

1 Equipe MRIM - Axe Traitement de Données et de Connaissances à Grande Echelle

1.1 Scientific Presentation

1.1.1 Research group members

The MRIM team currently includes 7 permanent researchers (Table 1), 8 PhD students, 2 post-docs and 1 invited researcher.

Table 1: MRIM Permanent Researchers

Name	First name	Position	Institution	Arrival date
Berrut	Catherine	Full Professor	UJF	Sep 1985
Chevallet	Jean-Pierre	Associate Professor	UPMF	Sep 2008
Denos	Nathalie	Associate Professor	UPMF	Sep 1998
Fauvet	Marie-Christine	Full Professor	UJF	Sep 2003
Lbath	Ahmed	Full Professor	UJF	Mar 2012
Mulhem	Philippe	Research Scientist	CNRS	Dec 2003
Quénot	Georges	Research Scientist	CNRS	Jan 1998

Marie-Christine Fauvet was on Long Term Sick Leave (CLM) from August 2012 to May 2013.

1.1.2 Group evolution in terms of members

During the period, the MRIM team had one departure (Éric Gaussier, Professor) and one arrival (Ahmed Lbath, Professor).

Prof. Yves Chiaramella, who retired in 2008, ended his emeritus in 2010.

Georges Quénot was promoted as Director of Research at CNRS in 2012. Catherine Berrut was promoted Professeur Classe Exceptionnelle in 2012, Ahmed Lbath, Professeur 1ère classe in 2010, Jean-Pierre Chevallet, Maître de Conférences Hors-Classe in 2011.

Jean-Pierre Chevallet and Philippe Mulhem are officially authorized to use 20% of their research time for scientific advisory for a company dedicated to indexing and retrieval of still images on servers and mobile devices.

1.1.3 Research description – themes

The research carried out in the MRIM targets Information Retrieval and Mobile Computing domains. While studies done in Information Retrieval are dedicated to satisfy users information needs from a huge corpus of documents, those which are conducted in Mobile Computing are dedicated to satisfy mobile users needs in terms of services taken from a corpus of services and then, composed altogether: in both domaines, users express their needs through queries, and the system gives back relevant documents or personalised services i.e., documents/services that match users' request.

An Information Retrieval system “merely informs on the existence (or non-existence) and whereabouts of documents relating to his request”, as Lancaster wrote in 1968. Nowadays hundreds of millions of people use Information Retrieval systems every day when they use a web search engine or search their emails, and thus Information Retrieval has become the dominant form of information access. For the last decade, the context of Information Retrieval has broadened not only to include all types of data (texts, images, video, etc.), but also to all types of users, and has been widely used through other applications (filtering, recommending systems, social networks, etc). International competitions in this field offer opportunities for both industrial and academic to compare the performances of their systems and approaches, and have now reach a high level in terms of quantity of data, in terms of tasks, in terms on type of corpus, etc.

The research developed in the MRIM group is dedicated to Information Retrieval and access, and Mobile Computing, through 4 main axes:

1. Formalization and models for Information Retrieval (see Section 1.2.1);
2. Development of operational models. The access to digital data includes many facets: the data itself (structure, type), the access to data (multilingual, remote access), and the individualized access (personalization, filtering - see Section 1.2.2). The work of the MRIM teams is therefore divided in the five following sub axes:
 - (a) indexing and retrieval of structured documents;
 - (b) multimedia indexing and retrieval (including still images and videos);
 - (c) contextual mobile information access;
 - (d) semantic and multilingual access to textual information;
3. Evaluation and production of resources. We work on the final evaluation of developed systems by participating at international campaigns but we also are members of advisory committees for international campaigns. We also product annotated resources needed for system evaluation (see Section 1.2.3).
4. Design and implementation of software systems which provide context-aware personalised services for mobile users. This axis has started with the arrival of Ahmed Lbath in the group in March 2012 as Professor at UJF (see Section 1.2.4).

1.2 Scientific and Technological Results

1.2.1 Formalization and Models for Information Retrieval

Modelling an Information Retrieval Systems (IRS) consists on building a formal description of an IRS module (Analysis, Indexing, Matching, or Ranking). This modelling activity is essential to understand the behavior of existing IRS, and it can be a way to proposed alternative and better IRS solutions. We work on modelling the matching and ranking IRS activity by using logical models. The use of logic for this modelling, rises from the following hypothesis:

A document is an answer to a query, if there exists a logical deduction chain that starts from the document and ends to the query.

This deduction chain can be a fuzzy one, i.e. a probability to deduce the query from the documents and used for ranking. We have proposed a new IR logic matching model using logical Boolean lattice mixed with a probabilistic function over this lattice. This modelling enables matching functions to be decomposed into a *direct* matching function (deduction from the document to the query), and a *reverse* matching function, that evaluate the strength of the deduction from the query to the document. Moreover we have shown that most IR matching function can be decomposed into these two more basic matching functions. This work as conducted to a PhD thesis (Mr. Abdulahhad), and as also being published as short paper in the top conference ACM-SIGIR 2013.

1.2.2 Development of operational models

Structured Documents One main difficulty for IR on structured documents is, due to computational cost, the impossibility to handle complex relationships between documents parts on large collections. The MRIM research dedicated to structured documents retrieval adapts and extends current state of the art theoretical works on language models of atomic documents for information retrieval. More precisely, we formalize propagation of terms occurrences between XML documents parts in a way to contextualize these parts according to the whole document and their structural neighbours. This leads to modification of document parts index, as a smoothing, yet keeping such extensions tractable for large corpus of documents. We experimented our proposal during the Inex 2009 campaign, and we obtained the best results on 3 evaluation measures on the 6 official measures. This work led to one international publication and 3 national publications (2 conferences and one journal).

Multimedia : Semantic Indexing of Video Documents Searching in image, video and audio collections has proper specificities, among which the “semantic gap” problem is the most challenging. The semantic gap refers to the “distance” between the signal samples (audio samples or pixels) of which raw multimedia documents are made of and the concepts and/or relations that make sense to human beings. Concept indexing or document categorization is very important for multimedia content-based search. The most common approach is based on supervised learning from labeled examples. Several challenges need to be addressed for efficient and practical multimedia content-based indexing and retrieval:

- the level of concept classification performance, still quite low on “wild” conditions (in the 0.2-0.3 range in a scale from 0 to 1);
- the generalization capabilities: the performance of the classifiers significantly degrades when they are used in domains different from those on which they were trained;
- scalability: classification methods need to be still operational when applied large numbers of documents, target concepts and content types.

We addressed these challenges using a generalized and sophisticated classification pipeline, working on all important stages: descriptor extraction and aggregation, descriptor optimization, classification for highly imbalanced datasets, fusion of classifiers, and re-ranking using the temporal and conceptual contexts. We also worked on the process of efficiently producing annotation using active learning and active cleaning approaches. We experimented these approaches in the context of the TRECVID¹ and MediaEval² international evaluation campaigns. In the 2013 issues, we got the second place within 26 participating groups at the TRECVID semantic indexing task and the first (resp. second) place within 5 (resp. 9) participants at the MediaEval subjective (resp. objective) violence detection task. This work led to five publications (of which two are to appear) in international journals, one as a book chapter, and several in national and international conferences.

Multimedia : Person Identification in Video Documents People are among the main elements of interest for users searching in video collections. Therefore, indexing who is appearing and who is speaking, or who is mentioned either in the speech or in the image track, is a major goal for content-based video indexing. Though the problem is similar to general concept indexing, quite specific techniques can be used to obtain the maximal performance in this very important practical case. In this domain, we focused on:

- written name extraction by developing and improving overlaid text recognition techniques;
- unsupervised naming of persons using written names, pronounced names or both;
- multimodal fusion for person identification in video documents.

We experimented these approaches in the context of the REPERE³ national evaluation campaign where we were often ranked first within the three participants. This work led to one publication in a national journal, and several in national and international conferences. A tool for overlaid text extraction⁴ has been made publicly available.

Multimedia and contextual mobile information access : Indexing and mobile device compatible Retrieval of Still Images The MRIM group is also involved in still images indexing and retrieval. We believe that spatial organization of images are important to be taken in account during indexing and retrieval, so we developed several approaches to include such spatial elements to represent images for information retrieval: integration of spatial locality in image annotation processes, integration of spatial relationships in vector space and extensions of language models in a way to integrate graph-based representations of images. The IOTA system has been developed on these works: it can index images on a server, and the retrieval can be performed on both servers and autonomously on mobile

¹<http://trecvid.nist.gov/>

²<http://www.multimediaeval.org/>

³<http://www.defi-repere.fr/>

⁴<http://mrim.imag.fr/johann.poignant/#download>

devices. Since 2012, part of these works are licensed to the Globe-VIP EyeSnap⁵ company, dedicated to indexing and retrieval of still images on servers and on mobile devices. At the academic level, this work led to two publications in international journals, one in a national journal, and several in national and international conferences. The applications developed with Globe-VIP led to prototypes tested in a museum.

Semantic and multilingual Access to Textual Information Semantic indexing consists in using explicit concepts as index instead of keywords, as in traditional IR. Transforming a natural language text into a sequence of concepts leads to specific problems due to the mapping processes from text to concepts. The weighting scheme for conceptual indexing, published in DEXA 2012, proposes a new counting function that is adapted to concept weighing, when one single portion of text can be mapped on several overlapping concepts. This work uses Medical knowledge from UMLS meta-thesaurus and is applied on medical test collections from the CLEF initiative. Finally, we have worked on the introduction of knowledge extracted from Wikipedia source, into a language model matching function. This conceptual indexing research is strongly linked with the logical modeling of the matching function.

Multilingual Access to Textual Information need the production of multilingual resources. These resources are efficiently automatically produced using parallel or comparable multilingual corpora. A reliable comparability measure of comparable corpora (Bo Li's PhD Thesis) is an important results in the automatic building of large multilingual terminological resources using existing Web comparable multilingual corpora.

1.2.3 Evaluation and production of resources

Evaluation is very important to measure the effectiveness and progresses of information indexing and retrieval methods. We participate in many national and international evaluations campaigns such as TREC, TRECVID, CLEF, INEX, VOC and MediaEval. We also contribute to the organization of such campaign by providing ancillary data and tools, by participating to advisory committees (TRECVID) or by directly organizing them (TRECVID semantic indexing task). These participations enable the positioning of the team among international research as previously mentioned.

Production of annotated resources is also very important both for system training and system evaluation. In the context of the Corpus project of the Quaero programme, we produced over 30 million of concept annotations in still images or in video shots. Most of them were publicly released for the TRECVID 2010-2014 semantic indexing tasks. Such annotations were produced using active learning and active cleaning approaches ensuring that each annotation made is both reliable and as useful as possible for system training.

1.2.4 Context-aware personalized services for mobile users

This part is dedicated to integration of web-based Information Systems. It aims at addressing issues that arise when designers want to assemble pieces of Information Systems accessible by the means of web services. Since the arrival of Ahmed Lbath (in March 2012), the studies involved in this axis are conducted as a part of a broader project whose main goal is to design and implement a software system which provides context-aware personalized services for mobile users. This project is partially funded by European Union's Erasmus Mundus project "Sustainable e-tourism, Erasmus Mundus Action 2 programme".

More precisely, the issues addressed are related to mobile computing, and more specifically to system design, software architecture, expansion of information systems, distributed and heterogeneous resource access and integration (indexing and querying in highly distributed and heterogeneous environments), quality of service requirements (in presence of limited bandwidth and service discontinuity), and information synchronization when switching between connected and disconnected mode. Challenges we attempt to respond are those related to: i) new models for user's privacy preservation in clustering and context aware recommendation (Mou Lei's PhD Thesis, funded by the European Commission), ii) semantic composition of recommended services into a composite service, and execution of the resulting composite service (Pathathai Na Lumpoon's PhD Thesis, funded by the European Commission), iii)

⁵<http://www.eyesnap.fr/>

discovery of services and recommendation (Isaac Caicedo’s PhD Thesis, supported by COLCIENCIAS–COLFUTURO Colombia (convocatoria 528 de 2011) scholarship, and also founded by the University of Córdoba in Colombia).

1.2.5 Publication statistics

Table 2 summarizes the publications made by the MRIM team during the period. This table does not include project deliverables.

Table 2: MRIM publications

	2009	2010	2011	2012	2013	2014	Total
International peer reviewed journal [ACL]	1	6	1	5	2	2	17
International peer-reviewed conference proceedings [ACT]	15	20	12	18	19	4	88
Short communications [COM] and posters [AFF] in conferences and workshops	1	0	1	0	0	0	2
Scientific books and chapter [OS]	0	1	2	1	0	2	6
National peer reviewed journal [ACLN]	1	2	0	0	1	1	5
National peer-reviewed conference proceedings [ACTN]	5	3	5	5	4	5	27
Book or Proceedings editing [DO]	1	0	0	0	0	0	1
Invited conferences [INV]	0	2	0	1	0	0	3
Doctoral Dissertations and Habilitations Theses [TH]	3	2	0	2	1	1	9
Other Publications [AP]	4	2	2	2	2	0	12
Total	31	39	23	34	29	16	170

The IOTA system developed by MRIM, has been deposited at APP (Agence Protection des Programmes) in 2010. It is licensed to the GlobeVIP⁶ startup.

1.3 Visibility and attractivity

Nominations and Publication awards

- C. Berrut, named to the “grade de chevalier de l’ordre des Palmes académiques”, 2011.
- É. Gaussier and S. Clinchant, best paper award at the CORIA 2010 conference.

Participations to national and international institutions

- C. Berrut: Vice-President for Human Resources at Université Joseph Fourier (2007-2012). Member of the Administrative Board of the Société Informatique de France⁷ since 2013.
- N. Denos: Expert at “Mission Numérique pour l’Enseignement Supérieur” at French Ministry of Higher Education and Research, in charge of C2i national certification (digital competencies). 2009-2014. Member of the France Université Numérique committee⁸
- A. Lbath: Expert at the national pedagogical commission at the Ministry of Education and Research (MESR), from 2007. Affiliate Scientist off Site at ITL Lab NIST in Washington DC metro, USA (under formalisation, starts in 2014).
- G. Quénot, responsible of the IRIM (Indexation et Recherche d’Information Multimédia) action of the GDR ISIS.

⁶<http://www.globevip.fr>

⁷<http://www.societe-informatique-de-france.fr/>

⁸<http://www.france-universite-numerique.fr/>

International Collaborations

- C. Berrut, is the leader of the PHC project with University de la Manouba in Tunisia (project funded by Ministry of research in Tunisia, and MAE/MENSR in France), she participates to Master lectures in Tunisia, from 2014.
- M. Dumas, Professor at University of Tartu, Estonia, collaborates regularly with M.-C. Fauvet (in terms of student exchanges and mutual invitations), from 2000. M. Dumas is Invited Professor by Grenoble INP in 2014.
- M.-C. Fauvet: Visiting professor at Queensland University of Technology, Brisbane, Australia from Sept. 2009 to Aug. 2010 (invited by Marcello La Rosa). Supervisor of a “co-tutelle” PhD thesis with the National University of Colombia in Bogota (collaboration with Helga Duarte Associated Professor at this University) from May 2012.
- J.-P. Chevallet supervisor of “co-tutelle” PhD thesis with: University de la Manouba of Tunis (collaboration with Pr C. Latiri and Pr. Y. Slimani), from 2014. University Sains Malaysia (USM) at Penang, from 2011.
- University of Addis-Abeba - Ethiopia, and he participates to Master classes in Ethiopia, from 2010.
- A. Lbath: Visiting Professor at ITL Lab NIST in Washington DC metro, USA . Sept 2012 to Oct 2013 and June 2014 to Sept 2014. Supervisor of a “co-tutelle” PhD thesis with the National University of Mongolia (with Dr Majig Mend Amir, Departement of Mathematics). Supervisor of a PhD thesis with the Chiang Mai Universty, Thaïlande (with Dr Dorn Pradorn, Departement of Media and Art). Supervisor of a PhD thesis with the ITL Lab, NIST, Washington Dc Metro, USA (with Dr A. Batou Division Chief at ITL). Supervisor of 2 PhD thesis with the UMQ University, Saudi Arabia and University of Purdue, USA (with Dr S Basalamh and Dr W Aref). Co-supervisor of a PhD thesis with the University Hassan 2, FST, Morocco (with Pr A Boulamkoul). In charge of a collaboration program between US federal agency and French universities started from May 2014.
- P. Mulhem gives a lecture at the Ecole d’été en Recherche d’Information in Tunis (2014).
- G. Pasi, Professor at University of Milan, Italy, is an invited Professor of Université Joseph Fourier in June 2014.
- G. Quénot, co-organization of the NIST/TRECVID benchmark on video retrieval.

Editorial boards

- C. Berrut: Editor of the Document Numérique Volume 17 – n° 1/2014 (special issue on Nouvelles Applications en Recherche d’Informations in 2014). Editor of the I3 special issue on Recherche d’Informations in 2014.
- A. Lbath: Coeditor of the book “Smart Cities and Location Based Services”, ISTE and Hermès Science Publishing Ltd, 2014
- G. Quénot: Co-editor of the Multimedia Tools and Application journal special issue on Content Based Multimedia Indexing published in 2010. Co-editor of the book “Fusion in computer vision”, Springer in 2014.

Organization of events, chairing conference/program committees, steering committees

- C. Berrut: Member of the steering committee of ESSIR (European Summer School on Information Retrieval) since 2003. Program chair of the poster session of ECIR 2009 (European Conference in Information retrieval). Program chair of the Conference en Recherche d’Information et Applications (CORIA) 2013.

- M.-C. Fauvet, member of the steering committee of the Web Intelligence Network funded by the French Rhône-Alpes region, from 2007 to 2010. Co-chair of the demo session of BPM 2013 in Beijing, China.
- J.-P. Chevallet, co-organizer and co-program chair of workshop “Recherche d’Information Sémantique” RISE⁹, from 2009.
- É. Gaussier and S. Aseervatham, organization in Athens with the NCSR Demokritos of the first PASCAL “Large scale hierarchical text classification” challenge, 2009.
- A. Lbath: Poster and Demo Chair of the ACM SIG Spatial Conference in 2013 and 2014, Poster and Demo Jury Chair in 2009, 2010, 2011 and 2012 Publication Chair of the 10th ACS/IEEE International Conference On Computer Systems And Applications, AICCSA 2013
- P. Mulhem, organisator of the French EARIA summer school in 2014.
- G. Quénot: Organisation Chair of the Content Based Multimedia Indexing (CBMI) workshop in 2010. Program chair of CBMI workshop in 2014.

Institutional evaluation committees

- C. Berrut: ANR (National Research Agency) in 2009, 2010, 2011. Fonds québécois de la recherche in 2009. STIC-AMSUD project in 2011. University of Leuven (Belgium) research projects in 2011. Evaluation of PES (Prime d’Excellence Scientifique) for the University of Marseille in 2013. Languedoc-Roussillon Research project (2013).
- J.-P. Chevallet: ANR (National Research Agency) in 2012, 2013. DIGITEO, in 2009.
- N. Denos: ANR (National Research Agency) in 2012, 2013.
- M.-C. Fauvet: ANR (National Research Agency) (Young Researcher Program) in 2013 ANR INS in 2011 and INFRA in 2013. Founding scheme “recherche sur la nature et les technologies”, Québec en 2011
- P. Mulhem: ANR (National Research Agency) Programme Blanc, in 2013. DIGITEO, in 2014.
- G. Quénot: ANR (National Research Agency) in 2011 and 2014. DIGITEO, in 2010.

Recruitment committees

- C. Berrut, member of committees for Professor and associate Professor recruitment in the Universities of Caen, Cergy, INSA Lyon, Marseille, Nancy, Paris VI, Rennes (Lannion), INSA Rouen, Saint-Etienne, Toulon, Toulouse.
- J.-P. Chevallet, member of committees for associate Professor recruitment in INP Grenoble (2010 and 2011), University of Lyon 1 (2012), University Paris 13 (2013).
- M.-C. Fauvet, member of committees for Professor and associate Professor recruitment in the Universities of Nancy (2011 and 2012), Rennes in 2011, and Grenoble INP in 2011. Member of Jury panels for the promotion and recruitment of “Ingénieur Recherche CNRS” (2009).
- A. Lbath, member of committees for Professor and associate Professor recruitment in the Universities of Grenoble, Lyon, Strasbourg, Dijon and Montpellier. Member of committee for Researcher recruitment in the CEMAGREF agency.
- P. Mulhem, member of committees for associate Professor recruitment in the Universities of INSA Lyon.
- G. Quénot, member of committees for associate Professor recruitment in the University of Bordeaux and in the University of Cergy-Pontoise.

⁹[HTTP://rise.imag.fr](http://rise.imag.fr)

1.4 Social, economical, and cultural impact

During the 2009-2014, the MRIM team was involved in a number of research projects

International projects

- Quaero programme¹⁰, 2007-2013, 1414 k€ for MRIM, 25 French and German participants. Quaero aimed at improving of key technologies for creation, retrieval, access and use of digital contents. In this project, MRIM had a significant contribution. It participated to the Quaero Core Technology Cluster (CTC) Project and was involved in all tasks related to image, video and multimodal processing, including concept indexing in images and videos, person identification in TV shows of journals, image clustering, filtering of image search engine results, and evaluation. We also participated to the Quaero Corpus project within which we managed the production of tens of millions of concept annotations in images or in video shots. Many of these were made publicly available to the community in the context of the TRECVID evaluation campaigns.
- CHIST-ERA Project CAMOMILE¹¹, 2012-2013, 250 k€ for MRIM and GETALP, participants: LIMSI, IMMI, LIG, CRP-GL (Luxemburg), ITU (Turkey) and UPC (Spain). Collaborative Annotation of multi-MOdal, multi-Lingual and multi-mEdia documents (joint project with GETALP).
- International RIFMS-FD project: this PHC-Utique project is funded by Ministry of research of Tunisia and MAE/MENSR in France (through Campus France), the overall budget is 65 k€, it lasts 3 years, from 2014 till 2016. The participants are the University de la Manouba (Tunis), LIRIS (Lyon), LORIA (Nancy). The responsables of the project are C. Berrut (MRIM-LIG, France) and Y. Slimani (Tunisia). The project allows 3 PhD co-supervisions between France and Tunisia, and senior/junior researchers exchanges. The project subject is multi-sources information retrieval and data mining.
- European Union's Erasmus Mundus project "Sustainable e-tourism, Erasmus Mundus Action 2 programme". 2011-2014, 2M€ for the European partners (450 K€ for UJF and MRIM), participants: five European Universities (France, Italy, Germany and UK) and six Asian Universities (Mongolia, Laos, Cambodia, Vietnam, Thailand and China). This project aims to support mobility of students, scholars and academic staff members from non-EU countries to EU countries. The research project is about addressing challenges on context aware mobile computing domain. One of the main goals is to design and develop a context aware mobile services framework with respect of privacy that is applied to mobile tourism.

National projects

- ANR project Metricc, 2008-2012, 147 K€ for MRIM and AMA, participants: Université de Bretagne Sud, Université de Nantes, LIG-UJF and three SMEs. METRICC aimed at exploiting the possibilities offered by comparable corpora, which include texts in different languages, in the context of three industrial applications: translation memories, interlingual information retrieval and multilingual information categorization.
- ANR project VideoSense¹², 2010-2013, 146 K€ for MRIM and GETALP, participants: Eurecom, ECL-LIRIS, LIG, LIF and Ghanni. The VideoSense project aims at automatic video tagging by high level concepts, including static concepts (e.g. object, scene, people, etc.), events, and emotions, while targeting two applications, namely video recommendation and ads monetization, on the Ghanni's media assess management platform.
- ANR project QCompere, 2012-2014, 81 K€ for MRIM and GETALP, participants: LIMSI, INRIA, GREYC, LIG, Vocapia and Yacast. The QCOMPERE project aims at building a consortium of Quaero partners for participating to the REPERE challenge. The project is organized in 6 task, with four tasks addressing a single modal questions: who is seen, who is speaking, whose name is written on screen, whose name is pronounced; a fifth task addresses the multimodal fusion and

¹⁰<http://www.quaero.org/>

¹¹<http://camomile.limsi.fr/doku.php>

¹²<http://www.videosense.org/>

the sixth task provides technology coordination and annotated data resources for the project (joint project with GETALP).

Regional projects

- Competitive Grant from Web Intelligence (Rhône-Alpes French Region). This project works as a network of excellence which involves research teams from the main universities located in Rhône Alpes. Y. Taher's PhD thesis was funded by this scheme (see [9] and Section 1.6).
- Clicide, project funded by Rhône-Alpes French Region, 2012-2013, 46 K€: indexing and retrieval of still images with mobile devices. This project, in collaboration with one of the leaders in audioguides, Ophrys, the Globe VIP company, and the Museum of Art of Grenoble, allowed us to show a real-sized proof of concept of mobile-based image retrieval to get multimedia information from a image query of paintings.
- Competitive Grant from ARC6 research action (Rhône-Alpes French Region). In this context, the informal collaboration between the Hubert Curien laboratory of Saint Etienne is strengthened by a joint PhD thesis on personalized socialized information retrieval, in collaboration with the Best Of Media group (www.bestofmedia.com). This thesis starts in 2014.

Consulting

- Research Contract with ClearBus (8K€ HT, one year in 2009). This collaboration was meant to conduct a state of the art on Digital Identity. M.-C. Fauvet and E. Gaussier were the 2 co-investigators.

University scale project

- APIMS projet funded by Pôle MSTIC (UJF) , 2009-2010, 12-month post-doc + 20 k€ for MRIM and partners: Apprentissage Parallèle pour l'Indexation Multimédia Sémantique. (joint project with GETALP, MESCAL and GIPSA/GPIG).

1.5 Team Organization and life

We have team meetings once a week, each Friday morning. These mostly consist in talk by invited speakers, and MRIM members about their recent work followed by discussions about how this work can be improved by or joint with the work of other team members. We also have talks or discussions that report about a conference issue, a particularly important paper or some methodology aspect or some useful tools. In order to share our knowledge, PhD students are co-supervised within the group.

As many research activities in the team involve very intensive computations and also require very large storage space, we have developed an internal facility including computing nodes (currently 176 cores) and storage servers (currently 180 Tbytes). Though we also have access for some projects to larger shared structures like Grid'5000 (national) or CIMENT (local), this facility gives us more flexibility, reactivity and storage. This is very important for the participation to evaluation campaigns that are themselves essential for our research. The facility is regularly upgraded whenever possible using contractual funding. It is shared between all MRIM members but also with GETALP members involved in joint projects.

Answers to the comments from the experts during the AERES previous evaluation.

The previous AERES evaluation of the MRIM team was quite positive two main problems were mentioned:

“L'évolution du personnel est plus problématique avec le départ de plusieurs membres dont les fondateurs historiques du groupe, celui du dernier professeur recruté sur une thématique plus théorique qui fonde une autre équipe, et un déficit global de recrutement. Pour maintenir le niveau d'activité, il faudra impérativement de nouveaux recrutements. Compte tenu de l'activité du groupe dans les campagnes d'évaluation, l'affectation d'un ingénieur chargé des logiciels et de l'aide aux campagnes paraît également nécessaire.”

The arrival of Ahmed Lbath has enriched the group and leads to new research axes. The recruitment of Lorraine Goeuriot as a permanent associate professor (started in september 2014) will emphasize our work on semantic retrieval, especially on the medical domain.

“Le projet est dans la continuation des activités actuelles et suit la même démarche. C’est un projet solide. Il est bien ancré dans la réalité de la RI moderne, mais il manque probablement un peu de souffle et d’originalité. L’équipe a bien réussi à prendre pied sur de nouvelles thématiques importantes (par exemple les documents structurés, la mobilité, la vidéo). Elle couvre actuellement une large gamme de problématiques classiques de la RI. Elle gagnerait peut être à réduire son périmètre thématique en focalisant ses efforts notamment en matière d’évaluation. Faut-il continuer à couvrir un large pan du domaine ? L’évolution et l’expansion de la RI remet en cause cette stratégie et il serait peut-être plus profitable à terme de se focaliser et de mettre en avant des aspects plus originaux pour le domaine.”

There are actually two concerns here: we might be doing too many different things and participating to too many evaluations. Concerning the first one, we focused more on core topics like formal models for IR, semantic indexing and multilingual IR. Concerning the evaluations, they are really necessary and important in our methodology and generally explicitly included in the projects we participate in (e.g. Quaero or QCompere). We currently both participate to them and contribute to the organization of some of them. We were however more selective and participated only to the most significant ones. We also developed re-usable tools for reducing the participation cost.

“L’activité services (SOA) est déconnectée des autres thématiques de l’équipe mais même si la situation devrait évoluer à terme, cela ne pose pas de véritable problème de cohérence.”

The following remark does not apply any more. In fact, all activity about services is now well integrated into MRIM themes.

1.6 Training through research, educational involvment

Five members of the team are professors or associate professors and mostly teach in the field of Databases, Languages and Programmation, and Information Retrieval or associated subjects like image or video processing for instance. The two full time researchers of the team are also involved in teaching on the same subjects, typically with volumes from 30 to 50 hours per year each. The teaching is done mostly at the master level.

PhDs

- 11 PhD students have defended their thesis since January 2009 (Qamar Ali Mustafa [], Yehia Taher [9], Le Xuan Hung [5], Thi Hoang Diem Le [4], Trong-Ton Pham [6], Rami Albatal [2], Li Bo [3] Bahjat Safadi [8], Johann Poignant [7], Karam Abdulahhad [1]) and Kiam Tan Lan []. Two of them got a position of associate professor, two of them are doing post-docs and all others found a position in the industry.
- There are currently 8 PhD students and 2 post-docs in the MRIM team. Three PhD students should defend their thesis by the end of the year.

Educational involvment

- J.-P. Chevallet, responsible of the Special Year (Année spéciale), of IUT Informatique at University Pierre Mendès France.
- N. Denos:
 - In charge of Digital Competences at Université Pierre Mendès-France. 2009-2014.
 - Production of 5 MOOCs on digital competencies for the general public (project manager and contribution as an author). 2012-2014.
- M.-C. Fauvet:

- Co-Director of the Master by Research in Computer Science (from 2007).
 - Director of Computer Science Programs at the School of Computer Science and Applied Mathematics at UJF from September 2003 to July 2009.
 - “Chargée de mission” for Information Systems for Education at UJF, since January 2011.
- A. Lbath:
 - Deputy Director of the IUT 1 of Grenoble at UJF (from sept 2011)
 - Head of Computer Science and Internet Programming Department at IUT1 of Grenoble, UJF (from 2006 to 2012).

1.7 Strategy and Research Project

In the next period, we plan to maintain the historical and successful approach of the team. In this approach, general and formal models for information retrieval are first proposed, operational models dedicated to various domains are then derived from them, and finally, these are integrated in a number of practical applications. This chain is constantly evaluated, whenever possible in the context of national and international challenges or benchmarks, and improved by taking into account the evaluation results. While this approach will be maintained for the existing domains, it will also be applied to new data and new challenges like: social networks, smart cities or generalized access to multimedia.

1.7.1 Formal models for information retrieval

We continue our research on formalizing information retrieval, which started in 1990 with Nie logical model. Our work in formalizing IR is deeply linked with efficient IR on large corpora. We work on 3 main axes : logical model linked to probabilistic models, generalized language models and efficient IR models.

- Logical models generally propose theoretical matching based on an implication linked to a specific logic. A major drawback of such models is that they rarely correctly integrate the calculus of the RSV function which allows to know the proximity of the document and the query. We propose a merge of a matching based on a logical deduction and an evaluation of this matching based on probabilities. We already investigated the Propositional Logical model linked to lattices and probabilities, and we will broaden this work through for example Description Logics.
- Language models are among the more successful IR models, they are classically based on ngrams. Our work tackles formal generalisation of language models that integrate all relationships between terms (sequence of course as in ngrams, but also semantic relations that might exist between terms, or topological relations between parts of image description). This impacts all aspects of these models, from the probability definition to the matching itself.
- Defining effective software implementations for theoretical formulas is our objective here. This work is essential to ensure that new proposals have an impact for systems managing large corpora. Such work may lead to classes of correspondence adapted to dynamic data, on very large corpus. But also it could lead to the analysis of power consumption needed in IR systems.

1.7.2 Access and retrieval on multimedia, social and professional data

Instantiation of formal models into operational models will be considered for multimedia indexing, text mining, semantic medical information retrieval, and personalized social information retrieval (not exhaustive).

Concerning the semantic indexing of multimedia contents, we plan to continue our work both on people and on general concepts, as this remains an important application and a problem far from being well solved. We will address the scalability issues both in terms of targets (thousands of people or concepts) and throughput. We have recently started work about the integration of deep learning methods with classical ones and we will further explore this promising direction. We shall also focus on an efficient

and effective use of parallel architectures (computing grids and GPUs) for automatic video indexing. This research will be conducted in particular in the context of the Camomile ANR project.

In the text mining domain, we aim at linking two important research domains: data mining and information retrieval. We want to integrate several factors such as semantics, spatiotemporal contexts, social, geographic dimension, etc. for provide a response, which takes into account several measures (relevance, similarity, space, time, social proximity, cultural ties, etc.). This research will be conducted in particular in the context of a CMCU project with Tunisia.

In the context of medical data processing, we aim at proposing systems able to process long and complex queries requiring an understanding of advanced concepts and real-world knowledge. We will develop approaches for Semantic Information Retrieval (SIR) systems that go beyond the surface form of both queries and documents in order to process meaning. This research will be conducted in particular in the context of the SIRUP project.

In personalized social information retrieval, we aim at exploiting social information related to a user in order to personalize an information retrieval system for him/her. This social information include virtual relationships maintained by the user as well as his/her explicitly defined centers of interest. This information is of a different nature and must therefore be representd adequately in order to integrate effeicacement in an IR system. This research will be conducted in particular in the context of an ARC6 project with the Hubert Curien laboratory of Saint Etienne.

1.8 Self assesment

Strength

The MRIM team is working in the field of Information Retrieval since 1983 and has always been recognized by national and international IR community. This is one of the few teams in the world that its research covers all the main aspects of Information Retrieval including theoretical IR models, search within text, images, video, multilingual and structured document, information filtering, collaborative search and recommendation systems. The methods developed by the team are largely evaluated and compared to other teams within national and international evaluation campaigns including TRECVID, MediaEval, VOC, INEX, CLEF. The MRIM team is leading the IRIM project of the ISIS Network whose goal is to federate research on Multimedia Content Indexing and Retrieval at the national level in order, among other objectives, to give a French answer to international evaluation campaigns. The MRIM team is also leading the semantic indexing task at TRECVID since 2010.

The MRIM team always works with many contractual activities. For example, in 2013, MRIM managed 6 contracts corresponding to 1.7 Million Euros. The MRIM team has a strong industrial link through a scientific collaboration over the long term with a company that integrates and commercialise IOTA system. Most of the professors and associate professors of the team are involved in their universities where they have high responsibilities (for instance, Catherine Berrut was the Vice-President of the University Joseph Fourier). Considering human resources, the departure of Éric Gaussier who moved to the LIG team AMA and the arrival of Ahmed Lbath leads to the same number of permanent people in the group. The arrival of Pr. Ahmed Lbath has reenforced the 4th axis and leads to opportunity of join research for mobile usage. The group will integrate a new associate professor in 2014, Lorraine Goeuriot, coming from Dublin City University.

Weakness

Information Retrieval research is highly driven by evaluation through international campaigns. Due to this strong orientation of our work, the MRIM team has a high number of conference publications and probably less journal publications. As a consequence of these needed evaluations, many of our permanent researchers spent a significant fraction of their research time in engineering works for participating in the evaluation campaigns.

The MRIM team particularly needs to recruit:

- 1 or 2 university lecturers in order to support research in the various axes of the team.

- an engineer, in particular for the assistance with the participation in the evaluation campaigns and to the development of systems on several platforms. To date, the participation in these campaigns is a difficult exercise and even if it allows the recognition of the team at a high level, the researchers or teacher-researchers of the team who carry out these validations devote much of their time to it instead on an ‘classical’ research activity often more recognized;

Since the end of the Quaero programme, the MRIM team has difficulties in participating to large scale european projects. Many teachers are heavily invested tasks related to teaching, which reduces all the research work.

Opportunity

Nowadays, everybody knows IR through search engines or electronic archives. The IR domain is now an effective industrial activity. A lot of big companies such as Microsoft, Yahoo, Google and Exalead, are making business with IR technologies and development. Many small companies or start-ups are also launched in this domain.

Huge amounts of electronic data are available, through Internet or archives. Audiovisual assets are being massively digitized or created. Access to new semantic social data is a strong potential for theoretical work developed in MRIM team.

With hardware development, approaches neglected because of their potential complexity can return to the front of the stage in the field.

Threat

If the niche in which the team works are also strategic issues for the industrial ”big players”, it is difficult for a team of some researchers to compete. On the other hand, if our work is important, it will be possible to have collaborations with companies. Moreover, the field becomes increasingly competitive since more and more academic partners come from other domains (e.g. machine learning, database or signal processing). The fast evolution of technology had a direct impact on implemented models.

References

- [1] K. Abdulahhad. *Modélisation de la Recherche d’Information par la Logique et les Treillis. Application à la Recherche d’Information Conceptuelle*. These, Université de Grenoble, May 2014.
- [2] R. Albatal. *Annotation automatique d’images à base de Phrases Visuelles*. These, Université de Grenoble, July 2010. Date de fin de rédaction 15 Mai 2010.
- [3] L. Bo. *Mesurer et améliorer la qualité des corpus comparables*. These, Université de Grenoble, June 2012.
- [4] T. H. D. Le. *Utilisation de ressources externes dans un modèle Bayésien de Recherche d’Information. Application à la recherche d’information multilingue avec UMLS*. These, Université Joseph-Fourier - Grenoble I, May 2009.
- [5] X. H. Lê. *Indexation des émotions dans les documents audiovisuels à partir de la modalité auditive*. These, Institut National Polytechnique de Grenoble - INPG; Institut Polytechnique de Hanoï, July 2009.
- [6] T.-T. Pham. *Modélisation et recherche de graphes visuels : une approche par modèles de langue pour la reconnaissance de scènes*. These, Université de Grenoble, Dec. 2010.
- [7] J. Poignant. *Identification non-supervisée de personnes dans les flux télévisés*. These, Université de Grenoble, Oct. 2013.
- [8] B. Safadi. *Indexation sémantique des images et des vidéos par apprentissage actif*. These, Université de Grenoble, Sept. 2012.

- [9] Y. Taher. *Un canevas pour l'adaptation et la substitution de services Web*. These, Université Claude Bernard - Lyon I, July 2009.