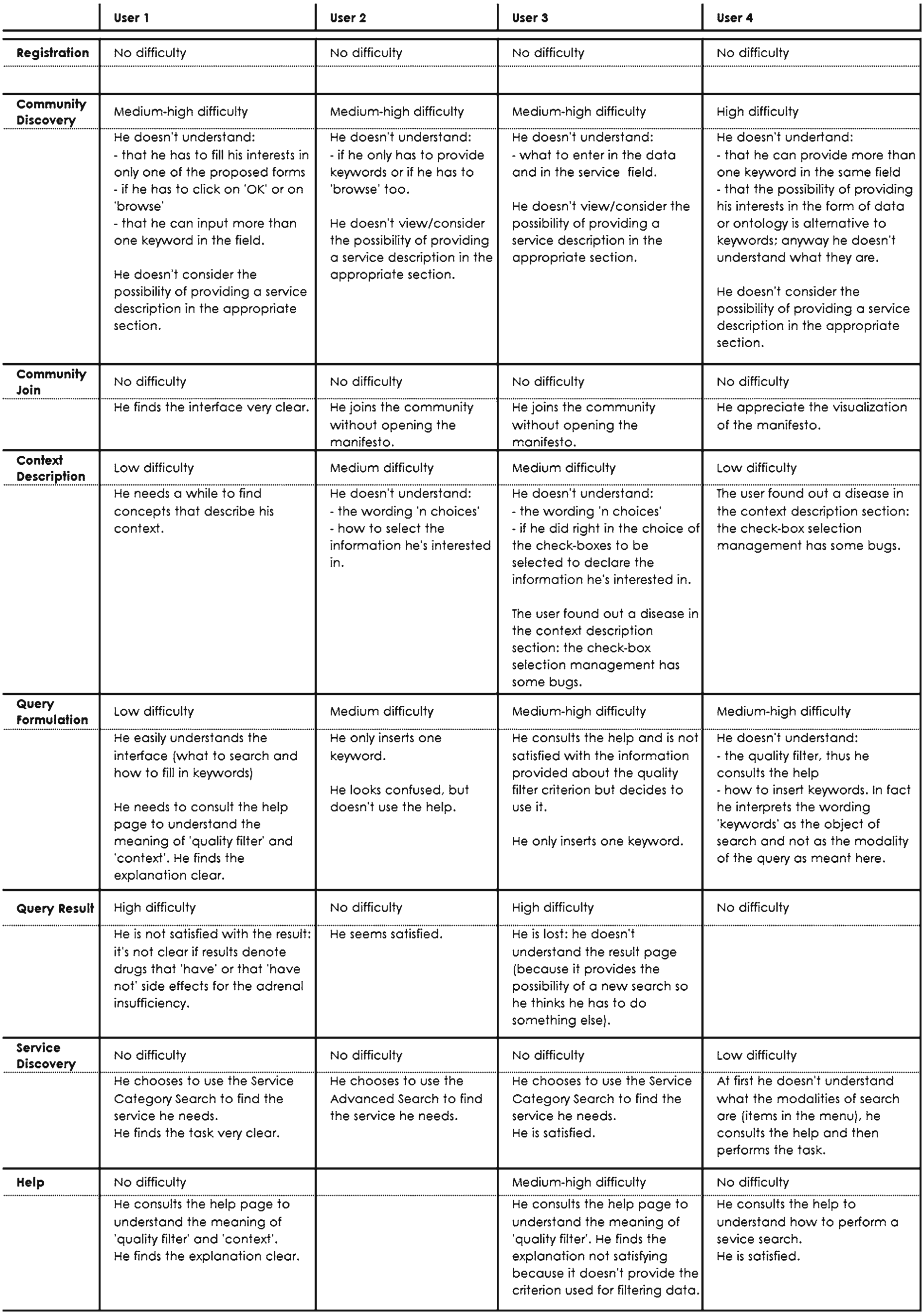
Users have been involved in a think aloud evaluation [18] in their work environment. During this evaluation activity, the users have been recorded during their usage of the GUI. The think aloud evaluation technique is a form of observation where the user is asked to talk through what he/she is doing as he/she is being observed; for example, describing what he/she believes is happening, why he/she takes an action, what he/she is trying to do. Think aloud has the advantage of simplicity, it requires little expertise to perform and can provide useful insights into the problems with interface. This kind of evaluation is often performed in the latest stages of development, where there is at least a working prototype of the system in place, but it can also be very useful in the earlier design stage, for example to capture user's requirements. We involved in the validation doctors belonging to different medical fields and with different expertise. To guide the user in the evaluation step, we have produced as supporting material a document that contains a brief description of the Esteem project, the advantages of using the Esteem platform instead of a search engine, the experimental tasks to be executed. The application scenario to be considered in the demonstration is the one described in Section 2. In particular, the users have been asked to join the Esteem network and subsequently look for a drug that treats malaria without side effects for adrenal insufficiency. As the drug is not available in their field hospital, the users have been also asked to find services to order the medicine and to require shipping of the drug to the location where they are operating. According to such a scenario, each user had to execute the following tasks:

- **Registration** to the system;
- **Community Discovery**, the user has to find the communities that match his/her interests by providing the keywords: drug, hospital, pharmacy and contagious disease;
- **Community Join**, the user has to join the communities Drugs, Tropical disease and Medical structures;
- **Context Description**, the user has to provide the description of his/her personal context;
- **Data Search**, the user has to obtain information about which drugs can treat malaria without side effects for adrenal insufficiency (this task is performed using the quality filter);
- **Service discovery**, the user has to find a structure that sells the drug and ships it to the location where he/she is operating (this task is performed using the context filter);
- **Help consultation**, the user has to consult the help.

8.2 Evaluation results

The information collected during the think aloud validation are reported in Figure 10. In the table we have reported, for each task, the behaviour of the user together with the level of difficulty encountered. We have classified the level of difficulty according to the following scale:

- **No difficulty**: the user easily performs the task;
- **Low difficulty**: the user needs a while to understand how to perform the task, but, after consulting the help, he/she properly performs the task;
- **Medium difficulty**: the task is inaccurately executed, but the user doesn't stop;
- **Medium-high difficulty**: the task is not correctly executed and/or the user needs a little tip to go on;
- **High difficulty**: the user needs the intervention of the expert to go on.



	User 1	User 2	User 3	User 4
Registration	No difficulty	No difficulty	No difficulty	No difficulty
Community Discovery	Medium-high difficulty	Medium-high difficulty	Medium-high difficulty	High difficulty
	He doesn't understand: - that he has to fill his interests in only one of the proposed forms - if he has to click on 'OK' or on 'browse' - that he can input more than one keyword in the field. He doesn't consider the possibility of providing a service description in the appropriate section.	He doesn't understand: - if he only has to provide keywords or if he has to 'browse' too. He doesn't view/consider the possibility of providing a service description in the appropriate section.	He doesn't understand: - what to enter in the data and in the service field. He doesn't view/consider the possibility of providing a service description in the appropriate section.	He doesn't undertand: - that he can provide more than one keyword in the same field - that the possibility of providing his interests in the form of data or ontology is alternative to keywords; anyway he doesn't understand what they are. He doesn't consider the possibility of providing a service description in the appropriate section.
Community Join	No difficulty	No difficulty	No difficulty	No difficulty
	He finds the interface very clear.	He joins the community without opening the manifesto.	He joins the community without opening the manifesto.	He appreciate the visualization of the manifesto.
Context Description	Low difficulty	Medium difficulty	Medium difficulty	Low difficulty
	He needs a while to find concepts that describe his context.	He doesn't understand: - the wording 'n choices' - how to select the information he's interested in.	He doesn't understand: - the wording 'n choices' - if he did right in the choice of the check-boxes to be selected to declare the information he's interested in. The user found out a disease in the context description section: the check-box selection management has some bugs.	The user found out a disease in the context description section: the check-box selection management has some bugs.
Query Formulation	Low difficulty	Medium difficulty	Medium-high difficulty	Medium-high difficulty
	He easily understands the interface (what to search and how to fill in keywords) He needs to consult the help page to understand the meaning of 'quality filter' and 'context'. He finds the explanation clear.	He only inserts one keyword. He looks confused, but doesn't use the help.	He consults the help and is not satisfied with the information provided about the quality filter criterion but decides to use it. He only inserts one keyword.	He doesn't understand: - the quality filter, thus he consults the help - how to insert keywords. In fact he interprets the wording 'keywords' as the object of search and not as the modality of the query as meant here.
Query Result	High difficulty	No difficulty	High difficulty	No difficulty
	He is not satisfied with the result: it's not clear if results denote drugs that 'have' or that 'have not' side effects for the adrenal insufficiency.	He seems satisfied.	He is lost: he doesn't understand the result page (because it provides the possibility of a new search so he thinks he has to do something else).	
Service Discovery	No difficulty	No difficulty	No difficulty	Low difficulty
	He chooses to use the Service Category Search to find the service he needs. He finds the task very clear.	He chooses to use the Advanced Search to find the service he needs.	He chooses to use the Service Category Search to find the service he needs. He is satisfied.	At first he doesn't understand what the modalities of search are (items in the menu), he consults the help and then performs the task.
Help	No difficulty		Medium-high difficulty	No difficulty
	He consults the help page to understand the meaning of 'quality filter' and 'context'. He finds the explanation clear.		He consults the help page to understand the meaning of 'quality filter'. He finds the explanation not satisfying because it doesn't provide the criterion used for filtering data.	He consults the help to understand how to perform a sevice search. He is satisfied.

Figure 10 Results of the think aloud test.

According to this classification, the tasks that caused none or low difficulty are registration, community join, service discovery and help consultation. These tasks have been executed correctly and users have expressed positive assessments both on the interface and on the proposed functionalities. An unexpected result was the difficulty encountered in the context description task, that was not executed so quickly, mainly because users did not feel confident with the concept of *context*. The tasks that have caused most problems to the users are the same that required the greatest attention in the design phase: community discovery, query formulation and query result.

The community discovery page presented difficulties in understanding that: (i) the modalities in which community interests can be represented (keywords, data set or ontologies) are mutually exclusive; (ii) in the keyword field it is possible to enter more than one keyword; (iii) keywords about data interests and keywords about service interests have to be provided separately. The query formulation page presented difficulties in understanding that more than one keyword is allowed and which aspects are filtered out to provide data with a high level of quality. Some users justified this because they intended the wording 'keywords' as the object of the query (instance) and not as the characteristics (attribute's values) of the searched object specified above. In the query result page some users pointed out that the result of the query is not satisfying because they can't understand if the provided results are the drugs that 'present' or that 'do not present' side effects for the adrenal insufficiency. A user got lost in this page because, after having got the results of the query, there is the possibility to perform a new search, so he feels like if there is something more to do. They would like an introductive sentence to clarify this aspect.

As we validated the GUI with few users we won't compute statistical indicators on such a small sample, but we'll restrict ourselves to report the frequencies of the answers. After having used the GUI, the doctors filled a satisfaction questionnaire concerning both the research topics of the project and the usability issues of the interface. The results of the user satisfaction questionnaire show that the users consider as very important the information exchange between medical operators and believe that the ESTEEM project could help improving the communication based on Web technologies. With reference to the proposed GUI, the users have a positive opinion of the interface although two of them find the interaction only intuitive on average. Details about the GUI evaluation and the complete version of the questionnaire are reported in [13]. The ESTEEM approach has been also evaluated in terms of traffic and scalability. In particular, a dedicated evaluation session has been specifically executed for the components responsible of overlay management, peer community formation and semantic routing. Detailed experimental results are provided in [12].

9 Related work

The ESTEEM approach results from the combination of several research areas (semantic resource discovery, virtual communities of peers, trust and data quality, context-aware delivery), properly integrated and extended to work in open and dynamic environments such as the P2P one and equipped with a user-friendly Web interface to satisfy usability requirements.

The recent growth of P2P applications has motivated the interest in general-purpose P2P overlay networks. P2P applications are based on structured or unstructured network topologies, the latter being more resilient to nodes joining or leaving the system, but also characterized by expensive resource discovery, based on flooding or random-walks routing techniques, raising also scalability issues. The Esteem system relies on an unstructured P2P network: (i) it is based on semantic communities, which emerge spontaneously and evolve through a shuffling-based Overlay Management Protocol (OMP) [36]; (ii) it implements advanced semantic routing techniques for both data and service discovery. In the literature, semantic routing techniques are being defined as a promising solution for improving the effectiveness of traditional query propagation strategies in P2P [19, 32]. However, such kind of techniques are not adequate when network dynamism and peer volatility are high, as occurs in most P2P sharing networks. A probabilistic adaptive method to implement informed search in unstructured P2P networks is described in [37]. In this approach, for each neighbor a peer maintains information on the probability that the neighbor can provide resources on a given topic. This information is dynamically adjusted and updated on the basis of the history of past searches. A peer decides to forward a request involving some topics to one of its neighbors if the probability of obtaining a positive answer is greater than a given threshold. The overall requirement of the routing policy is to maximize the number of retrieved results and minimize the number of peers involved in processing the request. This approach differs from the Esteem one mainly because: (i) it is devoted to data discovery only; (ii) peers and requests are described by conjunctions of topics and this way is not suitable for describing complex resources involving data and services; (iii) it does not admit that different peers use different ontologies, as considered in the Esteem scenario; (iv) each peer keeps knowledge about its logical neighbors in terms of probabilistic information, while in Esteem a peer establishes semantic links towards semantic neighbors selected on the basis of their content. Moreover, in Esteem, the notion of semantic community is used to replace the notion of P2P overlay and semantic affinity functions are being introduced to calculate peer proximity on a semantic basis rather than on a topological one. In most of the existing approaches, only basic community-oriented solutions are currently available and they usually rely on centralized formation techniques [8]. In some other cases, only metadata-based peer resource descriptions and string-based matching techniques are supported and they are not sufficiently expressive to be considered as really semantic-based approaches [3, 5, 35]. In [30] a mechanism based on communities is adopted to reduce the latency of search in content distribution networks. Target of the search are Web pages and the aim of the proposed mechanism is to reduce the latency of the page retrieval by pre-fetching pages that are supposed to be asked soon. The rationale behind this approach is that the higher is the number of links among a set of pages the higher is the probability that the pages are required in very close searches. To this purpose, pages are organized in a graph, where nodes are the Web pages and edges are the links between them. Pages are grouped into clusters, called *Web site communities*, on the basis of the density of connections between Web pages. When a Web page is required, also the other pages belonging to the same community are pre-fetched. This ensures fast content retrieval without requiring a-priori knowledge of request statistics. This approach is quite different from the Esteem one because: (i) it is devoted to Web page discovery only; (ii) communities are not set of peers dynamically aggregated through semantic comparison of their interests, but are defined as set of Web pages clustered on the basis of connections (i.e., links) between them. Esteem communities represent an original contribution as they allow to combine ontologies and ontology matching techniques with the notion of self-configuring semantic overlay. Furthermore, in contrast with most existing solutions, Esteem communities are *lightweight*, in the sense that membership is open and approval/rejection of a peer is not determined by the decision of a supervisor. Moreover, community maintenance is efficient due do 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.

The use of context-awareness in P2P communities aims at making more focused and efficient the mechanisms for joining and sharing knowledge among peers. Other *community-based* approaches exist, where the context definition is achieved in a distributed fashion [14, 26]. In both the approaches, context is not treated as a first-class citizen, but its treatment is somehow "hidden" within the system mechanisms. By contrast, the context model adopted by the Esteem approach is fully orthogonal, thus can be used in different application scenarios and be applied to data, service and peer context-based filtering.

Finally, the Esteem system has also been extended with a trust and quality filtering approach. Quality-aware data integration systems have been proposed in centralized data integration contexts (e.g., [24, 29]). As far as we know, the Esteem system is the first P2P integration system that takes data quality into account during the query processing phase. More specifically, the assumption that sources are free of errors is underlying current peer data management systems such as PIAZZA [20] and ORCHESTRA [21]. When considering P2P systems, in which the purpose is file sharing rather than data management, the concept of trust is instead widely adopted. However, in such systems trust is typically specified at the source level, while a finer granularity is necessary in order to achieve a quality-aware query processing. In the Esteem system, quality metadata are associated to the data exported by each peer, thus the best quality data are selected during the query processing phase.

10 Final remarks

In this paper we have presented the Esteem system for semantic cooperation in dynamic and multi-knowledge communities in open P2P environments, developed within the Esteem project. Key features of the Esteem system is to preserve the autonomous and spontaneous nature of peer community formation, while offering a rigorous and powerful approach to context-aware data/service discovery and sharing with additional constraints for trust and quality-based data management. This paper showed the capabilities of the Esteem system through a prototype, whose usage has been validated in a healthcare application scenario, both in terms of satisfaction for the offered functionalities and with respect to the usability issues of the interface. After the positive results of the first validation of the Esteem system both in terms of satisfaction for the offered functionalities, and with respect to the usability issues of the interface, we plan to improve the GUI and test it again with a higher number of users.

Acknowledgements The Esteem system has been developed in a PRIN Project funded by the Italian Ministry of Education, University and Research and coordinated by Dipartimento di Informatica e Sistemistica "A.Ruberti" at the Università di Roma "La Sapienza". The following persons contribute to the Esteem project with their work: Carlo A. Curino, Diego Milano, Giorgio Orsi, Antonella Poggi, Leonardo Querzoni, Denise Salvi, Sara Tucci.

In particular, the contribution of Devis Bianchini in the form of valuable coordination and editorial support was essential for composing this paper.

References

1. Aberer, K., Despotovic, Z.: Managing tust in a peer-2-peer information system. In: Proc. of the 10th Int. Conference on Information and Knowledge Management (CIKM'01), pp. 310–317, Atlanta, Georgia, USA (2001)
2. Aberer, K.A., et al.: Emergent semantics principles and issues. In: Proc. of the 9th Int. Conference on Database Systems for Advances Applications (DASFAA'04), pp. 25–38, Jeju Island, Korea (2004)
3. Agostini, A., Moro, G.: Identification of communities of peers by trust and reputation. In: Proc. of the 11th Int. Conference on Artificial Intelligence: Methodology, Systems and Applications (AIMSA'04), pp. 85–95, Varna, Bulgaria (2004)
4. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methods and Techniques (chapter 2). Springer, New York (2006)
5. Benatallah, B., Hacid, M.S., Paik, H.Y., Rey, C., Toumani, F.: Towards semantic-driven, flexible and scalable framework for peering and querying e-catalog communities. Inf. Syst. **31**(4–5), 266–294 (2008)
6. Bertolazzi, P., De Santis, L., Scannapieco, M.: Automatic record matching in cooperative information systems. In: Proc. of the ICDT Int. Workshop on Data Quality in Cooperative Information Systems (DQCIS'03), pp. 13–20, Siena, Italy (2003)
7. Bianchini, D., De Antonellis, V., Melchiori, M.: Flexible semantic-based service matchmaking and discovery. World Wide Web J. **11**(2), 227–251 (2008)
8. Bloehdorn, S., Haase, P., Hefke, M., Sure, Y., Tempich, C.: Intelligent community lifecycle support. In: Proc. of the 5th Int. Conference on Knowledge Management (I-KNOW'05), pp. 278–285, Graz, Austria (2005)
9. Bolchini, C., Curino, C., Quintarelli, E., Schreiber, F.A., Tanca, L.: Context information for knowledge reshaping. Int. J. Web Engineering and Technology **5**(1), 88–103 (2009)
10. Castano, S., Montanelli, S.: Semantically routing queries in peer-based systems: the H-link approach. Knowl. Eng. Rev. **23**(1), 1–22 (2007)
11. Castano, S., Ferrara, A., Montanelli, S.: Matching ontologies in open networked systems: techniques and applications. J. Data Semantics (JoDS) **5** (2006)
12. Catarci, T., et al.: Mock-up Prototypes and Data Collected During the Testing Phase. Deliverable DALL3, MIUR Esteem Project (2007)
13. Catarci, T., et al.: Integrated Mock-up Prototype and Experimental Results on the ESTEEM Application Scenario. Deliverable DALL4, MIUR Esteem Project (2007)
14. Chen, H., Finin, T., Joshi, A.: An intelligent broker for context-aware systems. In: Proc. of the Fifth Int. Conference on Ubiquitous Computing (UbiComp'03), pp. 183–184, Seattle, Washington, USA (2003)
15. Christensen, E., Curbera, F., Meredith, G., Weerawarana, S.: Web Services Description Language (WSDL) 1.1. World Wide Web Consortium (W3C). http://www.w3.org/TR/wsdl (2001)
16. Curino, C., Orsi, G., Panigati, E., Tanca, L.: Accessing and documenting relational databases through OWL ontologies. In: Proc. of Flexible Query Answering Systems (FQAS'09) (2009)
17. De Santis, L., Scannapieco, M., Catarci, T.: Trusting data quality in cooperative information systems. In: Proc. of the 11th Int. Conference on Cooperative Information Systems (CoopIS'03), pp. 354–369, Catania, Italy (2003)
18. Dix, A., Finlay, J.E., Abowd, G.D., Beale, R.: Human-Computer Interaction (3rd edn.). Prentice-Hall, Englewood Cliffs (2003)
19. Haase, P., Siebes, R., van Harmelen, F.: Expertise-based peer selection in peer-to-peer networks. Knowl. Inf. Syst. **15**(1), 75–107 (2008)
20. Halevy, A.Y., Ives, Z.G., Madhavan, J., Mork, P., Suciu, D., Tatarinov, I.: The Piazza peer data management system. IEEE Trans. Knowl. Data Eng. **16**(7), 787–798 (2004)
21. Ives, Z.G., Khandelwal, N., Kapur, A., Cakir, M.: ORCHESTRA: rapid, collaborative sharing of dynamic data. In: Proc. of Conference on Innovative Data Systems Research (CIDR'05), pp. 107–118, Asilomar, CA, USA (2005)

22. Lenzerini, M.: Data integration: a theoretical perspective. In: Proc. of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS'02), pp. 233–246, Madison, Winsconsin, USA (2002)
23. Motik, B., Horrocks, I., Sattler, U.: Bridging the Gap between OWL and Relational Databases. Journal of Web Semantics **7**(2), 74–89 (2009)
24. Motro, A., Anokhin, P.: Fusionplex: resolution of data inconsistencies in the integration of heterogeneous information sources. Information Fusion **7**(2), 176–196 (2006)
25. Mukherjee, C.S., Ramakrishnan, I.V.: Automated semantic analysis of schematic data. World Wide Web J. **11**(4), 427–464 (2008)
26. Ouksel, A.M.: In-context peer-to-peer information filtering on the Web. SIGMOD Rec. **32**(3), 65–70 (2003)
27. Pottinger, R., Bernstein, P.: Creating a mediated schema based on initial correspondences. IEEE Data Eng. Bull. **25**(3), 22–31 (2002)
28. Sattler, U., Calvanese, D., Molitor, R.: Relationships with other formalisms. Description Logic Handbook, pp. 137–177 (2003)
29. Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., Baldoni, R.: The architecture: a platform for exchanging and improving data quality in cooperative information systems. Inf. Syst. **29**(7), 551–582 (2004)
30. Sidiropoulos, A., Pallis, G., Katsaros, D., Stamos, K., Vakali, A., Manolopoulos, Y.: Prefetching in content distribution networks via web communities identification and outsourcing. World Wide Web J. **11**(1), 39–70 (2008)
31. Specia, L., Motta, E.: Integrating folksonomies with the semantic web. In: Proc. of the 4th European Semantic Web Conference (ESWC'07), pp. 624–639, Innsbruck, Austria (2007)
32. Staab, S., Tempich, C., Wranik, A.: REMINDIN: semantic query routing in peer-to-peer networks based on social metaphors. In: Proc. of the 13th Int. World Wide Web Conference (WWW'04), pp. 640–649, New York, NY, USA (2004)
33. The Esteem Project: Emergent Semantics and cooperaTion in multi-knowledgE EnvironMents. http://www.dis.uniroma1.it/~esteem/index.html (2008)
34. OASIS Universal Description, Discovery and Integration v3.0.2 (UDDI): http://www.oasis-open.org/committees/uddi-spec (2003)
35. Verma, K., Sivashanmugam, K., Sheth, A., Patil, A., Oundhakar, S., Miller, J.: METEOR-S WSDI: a scalable infrastructure of registries for semantic publication and discovery of web services. J. Inf. Technol. Appl. Manag. **6**(1), 17–39 (2005)
36. Voulgaris, S., Gavidia, D., van Steen, M.: CYCLON: inexpensive membership management for unstructured P2P overlays. J. Netw. Syst. Manag. **13**(2), 197–217 (2005)
37. Zhou, A., Xu, L., Dai, C.: Adaptive probabilistic search over unstructured peer-to-peer computing systems. World Wide Web J. **9**(4), 537–556 (2006)