International Journal of Cooperative Information Systems
© World Scientific Publishing Company

A B2B ARCHITECTURE AND PROTOCOL FOR RESEARCHERS COOPERATION

JAIME LLORET

Departamento de Comunicaciones, Universidad Politécnica de Valencia, Camino Vera, s/n
Valencia, 46022, Spain
jlloret@dcom.upv.es

JESUS TOMAS

Departamento de Comunicaciones, Universidad Politécnica de Valencia, Camino Vera, s/n
Valencia, 46022, Spain
jtomas@dcom.upv.es

MIGUEL GARCIA

Departamento de Comunicaciones, Universidad Politécnica de Valencia, Camino Vera, s/n
Valencia, 46022, Spain
migarpi@posgrado.upv.es

RAQUEL LACUESTA

Escuela Universitaria Politéctica de Teruel, University of Zaragoza, Ciudad Escolar, s/n
Teruel, Spain
lacuesta@unizar.es

Received 8 November 2011
Revised Day Month Day

Some works on the researchers cooperation's literature provide the key lines for building research networks and propose new protocols and standards for business to business (B2B) data exchange, but no one of them explains how researchers should contact and the procedure to select the most appropriate partner of a research enterprise, institute or university. In this paper, we propose a B2B architecture and protocol between research entities, that uses ebXML protocol. The contacts for cooperation are established based on some defined parameters and an information retrieval system. We explain the information retrieval system, the researcher selection procedure, the XML-based protocol and the workflow of our proposal. We also show the information that has to be exchanged to contact other researchers. Several simulations demonstrate that our proposal is a feasible architecture and may be used to promote the research cooperation. The main purpose of this paper is to propose an efficient procedure for searching project partners.

Keywords: Partners search system; research cooperation; entities interconnection; B2B; Workflow.

1. Introduction

Nowadays, the most used technology to put in contact people is social networks. A social network can be defined as a group of people connected by social relationships such as a co-working, information exchange and friendships which communication is performed using computer networks [1]. The main research on social networks is focused on the relationship between human-computer and computer-computer, although there are many works on social networks topologies. Many works study how two persons interact and how small groups perform online. These researches are focused on the analysis of relation patterns among people, organizations and their states [2 - 4]. These works can also be applied to diverse areas such as the workplace [5, 6] and virtual communities [7]. In these cases, relations are characterized by content, direction and strength. The content is the exchanged resource. A relation can be directed or undirected. Relations also differ in strength, which can be measured using many ways [8, 9]. Social networks and Internet has been very useful for the media users. They improve these networks by establishing relationships though platforms such as Myspace, Frienster, Facebook, Academia.edu, Docstoc, etc.

Academic and business people are interested on the research of social networks to understand better how enterprises work, how workers behave and how are the staff members' interactions. They try to understand dynamic relationships in order to improve the productivity and the spread of ideas in the enterprise [10]. Some Web 2.0 services take profit of the social network features for identifying communities of practice [11]. On one hand, Facebook [12-14] and Linkedin [15-16] are being used by the community research and development information service and by the research and education funding agencies to promote their open research calls and funding, while on the other hand, other social networks such as academia.edu (launched in September 2008, nowadays it has more than 211,000 registered users [17]), ResearchGate (with more than one million of registered users [18]), Epernicus [19] and Academic Family [20] are being used to put in contact researchers. Complex networks and social networks are self-organized by the actions of a large number of individuals. One of the premises of these businesses is that individuals might be only a few steps from a desirable business or social partner. In 2002, 60,000 participants in small world experiments using email chains gave an average of 4:1 links to bridge continents [21]. Holger Ebel et al., in [22], studied the networks composed of persons connected by exchanged e-mails. They demonstrated that these networks present both, small-world networks properties and scale-free behavior, as observed in social networks. These observations imply that the spreading of e-mail viruses is greatly facilitated in real e-mail networks.

In [23], Andrew McAfee et al. studied the patterns of relationship investments and how powerful firms can successfully build hierarchies. We can state that a university has similar structure than a company. The university is formed by departments, using a hierarchy structure. Moreover, universities, students, professors and researchers are familiar with social networks and collaborative works.

To establish commercial relations between companies we can use B2B standards [24]. As it is shown in reference [25], many companies have their business processes automated by workflow management systems (WFMSs) and have their information in distributed information systems shared by enterprise application integration (EAI) technologies. The current standards, such as XPDL (XML Process Definition Language), Wf-XML (BPM standard developed by the Workflow Management Coalition), WSCI

(Web Service Choreography Interface), and BPEL (Business Process Execution Language), and the technology, such as web services and message-oriented middleware, allow the B2B integration.

Nevertheless these standards only specify the messages between the entities, companies, supply chains or virtual enterprises, but not between the people working inside these enterprises. We are going to extend this standard to exchange information between universities and research institutes in order to establish connections directly between the researchers, but using a selection process. The main purpose of this paper is to propose a more efficient procedure than the existing systems for searching project partners, which is usually done by a human driven search in the World Wide Web or by using research and academic social networks.

The rest of the paper is structured as follows. Section 2 reviews the systems in existence used to find partners for cooperation in research projects. Section 3 presents the B2B architecture proposal, the information retrieval method to find candidates, the procedure to select the best candidate and its analytical description. Section 4 describes the proposed protocol and how it is implemented using ebXML protocol. In Section 5, simulation measurements show how the system performs in several scenarios. Moreover, in Section 6 we have validated the system showing its performance when there are several real searches. Finally, Section 7 gives our conclusion and future work.

2. Related Work

Besides the research and academic social networks and the groups created in the most well known social networks for this purpose, there are several systems to find partners for transnational cooperation or even interregional cooperation [26]. Most of them are based on electronic systems because partners are usually requested to other countries and it is the fastest way to find them.

One of them is through regional forums such as the Lithuania, Poland and Kaliningrad Region of the Russian Federation Neighborhood Programme [27], or through specific topic forums such as YOUTH IN ACTION Partner Searching Forum [28].

There is another type of system that uses a centralized database where each research group can register itself and can seek for partners to collaborate in a common project. Some of them are the CORDIS Partners Service [29] and the Ideal-ist partner search [30]. We can see search platforms to seek partners for specific projects (e.g. the parterns search web site for SME project [31]). There are associations that provide a service to put in contact companies and researchers [32]. Moreover, some projects, such as Collaboration Nets [33] from the University of Valladolid, have been appeared recently.

The most well-known search for partners system is to send e-mails to the universities and research institutes coordinators. Then, they select the most appropriate local research group or person and reply to the requester an e-mail showing their interest on research cooperation.

There are also distribution lists and news servers [34] that help to put researchers in contact, but no one of these systems is automated and there is not published any algorithm for optimal partner selection through the World Wide Web.

Currently, when someone is looking for a contact from other university, research institute or enterprise, without Internet, he/she has to follow a hierarchical structure. Thus, he/she will have to contact his/her department chair, research institute director, or department in charge of external relationships, which will contact the vice rector of external relations, in the case of a university, or the external relations person, in the case of an enterprise or a research institute. This external relations person will contact the external relations person of other universities, which will forward this request to the director of the appropriate departments, which will forward the request to the appropriate researchers.

But this procedure has many lacks. These are the followings:

- If the external relations vice rector or contact person has very few external contacts and agreements with other universities, research institutes or enterprises, the request is distributed to few destinations.
- When a request from another university, enterprise or research institute arrives and the external relations contact person forwards it to the appropriate department. Who decides which the most appropriate researcher is to forward the request (if there are some researchers working in the same topic).
- The chair of the department or the research institute will be a bottleneck if there are many requests.

It is needed a new system to encourage the collaboration between researchers from different universities, research institutes and enterprise research departments. There is no research partner selection such as the one proposed in this paper. In the following sections we detail our proposal.

3. Architecture Proposal

3.1. *Architecture outline*

Our architecture is inspired in a group-based interconnection architecture proposed by the same authors for computer network topologies [35]. Grouping nodes gives better performance to the group and to the whole system, thereby avoiding unnecessary message forwarding and additional overheads. Grouping nodes also diminishes the average network delay while allows to scale the network considerably. We have applied group-based networks to several research areas with success [26-38]. In this paper, each group could be an enterprise, a university or a research center or institute (from now we are going to call them research entities). Moreover, the amount of accessible information experimented in the last years made possible the use of information retrieval systems to search the most appropriate researchers. But, although some of this information can be accessed worldwide, because the entities provide it for free, in many of them, this information can only be accessed internally. Some examples of entities that maintain an updated database and provide it worldwide are shown in [39-42]. But most of them have an internal database (e.g. Senia Application of the Polytechnic University of Valencia [43])

In order to set up our architecture, we will suppose an agreement between several entities that allow sending and receiving searches about the research performed by their researchers. Every research entity joined to the architecture has a database with all their researchers structured in departments, research groups or research areas and areas of knowledge. Usually, all enterprises, universities and research institutes have this type of structure. Every time a researcher is registered to the database of an entity, he/she joins a department (or research group) and to an area of knowledge. In order to avoid having researchers without research topics, the first time each researcher has to write his/her research keywords (its research profile) in that database. Later, every time the researcher publishes a paper, a technical report, or has a new accepted grant, this information is added to the entity's database for this user (we have seen that it is regular procedure in many well known universities [39-41][43] and enterprises [42]). So we will suppose that each entity maintains and updated research database.

In our proposal, when a researcher needs a partner for his/her research project, he/she will send a request to the local entity database which will send the request to other research entities databases. Then, they will provide a list with the most appropriate researchers ordered by their capacity parameter (the capacity parameter will be defined later) which will be forwarded to the requester researcher.

3.2. *Architecture parameters*

We pretend to implement an architecture that could be used worldwide. Some parameters such as country and entity identifiers are needed. The country let us seek for researchers from the same country (for national projects) or for different countries (international projects), and the entity let us filter by universities, research centers, enterprises, etc. Moreover, the researcher, his/her area of knowledge and the list of topics where he/she is researching, have to be included. This list is needed in order to know the topics where the researcher is working on. We discuss how the list is built in the following section. Now, we are going to define the parameters and the information used in our proposal. All of them are based on identifiers, so messages will not be network layer dependent, thus they can be easily deployed over HTTP.

3.2.1. *ResearcherID*

It is the researcher identifier. It is unique for each researcher. It can be assigned sequentially to new researchers of an entity. A researcher can't join the architecture without a *researcherID.* Every time the researcher requests for some information, the *researcherID* is added to the message in order to know who is requesting the information.

3.2.2. *AreaID*

It is the identifier of the area of knowledge. Although this identifier is assigned to the researcher when he/she is registered to the architecture, equivalences between areas of knowledge from different countries may be created (e.g. many countries use the codes defined by UNESCO [44] as area of knowledge). Each country has well defined areas of knowledge for all their entities. This identifier is needed because any topic can be researched by different types of areas of knowledge. Any department or research group can be formed by researchers from several areas of knowledge.

3.2.3. *EntityID*

It is the entity identifier. When an entity joins the architecture it receives its *entityID* that will be permanent. It is unique for each entity. In our proposal we suppose that a researcher can not be member of more than one entity, but the system allows having researchers in more than one entity. The identifier *entityID* is only used when there are messages between different entities, not between researchers in the same entity.

3.2.4. *CountryID*

Every entity joined to the architecture has a *countryID*. When an entity joins the architecture its *countryID* is assigned. This identifier will allow the researcher to send requests to several entities of the same country at the same time. An international entity may have one *entityID*, but several *countryIDs*. The system will be able to distinguish an entity with several branches in different countries using the *entityID* and the *countryID* parameters.

3.2.5. *Topic list*

Every researcher has a list of topics which is built by several items. It is not just a list, but also links to the research keywords, published papers, technical reports, accepted grants, patents, etc., that is, every text that could provide information about the topics that are being researched by the researcher (this avoids knowledge withholding [45]). Researchers from different areas of knowledge (*areaID*) could research similar topics. The topic list is dynamic and usually increases over the time.

3.3. *Information retrieval searching method*

The main issue in our architecture is the procedure to find researchers from other entities, that are researching in the requested topic. In order to achieve this goal, we have used an Information Retrieval (IR) technique. Today IR is an open problem, however many advances have been produced in some fields [46].

In order to find the appropriate researcher to contact with, we define the following information retrieval elements:

- The query: The user query is a sequence of words.
- Information database: It is built using the research documents (papers, technical reports, research projects, patents, etc.). It is used to know the research areas where each researcher works. A suitable representation of documents is essential for an efficient IR. We select a first dimension representation based on probabilistic language models. As we will show later, we infer a document language model for each researcher using his/her works. Thus it will not be dependent on the type of engine, structure or architecture.
- Output: In IR, a query could provide more than a single result. Several researchers may match the query, perhaps with different degrees of relevance. The system assigns a score to every researcher and ranks researchers by this score.
- The user: In our specific IR framework, we can use the information provided by the query of the user in order to improve the system. That is, depending on the case, it is more probable to search for topics close to the topics of the requester researcher or vice versa.

The use of statistical approach to natural language area has been proved to be valuable in tasks such automatic speech recognition and machine translation. Nowadays, statistical methods are becoming the most outstanding approach to IR. The initial idea of probabilistic retrieval was proposed in 1960 [47]. Later, some authors [48, 49] proposed to use the methods used in statistical machine translation (SMT) for IR systems. Our research group has long experience in statistical machine translation, thus we have adapted our methods to IR [50-52].

Our IR proposal is based on the definition of a function that estimates the most likely researchers (D) given the query ($Q = (q_1, q_2, \ldots, q_m)$) sent by the user (U). That is, those D for which $\Pr(D|Q, U)$ is highest. Using Bayes' theorem, we can write equation 1.

$$\Pr(D|Q,U) = \frac{\Pr(Q|D,U)\,\Pr(D|U)}{\Pr(Q|U)} \qquad \mathbf{(1)}$$

Given a query, the denominator, $\Pr(Q|U)$, is fixed for all documents, then we can ignore it for the purpose of ranking documents. In this point, we have decomposed the initial equation into two terms:

- $\Pr(Q|D,U)$. It is a query-dependent term measuring the proximity of Q and D. Frequently, the query is independent of the user, so $\Pr(Q|D, U) \approx \Pr(Q|D)$.
- $\Pr(D|U)$. It is a query-independent term that measures the relevance of a document according the user information needs. It is also called the "prior" term.

So, equation 1 can be written as equation 2.

$$\underset{D}{\operatorname{argmax}} \Pr(D|Q,U) = \underset{D}{\operatorname{argmax}} \Pr(Q|D) \Pr(D|U) \tag{2}$$

In order to estimate Pr(Q|D) in equation 2 we use the well known language model approach [48]. In the simplest case, each query term is assumed to be independent of the other query terms, so that the query likelihood is given by equation 3.

$$\Pr(\mathrm{Q}|\mathrm{D}) = \prod_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{q_i}|\mathrm{D}) \tag{3}$$

Where p($\mathrm{q_i}$|D) is a statistical model over the terms estimated for each D research document. Zero probability can be assigned to query terms, so that language models must be smoothed. Thus we used the linear interpolation. Then, we estimate a language model using the information provided about the researcher from the database. The query language model has been used to model user preferences, the context of a query, synonymy and word senses in other previous proposals [53].

The first time each researcher is added to an entity's database, he/she has to write his/her research keywords in the university database. Later, every time the researcher publishes a paper, a technical report or has an accepted grant this information is added to the entity’s database (we have proved that it is a regular procedure in many universities [39-41][43] and well known enterprises [42]), that is, it will form the researcher topic list. This information will be used when researchers from the same entity or other entities are seeking for partners.

We have added the Pr(D|U) distribution, which tries to estimate the similarity grade between the requester and the researchers from other entities with closest topics. In order to estimate Pr(D|U), we used equation 4.

$$\Pr\,(\mathrm{D}|\mathrm{U}) \;=\; \Pr(\mathrm{D},\mathrm{U}) \;= \sum_{\mathrm{t}} \mathrm{p}(\mathrm{t}|\mathrm{D})\;\mathrm{p}(\mathrm{t}|\mathrm{U}) \tag{4}$$

We have included the terms provided by equations 3 and 4 in equation 2 and we have used a log-linear combination [51]. Then, equation 5 is obtained.

$$D' = \underset{D}{\operatorname{argmax}}\{\delta_1 \log \Pr(\mathrm{Q}|\mathrm{D}) + \delta_2 \log \Pr(\mathrm{D}|\mathrm{U})\} \tag{5}$$

It allows us to control the influence of each term. For example, our task query-dependent term may be more important than the prior term. We estimate the δ parameters using the minimum error rate criteria [52].

3.4. *Best candidate selection*

Now, let us suppose that a researcher needs to find partners from other research entities in order to start a proposal for a cooperative research project. This query should have several words that will be used to find the appropriate researcher. But, the main problem, when a search is started, is the availability of the researchers, that is, if the researcher is too busy, because he/she has too many projects or works running, he/she will have low availability for a new project. So, we must define several parameters in order to know which researcher will be the most appropriate for that request.

We define λ parameter as the researcher capacity (or availability). It depends on some parameters that are introduced manually and others that are estimated by the system. Those parameters are shown in table 1. Table 2 shows which parameters are introduced manually and which parameters are calculated by the system. This is our initial proposal, but their origin could be changed.

Table 1. Parameters used for λ parameter estimation.

Parameter	**Description**
N	Number of local collaborators working with him/her. They are obtained from a local list.
E	Equipment parameter. It is estimated by taking into account the equipment owned by the researcher. Higher parameter indicates that the researcher has more equipment to perform his/her work.
Max_P	Maximum number of collaborative projects. It could be limited by the entity or by the researcher.
P	Number of running collaborative projects. It is estimated by the system.
H	Hours per week dedicated to collaborative research projects. It could be limited by the entity or by the researcher.
S	Number of successful finished projects

Table 2. Origin of the parameters used for λ parameter estimation.

Parameter	**Introduced by the researcher**	**Introduced by the system**
N	X	
E		X
Max_P	X	
P		X
H	X	
S		X

λ parameter is defined by equation 6. It will be used to determine the best researcher to connect with.

$$\lambda = \frac{E \cdot \log_2 (N+1) \cdot H \cdot Max\_P}{(P+1)} + S \quad \textbf{(6)}$$

Where $0 \leq P \leq Max\_P$. E varies between 0.1 and 1 and it is given by the system (taking into account researcher's equipment infrastructure). We can introduce the hours of each collaborator, but we have seen that it does not affect too much in λ values. Figure 1 shows λ parameter values for different number of collaborators and running collaborative projects. We have fixed several parameters for this case: the researcher has a lot of research equipment (E=1), the researcher has five successful finished projects (S=5), the maximum number of simultaneous collaborative projects is 5 and the working time established to collaborate in research collaborative projects is 15 hours. We have limited the graph to 12 local collaborators. The graph shows that a researcher has quite more λ parameter when he/she has very few projects running so it is more feasible to be chosen for a new collaborative research project (which is our intention). Moreover, although many collaborators provide higher λ parameter values, the more number of projects running the lower λ.

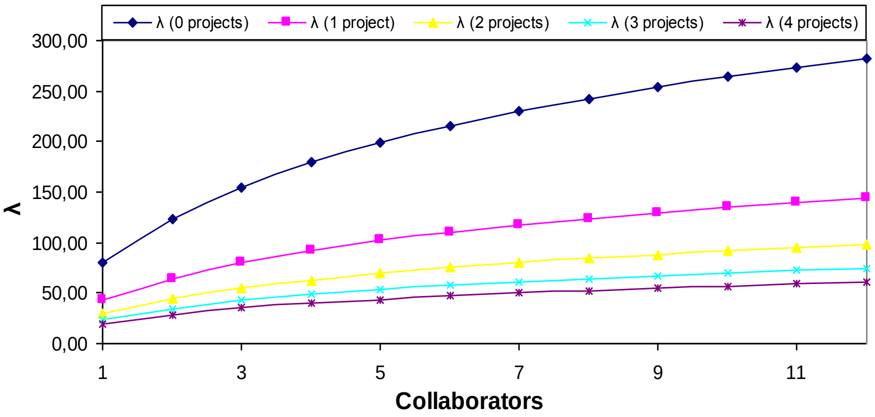


Fig. 1. λ parameter values.

3.5. *Information retrieval implementation and evaluation*

We have implemented a research prototype retrieval engine to test our approach. This engine is based on the statistical system described in [54-55]. It has two phases. The first phase is the corpus recompilation used to train the language models for each researcher. We use the information from the entity's database which contains the researcher's keywords, papers, technical reports and accepted grants. In order to extract the information from the entity's database, we use the data mining system described by us in [54]. Sentence fragmentation and tokenization is also performed by this software.

Punctuation marks and frequent words are eliminated in the tokenization phase. In the second phase, we estimate a language model from each researcher using the text of the previous phase. Unigram and bigram language model are trained. A linear interpolation smoothing technique is used [51]. In order to improve efficiency we pre-estimate the Cartesian product given in equation 4 between each pair of researchers.

In order to evaluate our IR system we prepared a set of data formed by 20 researchers from the same research area as an approximation of a real word retrieval. Four queries have been proposed for each researcher and a corresponding group, that matches the query, has been manually prepared. The query has an average length of 2.7 words, and a mean of 4.3 researchers are assigned to one query.

Two different experiments have been performed. The first one evaluates the influence of the new prior model proposed ($\Pr(D|U)$). For that purpose, δ_1 has been set to 1 in equation 5 and different values of δ_2 has been tested. Figure 2 shows the results. When δ_2 is set to 0 it means that no prior model is used. We can observe that we obtain a significant improvement using a value of $\delta_2 = 0.15$.

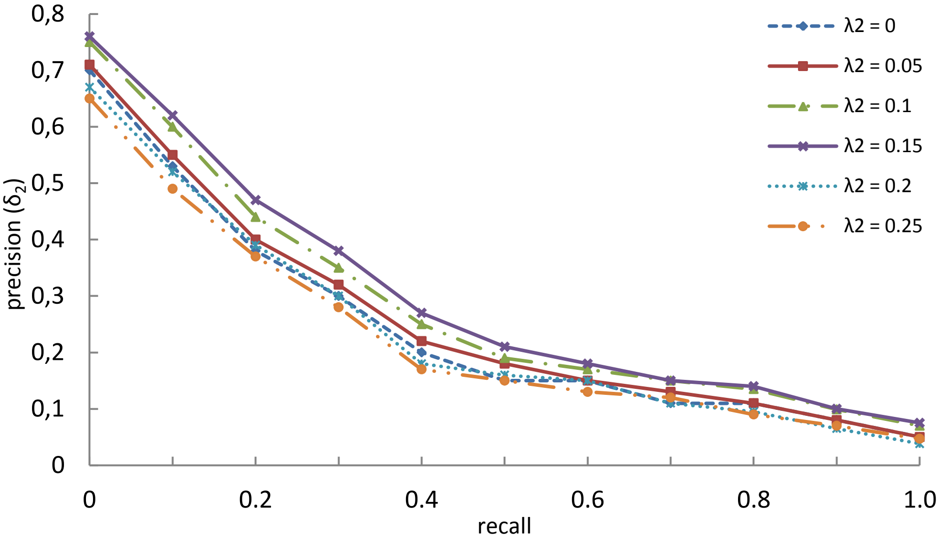


Fig. 2. Precision-recall curves for several values of δ_2 in equation 5.

The second experiment compares the use of unigram and bigram language model in equation 5 for $\Pr(Q|D)$. Figure 3 shows the results using $\delta_2 = 0.15$. Best results are obtained for unigram model. This result is given because there is few data to train the language models.

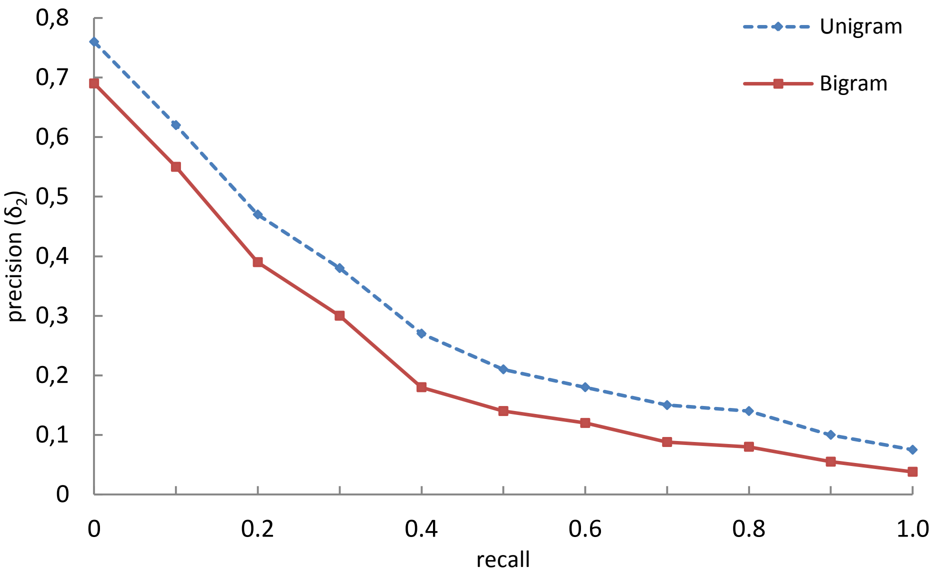


Fig. 3. Precision-recall curves for unigram and bigram language models.

3.6. *Analytical description of the Architecture*

We use graph theory to define the network topology. Let us suppose that all researchers in the system form the whole network. Let $G = (V, E)$ be a network of researchers, where V is a set of entities in the network, and E is a set of connections between researchers. Let k be a finite number of entities of V, so the whole network of researchers is $V=\cup(V_k)$. There is not any researcher in two or more entities ($\cap V_k = 0$). Each node has associated a λ parameter which indicates the capacity of the researcher to join new projects. Let's suppose $n=|V|$ (number of researchers in network) and k the number of entities of V. Equation 7 gives the number of researchers.

$$n = \sum_{i=1}^{k} |V_k| \qquad (7)$$

On the other hand, the number of connections (E_k) for each entity (V_k) will depend on the number of researchers inside the entity and their capacity. An entity can have connections with other entities (even with all other entities). The number of connections in the whole network $m=|E|$ depends on the number of entities (k), the number of researchers in each entity (r_k) and the number connections that a researcher has with researchers from other entities (l). We will suppose that there is a maximum limit in the number of projects of a researcher. A researcher could not be in more than max(l) projects simultaneously. Equation 8 gives the m value.

$$m = \frac{1}{2} average(l) \cdot \sum_{i=2}^{k} r_k \qquad \textbf{(8)}$$

Where *average(l)* is the average number of projects per researcher in the whole network. Figure 4 shows a particular case for 50, 100, 500, 1000 y 2000 entities as a function of the mean number of researchers in the entities (starting from a mean value of 50 researchers) with an average of 3 research cooperative projects for each researcher. We can observe that the increment is higher when more entities are joined to the network than when the mean value of researchers in the entities are incremented.

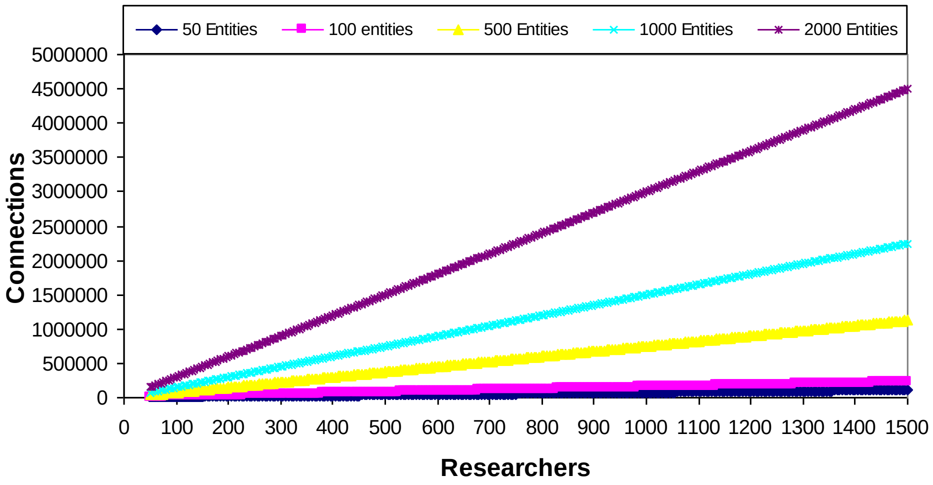


Fig. 4. Number of connections in the whole network.

4. Architecture Protocol

In order to exchange messages between research entities through Internet, we chose ebXML packets over HTML. First, ebXML servers must be registered in an ebXML registry using Collaboration Protocol Profiles (CPPs). When the researcher sends its request, if the request is only for local researchers, the request is sent to the local database. But when a researcher is looking for researchers from other research entities, the request is sent through a local smart proxy which uses ebXML packets to communicate with the databases of the other entities (using ebXML translators placed in their network). Figure 5 shows the described procedure.

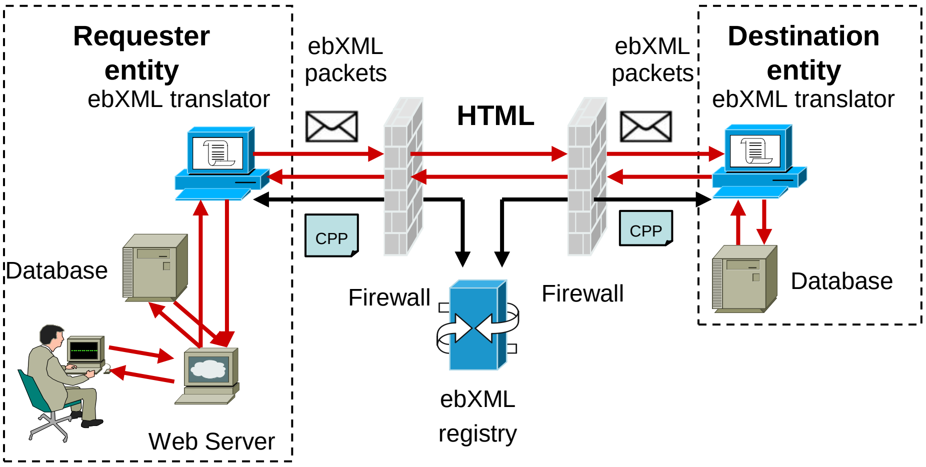


Fig. 5. Request Procedure.

When a researcher joins the architecture for its first time, it has to send (or update) its information in the local server through the *Update Profile* message. This message contains its area of knowledge and a list of research keywords.

When a researcher wants to start a new research collaboration project and needs to contact with a researcher from other entity, he/she opens the local application form in a web server and fills up the following fields:

1. Destination country (it could be the same country for national projects).
2. Destination entity (it could be universities, research centers, enterprises or all).
3. Area of knowledge.
4. Research topics.

When “all” is selected in the second field, the query is forwarded to all entities in the selected country (all entities have associated a country identifier) using the *Request for partners* message. The server forwards this request to the appropriate destinations using the *Request for partners (fwd)* message. When this message reaches the other entity database, the appropriate researchers are searched based on the area of knowledge identifier (first filter) and the topic list (second filter). Then, the database replies to the source entity server with a list of selected researchers ordered by their capacity parameter though the *Selected researchers* message. Finally, this list is forwarded and displayed in the requester’s researcher computer. When the requester chooses one of the researchers in the list, the agreement procedure is started by sending a *request for agreement* message directly to the e-mail of the selected researcher. Then, after exchanging a series of messages to reach an agreement, selected researcher can accept or reject this offer through a *reply agreement* message. We have simplified this process by adding just one message (reply agreement), but it is expected a negotiation process. A work that studies how individuals decide to collaborate with other groups is shown in [56]. Figure 6 shows the procedure from the search point of view. It also shows that there could be no response if there is not any researcher that matches the search parameters. We preferred to design our system avoiding sending unnecessary messages in order to avoid network overloads when there are many entities and searches running at the same time.

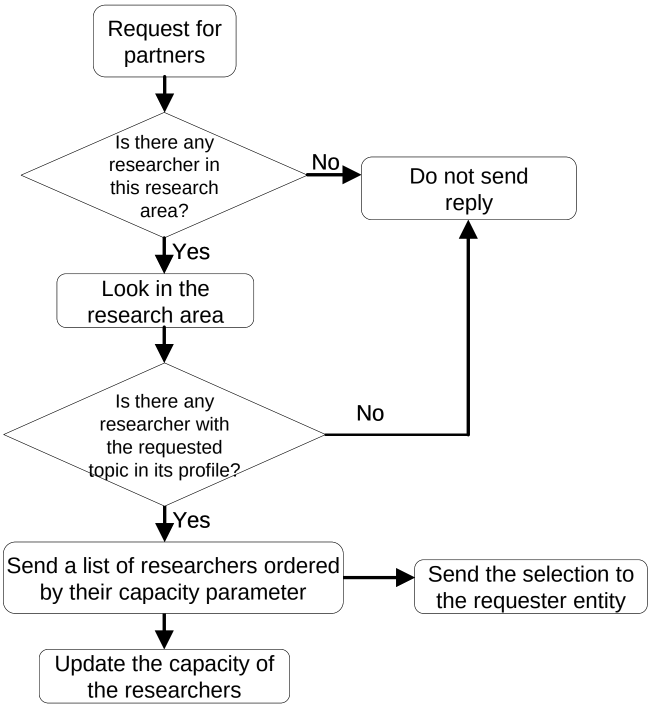

Request for partners
Is there any researcher in this research area?
No
Do not send reply
Yes
Look in the research area
Is there any researcher with the requested topic in its profile?
No
Yes
Send a list of researchers ordered by their capacity parameter
Send the selection to the requester entity
Update the capacity of the researchers

Fig. 6. Search algorithm.

Figure 7 shows the flow diagram of the aforementioned messages. Recall that the agreement phase could imply several messages allowing a negotiation process.

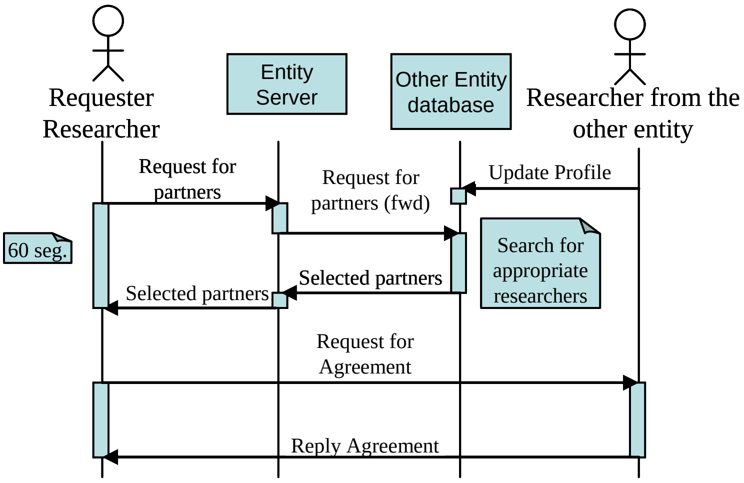
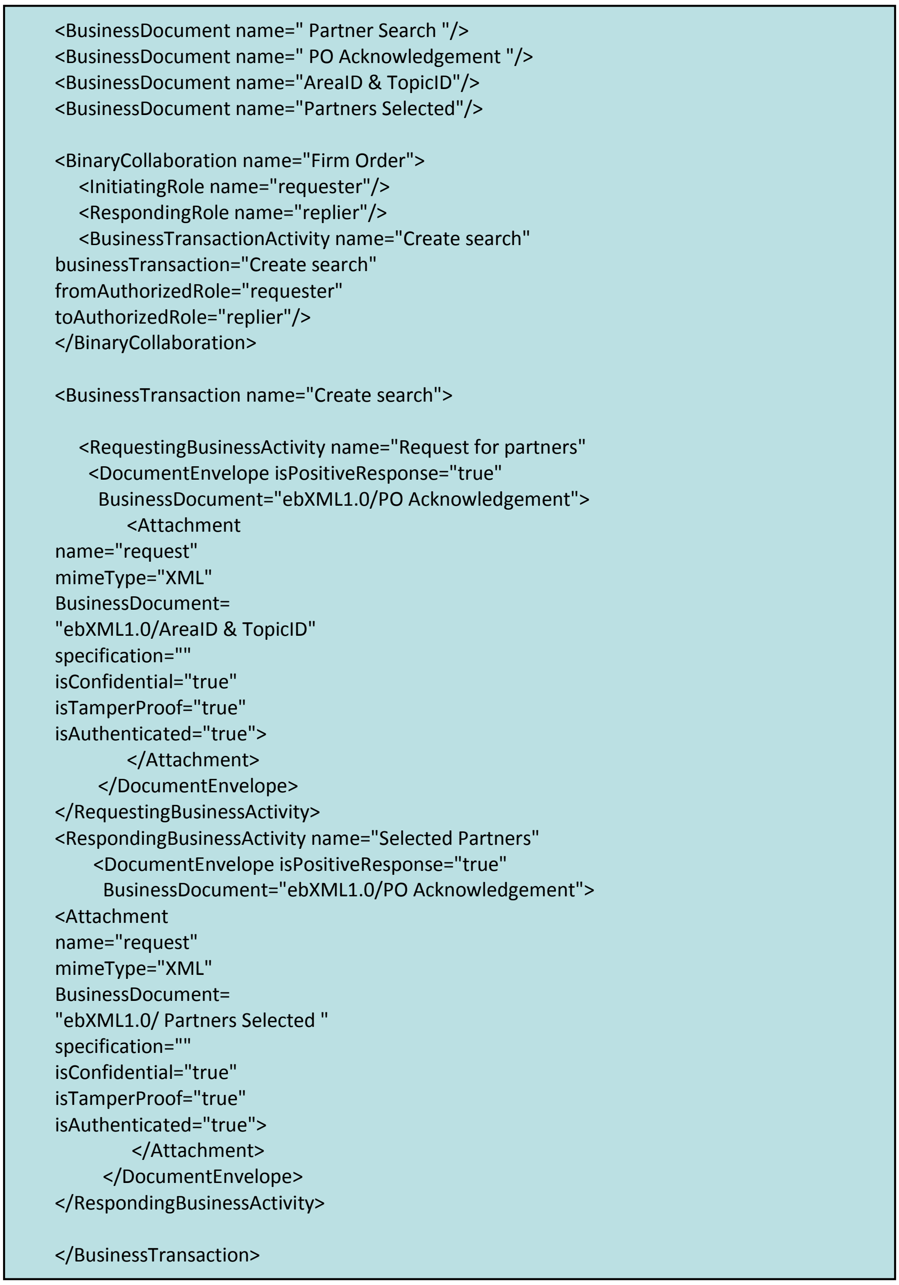


Fig. 7. Messages of the architecture.

Figure 8 shows a piece of the code for the entity-to-entity transaction in ebXML syntax. It is included the request for partners message and the selected researchers response. If there is not any researcher matching the search parameters, there is no response to avoid network overload. Documents exchanged in the transaction (with the area of knowledge and the research topic) are written in XML.



```
<BusinessDocument name=" Partner Search "/>
<BusinessDocument name=" PO Acknowledgement "/>
<BusinessDocument name="AreaID & TopicID"/>
<BusinessDocument name="Partners Selected"/>

<BinaryCollaboration name="Firm Order">
   <InitiatingRole name="requester"/>
   <RespondingRole name="replier"/>
   <BusinessTransactionActivity name="Create search"
businessTransaction="Create search"
fromAuthorizedRole="requester"
toAuthorizedRole="replier"/>
</BinaryCollaboration>

<BusinessTransaction name="Create search">

   <RequestingBusinessActivity name="Request for partners"
    <DocumentEnvelope isPositiveResponse="true"
     BusinessDocument="ebXML1.0/PO Acknowledgement">
         <Attachment
name="request"
mimeType="XML"
BusinessDocument=
"ebXML1.0/AreaID & TopicID"
specification=""
isConfidential="true"
isTamperProof="true"
isAuthenticated="true">
         </Attachment>
     </DocumentEnvelope>
</RequestingBusinessActivity>
<RespondingBusinessActivity name="Selected Partners"
     <DocumentEnvelope isPositiveResponse="true"
      BusinessDocument="ebXML1.0/PO Acknowledgement">
<Attachment
name="request"
mimeType="XML"
BusinessDocument=
"ebXML1.0/ Partners Selected "
specification=""
isConfidential="true"
isTamperProof="true"
isAuthenticated="true">
          </Attachment>
      </DocumentEnvelope>
</RespondingBusinessActivity>

</BusinessTransaction>
```

Fig. 8. Entity transaction in ebXML specification scheme.

4.1. *Tables in the entity database*

In order to send requests and select the best researchers, there has to be several tables in the database. They are described in the following subsections.

4.1.1. *Entities table*

This table is used to associate the entities joined to the architecture with their country and their database server IP address (or DNS name). It allows knowing the database to connect with in case of a search, while it also allows searching by country (the same or foreign countries), the same entity in different countries (in case of international entities) and even the type of entity (universities, research centers, enterprises…). Table 3 shows an example with its fields.

Table 3. Entities table.

Country	EntityID	Type of Entity	Database server address
Country 1	Entity 1	Type 1	Entity1.country1.org
Country 1	Entity 2	Type 2	Entity2.country1.org
Country 2	Entity 1	Type 1	Entity1.country2.org
…	…		…

4.1.2. *Researchers table*

This table is used to search researchers when a request is received from other entities (it could also be used bythe same entity). The table associates every researcher with its area of knowledge and topics. It has also a field where the estimation of the capacity parameter is stored in order to have faster searches. Table 4 shows its fields.

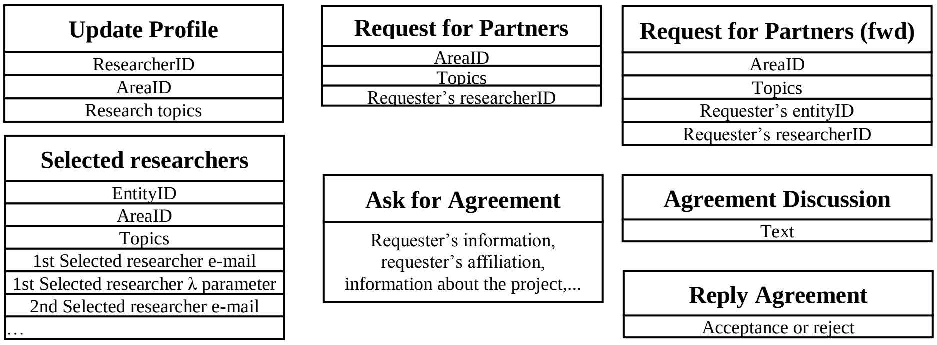
Table 4. Researchers table.

ResearcherID	Full Name	e-mail	AreaID	List of topics	λ
1	Name and surname 1	name1@entity1.org	1	Link to keywords database Link to papers database Link to tech. reports database Link to grants database …	λ_1
2	Name and surname 2	name2@entity1.org	2	Link to keywords database Link to papers database Link to tech. reports database Link to grants database …	λ_2
…	…	…	…	…	

All this information can be gathered easily from the research database entity just with simple database query messages (they exist in the most well known databases such as SQL and Oracle).

4.1.3. *Information exchanged in the protocol messages*

In order to run the architecture properly, messages must have the right information in each step. Figure 9 shows the messages needed and the information that must be inside each of them.



Update Profile
ResearcherID
AreaID
Research topics

Selected researchers
EntityID
AreaID
Topics
1st Selected researcher e-mail
1st Selected researcher λ parameter
2nd Selected researcher e-mail
…

Request for Partners
AreaID
Topics
Requester's researcherID

Ask for Agreement
Requester's information, requester's affiliation, information about the project,...

Request for Partners (fwd)
AreaID
Topics
Requester's entityID
Requester's researcherID

Agreement Discussion
Text

Reply Agreement
Acceptance or reject

Fig. 9. Information exchanged in the messages.

The reply agreement message is the shortest message. The messages that are larger are *selected researchers* message and *ask for agreement* message. The first one depends on the number of researchers found in the database matching the search parameters. The second one depends on the information that is wanted to be sent in the message in order to explain the project. *Agreement discussion* message could be very large or very short.

5. Simulations

The purpose of this section is to show the performance of the proposed architecture in several cases. Our protocol has been simulated in 2 scenarios with the purpose of demonstrating that our protocol overloads neither the entity network nor Internet, thus it is a lightweight protocol, and the servers can perform their tasks without the need of very high processing capacity and memory. In order to perform these measurements, we have used OPNET Modeler simulator [57].

5.1. *Test bench*

This sub-section presents the test-bench used to test our architecture. We have implemented it in two scenarios. We used an IP cloud to simulate the behavior of the interconnection of all entities in both scenarios. IP cloud parameters have been modified in order to obtain the more accurate measurements in our simulation. We added a delay of 2 seconds to the delivered packets. There has been a reject packet rate of 1%.

The computers used by the researchers have the following features: a PC with a processor Pentium IV, a RAM memory of 1024 MB and a FastEthernet connection to their local network. Moreover, the servers that store the databases have more processing capacity. In the simulation, we used Pentium IV servers with 4 processors, 4 GB of RAM memory with 2 point to point connections to Internet and 4 FastEthernet connections to the entity network.

In order to simulate the exchange of information we used a HTTP traffic profile called “searching” which was the one that fits better our case. The type of traffic was application protocol HTTP 1.1. The average time between entries (data requests) followed an exponential distribution of 10 seconds of average. The data requested had a constant size of 1200 bytes. This traffic was served following an exponential distribution.

5.2. *First test bench: One researcher requests 31 entities*

The first case shows the performance of the database when there is a pool of 16 researchers in one entity requesting for partners continuously. This request is sent to 31 database servers, that is, each searcher sends a call for partners to 31 entities. Figure 10 shows the described topology. We measured the time to receive the data. We also compared the traffic sent versus the traffic received (in bytes per second and in packets per second), and the load of the server in the outbound interface. All these measurements showed us the behavior of the whole network.

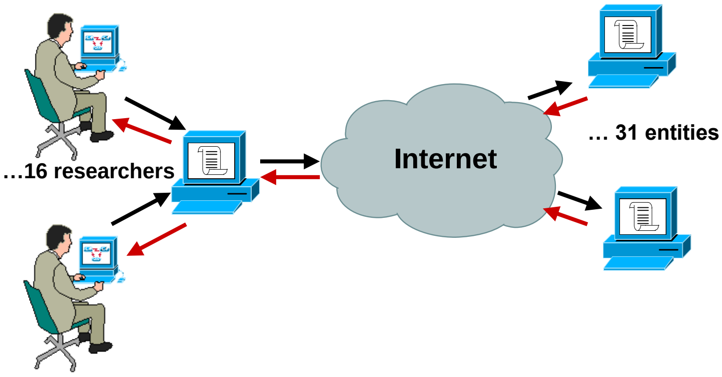


Fig. 10. 16 researchers searching.

Figure 11 shows the average response time received by the requesting computers. Although some packets in the beginning arrive at 2.5 seconds, the mean response time value is about 3 seconds.

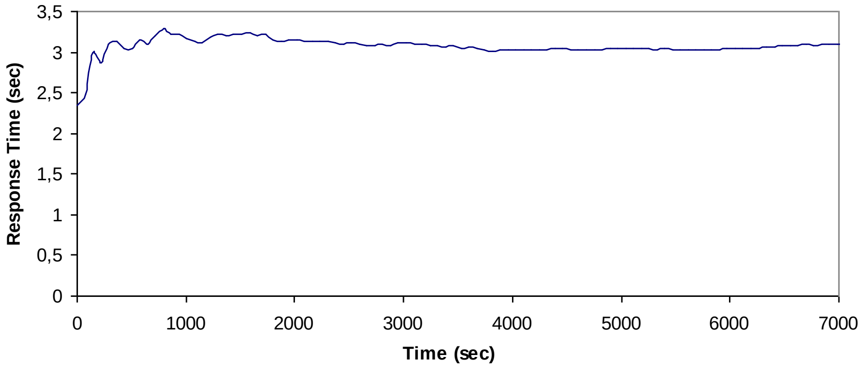


Fig. 11. Response time for 16 researchers.

When we compare the traffic sent versus the traffic received, we observe that the traffic received is just around 2.4 times more (see figure 12). It is not so much if we bear in mind that there are 31 database servers answering the reply. It happens because not all the entities databases have data that match the request, and because the replies are spread over the time.

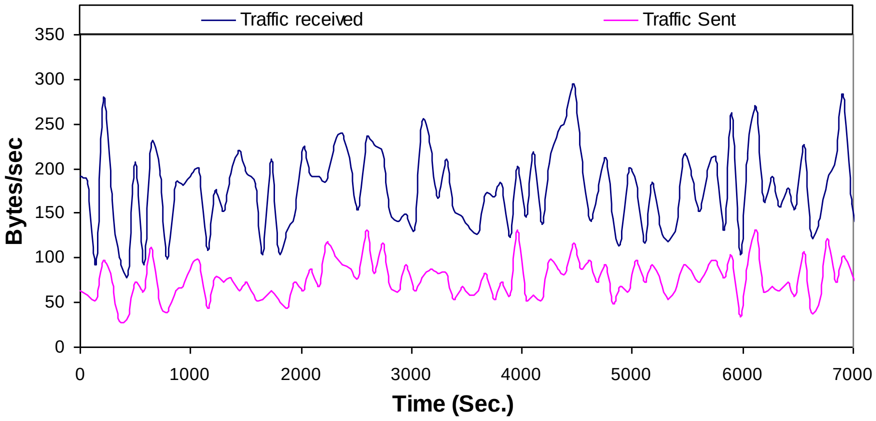


Fig. 12. Bytes/sec sent and received for 16 researchers.

Figure 13 shows the number of packets per second sent versus the number of packets per second received. In this case we observe that the number of packets per second sent is a little bit higher than the number of packets received. It usually happens when some entities do not reply because there are no partners that match the query.

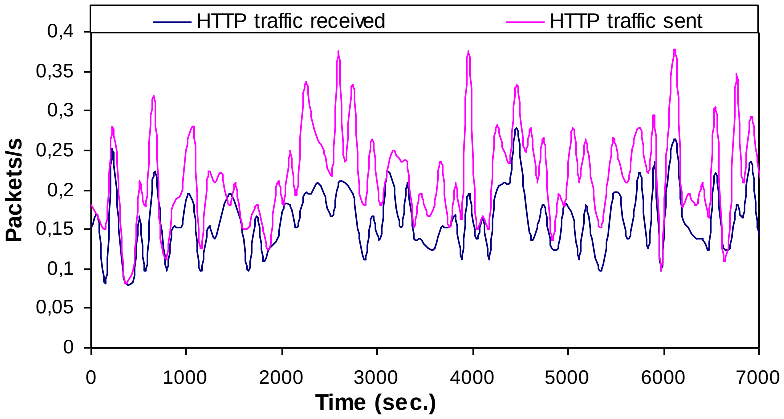


Fig. 13. Packets/sec sent and received for 16 researchers.

We also wanted to simulate the load of the server in the outbound interface of the entity. It is shown in figure 14. Although its average is between 4000 and 5000 bytes at the beginning of the simulation, it comes a little bit higher as time goes on.

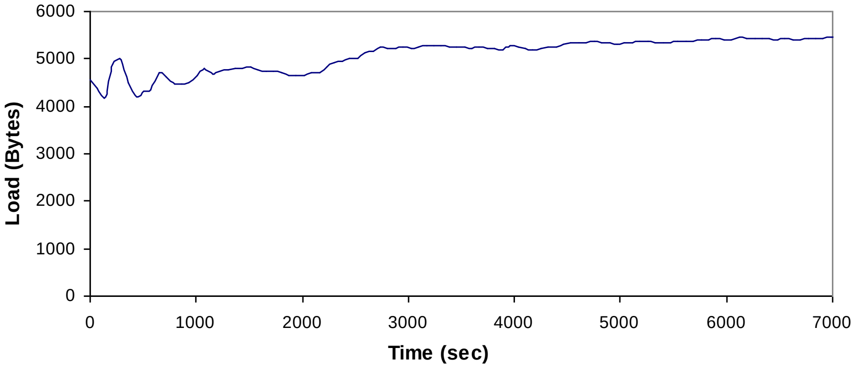


Fig. 14. Load (in Bytes) for 16 researchers.

5.3. *Second test bench: A network of 32 entities with 16 researchers inside each one*

In this scenario we sought to simulate a network with 32 entities where there are 16 researchers requesting at the same time in each entity. The topology used for our simulation is shown in figure 15. This simulation let us to know the performance of the network in an environment closer to the real life. Real deployments will have more entities and more researchers, but not all researchers are searching at the same time, so we believe that this amount is the most probable when there are many entities.

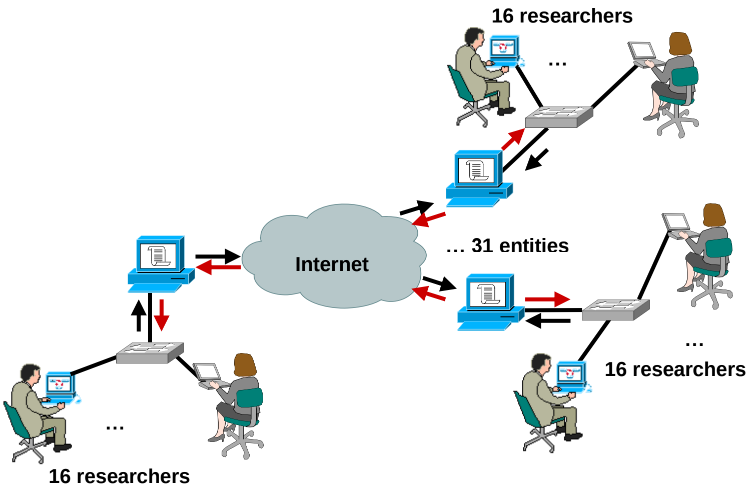


Fig. 15. A network of 32 entities.

In this case, we are going to simulate the average response time of a request and the number of bytes per seconds requested and received in the whole network.

In figure 16, we see the average response time when the 16 researchers of the 31 entities send requests continuously. The mean value is 7.17 seconds, but this value is not achieved until 3960 seconds approximately, so the system takes some time to be stabilized. This delay is given by the network delay and the amount of requests received in the server database.

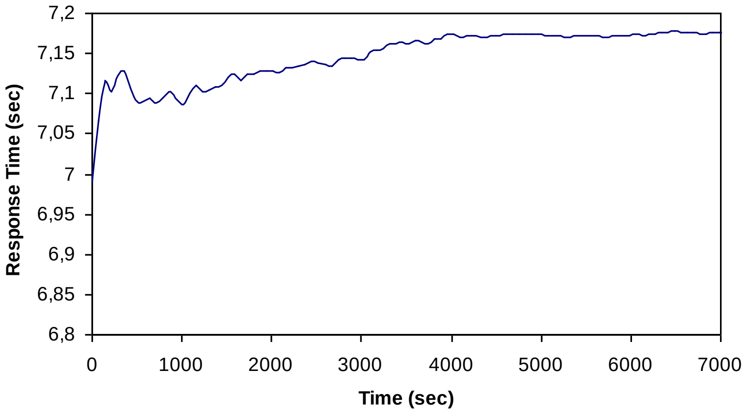


Fig. 16. Average response time for 32 entities.

In figure 17 we see the amount of HTTP traffic in the network (measured in bytes per second). We can see that the traffic sent and received is quite low. We observed that there was 50 Kbits/s for the traffic received by the researchers and 70 Kbits/s for the traffic sent by the researchers. In this case there is more traffic sent by the researchers because in this scenario there are many more researchers, but not all database servers have data that match the information requested.

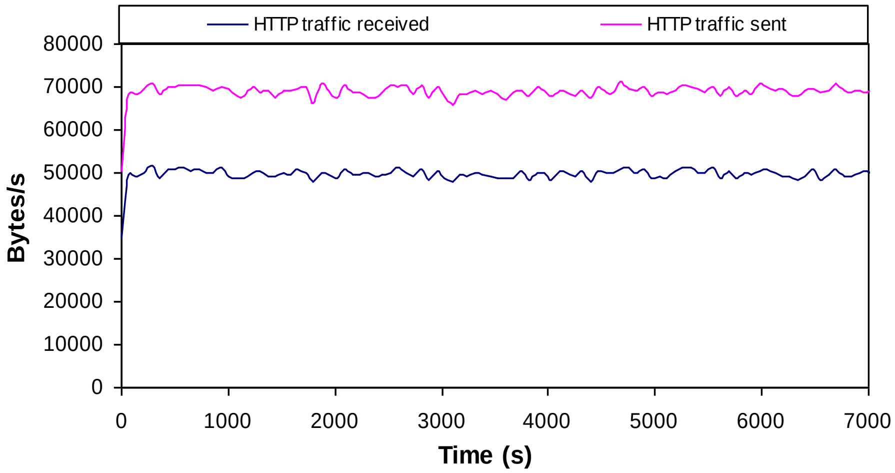


Fig. 17. Traffic received and sent for 31 entities.

6. System validation

In this section we validate our system by showing how it performs when some queries are sent to the system.

6.1. *Subjective end user validation*

In this section we validate the search method used to find the best candidate. As it is explained in sections 3.3 and 3.4, we propose the combination of an IR statistical approach with an analytic equation for the best candidate selection. Training and parameter estimation of the statistical models have been shown in section 3.5. In order to validate the entire search system on real data, we propose the following experimental framework.

Polytechnic University of Valencia has been selected to perform our test because it has an updated research database which is available to perform searches. It is provided through the Senia application [43]. Tests have been conducted within the area of knowledge “Sensor System Design” (UNESCO code 1203.25) and within this area, we have selected 20 researchers. For each researcher we have verified that all information required by the system to evaluate Equation (6) is included correctly. In order to train the information retrieval statistical models, we take into account all publications of each researcher and the projects where he/she is involved. δ_2 has been set to 0.15 and unigram model has been used as the language model.

We have proposed 5 questions in this area of knowledge, some very general and others quite more specific. The questions are shown in Table 5. For each one of these questions, an expert on this field has selected 6 researchers from the test and clasiffied them by degree of relevance. Then, these 5 questions have been introduced into our automatic system to compare them with the human results. In order to evaluate it we used two measures: accuracy (percentage of elements found in both outputs) and the Levenshtein distance [58] between the sequence output by the system and the reference sequence (the minimum number of insertion, deletion, substitution needed to transform one output into the other). The results are shown in Table 5.

Table 5. Evaluation results for the five questions performed in the test.

Query	**accuracy**	**Levenshtein distance**
sensors	50%	5
acoustic sensors	66%	3
aquatic sensors	83%	2
acoustic aquatic sensors	83%	1
acoustic sensors for Seabee classification	100%	1

We can see that when we perform specific searches the system provides very satisfactory results. In a more general search, results are poor. This may be due to the difficulty of discovering the relevant system of a researcher with the information provided and the difference of criteria between the automatic system and the human evaluator.

6.2. *Subjective end user validation*

In order to validate the end user usability and satisfaction with our system, we have carried out a set of subjective assessment tests. We have selected 16 researchers from different academic institutions; all of them with thorough knowledge of the research staff of the UPV. These evaluators used our system with the same query as in the previous test: acoustic sensors. After that, they filled in a survey. The questions and answers included in the survey are shown in Table 6. Results are shown in Figure 18. The average obtained for each question (from question 1 to 5) has been 4.25, 4.44, 4.5, 4.56 and 4.31 respectively. The maximum value has been 5 in all questions and the minimum mark has been 3 in all questions except question 3, which minimum value has been 4. These resuts demonstrates that researchers would use this system to look for partners in cooperative projects.

Table 6. Questions and answers included in the survey.

Questions
1. Do you think that the obtained list is properly sorted?
2. Do you consider that the speed of response is appropriate?
3. Was the system easy to use?
4. Do you consider useful this system to look for parents?
5. Would you use the system again?
Answers
1 - strongly disagree
2 - disagree
3 - indifferent
4 - agree
5 - strongly agree

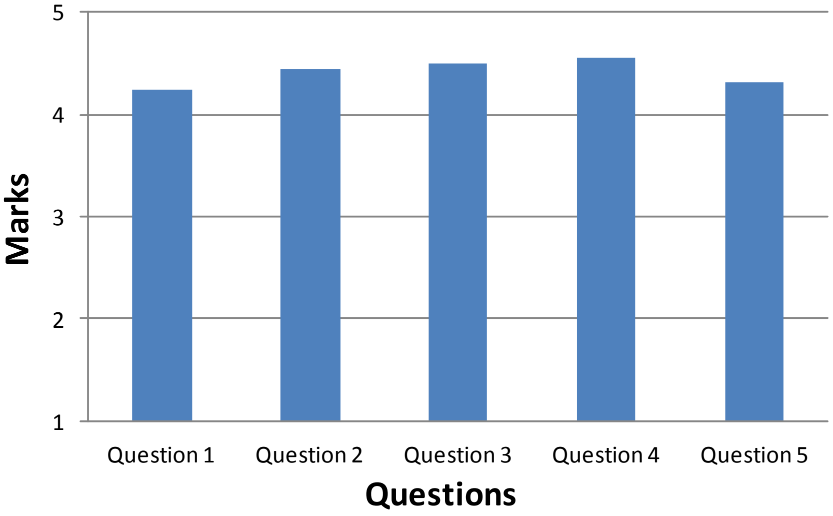


Fig. 18. Subjective end user validation results.

7. Conclusion and Future Work

Current search for partners systems that are being used for the cooperation of enterprises, universities and research institutes are based in a centralized scheme, in a human driven search in the World Wide Web or using research and academic social networks. This paper proposes a more efficient procedure than the existing systems for searching partners by using a B2B architecture that avoids a centralized structure and allows researchers cooperation using the databases of the universities, research institutes and enterprises. Our proposal allows the researcher to search for partners without any person mediating the search.

We have detailed the method to search and provide a list with the best candidates, given some search topics. This method is based on an information retrieval system.

The proposal has been simulated in a controlled environment with several entities from several countries. We can see that it is a feasible architecture with a lightweight protocol because there is low traffic sent through the network when the system is working. These excellent results let us implement it in Internet without problems. Moreover we have performed several searches in order to show the performance of the system and how several researchers see the obtained results. Obtained results show that researchers are confident with our system and they will like to use it.

If the requester always chooses the research partner with higher λ parameter from the other entity, the connections between researchers will tend to be balanced (there will not be researchers with quite more connections than others).

Now, we are adding online translators in order to join entities that use different languages. This translation system will translate areas of knowledge and research topics and will allowing send searches to entities with other type of languages without knowing their language.

References

1. Laura Garton, Caroline Haythornthwaite, Barry Wellman. (1997). Studying Online Social Networks. Journal of Computer-Mediated Communication, Vol 3. Issue 1. http://jcmc.indiana.edu/vol3/issue1/garton.html (Last Access August 12, 2011)
2. S. D. Berkowitz (1982). An introduction to structural analysis: The network approach to social research. Toronto: Butterworth..
3. B. Wellman and S. D. Berkowitz (1988). Structural analysis: From method and metaphor to theory and substance. Social structures: A network approach. Cambridge: Cambridge University Press. Pp. 19-61
4. S. Wasserman, and K. Faust, (1994). Social network analysis: Methods and applications. Cambridge: Cambridge University Press.
5. B. Wellman, J. Salaff, D. Dimitrova, L. Garton, M. Gulia, and C. Haythornthwaite, (1996). Computer networks as social networks: Virtual community, computer-supported cooperative work and telework. Annual Review of Sociology, Vol. 22. Pp. 213-238.
6. Janet Fulk and Charles W. Steinfeld. (1990). Organizations and communication technology. Newbury Park, CA: Sage.
7. B. Wellman, and M. Gulia (1997). Net Surfers Don't Ride Alone: Virtual Community as Community. Networks in the Global Village, Westview Press, Boulder, CO. Pp. 331-367.
8. P. Marsden, and K. E. Campbell, (1984). Measuring tie strength. Social Forces, Vol. 63. Pp. 482-501.
9. B. Wellman, and S. Wortley, (1990). Different strokes from different folks: Community ties and social support. American Journal of Sociology, Vol. 96. Pp. 558-588.
10. Knowledge@Wharton. Connecting the Corporate Dots: Social Networks Reveal How Employees and Companies Operate. June 14, 2006. Available at: http://www.networkmba.co.uk/kbuploads/7.pdf (Last access: August 12, 2011)
11. Stephen J. H. Yang, Jia Zhang, Irene Y.L. Chen. (2007). Web 2.0 Services for Identifying Communities of Practice through Social Networks. IEEE International Conference on Services Computing. Salt Lake City, USA. Pp. 130-137, July 2007.
12. CORDIS at Facebook: http://www.facebook.com/pages/CORDIS/139488262736435 (Last access: July 18, 2012)
13. Seventh-Framework-Programme at Facebook: http://www.facebook.com/pages/Seventh-Framework-Programme/109815379069290 (Last access: July 18, 2012)

14. US National Science Foundation at Facebook: http://www.facebook.com/US.NSF (Last access: July 18, 2012)
15. US National Science Foundation at Linkedin: http://www.linkedin.com/groups/US-National-Science-Foundation-77221 (Last access: July 18, 2012)
16. Seventh-Framework-Programme at Linkedin: http://www.linkedin.com/groups/FP7-Free-Partner-Search-3845753 (Last access: July 18, 2012)
17. Academia.edu website. At http://academia.edu/ (Last access: July 18, 2012)
18. Researchgate website. At http://www.researchgate.net/ (Last access: July 18, 2012)
19. Epernicus website. https://www.epernicus.com/network (Last access: July 18, 2012)
20. Academic Family website. http://www.academicfamily.com/ (Last access: July 18, 2012)
21. Lada A. Adamic and Eytan Adar. (2005). How to search a social network, Social Networks 27. Pp. 187–203.
22. Holger Ebel, Lutz-Ingo Mielsch, and Stefan Bornholdt. (2002) Scale-free topology of e-mail networks. PHYSICAL REVIEW E 66, 035103(R)
23. Andrew McAfee, Marco Bettiol, and Maria Chiarvesi, Electronic Hierarchies and Electronic Heterarchies: Relationship-Specific Assets and the Governance of Interfirm IT. Working Knowledge. February 2007.
24. Jae-Yoon Jung, Hoontae Kim, Suk-Ho Kang, Standards-based approaches to B2B workflow integration. Computers and Industrial Engineering archive. Volume 51, Issue 2. Pp. 321-334. October 2006.
25. Christoph Bussler, B2B Protocol Standards and their Role in Semantic. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. Vol 24, Issue 1. 2001
26. INTERREG IIIC Web site. Available at: http://www.interreg3c.net/sixcms/list.php?page=home_en (Last access: August 12, 2011)
27. Lithuania, Poland and Kaliningrad Region of Russian Federation Neighborhood Programme http://www.interreg3a.org/interregen/index.php (Last access: August 12, 2011)
28. YOUTH IN ACTION Partner Searching Forum. Available at: http://partnerji.mojforum.si/ (Last access: August 12, 2011).
29. CORDIS Partners Service Web site. Available at: http://cordis.europa.eu/partners / (Last access: July 18, 2012)
30. Ideal-ist Partner Search web site. Available at: http://www.ideal-ist.net/partner-search (Last access: August 12, 2011)
31. European NCP SME Network web site. Available at: http://www.ncp-sme.net/partner-search/ (Last access: July 18, 2012)
32. AETIC Web site. Available at: http://www.idi.aetic.es/es/inicio/corporativo/Presentacion/contenido.aspx (Last access: August 12, 2011)
33. Collaboration Nets Website, http://www.redesdecolaboracion.org/ (Last access: August 12, 2011)
34. Partner Search, APRE, Agenzia per la Promozione della Ricerca Europea. Available at, http://www.apre.it/formaAssist/Partnersearch.htm (Last access: April 20, 2010)
35. Jaime Lloret, Carlos Palau, Fernando Boronat and Jesus Tomas, Improving Networks Using Group-based Topologies. Computer Communications. Vol. 31 Issue 14. Pp. 3438-3450. September 2008.
36. Jaime Lloret, Miguel Garcia, Jesus Tomás and Fernando Boronat, GBP-WAHSN: A Group-Based Protocol for Large Wireless Ad Hoc and Sensor Networks, Journal of Computer Science and Technology. Vol. 23, Issue 3, Pp. 461-480, May 2008,
37. Jaime Lloret, Miguel Garcia, Diana Bri and Juan R. Diaz, Study and performance of a group-based content delivery network, Journal of Network and Computer Applications. Vol. 32, Issue 5, Pp. 991-999, September 2009.

38. Jaime Lloret, Miguel Garcia, Jesus Tomas and Sandra Sendra, A Group-based Architecture for Grids, Telecommunication Systems, Vol. 46, Issue, 2. 2011.
39. INRIA Lorraine website. Available at http://hal.inria.fr/view_by_stamp.php?label=INRIA-LORRAINE&action_todo=home&langue=en (Last access: July 18, 2012)
40. Indiana database for University research Expertise. Available at https://www.indure.org/ (Last access: August 12, 2011)
41. University of Otago. Available at http://www.otago.ac.nz/researchpublications/ (Last access: August 12, 2011)
42. Applied Crypto Group of Orange Labs. Available at http://crypto.rd.francetelecom.com/ (Last access: August 12, 2011)
43. Senia tutorial, Polytechnic University of Valencia, available at http://www.upv.es/entidades/ASIC/manuales/U0119783.pdf (Last access: August 12, 2011)
44. Proposed International Standard Nomenclature for Fields of Science and Technology, the United Nations Educational, Scientific and Cultural Organization (UNESCO). Available at http://unesdoc.unesco.org/images/0008/000829/082946EB.pdf (Last access: April 20, 2010)
45. Tung-Ching Lin and Chien-Chih Huang, Withholding effort in knowledge contribution: The role of social exchange and social cognitive on project teams, Information & Management, Vol. 47, Issue 3, April 2010, Pages 188-196.
46. A. Singhal (2001). Modern Information Retrieval: A Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Vol. 24, No. 4, Pp. 35-42.
47. M. E. Maron and J. L. Kuhns. On Relevance, Probabilistic Indexing and Information Retrieval. Journal of the ACM (JACM), v.7 n.3, Pp. 216-244, July 1960-
48. J. M. Ponte and W. Bruce Croft. A Language Modeling Approach to Information Retrieval. Annual ACM Conference on Research and Development in Information Retrieval. Pp. 275 – 281. Melbourne, Australia. 1998
49. A. Berger and J. Lafferty. Information retrieval as statistical translation. Annual ACM Conference on Research and Development in Information Retrieval. Pp. 222 – 229. Berkeley, California, USA. 1999.
50. J. Tomás and F. Casacuberta. Monotone Statistical Translation using Word Groups. Proceedings of the Machine Translation Summit VIII, Pp. 357-361. 2001
51. J. Tomás , J. M. Vilar and F. Casacuberta. The ITI Statistical Machine Translation System. Proceedings of the TC-Star Speech to Speech Translation Workshop. June 19-21. Pp.49-55.Barcelona, Spain. 2006
52. J. Tomás, J. Lloret and F. Casacuberta. Phrase-Based Alignment Models for Statistical Machine Translation. Pattern Recognition and Image Analysis. Lecture Notes in Computer Science Vol. 3523 Pp. 605-613 Springer-Verlag. 2005.
53. J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, Pp. 111-119, New York, USA. 2001
54. J. Tomás, J. Bataller , J. Lloret and F. Casacuberta. Mining Wikipedia as a Parallel and Comparable Corpus. 9th International Conference on Intelligent Text Processing and Computational Linguistics, 17- 23 February 2008.
55. J. Tomás, E. Sánchez-Villamil, J. Lloret and F. Casacuberta. WebMining: An unsupervised parallel corpora web retrieval system. Proceedings from the Corpus Linguistics Conference. July 14th -17th. Birmingham, U.K. 2005.
56. Ofir Turel and Yi (Jenny) Zhang, Should I e-collaborate with this group? A multilevel model of usage intentions. Information & Management. Volume 48, Issue 1, January 2011, Pages 62-68.

57. Opnet Modeler Website. December 5, 1988. Available at http://www.opnet.com/solutions/network_rd/modeler_wireless.html (Last access: August 12, 2011)
58. T. Okuda, E. Tanaka and T. Kasai. A Method for the Correction of Garbled Words Based on the Levenshtein Metric. IEEE Transactions on Computers. Vol. C-25 Issue: 2 Pp. 172 – 178. 1976.