Using ontologies of software : example of R functions management

Pascal Neveu, Caroline Domerg, Juliette Fabre, Vincent Negre, Emilie Gennari, Anne Tireau, Olivier Corby, Catherine Faron Zucker, Isabelle Mirbel

Zoè Lacroix
María-Esther Vidal

The Third International Workshop on **RE**source **D**iscovery

In Conjunction with the 12th International Conference on Information Integration and Web-based Applications & Services (iiWAS 2010)

Preface

This volume contains abstracts from the technical program of the Third International Workshop on REsource Discovery, held on November 5th,2010. After two successful venues in Linz, Austria, joint to IIWAS (2008) and Lyon, France, co-located with VLDB (2009), the third International Workshop on Resource Discovery (RED 2010) was gathering again with IIWAS in Pontoise, France.

Resource discovery is an exciting field of research where scientists of various communities meet to discuss a variety of topics. The workshop covers all challenges related to the definition, identification, localization, composition of resources including information sources such as a data repository or database management system (e.g., a query form or a textual search engine), links between resources (an index or hyperlink), or services such as an application or tool. Resource discovery systems allow the expression of queries to identify and locate resources that implement specific tasks. Because this problem is of particular interest to the bioinformatics community, many approaches have been designed to support biomedical applications and the analysis of workflows.

We received 24 submissions to the workshop and we composed an exciting program including two invited talks on Quality-Of-Service in the context of resource discovery respectively given by Joyce El Haddad on "*Optimization Techniques for QoS-Aware Workflow Realization in Web Services Context*" and Laure Berti-Equille on "*Assuring Quality of Service and Quality of Data: New Challenges for Service and Resource Discovery*". We accepted 15 papers organized in four sessions: resource discovery for composition, bioinformatics resource discovery, textual resource discovery, and Web service discovery. The workshop was concluded by a panel and open discussion on "*Challenges of Quality-driven Resource Discovery*".

We thank the 17 members of our Program Committee and panelists for their valuable contribution to the workshop. We are also grateful to iiWAS organizers for their kind support to going over limits to make this meeting successful. We kindly acknowledge the National Science Foundation for supporting student travel (grant IIS 0944126), and the Translational Genomics Research Institute and DID-USB for their support.

November, 2010

Zoè Lacroix
María-Esther Vidal

Table of Contents

Invited talks:

Research Accepted Papers:

Program Committee

Laure Berti-Equille, Université de Rennes 1, France.
Stephane Bressan, University of Singapore, Singapore.
Antonio Brogi, University of Pisa, Italy.
Yudith Cardinale, Universidad Simón Bolívar, Venezuela.
Barbara Catania, Universitá di Genova, Italy.
Camelia Constantin, Université Pierre et Marie Curie, France.
Oscar Corcho, Universidad Politécnica de Madrid (UPM), Spain
Valeria De Antonellis, Universitá degli Studi di Brescia, Italy.
Joyce El Haddad, Université Paris-Dauphine, France.
Marlene Goncalves, Universidad Simón Bolívar, Venezuela.
Birgitta Konig-Ries, Friedrich-Schiller-Universität Jena, Germany.
Maude Manouvrier, Université Paris-Dauphine, France.
Chantal Reynaud, Université Paris-Sud, France.
Marta Rukoz, Paris Ouest Nanterre La Défense University, France.
Miguel-Angel Sicilia, Univeristy of Alcalá, Spain.
F.Javier Zarazaga-Soria, Universidad de Zaragoza, Spain.
Lizhu Zhou, Tsinghua University, China.

Assuring Quality of Service and Quality of Data: New Challenges for Service and Resource Discovery (Extended Abstract)

Laure Berti-Équille
University of Rennes 1
France
berti@irisa.fr

Zoé Lacroix
Arizona State University
Translational Genomics Research Institute (TGen), U.S.A.
zoe.lacroix@asu.edu

Maria-Esther Vidal
Universidad Simón Bolívar
Caracas, Venezuela
mvidal@ldc.usb.ve

ABSTRACT

The growth of Internet technologies has unleashed a wave of innovations that are having tremendous impact on the way people and organizations interact with each other's, publish, share, and discover resources. In particular, the significant adoption of Web services and resource discovery technologies demonstrates the effective automation of business-to-business and interpersonal collaborations with new models for automated interactions among distributed and heterogeneous applications [1].

Many available Web resources including services and data sources provide overlapping or similar functionalities or contents, albeit with different levels of Quality of Service (QoS) and degrees of Quality of Data (QoD). And a choice needs to be made to determine which services are to participate in a given service composition or which resources can be adequately selected to satisfy the user's requirements both in terms of content and quality [2].

Quality related aspects relevant for service- and resource-based applications cover a broad field of research, including work on quality modeling and specification, QoS and SLA negotiation, as well as constructive and analytical quality assurance (e.g., testing, monitoring and static analysis) [3].

From an overview of related work in the areas of quality-aware service modeling and composition, resource discovery, and data quality management [4], we introduce and discuss recent methods and techniques for quality assessment. We also highlight new directions for assessing both quality of service and quality of content in data-centric service discovery. Finally, we discuss the new challenges have emerged through the shift from the traditional Web to Web 2.0 in the scope of quality of service and quality of data management.

Categories and Subject Descriptors D.2.11 **[Service-oriented architecture (SOA)]**: Algorithms

General Terms
Algorithms

Keywords Quality of Service, Data Quality, Resource Discovery

References

1. G. Alonso, F. Casati, H. Kuno, and V. Machiraju. Web Services: Concepts, Architectures, and Applications. Springer, 2004.
2. M. Goncalves, M.E. Vidal, A. Regalado, N. Yacoubi. Efficiently Selecting the Best Web Services. LNCS 6162. 2010.
3. Z. Lacroix, M.-E. Vidal C. Legendre: Customized and Optimized Service Selection with ProtocolDB. Globe 2009: 112-123M. Lohmann, L. Mariani, and R. Heckel. A Model-Driven Approach to Discovery, Testing and Monitoring of Web Services, pages 173 – 204. Springer, 2007.
4. L. Berti-Équille, Quality Awareness for Data Managing and Mining, *Habilitation à Diriger des Recherches, Université de Rennes 1*, Juin 2007

Optimization Techniques for QoS-Aware Workflow Realization in Web Services Context

Joyce El Haddad
LAMSADE CNRS FRE 3234, Université Paris Dauphine
Place de Lattre de Tassigny, 75775 Paris Cedex 16, France
elhaddad@lamsade.dauphine.fr

ABSTRACT

With the development of Web services technologies, a lot has been done to satisfy users' requirements through service composition. Service selection is an important part of the service composition problem and the Quality of Service (QoS) of the selected services has an impact on the QoS of the produced composite service. This paper is devoted to the presentation of some of the optimization techniques currently in use in Web service context.

1. SUMMARY

Of particular interest in the context of Web services is the problem of composition which consist in the realization of a workflow that creates the functionality of a new value-added service, called a composite service. As many provider might provide functionally-equivalent concrete services, multiple candidates might be available for a task in a composite service. To distinguish among these candidates, their non-functional properties are considered such as Quality of Service (QoS) properties (e.g., response time, reputation, performance, and availability). With the growing number of alternative Web services that are functionally similar but differ in QoS parameters, the QoS-aware service selection problem becomes a decision problem on the selection of concrete component services with regards to user functional and non-functional requirements.

In the literature many approaches have been proposed concerning the selection of the best set of concrete services in terms of QoS to execute the abstract tasks of the workflow while meeting the user's global QoS requirements, which is known to be a NP-hard problem. In these approaches, the service selection problem is formulated as a combinatorial optimization problem and optimal or near-optimal solutions are proposed.

The available techniques for solving combinatorial problems can be classified in two categories : exact and heuristics methods. Among the exact methods, we distinguished between linear programming based methods, mixed integer programming based methods and dynamic programming. For heuristics methods, we distinguished between constructive methods such as greedy algorithms, local search methods such as tabu search, simulated annealing, and population-based models such as ant colony optimization, evolutionary computation including genetic algorithms, and particle swarm optimization.

Several selection algorithms have been proposed to select service compositions with different composition structures and various QoS constraints. A taxonomy of theses solutions may be produced based on their objectives and the way the proceed. In this work, we first give a general classification of exact and heuristics methods for combinatorial optimization. According to this classification, we present a first class of state-of-the-art approaches that aim at determining the optimal service composition, using exact methods (e.g., global planning [4], dynamic programming [3]). These solutions have high computational cost, they can not provide a solution in a satisfying amount of time for large sized instances. To cope with this issue, we present other approaches proposing heuristics-based solutions (e.g., greedy algorithm [1], tabu search [2]) aiming to find near-optimal composition.

2. REFERENCES

[1] P. Bonatti and P. Festa. On optimal service selection. In *Proceedings of the 14th international conference on World Wide Web*, pages 530–538, New York, NY, USA, 2005. ACM.

[2] J. Ko, C. Kim, and I. Kwon. Quality-of-service oriented web service composition algorithm and planning architecture. *J. Syst. Softw.*, 81(11):2079–2090, 2008.

[3] T. Yu and K. Lin. Service selection algorithms for web services with end-to-end qos constraints. *Information Systems and E-Business Management*, 3(2):103–126, 2005.

[4] L. Zeng, B. Benatallah, A. Ngu, M. Dumas, J. Kalagnanam, and H. Chang. Qos-aware middleware for web services composition. *IEEE Trans. on Software Eng.*, 30(5):311–327, 2004.

Power Aware Cluster-based Service Discovery for MANETs

Bodoor Al-Fares
balfares@hotmail.com

Mznah Al-Rodhaan
rodhaan@ksu.edu.sa

Abdullah Al-Dhelaan
dhelaan@ksu.edu.sa

College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia

ABSTRACT

In large-scale Mobile Ad Hoc Networks (MANETs), one of the great design challenges is resource discovery due to the lack of infrastructure. Clustering the MANET into regions and forming virtual backbone will lead to a quick discovery of the services, but due to power constraint cluster head maybe out of reach. Re-computation of cluster heads and frequent information exchange will suffer high computational overheads. Therefore, it is clear that a more stable clustering architecture will lead to the performance improvement of the whole network. In this paper, we will improve the service discovery scheme proposed recently by L. Dekar et al. through increasing the stability of the network architectures. To stabilize the clustering architecture for a longer time, we elect the node with maximum power capability to be the cluster head, and deploy a secondary cluster head for each cluster head. This secondary cluster head, which is a cluster member node, is identified and assigned by its cluster head to be the backup cluster head. Since the backup for the cluster head is known, the cluster leadership is transferred with no need to invoke the clustering algorithm. This will increase stability of the network and decrease the clustering communication and computation overhead.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Retrieval models, Search process;
H.4.3 [Communications Applications]: Information browsing.

General Terms

Algorithms, Design.

Keywords

MANETS, Discovery, Service, Cluster, Ad hoc.

SUMMERY

In this paper, we propose a new resource discovery protocol that improves the resource discovery scheme proposed recently by L. Dekar et al. [1] which is a hypercubes-based resource discovery scheme for large scale mobile ad hoc networks. Our scheme starts by dividing the network into none overlapping clusters then hypercubes are constructed in each cluster of the network by considering closely its physical topology constraints and the hypercube multi-paths property. The constructed hypercubes are then connected to form a backbone. A Distributed Hash Table (DHT) is processed on these hypercubes to lookup and registers the resources.

Mobile devices are with power constraint, each cluster head acts as a coordinator in its cluster, which causes it to consume more power. Previously, the criteria of electing the cluster head did not consider such issue, thus there is no guarantee for cluster head to have more power which may cause it to rapidly deplete its power. The continuous re-computation of cluster heads and frequent information exchange will lead to high computation overheads. Moreover, leaving the cluster head to fail then responding to the failure will lead to more overhead; due to the frequent information exchange among the participating nodes. Therefore, it is clear that a more stable clustering architecture will lead to the performance improvement of the resource discovery.

In our scheme, we improved the resource discovery scheme by increasing the stability of the network architectures. To keep the clustering architecture stables for the longest possible time, we elect the node with the maximum power capability to be the cluster head. Afterwards, we deploy a secondary cluster head for each cluster. This secondary cluster head, which is member node in the cluster, is identified and assigned by its cluster-head to be the future cluster head. Whenever is needed, the cluster leadership is transferred with no need to invoke the clustering algorithm since the secondary cluster head is defined apriori. Such approach will increases the stability of the network and decreases the clustering communication and computation overhead.

Currently, we are conducting extensive simulation, using NS2 Simulator, to evaluate the performance of our proposed scheme and compare it with the existing resource discovery approaches and demonstrate its superiority.

REFERENCE

[1] L. Dekar and H. Kheddouci, "A Resource Discovery Scheme for Large Scale Ad Hoc Networks Using a Hypercube-Based Backbone", 2009 International Conference on Advanced Information Networking and Applications, 2009, pp.293-300

A Transactional-QoS driven Approach for Web Service Composition

Eduardo Blanco Yudith Cardinale María-Esther Vidal

{eduardo,yudith,mvidal}@ldc.usb.ve
Universidad Simión Bolívar, Departamento de Computación y T.I.
Apartado 89000, Caracas 1080-A, Venezuela

Joyce El Haddad Maude Manouvrier Marta Rukoz

{elhaddad,manouvrier,rukoz}@lamsade.dauphine.fr
LAMSADE CNRS UMR 7024, Université Paris Dauphine,
Place de Lattre de Tassigny, 75775 Paris Cedex 16, France

ABSTRACT

We approach the problem of WS composition in distributed platforms as an optimization problem by simultaneously considering functional, *QoS*, and transactional requirements. We propose a Petri-Net based approach named `PT-SAM`, and a utility function that combines *QoS* and transactional properties and guides the `PT-SAM` algorithm into the space of compositions that best meet the *QoS* and transactional criteria. Our experiments show that `PT-SAM` outperforms state-of-the-art solutions by identifying compositions that better meet the *QoS* and transactional criteria, while the composition time remains in the same order of magnitude.

1. SUMMARY

We address the problem of identifying WS compositions that best meet a user request. A user request represents functional, *QoS*, and transactional requirements; functionality is expressed as inputs and outputs, while a set of *QoS* restrictions and the risk level define the required aggregated quality and transactional property that the final Composite WS (CWS) has to satisfy. We formalize this optimization problem as the WS Composition problem, and provide a solution based on the Petri-Net formalism. Additionally, we extend the utility function previously proposed in [**?**] and the formalization of transactional properties defined in [**?**]. This enhanced utility function combines functional, *QoS*, and transactional properties to highly rank the WS compositions that best meet the user criteria.

CWSs are represented as Petri-Nets and the aggregated transactional property of a CWS is derived from the properties of its component WSs and the structure of the Petri-Net. The structure refers to whether there is a path between a pair of services or not (Sequential or Concurrent execution). We devise a Petri-Net unfolding algorithm as a composer guided by our utility function to only consider compositions that best meet the *QoS* and transactional criteria. As far as we know, only few approaches consider *QoS* and transactional properties at the same time, but none of them proposes a utility function able to highly rank the WS compositions that best meet the functional, *QoS*, and transactional criteria.

We conducted an experimental study to compare our approach with respect to a state-of-the-art solution named `SAM` [**?**]. The experiments were run on dataset of 5,000 services and a benchmark of 200 queries. Services were randomly annotated with two *QoS* parameters and transactional properties; values of *QoS* and transactional properties were assigned following a uniform distribution. We measured time to identify the best composition and the quality of the composition. We observed that `PT-SAM` identifies solutions where *QoS* parameter values are reduced by at least 50% while the transactional criteria are satisfied. Additionally, the composition time remains in the same order of magnitude as `SAM` and may be reduced by at least 50%. In the future we plan to compare `PT-SAM` with other existing transactional based approaches and extend our experiments with complex service datasets and benchmarks of queries.

2. REFERENCES

[1] S. Bansal and J. M. Vidal. Matchmaking of Web Services-Based on the DAML-S Service Model. In *II Internat. Joint Conf. on Autonomous Agents and Multiagent Systems*, pages 926–927, 2003.

[2] E. Blanco, Y. Cardinale, and M.-E. Vidal. chapter Aggregating Functional and Non-Functional Properties to Identify Service Compositions. IGI BOOK (Methodologies for Non-Funtional Requirements in Service Oriented Architecture), 2010.

[3] J. El Haddad, M. Manouvrier, and M. Rukoz. TQoS: Transactional and QoS-aware selection algorithm for automatic Web service composition. *IEEE Trans. on Services Computing*, 99(PrePrints):1–14, Nov. 2010.

Semantic Map for Structural Bioinformatics: enhanced service discovery based on high level concept ontology

Edouard Strauser[1]
e.strauser@gmail.com

Mikaël Naveau[1]
naveau.mikael@gmail.com

Hervé Ménager[2]
herve.menager@pasteur.fr

Julien Maupetit[1]
julien.maupetit@univ-paris-diderot.fr

Zoé Lacroix[3]
zoe.lacroix@asu.edu

Pierre Tufféry[1]
pierre.tuffery@univ-paris-diderot.fr

[1] MTi, INSERM UMR-S 973, Université Paris Diderot (Paris 7) and RPBS, Paris, France

[2] Groupe Logiciels et Banques de Données, Institut Pasteur,France

[3] Arizona State University and Translational Genomics Research Institute (TGen), Arizona, USA

ABSTRACT

International effort on structural bioinformatics has led, in the last decade, to developments that keep resulting in several hundred of new online services a year. Along with the progress of our knowledge, new directions for research are opened and new domains emerge while the level of data integration keeps growing, from molecule to complexes and complexes to assemblies. Although researchers usually accurately survey the novel methods in their field of expertise, it becomes more and more complex for bioinformaticians, and even more difficult for biologists to access an overview of available existing domains and associated methods. To address this problem, we have undertaken for several years, the development of the concept of a Structural Bioinformatics Semantic Map (SBMap) based on the coupling of a high level concept ontology and a low level map of the methods (services) as a mean to provide an efficient and meaningful service discovery tool. While the use of a domain ontology has proven to be welcome by the users, the number of concepts is rapidly too large to be tractable in a manner similar to that of an atlas, in which possibly the complete information is displayed and focus will allow the discovery of the relevant information. In this study, we address the problem and introduce and discuss a new history tracking scheme to reduce the visual complexity of explored graph. For more informations about the project, see `http://sbmap.rpbs.univ-paris-diderot.fr`

1. SBMAP OVERVIEW

The ontology is composed of concept classes and relationships expressed in OWL format. The catalog of services is a database where each service is represented with relevant informations such as the service URL, the input/output concept and the input/output data type. The service registration is a process opened to any potential user who wishes to register a service. The service entry interface is a form where users describe the services they wish to register. Once validated by a moderation committee that verifies the correctness of the description and its consistency with the existing data, it is incorporated in the map. Main component of the system, the visualization interface (SBMapViz) aims at allowing the exploration of the ontology and the services graphs, the discovery of services that connect two concepts, the retrieval to their characteristics, and the access to the selected service.

2. ONTOLOGY EXPLORATION CHANGES

After a few steps of node expansion, the exploration results in unseasy node identification. To overcome this weakness, the new release of SBMapViz has been revisited thanks to participative design workshops. Four new features have been developped to reduce the visual complexity and enhance the viewer user-friendliness: (i) **reduced node expansion**: this new navigation mode only displays the path of concepts clicked by the user; previous concepts childs remains hidden, (ii) **explicit node selection**: node (concepts or methods) selection can results from a direct text search, (iii) **dynamic colors**: concepts associated color gradient reveals the navigation history, and (iv) **full history management**: the user can rewind or fast-forward the navigation history and save or reload his session.

3. CONCLUSIONS AND PERSPECTIVES

Future works for SBMap will mainly focus on two points. First, a natural navigation session through the ontology will draw a path with more than a single service. Such a path could be translated in a standard XML grammar that describes a workflow and be executed in a workflow engine such as Mobyle. The second perspective of this work is to supply the SBMap registred web services with a collection of concept-related BioCatalogue services, allowing more complex workflows to be designed from the SBMapViz applet.

A Semantic Map of RSS Feeds to Support Resource Discovery on the Web

Gaiane Hochard
Arizona State University
Tempe, AZ-85281, USA

Zoe Lacroix
Arizona State University
Tempe, AZ-85281, USA
zoe.lacroix@asu.edu

Jordi Creus
U. Pierre et Marie Curie
Paris, France

Bernd Amann
U. Pierre et Marie Curie
Paris, France

ABSTRACT

Finding specific, valid, complete, and up-to-date information on the Web is a critical problem experienced daily by all users, regardless of their expertise. Many Web usage scenarios not only have to discover and identify web resources publishing high-quality information, but also must access these resources on a regular basis for detecting updates and new data. A key mechanism now offered by most web information sources to inform their clients about resource updates is RSS subscriptions. RSS is an XML-based format for syndicating news articles on the Internet. RSS services supply their subscribers with feeds of news items summarizing the headlines published by a particular web site. Yet the purpose of RSS goes beyond standard news publishing as it allows information providers on the Web to communicate to their subscribers all updates made on their Web site or in their data repository. It is a free and easy way to promote a site and its content without the need to advertise or create complicated content sharing partnerships. On the user's side, RSS feeds are a simple way to create a personalized information space monitoring a variety of web resources. Yet one of the weaknesses of RSS is that there is no easy method or an agreed-upon standard to locate and agregate feeds. Current RSS registries like GoogleReader[1], Syndic8[2] or Feedzilla[3] are mainly based on simple keyword search and a non-personalizable set of hierarchically organized categories for describing and retrieving feeds. Agregation is limited to the visual agregation of widgets in a web page (Netvibes, Feedzilla) or the creation of collections (Google Reader). The identification of RSS feeds relevant to a question of interest is challenging. The user must know which resources are relevant to answer the question, if they provide a RSS subscription, and compose mentally the various resources whose scope provide the complete information domain when combined. We present an approach to support semantic RSS feed discovery and agregation. Our proposed approach is the development of a semantic feed registry for RSS feeds that supports discovery queries as paths in terms of a domain ontology. This registry is based on a declarative RSS query language and a Semantic Map where RSS subscriptions are mapped to paths in some internal ontology. This approach not only provides a user-friendly access to RSS feeds of interest but it also allows the agregation of such feeds as answers to complex semantic questions. For example, ClinicalTrials.gov provides information on federally and privately supported clinical trials conducted in the United States and around the world. Users can subscribe to various RSS feeds which can be mapped to different concepts in some ontology such as clinical_trial, disease, therapy, and patient. The same ontology also might specify relationships between concepts in addition to the isa and is composed of hierarchies. These relationships may be exploited to annotate RSS feeds in a more precise way that the identi cation of concepts. For example, ClinicalTrials.gov documents the relationship Recruits between a clinical trial and a patient pro le. We will use this valuable information in the annotation process by mapping the resource to the paths, as linear expressions in terms of the domain ontology, it documents. Feed discovery queries aim at identifying existing resources that provide updates via RSS feeds on a topic of interest. The result of a discovery query is a set of path expressions which can be rewritten into a union of RSS feed agregation queries. The nal result is a new feed which aggregates all news published by the relevant RSS feeds. Future work include the implementation of this approach on top of RoSeS, a query-based RSS feed aggregation prototype.

Acknowledgement This research was partially supported by the National Science Foundation(NSF)[4] (IIS 0431174, IIS 0551444, IIS 0612273, IIS 0738906, IIS 0832551, and CNS 0849980) and by the Agence National de la Recherche (ANR-07-MDCO-011 RoSeS).

[1] http://wwww.google.fr/reader.
[2] http://www.syndic8.com.
[3] http://www.feedzilla.com.

[4] Any opinion, nding, and conclusion or recommendation expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Athena: Text Mining Based Discovery of Scientific Workflows in Disperse Repositories[1]

Flavio Costa[1,3], Daniel de Oliveira[1]
[1]Federal University of Rio de Janeiro
{flscosta, danielc}@cos.ufrj.br

Eduardo Ogasawara[1,2]
[2]Federal Center of Technological Education
ogasawara@cos.ufrj.br

Alexandre Lima[1], Marta Mattoso[1]
[3]Federal School Pedro II
{assis, marta}@cos.ufrj.br

ABSTRACT
Scientific workflows are abstractions used to model and execute *in silico* scientific experiments. They represent key resources for scientists and are usually enacted and managed by engines called Scientific Workflow Management Systems (SWfMS). Each SWfMS has a particular workflow format. This heterogeneity of formats poses a complex scenario for scientists to search or discover workflows in dispersed repositories for reuse. The existing workflows in these repositories can be used to leverage the identification and construction of families of workflows (clusters) that aim a particular goal. However it is hard to compare the structure of these workflows since they are modeled in different formats. One alternative way is to compare workflow metadata such as natural language descriptions (usually found in workflow repositories) instead of comparing workflow structure. In this scenario, we expect that the effective use of text mining techniques can cluster a set of workflows in families, offering to the scientists the possibility of finding and reusing them. This paper presents Athena, a cloud-based approach to support clustering workflows from disperse repositories using natural language descriptions, thus integrating these repositories and providing a facilitated form to search and reuse workflows.

Categories and Subject Descriptors
H.4 [**Information Systems Application**]: workflow management.

General Terms
Management, Performance, Experimentation.

Keywords
Scientific workflow, text mining, metadata

1. SUMMARY
Over the last years, scientific workflows became a *de facto* standard to model *in silico* scientific experiments [1]. Scientific workflows declaratively capture the activities of a scientific experiment and the dependencies between them. Such activities are represented as components that define the computations that should take place. Because scientific workflows are embodied knowledge of a scientific domain, scientific workflow modeling is a learning process where reuse techniques is a key issue. Although there are many available repositories for scientific workflows, like myExperiment [2], they do not provide necessary mechanisms to allow the effective search for existing workflows, especially in different formats to facilitate reuse. Since it is complex to structurally compare workflows in different formats and repositories, other alternatives should be considered. One possibility is to analyze the available natural language descriptions. However, another question arises: how do we analyze workflow descriptions that are represented using natural language? One option is to use Text Mining (TM) [3] techniques, to identify a set of workflows families.

This paper presents Athena, an approach that aims at searching workflows in different repositories, analyzing existing metadata and creating families of workflows with the same purpose using a cloud [4] infrastructure. In this way, Athena integrates different repositories allowing scientists to search for workflows in different formats and reuse them, avoiding unnecessary rework. By grouping these workflows, scientists are also able to model experiments using high level abstractions representations. Athena has been designed to operate over heterogeneous cloud environments, to be independent of the workflow repository (any repository can be coupled to Athena) and to be able to handle natural language descriptions. Athena architecture is composed by five main kinds of components: Web Crowler, Execution Broker, Pre-processing components, Cluster components and Integrated Resource Repository. Athena was evaluated in the Amazon EC2 environment. A first study was conducted to evaluate the viability of the architecture. It aimed at analyzing the comparison of workflow pairs and identifying clusters of workflows. All workflows used for this study were downloaded from the myExperiment site. We have evaluated the generated clusters with bioinformatics specialists in order to check if the workflows are grouped in a coherent form. In general, Athena produced coherent clusters that could be used to facilitate workflow search and reuse.

[1] This research was partially funded by CNPq and CAPES

2. REFERENCES
[1] M. Mattoso, C. Werner, G.H. Travassos, V. Braganholo, L. Murta, E. Ogasawara, D. Oliveira, S.M.S.D. Cruz, and W. Martinho, 2010, Towards Supporting the Life Cycle of Large Scale Scientific Experiments, *IJBPIM*, v. 5, n. 1, p. 79–92.

[2] D. De Roure and C. Goble, 2007, myExperiment – A Web 2.0 Virtual Research Environment., *In International Workshop on Virtual Research Environments and Collaborative Work Environments.*

[3] R. Feldman and J. Sanger, 2006, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

[4] D. Oliveira, F. Baião, and M. Mattoso, 2010, "Towards a Taxonomy for Cloud Computing from an e-Science Perspective", *Cloud Computing: Principles, Systems and Applications (to be published)*, Heidelberg: Springer-Verlag

A New Framework for Join Product Skew

Foto Afrati
National Technical University
of Athens, Greece
afrati@cs.ntua.gr

Victor Kyritsis
National Technical University
of Athens, Greece
vkyri@cs.ntua.gr

Paraskevas V. Lekeas
Applied Math Department,
University of Crete, Greece
plekeas@tem.uoc.gr

Dora Souliou
National Technical University
of Athens, Greece
dsouliou@mail.ntua.gr

ABSTRACT

Different types of data skew can result in load imbalance in the context of parallel joins under the shared nothing architecture. We study one important type of skew, join product skew (JPS). A static approach based on frequency classes is proposed which takes for granted the data distribution of join attribute values. It comes from the observation that the join selectivity can be expressed as a sum of products of frequencies of the join attribute values. As a consequence, an appropriate assignment of join sub-tasks, that takes into consideration the magnitude of the frequency products can alleviate the join product skew. Motivated by the aforementioned remark, we propose an algorithm, called Handling Join Product Skew (HJPS), to handle join product skew.

Keywords

Parallel DBMS, join operation, data distribution, data skew, load imbalance, shared nothing architecture

1. HANDLING JOIN PRODUCT SKEW

We propose a heuristic algorithm called HJPS that alleviates the Join Product Skew effect. Join product skew occurs when there is an imbalance in the number of Join tuples produced by each database processor. HJPS identifies the skew elements and assigns a specific number of processors to each of them. HJPS constitutes a refinement of previous proposed algorithms in the sense that the exact number of the needed processors is defined for each skewed value instead of duplicating or redistributing the tuples across all the database processors. Additionally HJPS is advantageous in the case of having Join Product Skew without having redistribution skew. HJPS is as follows: The number of the needed computations for the evaluation of the join operation, that identifies the total processing cost (TPC), is the sum of products of the number of tuples in both relations that have the same join attribute values. $TPC = \sum_{b_i \in D} |R_{b_i}| * |S_{b_i}|$. HJPS deals with the case of the binary join operation $R(A,B) \bowtie S(B,C)$ in which the join predicate is B. Let $D = \{b_1, ..., b_m\}$ be the domain of values associated with the join attribute B. We denote by $|R_{b_i}|$ ($|S_{b_i}|$) the number of tuples of the relation R (respectively S) with join attribute value equal to b_i, where $b_i \in D$. HJPS assumes that the quantities $|R_{b_i}|$, $|S_{b_i}|$ for every $b_i \in D$ are known in advance by either previously collected or sampled statistics. The number of the database processors is denoted by n. In the context of the parallel execution of the join operator, the ideal workload assigned to each processor, denoted by pwl, is defined as the approximate number of the joined tuples that it should produce in order not to experience the join product skew effect. Obviously, it holds that $pwl = TPC/n$. HJPS determines whether or not a join attribute value $b_i \in D$ is skewed by the number of the processors dedicated to the production of the joined tuples corresponding to this value. The quotient of the division of the number of joined tuples associated with the join attribute value b_i (which is equal to $|R_{b_i}| * |S_{b_i}|$) by pwl gives the number of the processors needed to handle this attribute value. In the case that the result of the division, denoted by vwl_{b_i}, exceeds the value of two, HJPS considers the join attribute value as skewed. The latter is inserted into a set of values, denoted by SK. Let $SK = \{b_{a_1}, b_{a_2}, b_{a_3}, ..., b_{a_l}\}$ be the set of the skewed values. HJPS iterates over the set SK. In particular, for the value b_{a_1}, suppose that the number of the needed processors is equal to $vwl_{b_{a_1}}$. The algorithm takes a decision based on the number of tuples with join attribute value b_{a_1} in relations R and S. If $|R_{b_{a_1}}| > |S_{b_{a_1}}|$, the tuples of the relation R are redistributed to the first $vwl_{b_{a_1}}$ processors while all the tuples from the second relation are duplicated to all of the $vwl_{b_{a_1}}$ processors. In order to decide which of the $vwl_{b_{a_1}}$ processors is going to receive a tuple of the relation R with join attribute value b_{a_1}, the algorithm applies a hash function on a set of attributes. On the contrary, if it holds that $|R_{b_{a_1}}| < |S_{b_{a_1}}|$, all the tuples from the relation R with join attribute value equal to b_{a_1} are duplicated to all of the $vwl_{b_{a_1}}$ processors while the tuples of the relation S are distributed to all of the $vwl_{b_{a_1}}$ processors according to a hash function. The same procedure takes place for the rest of the skewed values. The remaining tuples are redistributed to the rest processors according to a hash function on the join attribute.

Using Ontologies of Software, Example of R Functions Management (Abstract)

Pascal Neveu
INRA, UMR 729 MISTEA
F-34060 Montpellier, France
Pascal.Neveu@supagro.inra.fr

Caroline Domerg
INRA, UMR 759 LEPSE
F-34060 Montpellier, France

Juliette Fabre
INRA, UMR 759 LEPSE
F-34060 Montpellier, France

Vincent Nègre
INRA, UMR 759 LEPSE
F-34060 Montpellier, France

Anne Tireau
INRA, UMR 729 MISTEA
F-34060 Montpellier, France

Olivier Corby
INRIA Sophia Antipolis
F-06902 Sophia Antipolis, France

Catherine Faron Zucker
Université Nice-Sophia Antipolis, I3S, UMR 6070
F-06903 Sophia Antipolis, France

Isabelle Mirbel
Université Nice-Sophia Antipolis, I3S, UMR 6070
F-06903 Sophia Antipolis, France

Promote, sustain and make available scientific resources, as computer programs, for multidisciplinary research teams is often a real difficulty. We propose an ontology-based approach to manage, share and promote software programs in a research community. Our proposition takes into account relations between programs: how they are linked, how they could collaborate, follow on and be retrieved. We have developed a new kind of software repository for a team of biologists, statisticians, agronomists and geneticians.

In the research laboratory LEPSE specialized on studying plant responses to environmental stresses, dozens of R[1] functions are produced every year. As a result, there is an important turn-over of function authors and users which calls for understanding, sharing and re-using these functions. In this context, we have initiated a development to organize and promote these functions through the development of a knowledge-based repository.
Given the great diversity of R functions, we have decided to index them with some formalized knowledge describing them, in order to retrieve them by formal reasoning. For this purpose, we developed an ontology providing a controlled and structured vocabulary that captures the concepts and properties necessary to describe R functions. This ontology comprises concepts and properties to describe functions like "Intention", "Argument" and the relations between functions like "hasCall" or "couldBeUsedAfter".
As a result, functions can be retrieved according to a wide range of criteria: author, graphics type, intentions, function calls –more generally, it is relevant to generate the call graph of one function to understand it–, functions from which they are adapted –this makes easier the maintenance of the repository–, functions used after or before –this helps to construct chaining of treatments–, etc..
To formalize both the ontology and the annotations of R functions, we adopt the Semantic Web models: the annotations are represented into the Resource Description Framework (RDF) and the ontology in the Ontology Web Language (OWL). As a result we are able to semantically retrieve R functions by expressing queries in the SPARQL language. We have developed a Semantic Web application for the repository, annotation and search of R functions. It relies upon the Corese engine dedicated to ontological query answering on the Semantic Web.
The architecture of our application is based on a Web Service and allows to: (i) Upload, download and update R functions or RDF annotations, (ii) Retrieve functions by processing SPARQL queries over RDF annotations. The Web Service allows to be used by heterogeneous clients for different purposes (download to R session, versioning, etc.).
We have also developed a user-friendly Web interface with dynamic pre-filled forms. Our application provides an environment for (1) *create and edit annotation*: a Web user interface allows authors to upload R functions and to describe them in a few minutes; and (2) *powerful search*: based on a SPARQL queries generator, users can find and get R functions with a global and accurate understanding and receive suggestions to support their search.

To conclude, we have built a semantic repository of annotated R functions to centralize and share R functions for biologists. It capitalizes expert know-hows that would otherwise often be lost or become non-usable because of a lack of documentation and description. We are convinced that this kind of repository developed for the LEPSE could benefit a much wider community of R function authors and users and be adapted to handle other programming languages.

[1] R is a software language for statistics and graphics

iiWAS2010 , 8-10 November, 2010, Paris, France

Bioinformatics applications discovery and composition with the Mobyle suite and MobyleNet (Abstract)

Hervé Ménager[1]
hmenager@pasteur.fr

Vivek Gopalan[2]
gopalanv@niaid.nih.gov

Bertrand Néron[1]
bneron@pasteur.fr

Sandrine Larroudé[1]
slarroud@pasteur.fr

Julien Maupetit[3]
julien.maupetit@univ-paris-diderot.fr

Adrien Saladin[3]
adrien.saladin@gmail.com

Pierre Tufféry[3]
pierre.tuffery@univ-paris-diderot.fr

Yentram Huyen[2]
huyeny@niaid.nih.gov

Bernard Caudron[1]
caudron@pasteur.fr

[1] Groupe Logiciels et Banques de Données, Institut Pasteur, France

[2] Bioinformatics and Computational Biosciences Branch, OCICB NIAID, NIH, USA

[3] MTi, INSERM UMR-S 973 and RPBS, France

ABSTRACT

Performing bioinformatics analyses requires the selection and combination of tools and data to answer a given scientific question. Many bioinformatics applications are **command-line** only and researchers are often hesitant to use them based on installation issues and complex command requirements. The **Mobyle framework** aims at providing a usable interface that offers an access to many bioinformatics tools within a single integrated environment. It includes a complete set of tools that cover the description, publication, and execution of bioinformatics software in a full-web environment.

However, the traditional approach of setting up centralized resource centers to aggregate tools and data does not scale well with the growing diversity of the methods to operate, which can exceed the skills of the staff attached to such centers. The **MobyleNet** project is an initiative to create a network of smaller platforms with specific skills. The services are published on a framework that favors interoperability between the sites, enabling the seamless integration of resources to run cross-domain protocols.

1. MOBYLE SERVICE DISCOVERY

The description of the services published by Mobyle covers multiple aspects of the tool or service, describing (1) the scientific task it performs, (2) how to use it (from a user point of view), (3) the user interface of the submission form and the presentation of the results, (4) how to call it (how the system should instantiate it and capture the results). This description, stored as an XML file, provides an abstract description of what it achieves, using both natural language descriptions and controlled vocabularies that provide the required metadata which is exploited to facilitate service discovery. It contains the required information to generate components such as a search-able classification tree and a search-engine accessible sitemap. Furthermore, the syntactic and semantic description of the data and parameters in the system provides a solution to guide users in the construction of interactive as well as pre-defined bioinformatics workflows by performing automatic conversions between semantically compatible (i.e., conveying the same information) data formats.

2. MOBYLENET RESOURCE SHARING

The technical basis of MobyleNet is the Mobyle Network functionality. It provides Mobyle server administrators the possibility to share resources, exporting and importing programs between servers: while users still access the same portal, the jobs corresponding to some programs can be remotely executed and stored. The administrator of a server can (1) import programs from other servers, allowing these programs to be displayed as available (though remotely-executed) programs, or (2) export programs, i.e. allow administrators from other Mobyle servers to import them into their portal.

3. CONCLUSION

We presented here the Mobyle software suite, and its derived project, the MobyleNet initiative. Both provide solutions to the question of resource integration in bioinformatics, including mechanisms that facilitate the discovery of new services, which is a key element to address when publishing bioinformatics resources to biologists.

One-Class Classification for Finding Interesting Resources in Social Bookmarking Systems (Extended Abstract)

Daniela Godoy*
ISISTAN Research Institute,
UNICEN University
Also at CONICET, Argentina
CP 7000, Tandil, Bs. As.
dgodoy@conicet.gov.ar

ABSTRACT

In this work two one-class classification approaches were empirically evaluated and compared for the task of identifying interesting resource in social tagging systems. One-class SVM and Rocchio classifiers were used to learn the user interests starting from different sources, such as the full-text of resources and their social tags.

Categories and Subject Descriptors

H.3.3 [**Information Search-Retrieval**]: Information Filtering-*social tagging systems, one-class classification, folksonomies*

1. INTRODUCTION

Finding interesting information in the massive amount of freely accessible, user contributed and annotated Web resources existing in folksonomies is becoming a time consuming and difficult task for users. To alleviate this problem, both content and social tags associated with resources annotated by a user can be used to build a user interest pro le that, in turn, can be applied to filter further incoming information from tagging systems (e.g. RSS feeds).

In this work we evaluate two classification algorithms as a means to distinguish relevant resources. Tag-based classifiers are learned using the Web resources users annotate and have in their personomies as positive examples of their interests. This is a special case of classification, known as one-class classification, in which it is necessary to determine whether an example belongs to a target class (interesting) when only examples of the target class are given.

2. PROPOSED APPROACH

User actions of assigning tags to resources are a strong indication of relevance about its content. Consequently, positive examples of the user interests can be easily collected from folksonomies. On the contrary, identify representative negative examples or non-interesting resources is more complex since users might not tag a potentially interesting resource because of multiple reasons.

For the task of determining whether a resource is inter-sting for a user basing training only on positive examples one-class SVMs (Support Vector Machines) and Rocchio were compared. Schölkopf et al. [2] extended the SVM methodology to handle training using only positive information. Essentially, one-class SVM consists in learning the minimum volume contour that encloses most of the data and it was proposed for estimating the support of a high-dimensional distribution. Rocchio [1] algorithm is a materialization of prototype-based classifiers, which represent each class in terms of a prototype vector in the same dimensional space as documents, making it feasible to estimate the similarity between documents and prototypes of classes.

3. RESULTS AND CONCLUSIONS

Experimental results obtained with a set of personomies gathered from *Del.icio.us* site showed that tag-based classifiers outperformed full-text classification in identifying interesting resources. Tag-based classifiers trained using the top 10 tags assigned to resources showed better performance than those learned using the full tagging activity associated to them. Thus, top 10 tags offers good performance levels and an important reduction in leaning complexity given the smaller size of the resulting dimensional space.

Both frequency-based (number of users that employ a given tag) and binary representations (occurrence or non-occurrence of a given tag) of the resulting tag vectors were considered in the experiments. In this regard, SVM and Rocchio showed distinctive behaviors. Rocchio outperforms SVM when frequency vectors are considered, more likely be-cause is a method coming from information retrieval area, whereas SVM improves Rocchio results over binary vectors.

References

[1] J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, The SMART Retrieval System, pages 313{323. Prentice Hall, 1971.

[2] B. Scholkopf,• J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. Neural Computation, 13(7):1443{1471, 2001.

***This research was supported by CONICET under grant PIP No114-200901-00381.**

A user-centric classification of tools for biological resource discovery and integration on the Web

Cartik R Kothari
Arizona State University
Tempe, AZ-85281, USA
cartik1.0@gmail.com

Preetika Tyagi
Arizona State University
Tempe, AZ-85281, USA
ptyagi3@asu.edu

Jeff Kiefer
Translationall Genomics Research Institute (TGen)
Phoenix, AZ-85004, USA
jkiefer@tgen.org

Zoe Lacroix
Arizona State University
Tempe, AZ-85281, USA
zoe.lacroix@asu.edu

Rida Bazzi
Arizona State University
Tempe, AZ-85281, USA
bazzi@asu.edu

ABSTRACT

Recent studies have revealed there are many more extant biological resources such as web services and data repositories than the actual number of such resources that are being leveraged by life scientists in their work. Life scientists typically use those resources that they are familiar with to implement protocols in efficient workflows, which may not be the best possible way to meet the requirements of the protocol. They have difficulties to identify the most suitable resources to design their workflows and maintain them as new resources or new versions are becoming available.

Tools that facilitate biological resource discovery and integration address this issue by helping scientists choose and compose the appropriate resources for their protocols. This paper defines five different criteria to classify the available tools for resource discovery with an emphasis on user interaction. Each category is further divided into sub-categories for finer grain classification. A number of leading scientific resource discovery tools are introduced and classified according to the proposed criteria in this work. An example is also provided to show how a scientist can use these criteria to select an appropriate resource discovery tool for his work.

Categories and Subject Descriptors

D.2.11 [**Service-oriented architecture (SOA)**]: :Algorithms

Keywords

Web Service, Data source, Ontology, Semantic Web, Semantic Map, Resource Discovery, Bioinformatics

Acknowledgement This research was partially supported by the National Science Foundation (grants IIS 0431174, IIS 0551444, IIS 0612273, IIS 0738906, IIS 0832551, and CNS 0849980).

InSciTe®: Technology Intelligence Service Based on Semantic Web and Text Mining Technologies

Mikyoung Lee, Seungwoo Lee, Hanmin Jung, Pyung Kim,
Taehong Kim, Dong Min Seo, Won-Kyung Sung
Korea Institute of Science and Technology Information (KISTI)
335 Gwahangno, Yuseong-gu, Daejeon, KOREA 305-806
+82-42-869-1783
jhm@kisti.re.kr

ABSTRACT
Technology intelligence refers to activity for supporting an organization's decision-making process by collecting and forwarding information on new technologies. We have developed a Technology Intelligence Service named InSciTe, into which Semantic Web and text mining technologies are combined, in order to help users to make decisions necessary to establish an R&D direction for new research domain. It analyzes relations among technologies, research agents and research results and provides services in the viewpoint of discovery, combination, and comparison.

Categories and Subject Descriptors
H.3.4 [**Semantic Web**], I.2.7 [**Natural Language Processing**]

General Terms
Experimentation.

Keywords
InSciTe, Technology Intelligence Service, Semantic Web, Text Mining.

1. InSciTe as a Technology Intelligence Service
InSciTe, a technology intelligence service, supports decision-making processes to establish R&D strategy. It is designed to help users to analyze patents and papers and to achieve the goal. It also aims at building fast and automated knowledge, developing effective services conducive to making decisions and enhancing information accessibility in conjunction with Linked Data.

The service is based on OntoFrame, which is a Semantic Web-based service platform providing implementation infrastructure. OntoFrame aims to search information and discover implicit knowledge in the information for helping users to achieve their needs efficiently [2]. It includes a semantic knowledge management tool named OntoURI, a commercial search engine, and a reasoning engine named OntoReasoner. The ontology instances populated by OntoURI are stored and inferred using OntoReasoner, which performs rule-based reasoning based on RDF Semantics and partial OWL Semantics in ways of forward-chaining.

We managed to extract names of technologies and their relations from 450,000 papers and patents in the field of green technology. We have also extracted various relations among the technologies such as element technology, similar technology, competing technology, etc. from PubMed data, NDSL, and Wikipedia. These extracted technologies and their relations are compiled into OntoFrame in RDF format and further utilized for various technology-related services.

Main features of InSciTe are as follows. First, it has a through process of extract-transform-load (ETL) and -analysis by combining text mining and Semantic Web technologies. Second, InSciTe allows verification of search and analysis results by using reasoning verification function of OntoReasoner. Third, InSciTe offers enhanced information accessibility through connection with Semantic Web-based open sources represented by Linked Data. Fourth, InSciTe can present multifaceted viewpoints such as academic and business views by blending heterogeneous sources such as papers and patents.

InSciTe can be categorized into a composite service of technologies, research agents, and research results. Technology-Agent Map service is a representative composite service, which allows users to compare research results and relations of competition/cooperation of research agents (researchers, institutions, and nations) for the technologies searched and expanded by the users. Technology-centric services include technology browser which visualizes the relations among element technologies, similar technologies, and competing technologies, technology performance graph which allows users to discover a given technology's current level of maturity, and irrupted technology/agent trend services which show entry timings of given technologies and agents. Meanwhile, agent-centric services are designed to help users to figure out the competitive/cooperative relations between research agents for a given technology by grouping cooperative research agents. These services also allow users to see market trends of a technology. For instance, users can find out whether the market for a technology is led by a sole, representative agent or is being developed through mutual, balanced cooperation.

2. REFERENCES
[1] Lee S., Lee M., Kim P., Jung H., Sung W. 2010. OntoFrame S3: Semantic Web-Based Academic Research Information Portal Service Empowered by STAR-WIN. In *LNCS6089*.

Combining Uncorrelated Similarity Measures for Service Discovery (Abstract)

Fernando Sánchez-Vilas
Univ. Santiago de Compostela
fernando.sanchez@usc.es

Manuel Lama
Univ. Santiago de Compostela
manuel.lama@usc.es

Juan C. Vidal
Univ. Santiago de Compostela
juan.vidal@usc.es

Eduardo Sánchez
Univ. Santiago de Compostela
eduardo.sanchez.vila@usc.es

1. SUMMARY

Web services are a set of operations that are described in an XML format to facilitate its automatic publication, discovery and composition. Particularly in the service discovery problem a number of web services that match a user request must be selected and ranked. This is so important as the number of web services increases, because the manual search of services is not feasible.

The most promising approaches that deal with the service discovery problem are hybrid solutions that apply both semantic and syntactic similarity metrics to determine whether the service matches the user request or not. These approaches consider that the combination of complementary similarity metrics will improve the matching between services and the request. However, in these proposals *(i)* there is not an analysis to decide the more appropriate similarity metrics, that is, selected metrics are used without a clear reason; and *(ii)* there is not an analysis to decide to which element of the service description each selected similarity metric should be applied. In our opinion a deeper analysis of the similarity metrics to be applied is required.

To solve these two drawbacks we have developed an OWLS matchmaker that deals with the service name, description, and inputs/outputs as the elements to compare an OWLS service with the user request. To develop this matchmaker we have carried out the following steps:

1. A deep study of the similarity metrics available in the literature has been carried out in order to make an appropriate selection of the metrics to be applied. Thus, semantic, syntactic and structural-based metrics have been considered.
2. A careful selection of the elements of the OWL-S service profile over which these metrics should be applied has been made. While services names and descriptions are compared one to one, each service input/output is matched with a request one. In order to obtain a global similarity for inputs and other for outputs we present equations that allow us to combine the similarity of these matched pairs.
3. Redundant similarity measures are eliminated based on their correlations. It is necessary to minimize the number of similarity calculations to be made, since for each user request, a comparison with all the services available in the repository must be carried out. Furthermore, this step allow us to select only the similarity metrics that contribute to the solution.
4. Finally, the uncorrelated similarity metrics applied to services and requests are the inputs of a neural network (*multilayer perceptron*) trained with the 80% of the user requests. For each pair of user request and repository service, this neuronal network classifies the service in relevant or not relevant for the given request.

In this paper we used OWLS-TC version 3 as the web service repository to validate our proposal. The results obtained show that our approach clearly outperforms OWLS-MX, but only has a small improvement respect to iMatcher. The reason for this slight improvement can be found in the use of two unfolded metrics in our matchmaker, while iMatcher uses only one. Unfolded metrics are syntactic metrics based on the ontology structure that describe the inputs and outputs of the services. When these kind of metrics are used in light-weight ontologies we obtain very valuable results, which could be even better than the results obtained with semantic reasoning. Furthermore, as final result of our experimentation we show that unfolded metrics can not be overcome adding other metrics not based on the ontology structure.

2. ACKNOWLEDGMENT

Authors wish to thank the Xunta de Galicia and the Ministerio de Educacion y Ciencia for their financial support under the projects 09TUR001E and TSI2007-65677C02-02 respectively.

Challenges of Quality-driven Resource Discovery

Bernd Amann
Université Pierre et Marie Curie
LIP6, 4 Place Jussieu
75252 Paris cedex 05, France
+33 (0) 1 44 27 70 09
Bernd.Amann@lip6.fr

Zoé Lacroix
Arizona State University
and
Translational Genomics Research Institute (TGen)
445 N. Fifth Street, Phoenix AZ 85004, U.S.A.
+1 602 343 8679
Zoe.lacroix@asu.edu

ABSTRACT
This panel summarizes some of the challenges addressed at the RED 2010 panel. The panel follows two invited presentations that address the problem of quality in the context of resource discovery. They are *Assuring Quality of Service and Quality of Data: New Challenges for Service and Resource Discovery* by Laure Berti-Equille and *Optimization Techniques for QoS-Aware Workflow Realization in Web Services Context* by Joyce El Haddad. The questions discussed by the panelists cover modeling issues, formats, languages, semantics, applications, and benchmarks.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: Standards; User issues

Keywords
Quality metrics, Quality of Service, Resource Discovery

1. Quality for Resource Discovery
First the panelists share their views on resources, what they are, what quality measures can be assigned to them, and the problem of resource discovery. As discussed in our previous workshops, a resource can be many things depending on the user or expert including an agent, an application, a service, a data source, etc. [1] The problem of resource discovery itself can be approached in various ways each with its own quality measures.

2. Traditional QoS Approaches
Although resource discovery raises specific issues, many existing approaches designed for Quality of Service may be exploited and revised for that purpose.

3. Workflow-driven Quality
More recent work look at a resource in the context it is used, often composed with other resources in a workflow. The panelists discuss issues specific to service composition, workflow optimization, continuous information discovery (RSS), resource discovery for data integration, etc.

4. Quality Dimensions
Existing models and formats to represent resource often lack the mechanisms needed to capture and publish the metadata required to express quality. The panelists discuss the metadata needed to support quality measures and identify the limitations of existing formats.

5. Discovery Language and Customization
Users expect to be able to express queries that support resource discovery and specify the measures that best capture the quality measure they expect to optimize. Resource selection and quality have a significant impact on the provenance of data.

6. Benchmarks
The design of benchmarks is critical to evaluate and compare different solutions. These benchmarks raise similar questions from a different point of view (performance, data sets, queries) independently of a particular solution.

7. Applications
The problem of resource discovery covers a large spectrum of applications. Some domains such as the biomedical domain [2] seem to express a particular interest in the development of solutions.

8. REFERENCES
[1] Lacroix, Z. 2010. Editorial. *International Journal on Metadata, Semantics, and Ontologies*, Special Issue on Resource Discovery 5, 3, July 2010, 167-169.

[2] Lacroix, L., C. R. Kothari, P. Mork, R. Rifaieh, M. Wilkinson, S. Cohen-Boulakia, and J. Freire. 2009 Biological Resource Discovery, in *Encyclopedia of Database Systems*, Ling Liu and M. Tamer Ozsu (ed.), Springer, 220-223.