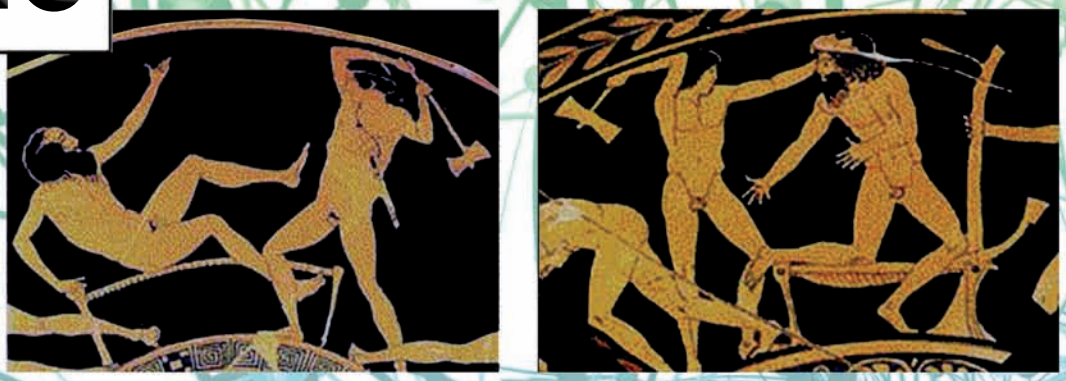
Actualité de l'Institut de Recherche en Informatique de Toulouse

Janvier 2009

16

Numéro

noir SUR blanc



Procuste, symbole de l'uniformité

Éditorial

Luis Fariñas del Cerro
Directeur de l'IRIT

Les données, quelle que soit leur forme et leur localisation, constituent un élément essentiel autour duquel gravitent de plus en plus un grand nombre d'activités humaines.

L'informatique a, depuis ses débuts, été un acteur déterminant de cette évolution. Les systèmes d'information, comme lieu de rencontre des données et des traitements, ont constitué un objet de recherche qui a mobilisé les chercheurs de l'IRIT pour les situer parmi les équipes d'excellence dans ce domaine.

La disponibilité, au niveau planétaire, de grandes masses de données dues au développement du web, a fait émerger de nouveaux défis scientifiques et technologiques, comme celui de retrouver des informations implicites et enfouies au sein de collections d'objets hétérogènes. Ces objectifs supposent de se confronter à des situations non-classiques en Recherche d'Information, telles que les structures de données inconnues ou incomplètes (« semi-structurées ») ou la part grandissante du contexte (implication des utilisateurs, contraintes des tâches…)

L'équipe SIG (Systèmes d'Information Généralisés) s'intéresse depuis plusieurs années au domaine de la recherche d'information dans des systèmes textuels, documentaires et décisionnels. Elle a montré la pertinence de ses travaux avec beaucoup de succès, vu le grand nombre d'applications de la plate-forme Tétralogie sur la veille stratégique, ou la qualité de résultats dans des campagnes d'évaluation internationales comme TREC (Text RETrieval Conference).

Parmi les axes stratégiques définis à l'IRIT, l'axe « masses de données et calculs » constitue un développement naturel des travaux de l'équipe SIG, de même que la plate-forme OSIRIM : Offre de Service pour l'Indexation et la Recherche d'Information Multimédia qui se met en place au sein du laboratoire ; celle-ci permettra à nos équipes, mais aussi à des chercheurs d'autres structures, d'aller plus loin dans leurs recherches dans d'excellentes conditions d'expérimentation.

Directeur de la publication : Luis Fariñas del Cerro **Directeur adjoint de la publication :** Jean-Luc Soubie **Secrétariat de rédaction :** Véronique Debats **Comité de rédaction :** Dominique Bertrand, Cédric Beucher, Vincent Charvillat, Gérard Padiou, Pascal Sainrat, Jacques Virbel **Conception et création de la maquette :** Ludovic Chacun **Ont collaboré à ce numéro :** les membres de l'équipe SIG

Contact de la rédaction : 05 61 55 65 10 - nsb@irit.fr - www.irit.fr
118 Route de Narbonne - 31062 Toulouse cedex 9

ISSN 1620-2937

SIG, Systèmes d'Informations Généralisés

L'équipe a été créée en 1971 et se dénommait CSI pour «Conception de Systèmes d'Information». Ses thématiques de recherches se sont centrées sur les systèmes d'information, classiques ou décisionnels et les bases de données.

Elle devient l'équipe SIG (Systèmes d'Informations Généralisés) en 1981.

Le qualificatif «généralisé» faisant référence à la représentation uniforme, générique, construite, sur des informations variées (textes, sons, images…) et pas uniquement sur des données structurées. Ses problématiques de recherche actuelles sont centrées autour des fonctionnalités de Recherche d'Information (RI) dans ces systèmes (textuels, décisionnels, documentaires…).

On a tendance à caricaturer ces problématiques par la métaphore de la «Recherche d'une aiguille dans une botte de foin». En fait elles sont beaucoup plus complexes, car les recherches ne se font pas uniquement dans une collection unique et homogène («botte de foin») mais dans un ensemble de collections, chacune ayant ses caractéristiques spécifiques différentes d'une collection à l'autre. Lorsque la collection est homogène, la problématique de RI se ramène à retrouver un objet spécifique, ce qui peut se faire simplement lorsqu'on arrive à caractériser les propriétés de ce dernier: en l'occurrence avec la botte de foin, on utilise un détecteur de métaux pour localiser l'aiguille. La complexité des problématiques de la RI provient de la nécessité d'explorer un ensemble de collections hétérogènes.

Tout d'abord, il y a hétérogénéité sémantique en fonction de la thématique attachée à la collection. Il y a ensuite hétérogénéité structurelle; les collections pouvant être non structurées, partiellement ou totalement structurées avec ou non des interconnexions entre elles (on parle respectivement de bases textuelles, hypertextes, bases de données, hyperbases, entrepôts de pages Web).

Les informations mémorisées peuvent aussi présenter des caractéristiques hétérogènes: hétérogénéité de format, de type (textes, images, sons…). Dans ce contexte, pour garantir la généricité, il est nécessaire de proposer des approches basées sur une vue unifiée des ces ensembles de collections hétérogènes pour pouvoir retrouver efficacement l'information recherchée.

Cette complexité s'accroît avec le fait que le système de recherche doit être capable de capter l'intention de l'utilisateur afin d'orienter l'exploration des collections. Son comportement sera différent selon que la recherche ne concerne que des éléments bibliographiques, informationnels ou qu'elle s'intègre dans un processus décisionnel.

Dans ce dernier cas, le système devra agréger et synthétiser un ensemble d'informations extraites des collections en fonction de la finalité de la requête de recherche.

Il est donc nécessaire de concevoir des approches qui ramènent l'intention de l'utilisateur au cœur de la fonction de recherche afin de permettre au système de s'y adapter.

L'équipe SIG

Claude CHRISMENT
Professeur à l'UPS
Responsable de l'équipe SIG
Claude.Chrisment@irit.fr

L'équipe SIG (51 personnes, 22 enseignants-chercheurs dont 9 habilités, 22 doctorants, 7 post-doctorants ou invités) est structurée en quatre composantes ayant pour objectif d'explorer certaines facettes des problématiques précédentes.

- ***RI-EVI : Exploration et Visualisation d'Information***
  Responsables : B. Dousset et J. Mothe
- ***RI-RFI : Modèles adaptatifs pour la recherche d'information***
  Responsable : M. Boughanem
- ***ED : Conception de systèmes d'informations décisionnels (SID)***
  Responsable : G. Zurfluh
- ***DDSS : Documents, Données Semi-Structurées et usages***
  Responsables : F. Sedes et C. Soulé-Dupuy

Les composantes RI-EVI et RI-RFI font partie d'une composante plus générale RI centrée sur la Recherche d'Informations

Recherche d'information, exploration et visualisation d'information (RI-EVI)

La recherche d'information (RI) est présente dans de nombreuses tâches quotidiennes : s'enquérir de la météo du jour pour choisir la façon de s'habiller ou lire dans le journal le résultat des dernières élections font intervenir la RI.

La RI est également largement présente dans la vie professionnelle par exemple lorsque nous nous informons sur les avancées d'un domaine, les nouveaux produits ou alliances, les stratégies des concurrents, ou plus simplement que nous recherchons un document, un rapport.

La RI en tant que domaine de recherche s'intéresse à définir des méthodes et développer des systèmes d'aide à l'accès à l'information lorsque celle-ci est disponible sous format électronique.

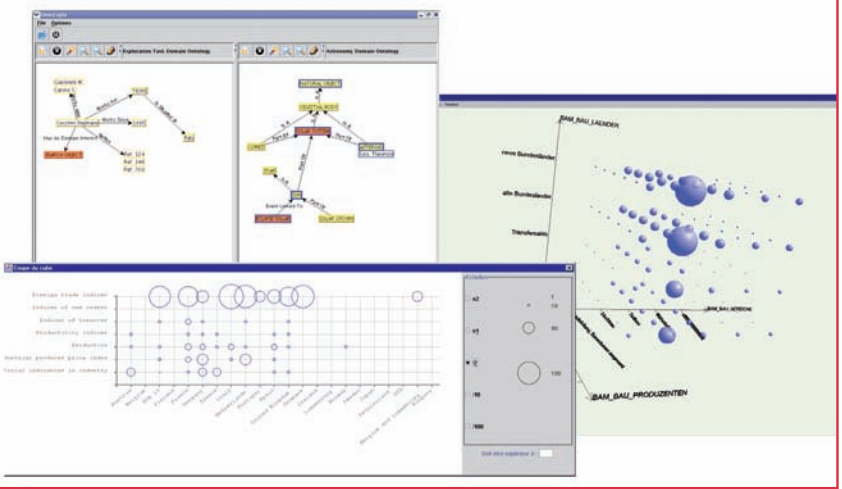
Ces systèmes doivent permettre à l'utilisateur de retrouver les informations qui répondent à son besoin qu'il exprime via une requête ; à charge au système de «comprendre» le besoin exprimé pour répondre au mieux. Pour ce faire, le système annote et indexe au préalable les documents qu'il gère : il s'agit d'éliciter une représentation des informations rendant ensuite possible la mise en correspondance avec les requêtes des utilisateurs. L'indexation la plus simple consiste à extraire des mots ou chaînes de caractères des documents. Si les mots de la requête se retrouvent dans le document, celui-ci est considéré comme potentiellement pertinent.

L'indexation intelligente est au cœur de nos recherches. Le groupe RI-EVI propose des méthodes permettant de s'affranchir de la terminologie ou des variantes des termes utilisés dans les textes. Ce ne sont pas des mots qui sont choisis comme termes d'indexation mais les concepts associés à ces mots. Les méthodes d'indexation s'appuient donc sur une représentation externe de la connaissance d'un domaine qui lie terminologie et concepts. Un domaine peut être représenté par des listes terminologiques, des règles de normalisation, des hiérarchies de concepts, des ontologies, des graphes ou des réseaux sémantiques. Ceci permet de représenter ainsi à la fois le contenu sémantique des documents mais également des informations de haut niveau sur le document (méta-données) telles que son auteur ou sa date de publication.

La notion de tâche est une des composantes du contexte de la recherche (on parle de RI contextuelle) (Figure 1). Rechercher les documents qui parlent de réchauffement climatique est une chose, extraire les apports de tel ou tel document, les évolutions qu'a connu le domaine, les acteurs principaux du domaine et leurs corrélations, implique des

« Une représentation externe de la connaissance d'un domaine qui lie terminologie et concepts »

Figure 1



OntoExplo – DocCube : Exploration de corpus en Astronomie via des ontologies supportées par des représentations visuelles. L'exploration est guidée par la tâche de recherche que l'on souhaite accomplir.

Sommaire :

mécanismes autrement compliqués. Le premier type de tâche se nomme la recherche *ad hoc* alors que la seconde relève de la découverte de connaissance ou fouille de données textuelles et permet la veille scientifique, économique ou technologique. Pour répondre au mieux aux divers besoins des utilisateurs, il faut les identifier. Le groupe RI-EVI a développé une méthode d'analyse de besoins en intelligence économique et territoriale ainsi que des méthodes qui permettent de reformuler automatiquement une requête qui n'a pas donné satisfaction, de gérer des documents longs en permettant une recherche de composants ou paragraphes les plus pertinents (basé sur XML), de détecter les éléments nouveaux apportés par tel ou tel document ou d'extraire et présenter les corrélations qui existent entre les éléments d'information. L'aspect contextuel se décline également au sein d'une même tâche. Il s'agit là de caractériser les contextes et d'y associer les mécanismes les plus adaptés pour assurer une recherche qui s'adapte au contexte rencontré.

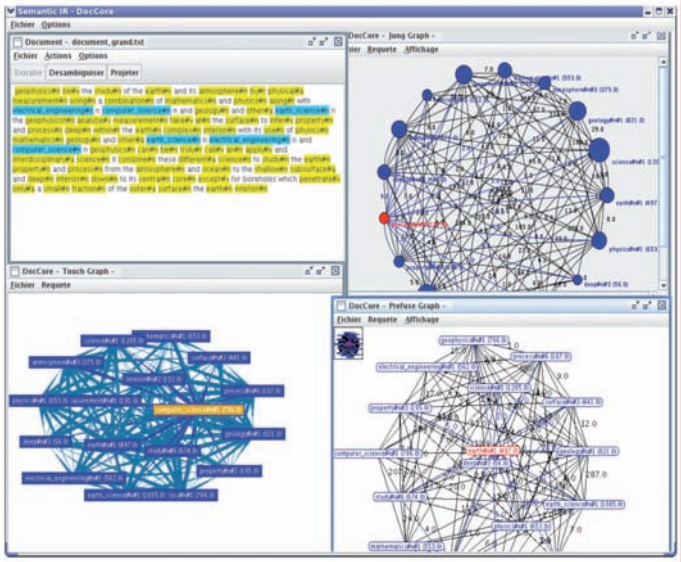
Recherche et Filtrage d'Information (RI-RFI)

La thématique Recherche et Filtrage d'Information s'intéresse à la recherche d'information dans les documents textuels semi-structurés ou non structurés.

Les travaux visent à la mise en œuvre de systèmes de recherche d'information (SRI) fondés sur des modèles théoriques éprouvés et s'appuient sur des approches qui allient théories et expérimentations. Les problématiques actuelles, outre la prise en compte de l'hétérogénéité des informations, concernent l'adaptativité des SRI à leurs utilisateurs et au contexte des tâches de recherches. On parle des problématiques de personnalisation ou de contextualisation de la recherche.

L'objectif de la personnalisation de l'information est d'intégrer l'utilisateur dans le processus global d'accès à l'information en vue d'adapter les différentes étapes du processus de RI (recherche, filtrage, navigation, visualisation) au contexte de l'utilisateur.

Figure 2 *L'amélioration de la précision des réponses relève d'une meilleure prise en compte de la sémantique des requêtes.*



DocCore: prototype de recherche conceptuelle d'informations via des visualisations graphiques de réseaux de concepts

L'équipe SIG à travers quelques prototypes ou systèmes

- Réalisation des systèmes bureautiques Microburo, puis Top-Buro ainsi que la création de la Start-up 'Burogiciel' qui l'a diffusé.
- Mise en œuvre du projet 'BIG' (années 80) avec le service 'Documentation Après-Vente' de l'Aérospatiale, ainsi que AEROFORMATION pour structurer des cours de formation (contexte EAO) pour l'apprentissage au pilotage. Cela s'est matérialisé par la réalisation d'un prototype sur ICL Perq faisant intervenir une des premières tablettes graphiques pour l'interaction multimédia.
- Déploiement du système bureautique intégré à grande échelle BURHAU implanté par l'équipe à l'URSSAF de la Haute-Garonne.
- Déploiement des systèmes d'information :
  - AQUEDUC (AQUitaine EDUCation) déployé dans les lycées de la région Aquitaine et accessible via des bornes d'information spécialisées,
  - INFODIAB : Système d'Information pour les Diabétiques déployé au CHU de Rangueil et accessible par Minitel.
- Collaboration avec Matra (Laboratoire mixte ARAMIIHS) sur la gestion de documentation électronique pour l'aide à la maintenance de systèmes logiciels avec la réalisation d'un moteur de réécriture de textes XREP ainsi que pour la mise en œuvre du concept d'hyperbase dans une application proposée par le CNES au travers du prototype BDE-SP (hypérisation de documents et consultation 'active' par interprétation dynamique de formules).
- Spécification des langages graphiques HQL (Hypertext Query Language), OHQL et réalisation du prototype GRAPHIC-OLAP.
- Réalisation du moteur de Recherche d'Information MERCURE.
- Réalisation du moteur de recherche d'information centré sur des ontologies ONTO-EXPLO.

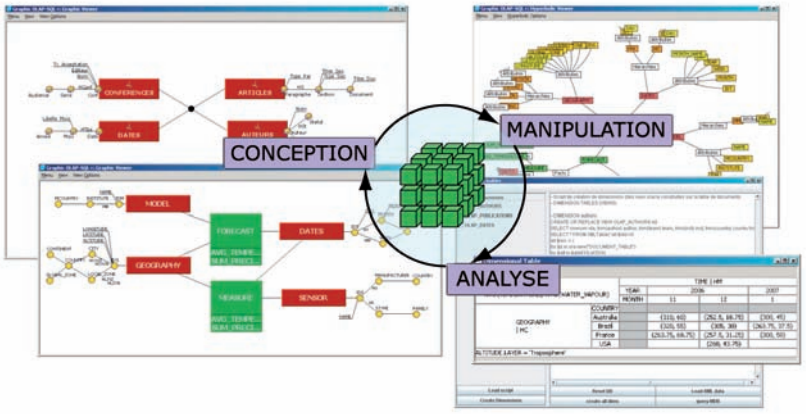
Ce dernier intègre ses centres d'intérêts, ses préférences, son environnement cognitif et spatio-temporel, les réseaux sociaux auxquels il appartient.
Pour capter la sémantique véhiculée par l'information analysée des ressources préexistantes, telles que les ontologies, les thésaurus et les métadonnées, peuvent être utilisées. L'ensemble de ces travaux est consolidé par une importante démarche de validation et d'évaluation expérimentale dans le cadre de campagnes d'évaluation internationales.
Les expérimentations sont conduites sur la plate-forme OSIRIM dont la composante textuelle a été élaborée par l'équipe.
Elle a pour objectif de fédérer des outils et des ressources de référence (corpus d'évaluation, outils d'analyse) dans le cadre de procédures d'évaluation. Elle s'appuie sur une architecture matérielle financée par la Région Midi-Pyrénées combinant espace de stockage d'une centaine de téraoctets et puissance de calcul fournie par un cluster d'une dizaine de nœuds bi-processeurs.
Les principaux résultats de recherche obtenus concernent la proposition de modèles adaptatifs de recherche d'information: MERCURE, classé parmi les meilleurs moteurs dans le cadre du programme d'évaluation international TREC (1997). Ce moteur représente aujourd'hui (1) le noyau sur lequel s'articulent incrémentalement d'autres travaux de l'équipe, (2) le modèle FILTRE permettant le filtrage personnalisé et adaptatif de l'information, (3) le modèle XFIRM pour la recherche flexible dans des documents semi-structurés, (4) les modèles DocTree et DocCore (Figure 2) pour la représentation conceptuelle et sémantique des documents, (5) le modèle SiRIX pour l'accès personnalisé à l'information.
Toutes les compétences capitalisées se sont concrétisées en 2006/2007 par notre participation au projet QUAERO, destiné à développer des outils intégrés de gestion de contenus informationnels.

Entrepôts de données (ED)

La thématique « Entrepôts de données » propose des solutions pour la conception, le développement et l'utilisation de Systèmes d'Information Décisionnel (SID). De tels systèmes d'information visent à extraire les données d'une organisation et à les présenter de façon à faciliter les prises de décisions.
De nos jours, ces systèmes reposent le plus souvent sur différents espaces de stockage pour les données décisionnelles: un entrepôt de données d'une part et des magasins de données d'autre part. Un entrepôt est un espace permettant de stocker les données nécessaires aux prises de décisions d'une organisation ainsi que leurs évolutions dans le temps. Un magasin est un extrait d'un entrepôt dédié à un métier ou une fonction particulière. Dans la mesure où chaque magasin est manipulé par un décideur, il est alors nécessaire d'offrir des modèles de données et des outils d'analyse décisionnelle facilitant l'accès et la manipulation des données.
Les modèles d'entrepôts permettent une représentation uniforme, non redondante, historisée et fiable, des données décisionnelles issues de diverses sources (bases de données relationnelles, fichiers au format XML...).
Les modèles de magasins reposent pour l'essentiel sur une représentation multidimensionnelle des données. Cette modélisation permet de représenter les sujets d'analyse (ventes par exemple) en fonction d'axes d'analyse (temps, produits, clients par exemple).
Les méthodes de conception s'appuyant sur ces modèles ont pour objectif de proposer des solutions pour la conception et le développement de SID. Ces méthodes doivent permettre de concevoir, d'implanter, d'alimenter et d'actualiser les différents espaces de stockage d'un SID. Elles font appel à des concepts, des formalismes et une démarche tenant compte des particularités de type de systèmes. Les travaux sur les langages décisionnels visent à proposer des langages de haut niveau dédiés aux différents utilisateurs (administrateurs, décideurs). Ces langages permettent de définir, de manipuler et de contrôler les données décisionnelles. Les propositions d'ED reposent sur des bases formelles (algèbre) et ont abouti à des langages textuels et graphiques.

« Offrir des modèles de données et des outils d'analyse décisionnelle. »

Figure 3



Prototype Graphic-OLAP
Conception de schémas multidimensionnels en constellation
Manipulation tabulaire via le cube de données et des requêtes graphiques
Analyse des données multidimensionnelles

En matière de conception, ces propositions ont été mises en œuvre et évaluées dans un cadre industriel en collaboration avec la société I-D6. Les résultats de cette confrontation ont permis d'ajuster nos procédures et algorithmes et ont abouti à un catalogue de patrons pour capitaliser le savoir-faire des experts.

En ce qui concerne la modélisation et la manipulation de données décisionnelles, ces propositions ont été déployées au sein de prototypes permettant l'implantation et l'analyse multidimensionnelle OLAP (On-Line Analytical Processing) de données décisionnelles (Figure 3). Elles ont été validées dans des contextes comme l'analyse de données médicales. Par exemple, la collaboration avec, entre autres, l'association de médecins OUTCOME-REA a permis de proposer une solution pour prédire la survenue de complications nosocomiales. Elle se prolonge par une participation au projet IAPA (Infrastructure d'Accès, de Partage et d'Analyse de données biomédicales) soutenu par l'IRIT en partenariat avec l'Institut Claudius Regaud afin de faciliter l'aide au diagnostic.

Documents, Données Semi-Structurées et usages (DDSS)

Depuis son avènement, Internet révolutionne l'informatique « grand public ». HTML est le langage du Web, même s'il encapsule des fichiers (.doc, .pdf, .jpg, .gif), du son, de la vidéo. Des milliards de pages existent actuellement: on parle de web public/privé, statique/dynamique, visible/caché… Il est devenu le support naturel pour l'information distribuée, à destination d'êtres humains et de plus en plus, pour des applications plus ciblées. On assiste en effet à une explosion du développement d'applications distribuées sur le Web: le B2C, B2B, les bibliothèques et fonds documentaires en ligne, le G2C, en sont quelques exemples. HTML n'est pas adapté pour ces applications, qui ont besoin de typage pour représenter la structure des données. On ne peut pas se contenter d'obtenir des ensembles de pages comme avec les moteurs de recherche du Web, les modèles documentaires sous-jacents ne proposant pas assez de structure. Se pose alors la question de savoir: quel modèle de données semi-structurées et quelles structures peuvent s'avérer pertinents?

Faut-il envisager un typage comme dans les Bases de Données (BD)? Pour caractériser les modèles dits « semi-structurés », il faut appréhender les caractéristiques contextuelles suivantes:

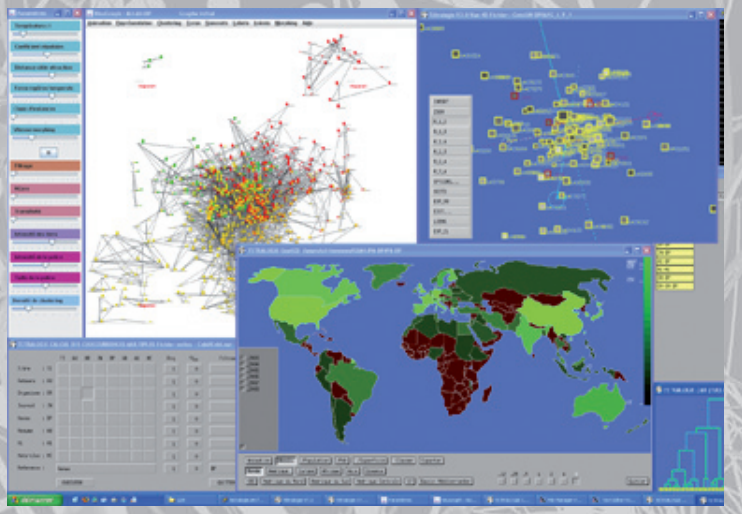
- on ne connaît pas bien la structure des données (contexte du Web) alors que dans une BD, on connaît et on fixe *a priori* la structure des tables et la sémantique des colonnes,
- la structure est irrégulière,
- dans les collections, certaines données sont manquantes, d'autres « supplémentaires » (annotations),
- on constate des variations de type, d'unités (dollar vs. euro), standards différents pour les adresses,
- la structure peut être implicite: il faut analyser les contenus pour la découvrir, c'est-à-dire l'éliciter,
- la structure peut n'être que partielle,
- une partie des données peut être sans structure: plein texte, images, son,
- le typage peut n'être qu'indicatif et/ou dépendre du contexte,
- on tolère des données non strictement conformes à un type spécifié *a priori*,

La plate-forme Tétralogie

La plate-forme Tétralogie est dédiée à la veille stratégique (fouille de données pour intelligence économique).

Elle est alimentée par des données textuelles ou factuelles issues de bases bibliographiques en ligne, de CD/Rom, d'Internet ou de toute autre source informatisée. Par l'intermédiaire de méthodes statistiques, d'analyse exploratoire des données et d'analyse relationnelle, elle permet de mettre en évidence, dans des temps très courts, des éléments d'information stratégique jusque là inexploitables comme: l'identité des acteurs, leur notoriété, leurs relations, leurs lieux d'action, leur mobilité, l'émergence et l'évolution des sujets et des concepts, la terminologie, les domaines porteurs, que lire et où publier, avec qui collaborer.

De nombreux outils de visualisation interopérables et distribués permettent de conduire ces analyses d'information à plusieurs et à distance via le réseau.



Site internet: http://atlas.irit.fr

Cette plateforme est complétée par le serveur web XPlor qui permet de mettre en ligne les résultats de ces analyses stratégiques et qui offre, à l'utilisateur, la possibilité de naviguer dans l'information relationnelle et d'en tirer des graphes statistiques sur mesure.

- le schéma est souvent élaboré *a posteriori* par rapport à la collection : le type se déduit à partir des données,
- le schéma, souvent peu concis et complexe, est parfois ignoré par les requêtes,
- la formulation des requêtes est aussi peu conventionnelle : par mots clés, navigation…
- le schéma évolue très rapidement, ce qui est souvent une raison pour ne pas adopter une approche BD qui repose sur une structure trop rigide.

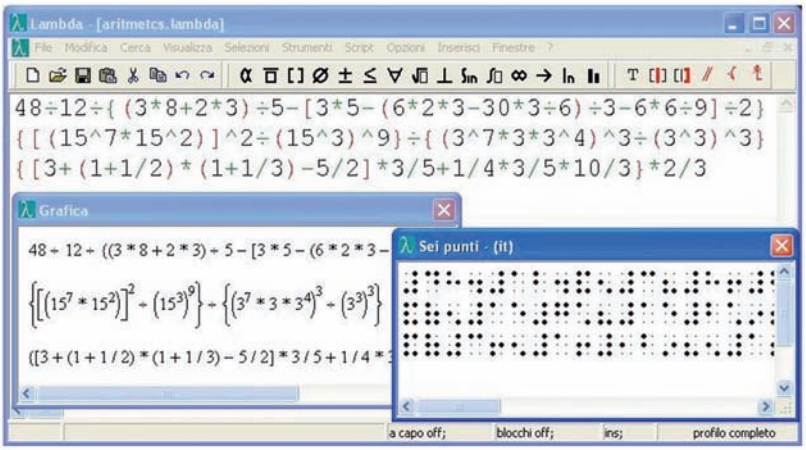
Deux mondes se rejoignent dans les modèles semi-structurés, celui de la gestion de documents et celui de la gestion de BD et de Systèmes d'Informations au sens large.

Le spectre de l'information et des données semi-structurées est abordé dans cette composante à travers les documents multimédias, multi-structurels, avec des cas particuliers comme les partitions musicales ou les expressions mathématiques (Figure 4). Les mécanismes entrant en jeu relèvent de l'annotation, simple ou via les métadonnées, la gestion et la prise en compte des profils et contextes, via l'adaptation et les usages.

Plus largement, l'activité se focalise sur l'intégration de nouveaux services, par exemple la géolocalisation au travers d'interfaces haptiques, dans le champ d'application des systèmes pervasifs

« Les mécanismes relèvent de l'unification de la représentation des données semi-structurées via des modèles pivots. »

Figure 4



Éditeur de formules mathématiques Lambda 1.1 manipulées via des interfaces Braille et à synthèse vocale.

Projet LAMBDA : Linear Acces to Mathematics for Braille Devices and Audio-Synthesis
www.lambdaproject.org

DES PROJETS EN COURS ET COOPÉRATIONS

CONTRAPUNCTUS : Unification des formats de numérisation des partitions musicales en Braille via un format XML standardisé pour en faciliter l'accès. Contrat Européen IST 2005-034226 [2006-2009]. www.punctus.org

LINDO : Large scale distributed INDexation of multimédia Objects. Programme EUREKA-ITEA2 [2007-2010]. www.lindo.itea-eu

QUAERO : Outils intégrés de gestion des contenus multimédias. (Projet OSEO, www.oseo.fr - Établissement public, aide à l'innovation) [2007-2011]. www.quaero.org

OSIRIM : Offre de Service pour l'Indexation et la Recherche d'Information Multimédia. Contrat de Plan État Région [2007-2013]. www.irit.fr/OSIRIM Prolonge les travaux sur la plateforme RFIEC dédiée à la recherche d'information textuelle. www.irit.fr/RFIEC

DYNAMO : DYNAMic Ontology for information retrieval. Projet ANR [2008-2011].
L'objectif du projet est de permettre l'extraction automatisée d'ontologies à partir de corpus et l'indexation sémantique de contenus avec la mise à jour dynamique de l'ontologie à l'ajout des documents et des annotations des documents associés. Coopération avec les équipes I3C, SMAC de l'IRIT, ACTIA et ARTAL.

CAVALA : Méthode de suivi et d'évaluation des politiques régionales de développement économique. Contrat région [2007-2009].

CERISE : Concurrent EngineeRing des Interfaces de SystèmEs. Contrat Région [2007-2009]

IAPA : Infrastructure d'Accès, de Partage et d'Analyse de données Biomédicales. BQR UPS. Démarrage en Janvier 2007.

Structures :

FREMIT : Programme Pluri-Formations de Recherche en Mathématiques et Informatique de Toulouse entre l'Institut de Mathématiques (IMT) et l'IRIT www.irit.fr/-Programme-FREMIT-

ERT 34 et PReF : Plateforme Recherche Formation, hypermédias et apprentissage. Collaboration avec l'IUFM Midi-Pyrénées depuis 2003.

Jacques SAVOY

Jacques Savoy est professeur d'informatique à l'Université de Neuchâtel (Suisse).

Après avoir conçu et réalisé le premier système hypertexte francophone pour sa thèse de doctorat en 1987, il a été professeur adjoint d'informatique à l'Université de Montréal.

Lors de son séjour au Québec, il s'est intéressé aux problèmes liés au traitement de la langue naturelle et plus particulièrement à la recherche documentaire. Depuis 1993, il enseigne à l'Université de Neuchâtel. Ses activités de recherche concernent la recherche d'information sur le Web et dans les systèmes distribués.

Depuis une dizaine d'années, son équipe propose et évalue des systèmes de recherche de l'information pour diverses langues naturelles (comme le français, allemand, espagnol, suédois, hongrois, finnois, russe, bulgare, etc.) mais également touchant des langues asiatiques (chinois, japonais, coréen, bengali, marathi, persan, etc.).

Ses travaux abordent également les problématiques de la recherche bilingue (traduction automatique) ou multilingue. Son équipe participe régulièrement à diverses campagnes d'évaluation comme CLEF, NTCIR ou TREC (dans des domaines spécifiques comme la médecine, le blog, ou la recherche d'opinions).

Il est auteur d'une centaine d'articles du domaine et co-chairmann de SIGIR 2010 à Genève.

Noir sur Blanc: Comment avez-vous connu l'équipe SIG?

Jacques SAVOY: J'ai connu l'équipe à travers sa participation aux campagnes TREC dédiées à des comparaisons de performances de moteurs de recherche d'information.

NsB: Comment positionnez-vous l'équipe SIG par rapport aux grandes thématiques actuelles de recherche du domaine?

J. S.: J'identifie comme grandes thématiques:
- la prise en compte des usagers pour les ramener au cœur des systèmes qu'ils utilisent,
- la prise en compte du contexte des tâches de recherche incluant la nature de la tâche, le profil de l'utilisateur, les facteurs spatio-temporels…
- la capacité à pouvoir sélectionner ou configurer le bon système de recherche en fonction des caractéristiques de la tâche,
- traiter de manière transparente le «structuré» et le «semi-structuré»,
- la production d'outils de visualisation synthétique de collections qui intègrent de plus en plus l'expertise de ceux qui les utilisent.

Je considère que les préoccupations de l'équipe SIG rentrent parfaitement dans ces différentes thématiques avec une bonne complémentarité de ses composantes.

NsB: Quelle est votre perception de l'évolution des activités de l'équipe?

J. S.: SIG a acquis sa visibilité avec ses excellents résultats obtenus à TREC pour devenir très vite un des leaders du domaine. Cette montée en puissance au niveau national et international se traduit par les différentes participations aux campagnes TREC, CLEF, INEX…

C'est le premier groupe français ayant pris part à TREC avec de bons résultats obtenus à TREC6.
Par ailleurs, cette équipe est à l'origine de la mise en place de l'association ARIA (dédiée à la RI) et de la conférence francophone CORIA.
La notoriété de l'équipe se traduit aussi par son rôle éditorial dans le Journal International IRJ (Information Retrieval Journal), l'organisation de la grande conférence du domaine, ECIR 2009.
L'équipe SIG est sur la carte de la RI mondiale.
Une singularité de l'équipe SIG est sa composition exclusive d'enseignants-chercheurs ce qui explique des retombées au niveau formation (actions de formation et projets utilisant Mercure, Tétralogie… avec des partenaires industriels) avec la rédaction de nombreux ouvrages pédagogiques en Bases de Données, Recherche d'Information, Intelligence Economique…

NsB: Comment imaginiez-vous un «futur» pour l'équipe SIG?

J. S.: J'imagine ce futur à deux niveaux:
- au niveau formation, l'équipe dispose d'un «background» pour la mise en place de formations sur ses thématiques d'envergure européenne,
- au niveau recherche, elle doit toujours penser à mettre «l'humain» au centre de ses préoccupations,

et adosser ses résultats de recherche à la mise en œuvre d'outils au service de l'humain (la personne «usager» et son environnement).
Par exemple, la combinaison de l'outil de fouille Tétralogie avec les capacités de perceptions de l'œil humain ou des interfaces haptiques devrait être une piste pertinente pour accroitre la puissance exploratoire des collections.

À LIRE

J. Savoy S. Abdou, **Searching in Medline: Query expansion and manual indexing evaluation,** IPM: Information Processing Management, 44(2): 781-789, 2008

J. Savoy, **Comparative Study of Monolingual and Multilingual Search Models for Use with Asian Languages.** ACM-Transactions on Asian Language, Information Processing, 4(2), 2005

Journées Francophones d'Informatique Graphique 2008, Toulouse

L'équipe VORTEX de l'IRIT a organisé du 19 au 21 novembre, à la Manufacture des Tabacs à l'UT1, les 21es Journées Francophones d'Informatique Graphique 2008.

L'informatique graphique s'est structurée à partir des communautés regroupées autour de la programmation et de l'algorithmique mathématique. Les premières journées ont été des Journées AFCET-GROPLAN avec des publications dans BIGRE+GLOBULE en 1988 à Toulouse. Depuis cette date, toutes les équipes françaises travaillant dans le domaine ont hébergé ces journées sous différentes dénominations.

L'Informatique Graphique a beaucoup évoluée mais les thématiques scientifiques sont toujours centrées autour de la modélisation, du rendu de l'animation et de l'interaction, en passant du graphique interactif à la réalité virtuelle. Les liens avec le traitement d'image et la vision sont plus fréquents et prennent de multiples aspects dans la reconstruction, la visualisation et la fusion réel - virtuel avec la réalité augmentée.

Ces journées ont largement donné la parole aux jeunes chercheurs, mais aussi aux chercheurs plus confirmés ainsi qu'aux industriels. Elles ont été l'occasion de faire se rencontrer les diverses facettes de tous ces domaines, à travers les échanges entre les chercheurs qui les animent, et qui, en partageant leurs intérêts scientifiques, leur passion et leur enthousiasme ont fait de ces trois jours des moments riches, studieux, chaleureux, mais aussi festifs.

INFORSID'09 :

Informatique des Organisations et Systèmes d'Information et de Décision

L'édition d'INFORSID 2009 (Informatique des Organisations et Systèmes d'Information et de Décision), organisée par l'IRIT et l'Université de Toulouse, se tiendra du 26 au 29 mai 2009 à la Manufacture des Tabacs de Toulouse. Son objectif est de rassembler la communauté scientifique francophone en Bases de Données et en Systèmes d'Information pour faire un état de l'art des recherches actuelles et faire émerger de nouvelles problématiques.

Le congrès est organisé sur trois journées et demi : une journée avec des ateliers thématiques (workshops), deux journées et demi consacrées aux présentations scientifiques.

Cette année, conjointement à INFORSID 2009, un forum Jeunes Chercheurs est organisé.

Programme détaillé sur le site :
www.irit.fr/inforsid09

ECIR'09 :

European Conference on Information Retrieval

L'équipe SIG de l'IRIT organise au centre de congrès Pierre Baudis à Toulouse du 6 au 9 avril 2009 la 31e édition de la conférence ECIR'09.

ECIR est la principale conférence européenne dans le domaine de la recherche d'information (RI) et elle est l'une des conférences majeures au niveau international.

Elle offre aux chercheurs du monde entier un forum pour partager leurs travaux sur les fondements théoriques et pratiques dans le domaine de la RI qui connaît depuis quelques années une évolution sans précédent avec en particulier la croissance des applications liées au web. Le domaine est lui-même en pleine mutation et s'ouvre en permanence à de nouvelles thématiques.

Outre la conférence principale, ECIR propose également des manifestations scientifiques regroupant des ateliers thématiques, des tutoriaux ainsi que des posters.

ECIR connaît depuis quelques années un réel engouement, à en croire le nombre de participants, qui est passé en quelques années de 100 à 250. Elle est parrainée de manière régulière par plusieurs organismes dont ACM-SIGR (Special Interest Group of Information Retrieval), Google, Yahoo, Microsoft.

Programme détaillé sur le site :
http://ecir09.irit.fr

Manifestations passées

30 sept. - 2 oct. 2008
WMNC'2008
IFIP International Wireless and Mobile Networking Conference
ISAE-ENSICA, Toulouse
www.irit.fr/WMNC2008

19 - 21 nov. 2008
Journées AFIG 2008
21es Journées de l'Association Francophone d'Informatique Graphique
Manufacture des Tabacs, Toulouse
www.irit.fr/afig08

20 novembre 2008
La Fête de la Science à l'IRIT
sur le thème de l'Europe
Expositions et conférences

1er - 2 décembre 2008
Réunion annuelle de l'ILIKS
Interdisciplinary Laboratory on Interactive Knowledge Systems
IRIT

4 décembre 2008
Visite de l'IRIT par des étudiants du département informatique de l'ENS Cachan

8 - 9 décembre 2008
Journées FREMIT
Programme Pluri-Formations de Recherche en Mathématiques et Informatique de Toulouse
IRIT

10 - 12 déc.2008
ICLAN 2008
International Conference on the Latest Advances in Networks
ENSEEIHT,Toulouse
www.iclanconf.org/2008

26 - 28 janvier 2009
AFADL'2009
Approches Formelles dans l'Assistance au Développement de Logiciels
ENSEEIHT,Toulouse
www.irit.fr/afadl09

Manifestations à venir

12 - 13 mars 2009
GTMG '09
Groupe de Travail Modélisation Géométrique
IRIT
www.irit.fr/gtmg09

18 - 20 mars 2009
CORESA
COmpression et REprésentation des Signaux Audiovisuels
IAS, Toulouse
http://coresa2009.enseeiht.fr/

6 - 9 avril 2009
ECIR '09
31st European Conference on Information Retrieval
Centre de congrès Pierre Baudis
http://ecir09.irit.fr

26 - 29 mai 2009
INFORSID 2009
Informatique des Organisations et Systèmes d'Information et de Décision
Manufacture des Tabacs, Toulouse
www.irit.fr/inforsid09

9 - 11 juin 2009
CJCSC 2009
8e Colloque des Jeunes Chercheurs en Sciences Cognitives
Manufacture des Tabacs, Toulouse
www.irit.fr/cjcsc09

9 - 11 septembre 2009
RenPar'19
Rencontres francophones du Parallélisme
Sympa'13
Symposium en Architecture de machines
CFSE 7
Conférence Française sur les Systèmes d'Exploitation
Manufacture des Tabacs, Toulouse
www.irit.fr/Toulouse2009

Vous pouvez retrouver l'agenda complet sur www.irit.fr/-Agenda-

Campagne ANR 2008

La campagne de l'ANR pour l'année 2008 s'est conclue pour l'IRIT par l'obtention de cinq nouveaux projets : trois dans le domaine des sciences de l'information et de la communication et deux dans le domaine du vivant (programme Technologies de la Santé - Tecsan).

Pour le vivant, les deux projets sont les suivants :

- le projet Palliacom dont le responsable scientifique pour l'IRIT est Nadine Vigouroux. Ce projet a pour but la production d'un communicateur pour les personnes privées des facultés de communications ordinaires (parole, écrit).
- Le projet NAVIG dont le responsable scientifique pour l'IRIT est Christophe Jouffrais. Ce projet a pour but de proposer aux déficients visuels un dispositif leur permettant de se repérer de manière efficace dans l'espace tout en évitant les obstacles.

L'IRIT marque ainsi sa volonté de travailler dans l'informatique médicale (le dossier médical électronique notamment) pour répondre aux attentes sociétales accrues de nos concitoyens.

Zoom sur...

Le projet NAVIG

Le projet Navigation Assistée par Vision embarquée et GNSS (NAVIG) vise à concevoir un dispositif de suppléance pour les déficients visuels afin de leur permettre de se repérer plus facilement dans l'espace tout en évitant les obstacles. Ce projet est coordonné par l'IRIT, en la personne de Christophe Jouffrais, Chargé de recherche au CNRS. Les autres partenaires du projet sont : le Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), le Centre de Recherche Cerveau et Cognition (CERCO), les entreprises Spikenet (travaillant dans l'ingénierie électronique de la vision artificielle), Navocap (travaillant dans l'ingénérie de la mobilité) et l'Institut des Jeunes Aveugles. La valeur ajoutée de ce dispositif repose sur le fait que l'individu guidé aura une précision spatiale compatible avec la marche contrairement au GPS ou aux autres systèmes de navigation d'une précision moindre. Le couplage de la géolocalisation avec un logiciel de reconnaissance d'objets performant permettra d'obtenir cette reconnaissance rapide des objets, leur identification et la localisation auditive de ces derniers.

Les principaux verrous technologiques à lever se situent au niveau de la vision artificielle, du guidage auditif et de la fusion des données issues de la vision embarquée et de la géolocalisation. Le but est de fournir un dispositif permettant d'améliorer l'autonomie des utilisateurs en complément de la canne blanche et du chien guide. Ce projet financé par la Région Midi-Pyrénées et l'Agence Nationale de la Recherche, bénéficie également du soutien du Grand Toulouse et a été labellisé par le pôle de compétitivité Aerospace Valley.

Site du projet : http://navig.irit.fr

Contact : *secteur Valorisation*
valo@irit.fr - 05 61 55 76 81

L'Europe unit ses forces dans le domaine de l'indexation et de l'accès à l'information : LINDO, un projet ITEA 2

ITEA 2 est un programme EUREKA de R&D industrielle dans le domaine des technologies logicielles embarquées et distribuées.

Dans le contexte d'ITEA 2, de grands industriels et centres de recherche collaborent aujourd'hui dans le cadre du programme LINDO (Large scale distributed INDexation of multimedia Objects) qui réunit des partenaires français, belges et espagnols. Les principaux objectifs du consortium consistent à administrer des volumes toujours croissants de contenus multimédias et à rechercher la stricte information nécessaire répondant à la requête d'un utilisateur.

L'originalité de la recherche est le développement d'une architecture générique, dans laquelle non seulement le stockage des contenus multimédias est distribué, mais également l'indexation, répartie sur différentes unités de stockage, hétérogènes, éloignées géographiquement, de capacités diverses. Le but du projet n'est pas de développer encore un nouveau modèle ou moteur d'indexation mais plutôt d'encapsuler n'importe quel moteur ou extracteur existant et de l'intégrer dans la dite architecture générique.

La validation des résultats obtenus sera réalisée par le biais de démonstrateurs en temps réel dans des contextes d'applications tels que la vidéosurveillance et les archives du domaine médical.

La contribution de Florence Sèdes, Ana-Maria Manzat et Sébastien Laborie (équipe SIG) dans LINDO porte sur l'organisation et l'indexation des contenus multimédias.

LINDO a débuté le 1er novembre 2007 sous la coordination de Thales Security Systems et s'achèvera le 31 octobre 2010.

Site du projet :
www.lindo-itea.eu

Contact : *Affaires Européennes*
fourcade@irit.fr - 05 61 55 74 48