Réseau thématique franco-brésilien GDRI–INCT

International Research Group (GDRI) proposal for the «Centre national de la recherche scientifique»

«Innovative Research Issues on Web Science»
GDRI "WebScience

Abstract
The World Wide Web (Web for short) is becoming a very complex system which evolves at an astonishing rate. The need to understand it as an "entity" and to understand its evolution motivated the advent of a new research domain, called Web Science. With this motivation, the first objective of the GDRI "**Innovative Research Issues on Web Science"** is to create a French-Brazilian research network aiming at scientific cooperation and exchange between French and Brazilian researchers interested in the field of Web Science. This network will promote actions for bilateral collaborations and exchanges of students and visiting scholars, PhD students co-supervising, knowledge sharing, international workshops and conferences organization. The second purpose of the GDRI is to contribute to the development of models and software for Web-wide applications related to searching, retrieving and managing the variety of resources (social, content) stored in hundreds of millions of Web sites.

Résumé
Le World Wide Web (Web) devient de plus en plus un système complexe évoluant à un rythme étonnant. La nécessité de le comprendre et de comprendre son évolution a conduit à l'apparition d'un nouveau domaine de recherche, appelé « Science du Web ».
Notre premier objectif dans le GDRI « Recherches innovantes en Science du Web » est de créer un réseau de recherche franco-brésilien visant à la coopération scientifique et les échanges entre chercheurs français et brésiliens intéressés à ce domaine en pleine expansion. Ce réseau se veut un outil pour promouvoir des actions de collaborations telles que les échanges bilatéraux d'étudiants et de chercheurs invités, le co-encadrement d'étudiants, le partage des connaissances, l'organisation de conférences et d'ateliers. Le second objectif est de contribuer à l'élaboration de modèles et d'outils pour les applications Web liées à la gestion et à la modélisation de ces quantités gigantesques de ressources (sociales, informationnelles) hétérogènes stockées ou déployées sur des millions de sites Web.

1 General objective and motivations

The objective of the GDRI-WebScience is to create a bilateral French-Brazilian network aiming at promoting a scientific cooperation and exchange between French and Brazilian researchers interested in the Web Science field. By means of bilateral academic conferences, exchange of students and visiting scholars, PhD students co-supervising, short-term visits and international workshops and conferences organization, the GDRI is aimed to strengthen the exchange and cooperation between the two countries, promote the Brazilian-French cooperative research in Web and enhance the coordination between the laboratories and research institutes of the two countries.

In term of scientific objectives, the GDRI will contribute to the development of Web Science in a variety of ways which covers the problems of defining models and developing software for Web-wide

applications, for searching, retrieving and managing data (multimedia content, data, metadata, social content) stored in hundreds of millions of Web sites. The research program concern several research challenge organized on three axes detailed in the research program section.

2 Research program

The World Wide Web (Web for short) has become a part of our everyday lives. People access at home, on the move, on their phones and on TV, search, create, trade, communicate and share information on any subject. It revolutionized all the society, i.e. the media, commerce, banking, health care, politics and almost anywhere where the communication serves a purpose. The Web is therefore becoming a very complex system; it is more than a sum of its Web pages. It evolves at an astonishing pace. The need to understand it and understand its evolution motivated the appearance of a new research domain, called Web Science. The Web science is defined by (Berners-Lee et al., 2006), as: "*Web science is about more than modeling the current Web. It is about engineering new infrastructure protocols and understanding the society that uses them, and it is about the creation of beneficial new systems. It has its own ethos: decentralization to avoid social and technical bottlenecks, openness to the reuse of information in unexpected ways and fairness. It uses powerful scientific and mathematical techniques from many disciplines to consider at once microscopic Web properties, macroscopic Web phenomena, and the relationships between them. Web science is about making powerful new tools for humanity, and doing it with our eyes open.*"

In this domain, the Web is the primary object of study and ceases to be viewed as a mere technology, based on computers, that helps people to communicate and interact on a global basis. As such, it involves not only research on computing and technological aspects, but also on social and economic issues. More precisely, it is the science that investigates all issues around decentralized information systems, covering people, software and hardware, and their multiple, complex interactions.

The GDRI-WebScience will contribute to the development of Web Science in a variety of ways which cover the problems of defining models and developing software for Web-wide applications, for searching, retrieving and managing data (multimedia content, data, metadata, social content) stored in hundreds of millions of Web sites. The research program is organized into three axes:

- Social Web harnessing
- Management of Web Data and Web content (searching Web data and organizing Web content)
- Software for Web applications

2.1 Social web harnessing

The evolution of the Web from a technology platform to a social milieu is changing the way users consume content online. As the barriers between content producers and consumers fall and users are given the ability to express their opinions, large amounts of social contents are being made available. With that comes the need to organize, mine and manage that content in effective and scalable ways. In this context we plan to investigate this axis on five main research directions.

1. Defining tools for the social web of news;
2. Exploiting social content and user context evidences to improve information retrieval;
3. Opinion detection and dynamics of representations;
4. Socio-pragmatics of human interaction through Web applications
5. Understanding the Web

2.1.1 The Social Web of News

News agencies such as Al Jazeera, CNN, BBC and Fox News have been relying on citizen journalists to report the latest events from around the globe using a variety of social media. Yet, the integration of traditional news with social media is in its infancy. We argue that while social media bring traffic to traditional news sites and increase their appeal, their survival depends on their tighter integration. A successful integration relies on two key challenges. On the one hand, it is becoming crucial to design tools to help reporters and journalists the benefit from the continuous flow of information coming from a variety of sources such as Twitter and Facebook. On the other hand, the need to understand user engagement in the context of social media is key to maintaining a sustained audience. The Social Web offers unprecedented opportunities to characterize networked audiences and their ephemeral interest in content and leverage that to offer a better content sharing experience. We plan to address several research challenges that arise from those opportunities:

- Capturing Networked Audiences: design algorithms to quickly determine ephemeral audiences gathered around topics of interest.
- Understanding User Engagement: formalize user engagement beyond traditional click-through-rate by accounting for social sharing.
- Integrating Social Media with Traditional News: design news recommendation algorithms that incorporate popularity, trends and polarization.
- Enabling Experience Sharing: develop new models for tracking user actions online and use novel graph traversal algorithms to enable social experience sharing.

2.2 Exploiting social content and user context evidences to improve Information Retrieval

Information Retrieval (IR) allows people to retrieve documents according to their textual contents. With the advent of hypertexts and its prominent use on the Web, several IR models intended to exploit such links so as to improve the effectiveness of search results. In these works, a hyperlink is understood as a vote that a reader would emit regarding the targeted document. As a result, websites much appreciated by readers get referenced by many votes/hyperlinks that are used to improve the search process afterwards.

The Web changed the way people communicated. From a 1-to-many relationship where an author disseminated information to many readers, now readers are empowered in a many-to-many relationship where both authors and readers can voice their views and opinions. As a result, the reader is not a passive consumer of information anymore, but he/she progressively evolves towards a consumer/producer behavior. Several initiatives support these so-called Social Medias, such as:

- Participative encyclopedias (e.g., wikipedia.org)
- Social bookmarking platforms (e.g., connotea.org from the Nature Publishing Group)
- Public debate platforms (e.g., agoravox.fr)
- Photo sharing platforms (e.g., flickr.com Micro)blogging platforms (e.g., blogger.com, twitter.com)

By extracting, modeling, mining, and processing these User Generated Contents stemming from Social Medias, we intend to address work related to the following issues:

- What kind of information should be derived from Web 2.0 platforms in order to improve the IR process?
- How user context can be derived and modeled?
- How to extend existing IR models for taking into account such materials? Should we design a new IR model for handling these pieces of evidence?
- How to formally validate their benefits from the user standpoint? To what extent do these materials help in improving the satisfaction of end-users?

2.2.1 Opinion detection and dynamics of representations

The publication of opinions through various ways contributes to the production and dissemination of mental "images" related to various entities, such as politicians, artists, companies, or their products. Their "image" is a structured and dynamic representation which can be seen both sides: the representation an entity tries to assign to itself, and how a person or a group of persons perceive this entity. The way the Web is used as an active channel to disseminate, and amplify such representations, has a strong influence on real life. Retrieving (detecting) useful post related to these entities, modeling their representations and their dynamics, analyzing their trends, is an example of the difficult challenges we attempt to face through collaborative works between the partners of the project.

2.2.2 Socio-pragmatics of human interaction through Web applications

The Web has a large impact on our daily life. It revolutionized all the society, i.e. the media, commerce, banking, health care, etc. Our goal in this topic is investigate the social, political and economic aspects of the Web. It will discuss how applications such as risk prevention, effective mobility, can benefit from the web.

- Risks and Flood hazards. The need for the diffusion of risk information and knowledge on hazards characteristics is becoming increasingly required by several stakeholders and risk managers. They are therefore interested to research the best ICT tools (by Smartphone, Android, Web site, and so on) to inform in a short time citizens and tourists on the risk they are endangered. As many applications are developed nowadays (as peer-to-peer applications), we need to find the best “information network” (between population, local actors and risk managers at different political levels) very quickly because only the mayor-elects can legally inform their population. This also leads to many scientific novel and interesting issues: - optimal watching and spatial devices to prevent the crisis and to reduce the human vulnerability, reliable computational and expert systems for monitoring event progress.
- Mobility, territory, urbanity and the Web. The Web provides huge amounts of data that are still only under exploited. In particular, sensors in our vehicles provide real-time data that can be exploited in different ways, for instance for a better understanding of travel within a city (and then optimize the traffic lights), to better think about the location of economic activities, and so on. Therefore, the exploitation of these data leads to new practical applications but also to new scientific questions for planning and optimization. We are here concerned with the related combinatorial optimization problems.

2.2.3 Understanding the Web

Our focus in this project is to capture the effect of interactions between agents whose strategies involve creation, diffusion and removal of digital contents, namely, search engines, public institutions, service providers and private citizens. All such agents compete or cooperate by either spreading or inhibiting the diffusion of contents in order to attain some utility. This study of complex network allows the understanding of how information propagates in the network. In fact, it is known that there exists a strong relation between complex networks and diffusion of epidemics. However, our study in complex networks is rather original, since it brings in several new aspects in the formation of complex networks. One important facet to look at is indeed performance: a sample case is the speed of diffusion of contents and the speed of formation of novel links to those contents among different social networks and possibly pages ranked by search engine. This is a domain that has not been fully investigated so far.

2.3 Management of web data and web content

The “Management of Web Data and web content” axis will address the question of access and management of heterogeneous, distributed data sources. The Web is among the largest repositories of information, which is getting increasingly large and complex, and users’ demands get harder to satisfy. The complexity of information is related mainly to its volume which grows every day and the heterogeneity of its contents (text, image, video, news, social content, ...). In addition to these materials, a large number of pages are seen as data containers which may have different forms, including HTML tables, HTML lists, and back-end Deep Web databases (such as the books sold on Amazon.com). Managing and exploiting these data together or not with page content may serve several challenging applications such as data integration, data retrieval, question answering, aggregated search and so on.
In addition to the management of the web contents, preserving the invaluable resources (scholarly, cultural, scientific economic) being lost for future generations and future applications and systems, is one of our main responsibilities. Therefore designing effective systems for archiving useful information contained in Web pages is necessary for preserving the knowledge accumulated as time passes.

This axis will focus on three challenging topics:

1. Enabling data integration from web sources, users are not limited to data that has been prepared for integration (such as already available in XML), but new data can be generated and extracted from web sources
2. Improving Web search by providing structured data retrieval instead of information retrieval. This new generation of search engines will be able to extract data (or any items, text fragment, image) from documents and aggregate them to better answer the information need.
3. Archiving useful information for future applications and generations.

2.3.1 Data integration

Data management had witnessed several major technological advances during in the last decades. One of them is the emergence of data integration systems, providing access to distributed, autonomous and heterogeneous data sources. Supported by the emergence of new technologies such as Web 2.0, Cloud Computing and Web Services, information available on the Web is increasing every day.
Besides, the nature of data to be integrated evolved from structured data to various kinds of data (unstructured, sensor, geographical and multimedia, etc). Data integration tasks are made more complex and data integration systems evolved from data warehouses or mediation systems to web warehouses or Peer-to-Peer (P2P) systems, which are massively distributed and highly volatile systems for sharing large amounts of resources. In this context, ensuring that the users get relevant and meaningful answers with the required level of quality is a critical factor of success of these systems. This raises several research problems, among which we plan to address the followings:

- Quality-Driven Data Integration: Our goal is to establish the fundamentals concerning information quality evaluation and improvement during the life cycle of flexible and dynamic information integration environments. We are interested in studying the ways in which quality can drive data integration tasks as well as data transformations and multidimensional operations. We will address the specification of relevant Information Quality (IQ) criteria and their use in different data transformation and data integration processes such as schema matching, query processing or schema mapping.
- -Semantic Data Integration: As data integration systems become more complex, a key issue is to understand the content of the participating resources and the possible semantic links between them. We are interested in providing tools allowing the use of the knowledge stored in ontologies for integrating and querying data. Our goal is to improve data integration tasks such

as query reformulation or schema matching by introducing information related to the meaning of the participating resources.

- Using Context in Data Integration: Considering the context as the set of elements surrounding a given entity of interest (e.g. a user or a query), we are interested in taking this kind of information into account in order to improve data integration processes. Our goal is to propose suitable representation models for this context and to propose context-aware data integration algorithms.

2.3.2 Aggregated search and data retrieval

Traditional search engines which are tuned to return a ranked list of relevant documents, work well for queries searching for text web pages. Nevertheless, with the growth of information sources and their heterogeneity users' demands get more complex and harder to satisfy. Such demands (queries) concern for instance users searching for data, for example the query "List of European countries with their GDP" or a more general query such as "Chinese restaurants in Paris". This later needs to be answered with a list of Chinese restaurants and their attributes (properties) such as address, email, phone number, menu, etc. Within these queries we distinguish queries on entities (e.g. Paris, France, Hotel Bellagio, Nokia e72 …) and queries on classes of entities (French wines, Chinese restaurants in Paris, European countries …). Current search engines do not answer properly these queries. A better answer can be an aggregated document that is composed of different information nuggets such as relational data (classes, instances, attributes), but also video, images, news articles, etc. For instance a query "trip to Paris" might be answered by presenting a map, a list of museums, a list of places to visit, weather forecast and so on.

In order to answer these queries in an effective way, one needs to extract, aggregate and organize information nuggets coming from different web pages and sources. We will first focus on nuggets in the form of relational data (classes, instances, attributes), to enable more flexible and organized answers. Then, we aim to generalize our approach to integrate any type of information nugget. In this context, we plan to address the following research challenges:

- Understanding user needs: analyze what are the queries that require aggregation and formalize the expected properties that will help aggregation
- Result generation: formalize the relationships between pieces of information to build the aggregated result, and define aggregated information retrieval models that measure the relevance of an aggregate. Current clues concern cross-vertical aggregated search models, relational aggregated search, document generation, …
- Evaluation methodology: design methodologies that will allow to evaluate the effectiveness of aggregated retrieval models and to compare them.
- Application exploration: extend applications mainly addressing Web search to contextual and domain-specific approaches

2.3.3 Web archiving

Web archiving aims at preserving useful information contained in Web pages for future generations. It has been studied over the last two decades and nowadays, many institutions (such as national libraries) in the World are involved in creating and maintaining Web archives. The main issues in Web archiving include

- corpus selection,
- efficient crawling,
- storing and indexing page versions,
- information retrieval in archives,

- archive preservation over time.

We plan to address the issues mentioned above, with a particular interest on archive quality. Web archive quality include

- spatial completeness (how archived pages "cover" the real Web),
- temporal completeness (how and when to capture the "best" page versions with limited resources - bandwidth, storage,..- to get an accurate history of pages in the archive)
- and temporal consistency (do stored versions of different pages have appeared simultaneously on the real Web).

Some of these works have been led in the context of the French ANR Cartec project, and are continued within the Scope FP7 European project.

2.4 Software for web Applications

The “Software for web applications” focuses on engineering issues; it concerns specific questions on the design, development and deployment of large distributed applications on the Web. Existing and developing approaches to leveraging the Web have to be extended into new Web environments as they are created (such as P2P networks, cloud for example). Services are a key area where our engineering models of the Web need to be engaged and extended. We plan to address two main challenges that arise from those opportunities.

Service composition: synthesis, verification and repair: One of the ultimate goals of Service Oriented Architecture and its supporting technologies, e.g., web services, is to enable rapid low-cost development and easy composition of distributed applications, a goal that has a long history strewn with only partial successes. However, an important bottleneck that hampers a wide use of Web services technology in real life applications lies in the complexity of the composition task as well as the high level of expertise needed in order to specify it. The problem is even more complicated when data and/or non-functional aspects, which constitute important dimensions of business process semantics, are taken into account. For example, implementing a service composition using WSBPEL is a lengthy, costly, and high-risk process, especially if additional packages such as WS-Transactions and/or WS-Security are used to manage transactional and/or security aspects. In particular, our aim is to investigate the following problems related to service composition:

- Automatic synthesis of data centric web services (or artefact-based business processes).
- Automatic failure recovery in web service composition.

Business Processes on the Cloud: Cloud computing brings a paradigm shift to the computing world, by revolutionizing the ways computer processing, storage and software delivery are achieved. It frees organizations from large IT capital investments, and enables them to plug into extremely powerful computing resources over the network. The underlying paradigms such as elasticity and pay-for-use have the potential to profoundly transform the nature of organizational IT strategies, technology infrastructure, and business models. Our aim is to investigate issues underlying business processes delivery over the clouds. We will study both methodological issues to facilitate a migration of concrete workflows into a cloud environment as well as technical issues related to infrastructure configuration and optimization techniques to enable effective business process executions and monitoring over the cloud.

3 Program activities and expected results

In the light of the above challenges, the network will promote bilateral initiatives aiming at strengthening the exchange and cooperation on a shared topic between various partners of the two countries. The program activities will comprise:

- A series of seminars in the topics of the project. We plan to organize one to two seminars per year.
- Organization of workshops in Brazil and France.
- Building cooperative projects in response to call for proposals from the European Union (EU), ANR or other funding agencies
- Co-supervising PhD students from both sides of the continent to reinforce our collaboration. The network will also help PhD students to attend these different meetings for promoting their research

In addition to the above points allowing the strengthening of the cooperation, some expected scientific results are:

– Social Web harnessing
  - Algorithms for large-scale social analytics on news
  - Scalable news recommendation algorithms based on opinion diversification
  - Information retrieval model integrating evidences surrounding users and information.
  - Models of the future Web graph, where content is highly dynamic and personalized
– Management of Web Data and Web content (searching Web data and organizing Web content)
  - Models and algorithms suitable for using contextual information during data integration
  - Relational retrieval model based on data extracted from HTML table
  - Efficient crawling methods to improve archives quality
– Software for Web applications
  - Definition of a formal framework and algorithms to handle data-centric web service synthesis
  - Algorithms to optimize the execution of business processes on the Cloud

4 Consortium organization

The project is coordinated by Mohand Boughanem for the French side and Carlos José Pereira de Lucena for the Brazilian side. It will be managed by a steering committee (SC). This section describes the structure of the SC and the project governance.

4.1 The Steering Committee

The Steering Committee (SC) is responsible for discussing and preparing proposals for strategic decisions. It is the official governing body, dedicated to technical and administrative management of the project. This committee will meet at least once a year all the duration of the GDRI.

Activities

- Technical management of the project: monitoring the progress of the project, decisions making concerning the ongoing on actions
- Administrative management of the project: drafting / validation reports and minutes.
- Decision on the allocation of resources (missions, organization of meetings / seminars / workshops).
-

Members

- The SC is chaired by the coordinator(s) of the project.
- Each Partner of the project has one representative member

4.2 Project management

Throughout the project, meetings will be held to discuss and to follow the ongoing and future actions of the project. Meeting minutes will be noted, they will serve as the official written record of the committee's work. As far as possible, the use of audio or video conferences between France and Brazil will be recommended for most of the meetings of the steering committee of the GDRI.

4.2.1 Internal Communication

Resources such as wiki, mailing lists will be used during the project. All the partners will have the access and the right to view, edit and modify the site contents. The wiki site will contain, the agendas of actions (meetings, seminars, workshops), the minutes, the contacts of the partners, the ongoing discussions).
The Wiki will be integrated into the private part of the project site and will also incorporate a file sharing system. Beyond this initiative, other conventional means of communication will be used: meetings (vision conference), mailing list, etc.

4.2.2 External Communication

To communicate the progress, the results and the actions carried out in the project, various means of communications will be established.

- A website will be set up at the beginning of the project. This site will serve as a portal for project partners, as well as for visitors potentially interested in the project. This site will include, the institutional partners, the list of participants, the project description, the latest news, the list of publications and other enhancement activities (conferences, articles, ...) related to the project,
- Organization of seminars, workshops and conferences. These initiatives will be organized both in face to face meetings or via videoconferencing

5 Requested Budget

As the goal of the project is to increase collaboration between Brazilian and French researchers and institutions, the **annual budget** is focused on scientific exchanges:

Annual budget (in euros)				
Type of expense		**Cost per item**	**Total per year**	**Total for 4 years**
Missions	12	2000	24000	96000
Fonctionnements		1000	1.000	4000
			25000	**200000**

The participating institutions will provide the necessary infrastructure, equipment and consumables for the local project activities.

6 Available infrastructure

The project will benefit from the infrastructure already available at the participating laboratories. This includes multiple laboratories, public or thematic, classrooms, auditoriums, administrative areas and other facilities. All institutions will offer space for the installation of an interaction room, with video-conferencing facilities.

7 Laboratories and Partners involved in the GDRI

The Institut de Recherche en Informatique de Toulouse - IRIT
IRIT is a Joint CNRS Research Unit (Unité Mixte de Recherches UMR),it is one of the biggest French laboratories in computer science, It counts 250 faculty members (affiliated to the Toulouse Universities and engineering schools) and 240 PhD students. IRIT is one of the leading computer science laboratories in France and plays an essential role in terms of scientific animation and structuring. The 19 research groups of the laboratory are dispatched in seven scientific themes covering all the computer science domains.

Agorantic
The Federative Sructure Agorantic has been created in 2010 by the University of Avignon in order to mobilize and bring together its best research teams on scientific themes and projects, associating on the one hand Information and Communication Science and Technology (ICST), and on the other hand Human and Social Sciences (SHS). Specific knowledge and expertise are shared in the context of these interdisciplinary projects so that the participants learn new skills and promote the emergence of a Center of excellence based on the following key words: Culture and Communication, Computer Science and Information, Territories and Cultural Heritage, and Society.

LIG
The LIG (Laboratoire d'Informatique de Grenoble, http://www.liglab.fr/) is a joint CNRS research laboratory (UMR), federating researchers from the Grenoble area universities of UJF, Frenoble INP and UPMF, as well as the French national research organizations CNRS and INRIA. The LIG, created in 2007, brings together 200 academic researchers and 240 doctoral and post-doctoral students supported by 65 administrative and technical staff. Research activities are structured around 23 autonomous research groups covering a large spectrum of computer science and a scientific common project around Ambient Computing, based on the development of concepts, methods, software and tools for ubiquitous data and services.

LIP6
LIP6 is a research laboratory in computer science of University Pierre & Marie Curie and CNRS (UMR 7606). With 188 permanent researchers and 244 PhD students, it is a major research laboratories in France.The laboratory covers a broad spectrum of activities grouped in 5 departments : 1. Scientific Computing, 2. Decision making, optimization problèmes in artificial intelligence and operational research, 3. Databases and machine learning, 4. Networks and Distributed Systems 5. Systems On Chips

LIMOS
The LIMOS (Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes), UMR CNRS 6158, is a public research center, joint initiative of the French National Council for Scientific Research (CNRS) and the universities Clermont Ferrand I (University of Auvergne), Clermont Ferrand II (University Blaise Pascal) and the French Institut for Advanced Mechanics (IFMA). The LIMOS includes more than 140 individuals with 74 researchers and more than 60 doctorate students and post-doctorates. The research activity of the laboratory covers a wide range of topics including enterprise integration, large-scale data processing, control and optimization of complex systems, graphs and algorithms and datamining.

PRISM
The PRiSM laboratory at Versailles University is a Joint CNRS Research Unit (UMR 8144) composed of ten teams working in the areas of parallelism, networks, systems and modelling. Fifty permanent

researchers and more than seventy PhD students work at PRiSM. The Advanced Modelling of Information Systems (AMIS) group at PRiSM, involved in this project, has a long experience in databases and information system design. Recently, the group was involved in many projects on data warehousing, data integration, data personalisation and data quality.
PRiSM has a strong cooperation with the Federal University of Pernambuco (Universidade Federal de Pernambuco, UFPE), in Recife, Brazil. Professor Ana Carolina Salgado spent several short visits in our group as well as a sabbatical year in 2007-2008. Zoubida Kedad co-supervised two PhD students from this university, Carlos Pires and Damires Souza, who defended their PhD thesis in April 2009. The two teams participated in a joint research project from 2008 to 2010 in the context of the STIC-AMSUD program. The main research topics of this project were data quality and data integration.

Brazilian institutions involved in the project

- Departamento de Informática - PUC-Rio (Project Coordination)
- Programa de Engenharia de Sistemas e Computação – UFRJ
- Instituto de Computação – UNICAMP
- Instituto de Computação – UFF
- Departamento de Computação – UFC

8 Main publications of the French partners in relation with the project

1. Arlind Kopliku, Mohand Boughanem, Karen Pinel-Sauvagnat. *Towards a framework for attribute retrieval. Proc. of Conference on Information and Knowledge Management (CIKM 2011)*, *Glasgow, UK*, *24/10/11-28/10/11*, ACM, (online), 2011.

2. Lynda Tamine, Mohand Boughanem, Mariam Daoud. *Evaluation of contextual information retrieval: overview of issues and research. Knowledge and Information Systems*, Springer, Vol. 24, p. 1-34, 2010.

3. Malik Muhammad Saad Missen, Mohand Boughanem, Guillaume Cabanac. *Opinion Finding in Blogs: A Passage-Based Language Modeling Approach (short paper).Proc. Int. Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO 2010)*, *Paris, France*, *28/04/10-30/04/10*, p. 148-152, 2010.
4. Guillaume Cabanac: Accuracy of inter-researcher similarity measures based on topical and social clues. Scientometrics 87(3): 597-620 (2011).
5. Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, Pierre Senellart. Web Data Management, Cambridge University Press, to appear, 2011
6. Manuel Atencia, Jerome Euzenat, Giuseppe Pirro and Marie-Christine Rousset. Alignment-based Trust for Resource Finding in Semantic P2P Networks. Proceedings of ISWC 2011 (10th Int. Semantic Web Conference)
7. Remi Tournaire, Jean-Marc Petit, Marie-Christine Rousset, and Alexandre Termier. Discovery of Probabilistic Mappings between Taxonomies: Principles and Experiments. Journal of Data Semantics (JoDS), Vol. 15, pages 66-101.
8. Combining a Logical and a Numerical Method for Reference Reconciliation. Fatiha Sais, Nathalie Pernelle and Marie-Christine Rousset. Journal of Data Semantics, Vol. 12, 2009 .
9. Senjuti Basu Roy, Sihem Amer-Yahia, Ashish Chawla, Gautam Das, Cong Yu. Constructing and Exploring Composite Items. ACM SIGMOD Conference 2010, pages 843-854.
10. Mahashweta Das, Sihem Amer-Yahia, Gautam Das, Cong Yu. Meaningful Interpretations of Collaborative Ratings. Proceedings of Very Large DataBases 2011.
11. Julien Ponge, Boualem Benatallah, Fabio Casati, Farouk Toumani: Analysis and applications of timed service protocols. ACM Trans. Softw. Eng. Methodol. 19(4): (2010).
12. Laurent d'Orazio, Claudia Roncancio, Cyril Labbé: Adaptable cache service and application to grid caching. Concurrency and Computation: Practice and Experience 22(9): 1118-1137 (2010).
13. Laurent d'Orazio, Sandro Bimonte: Multidimensional Arrays for Warehousing Data on Clouds. Globe 2010: 26-37.
14. Hélène Jaudoin, Frédéric Flouvat, Jean-Marc Petit, Farouk Toumani: Towards a Scalable Query Rewriting Algorithm in Presence of Value Constraints. J. Data Semantics 12: 37-65 (2009).
15. Ramy Ragab Hassen, Lhouari Nourine, Farouk Toumani. Protocol-Based Web Service Composition. ICSOC 2008: 38-53.
16. Boualem Benatallah, Mohand-Said Hacid, Alain Léger, Christophe Rey, Farouk Toumani: On automating Web services discovery. VLDB J. 14(1): 84-96 (2005).

17. Myriam Ben Saad, Zeynep Pehlivan, Stéphane Gançarski. Coherence-oriented Crawling and Navigation for Web Archives using Patterns. In TPDL '11: Proceedings of the 15th Int. Conference on Theory and Practice of Digital Libraries (formerly ECDL: European Conference on Digital Libraries), Berlin, Germany, September, 2011.
18. Myriam Ben Saad, Stéphane Gançarski. Improving the Quality of Web Archives through the Importance of Changes. In 22st Int. Conference on Database and Expert Systems Applications (DEXA 2011), Toulouse, France, August 2011.
19. Myriam Ben Saad, Stéphane Gançarski. Archiving the Web using Page Changes Patterns: A Case Study. In ACM/IEEE Joint Conference on Digital Libraries (JCDL 2011), Ottawa, Canada, June 2011. Best Student Paper Award.
20. Zeynep Pehlivan, Anne Doucet, Stéphane Gançarski Changing Vision for Access to Web Archives TWAW, Hyderabad, Inde, March 2011.
21. Zeynep Pehlivan., Myriam Ben Saad, Stéphane Gançarski.Understanding Web Pages Changes. DEXA (1) 2010: 1-15
22. Myriam Ben Saad, Stéphane Gançarski: Using visual pages analysis for optimizing web archiving. EDBT/ICDT Workshops 2010. Best contribution award.
23. Souza D., Pires C.E., Kedad Z., Tedesco P., Salgado A.C., A Semantic-Based Approach for Data Management in a P2P System, in Transactions on Large-Scale Data and Knowledge Centered Systems (TLDKS), special issue on Data and Knowledge Management in Grid and P2P Systems, LNCS, Vol. III, LNCS 6790, 2011.
24. Carlos Eduardo Santos Pires, Rocir Marcos Leite Santiago, Ana Carolina Salgado, Zoubida Kedad, Mokrane Bouzeghoub, “Ontology-based Clustering in a Peer Data Management System”, accepted paper, to appear in Int. Journal of Distributed Systems and Technologies (IJDST), 2011.
25. Berti L., Comyn I., Cosquer M., Kedad Z., Nugier S., Peralta V., Si-Said Cherfi S., Thion V., Assessment and Analysis of Information Quality: a Multidimensional Model and Case Studies, accepted paper, to appear in Int. Journal of Information Quality (IJIQ), 2011.
26. Lemos F., Bouadjnek M., Bouzeghoub M., Kedad Z., Using the Qbox Platform to Assess Quality in Data Integration Systems, Ingénierie des Systèmes d'Information, vol. 15, n°6, 2010, pp 105-124.
27. Carlos Eduardo Pires, Paulo Sousa, Zoubida Kedad, Ana Carolina Salgado, Summarizing Ontology-based Schemas in PDMS, DESWeb 2010, 1st Int. Workshop on Data Engineering meets the Semantic Web, In conjunction with ICDE 2010, 5-6 March 2010, Long Beach CA, USA
28. V. Peralta, V. Goasdoué-Thion, Z. Kedad, L. Berti-Equille, I. Comyn-Wattiau, S. Nugier, S. Sisaïd-Cherfi. Multidimensional Management and Analysis of Quality Measures for CRM Applications in an Electricity Company, In Proceedings of the 14th Int. Conference for Information Quality (ICIQ 2009), November 7-8, 2009, Potsdam, Germany
29. V. Moriceau, E. San Juan, X. Tannier, P. Bellot, “QA@INEX 2009 : A common task for QA, focused IR and automatic summarization systems”, Focused Retrieval and Evaluation (INEX 2009) - Lecture Notes in Computer Science LNCS 6203 - Elsevier, LNCS, 2010
30. Young-Min Kin, P. Bellot, E. Faath, M. Dacos, “Automatic Annotation of Bibliographical References in Digital Humanities Books, Articles and Blogs”, BooksOnline 2011 at CIKM 2011, Microsoft Research Ed., ACM Press, Glasgow, Scotland, 2011.
31. R. Deveaud, F. Boudin, P. Bellot, “LIA at INEX 2010 Book Track”, INEX 2010 - Lecture Notes in Computer Science LNCS - Elsevier, LNCS 6932, p. 118-127, 2011.
32. Laurianne Sitbon, Patrice Bellot, “Topic segmentation using weighted lexical links (WLL)”, ACM SIGIR 07, ACM Press, Amsterdam (Pays-Bas), p. 737-738, 2007
33. Acuna Agost R., Michelon P., Feillet D., Gueye S. A MIP-based local search method for the railway rescheduling problem, Networks, 57(1), 69-86, 2011.
34. Gomes, M. J. N. ; Xavier, Adilson ; Xavier, Airton Fontenele Sampaio ;Philippe Michelon ; Maculan, Nelson . Integração de Sistemas Computacionais e Modelos Logísticos de Otimização para a Prevenção e Combate ao Dengue. Pesquisa Operacional, V. 28, 1-27, 2008.
35. M. Diallo, S. Gueye, P. Berthomé “Sensitivity Analysis on the all pairs q- route flows in a network”, 2009 Int. Transaction in Operational Research 17(1) (2009), pp. 103-117
36. Douvinet J., Delahaye D., Langlois P. (2009) Use of geosimulations and the complex system theory to better assess flash flood risks in the Paris Basin watersheds (France). Proc. of the 3rd Int. Conference on Complex Systems and Applications (ICCSA?09), Le Havre, June 29 July 02, 2009.
37. ROJAS-MORA J., GIL-LAFUENTE J., JOSSELIN D., accepted, On the absolute value of trapezoidal fuzzy numbers and the Manhattan distance of fuzzy vectors, 10 pages, Int. Conference of Fuzzy Computation Theory and Applications (FCTA 2011) 24-26 October 2011 Paris, France http://www.fcta.ijcci.org//home.asp
38. FOLTÊTE J.-C., GENRE-GRANDPIERRE C., JOSSELIN D., 2010, Impact of Road networks on Urban Mobility, in Modelling Urban Dynamics, (Eds. Theriault M. & Des Rosiers F.), Chapter 5, Wiley , NY, 26 pages.
39. CILIGOT-TRAVAIN M. et JOSSELIN D., 2009, Impact of the norm on optimal locations. Computational Science and its applications – ICCSA 2009. Seoul, Korea, June-July 2009, Proceedings, Part I. Osvaldo Gervasi, david Taniar, Beniamino Murgante, Antonio Laganà, Youngsong Mun, Marina L. Gavrilova (Eds.), Lecture Notes in Computer Sciences 5592. pp. 426-441. Springer, Germany.
40. Essaïd Sabir, Rachid El-Azouzi, and Yezekael Hayel, "Towards Sustaining Partial Cooperation in Slotted Aloha-like Protocols Using Hierarchy and Lossy Channel", accepted in Int. Journal of Computer Communications 2011.

41. Veeraruna Kavitha, Eitan Altman, Rachid Elazouzi, Rajesh Sundareshan, "Opportunistic scheduling in cellular systems in the presence of non-cooperative mobiles" accepted in IEEE Transactions on Information Theory, 2011.
42. ETHIS E ., “Le public de théâtre : un état des lieux” in Le journal du CNRS, mars 2009.
43. MALINAS D. et ZERBIB O. « Happy together : le festival de Cannes en Photographies. Témoigner un affect cinématographique », in 20 ans de sociologie de l’art : Bilan et perspectives, Tome I, L’Harmattan, Collection Sociologie des Arts - Logiques sociales, Paris, Chicoutimi, 2007.