Rapport final du projet ANR Contint 2010 - CAAS (ANR-FORM-090601-01-01)

Josiane Mothe, Patrice Bellot, Eric San Juan, Ludovic Tanguy

Compte-rendu de fin de projet

Projet ANR-10-CORD-001-01

Acronyme et/ou nom du projet

Programme CONTINT 2010

A Identification

Acronyme du projet	CAAS
Titre du projet	Analyse Contextuelle et Recherche d'information **C**ontextual **A**nalysis and **A**daptive **S**earch
Coordinateur du projet (société/organisme)	Josiane Mothe Institut de Recherche en Informatique de Toulouse Université de Toulouse, Université P. Sabatier
Période du projet (date de début – date de fin)	15 déc. 2010 14 juin 2014
Site web du projet, le cas échéant	www.irit.de/caas

Rédacteur de ce rapport	
Civilité, prénom, nom	Mme Mothe Josiane
Téléphone	05 61 55 64 44
Adresse électronique	Josiane.mothe@irit.fr
Date de rédaction	30/06/2014

Si différent du rédacteur, indiquer un contact pour le projet	
Civilité, prénom, nom	
Téléphone	
Adresse électronique	

Liste des partenaires présents à la fin du projet (société/organisme et responsable scientifique)	IRIT LIA CLLE LSIS (un des membres du projet a rejoint le LSIS en sept 2014)

B Résumé consolidé public

B.1 Résumé consolidé public en français

CAAS : Adapter les moteurs de recherche au contexte

Recherche d'information contextuelle : adaptation et sélection

Au lieu de traiter toutes les requêtes soumises à un moteur de recherche selon le même processus, la recherche d'information contextuelle vise à considérer des éléments autres que les seuls termes de la requête pour sélectionner les informations à restituer. Il s'agit donc de mieux prendre en compte le besoin d'information spécifique des utilisateurs pour apporter des réponses plus pertinentes, en fonction des requêtes.

Ce projet s'est plus particulièrement intéressé à la prise en compte des types de requêtes, des types de documents interrogés et aux paramètres du moteur de recherche pour optimiser les réponses du système. Nous avons réalisé une étude fine de sessions de recherche en nous intéressant aux reformulations manuelles des utilisateurs pour extraire des termes pivots dans les formulations. Nous avons également montré qu'il était possible de déterminer quelles requêtes devaient être désambiguïsées avant d'être traitées par le moteur ; de la

même façon d'autres requêtes gagnent à subir des traitements additionnels d'expansion par exemple et cela en s'appuyant sur différentes ressources ; ressources qui peuvent être choisies de façon contextuelle.

Exploration et apprentissage pour la contextualisation des moteurs de recherche
Une analyse manuelle des formulations de requêtes a été conduite, aidée par la mise en base de données orientée graphe des formulations des requêtes. Par ailleurs, des méthodes d'apprentissage automatique (machine à vecteur de supports) et des méthodes d'exploration de données (arbres de décision, classification, analyses factorielles) ont été utilisées afin d'apprendre les meilleurs paramètres à utiliser en fonction des requêtes ou de leur type. Nous nous sommes intéressés à la fois aux types de documents et aux types de requêtes. Les caractéristiques des requêtes et celles des documents pour leur typage correspondent à des indicateurs linguistiques et statistiques, soit que nous avons définis, soit issus de la littérature. Ils ont été extraits automatiquement des collections que nous avons utilisées par des outils que nous avons développés à base de règles pour certains, existants pour d'autres (treetagger, wordnet). Nos recherches se sont appuyées sur des collections internationales de référence (TREC, CLEF, INEX tâche *Tweet contextualization* que nous avons créée) mais également sur des collections fournies par Nomao et le CLEO.

Ressources, prototypes et résultats majeurs :

- Une collection de résultats de configurations de moteurs de recherche (80 000 configurations) sur les collections internationales TREC7 & 8 adhoc et l'outil permettant de les générer à partir de la plate-forme Terrier.
- Détermination des paramètres les plus influents dans la recherche d'information
- Un outil permettant d'extraire les références bibliographique de documents (bilbo).
- Un cadre d'évaluation et la mise en œuvre d'une campagne d'évaluation internationale sur la contextualisation de textes courts (tweets) : 3 campagnes dans le cadre d'INEX.
- Un système de contextualisation de textes courts à partir d'une variété de ressources de référence (WikiPedia, New York Times intégral, large sélection du Web, dépêches d'actualité).
- Un couple d'outils crawler/indexeur XML sur une architecture YeSQL.
- Une base de parcours d'utilisateurs sur le site revues.org (découpage en session, actions utilisateurs, requêtes) et des ressources langagières dérivées (reformulations)
- Une base sémantique distributionnelle construite sur la collection indexée du site revues.org.

Production scientifique
Au cours du projet, nous avons produit X articles dans des revues internationales, Y dans des revues nationales. Les résultats ont également été publiés dans Z revues nationales et W conférences nationales. L'ensemble de ces éléments se trouve à l'adresse suivante : www.irit.fr/CAAS
Par ailleurs, nous avons élaboré, en lien avec d'autres collègues français, une tâche de contextualisation de tweets qui a été intégrée aux labs INEX et pour lesquels nous avons eu une trentaine de participants internationaux sur les 3 années de durée de cette tâche.
Nous avons également organisé un atelier sur la contextualisation des messages courts (CMC) dans le cadre de la conférence nationale EGC 2012 à Toulouse.

Illustration

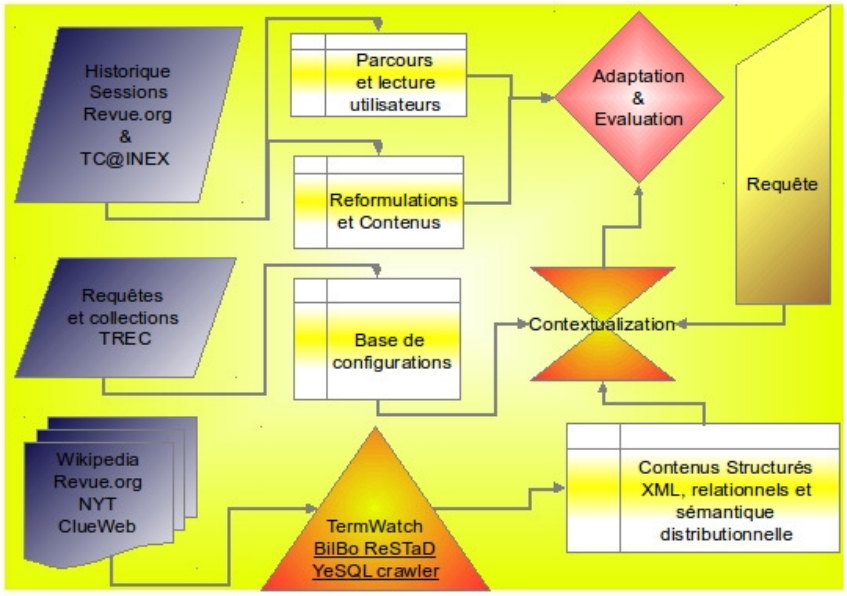

Illustration du projet CAAS : ressources et modules développés

En bleu les ressources réunies et analysées par ce projet. En dégradé jaune blanc, leur structuration par le projet, celles-ci sont disponibles pour la communauté à des fins de recherche. En orange apparaissent les modulées automatiques d'extraction (triangle) et de croisement (contextualisation) auxquels l'ensemble des participants ont contribué selon différentes approches. Les outils spécifiques au projet sont soulignés. En jaune gris la requête considérée comme entrée à nos systèmes. Si la contextualisation de la requête peut être faite automatiquement, l'adaptation au besoin spécifique et à la session de l'utilisateur ne peut être générique. Celle-ci dépend du scénario utilisateur, d'où sa représentation par un losange rose nécessitant un arbitrage en itération avec l'usager.

Informations factuelles

Le projet CAAS est un projet de recherche fondamentale (basic research) coordonné par Josiane Mothe, de l'institut de recherche en informatique de Toulouse – IRIT - UMR 5505 en partenariat avec le laboratoire Cognition, Langue, Langage, Ergonomie – CLLE – URM 5263 (Ludovic Tanguy), le Laboratoire Informatique d'Avignon – LIA - (Eric San-Juan) et le laboratoire des Sciences de l'Information et des Systèmes - LISI – UMR 6168 (Patrice Bellot). Le projet a débuté en décembre 2010 et s'est terminé le 15 juin 2014. Il a bénéficié d'une aide de xx € pour un coût global de xx €.

B.2 Résumé consolidé public en anglais

CAAS : Contextual Analysis and Adaptive Search

Contextual information retrieval : adaptation and selection

Instead of treating all the queries that are submitted to a search engine using the same process, the contextual information retrieval considers factors other than just the query terms to select the information to be retrieved. In this way, the specific user's need is taken into account in order to better answer, depending on the query.

This project focused on query types, types of queried documents as well as on the parameters of the search engine to optimize system answers. We studied in detail query sessions, considering query reformulations users made in order to extract pivot terms in the formulations. We also showed that it was possible to determine which queries should be disambiguated before being processed by the engine; in the same way other queries can be better treated applying additional treatments such expansion. For this, we rely on different resources which can be contextually selected.

Mining and learning for contextualization of search engines

A manual analysis of query formulations has been conducted, aided by the development of graph-oriented database from query logs. In addition, machine learning methods (support vector machine) and data mining methods (decision trees, classification, factor analysis) have been used to learn the best parameters to use on a query-based way. We worked both on documents types and query types. The characteristics of queries and those of documents we used to type them correspond to linguistic and statistical features. We define some of these features, others were taken from the literature. They were automatically extracted from the collections that we used by tools that we have developed based on some rules or using existing tools (TreeTagger, wordnet). Our research relied on international reference collections (TREC, CLEF, and INEX task Tweet contextualization that we created) but also on collections provided by "real" serach engines such as Nomao and CLEO.

Main resources, prototypes and results :

- A collection of runs resulting from various search engines configurations (80000 configurations) on international collections TREC7 & 8 adhoc ; and a tool to generate these runs from the platform Terrier,
- Determination of the most influential parameters in the information retrieval process,
- A tool to retrieve bibliographic references documents (bilbo).
- An evaluation framework as well as its implementation : an international evaluation campaign on the contextualization of short texts (tweets): 3 campaigns within INEX.
- A system to contextualize short texts from a variety of reference resources (WikiPedia, New York Times full, wide selection of web news).
- A couple of tools crawler / XML Indexer based on the YeSQL architecture.
- A database of users visit to the site revues.org (definition of sessions, users' actions, queries) and derived language resources (reformulations)
- Distributional semantic database built on the indexed collection from the revues.org site.

Scientific production

During the project, we produced X articles in international journals, Y in national journals. The results were also published in national journals Z and W national conferences. The detailed list can be found at the following address: www.irit.fr/CAAS

In addition, we have developed, in conjunction with other French colleagues, the "tweets contextualization" task that has been integrated into INEX labs and for which we had thirty international participants on three years of duration of the task.

We also organized a workshop on the contextualization of short messages (CMC) at the 2012 EGC conference in Toulouse.

Illustration

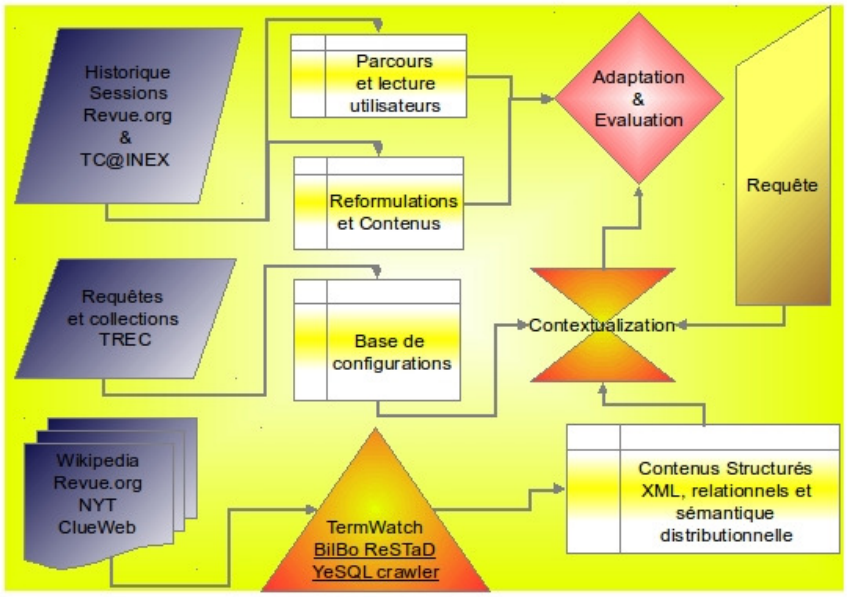

Illustration of CAAS project : developped resources and modules

Highlighted in blue are resources the project collected and analyzed. Yellow white marks-up the data structuration we made during the project, the data are available to the community for research purposes. Orange shows the automatic extraction modules (triangle) and crossover (contextualization) that all participants contributed to using different approaches. Tools specific to the project are outlined. Gray-yellow marks-up the query which is considered as an input to our resulting system. While the contextualization of the query can be automatic, adapting to the specific needs and to the user can not be generic. It depends on the user scenario, which is represented by a pink diamond and which requires arbitration.

Factual data

CAAS is a basic research project coordinated by Josiane Mothe from the Institut de Recherche en Informatique de Toulouse - IRIT - UMR 5505 in partnership with the Cognition, Langue, Langage, Ergonomie Lab - CLLE - UMR 5263 (Ludovic Tanguy), the Laboratoire Informatique d'Avignon - LIA - (Eric San Juan), and the laboratoire des Sciences de l'Information et des Systèmes - LISI - UMR 6168 (Patrice Bellot). The project began in December 2010 and ended June 15, 2014. It was funded by € for a total cost of approximatively xx €.

C Mémoire scientifique

Mémoire scientifique confidentiel : ~~oui /~~ non

C.1 Résumé du mémoire

CAAS : Adapter les moteurs de recherche au contexte

Recherche d'information contextuelle : adaptation et sélection

Au lieu de traiter toutes les requêtes soumises à un moteur de recherche selon le même processus, la recherche d'information contextuelle vise à considérer des éléments autres que les seuls termes de la requête pour sélectionner les informations à restituer. Il s'agit donc de mieux prendre en compte le besoin d'information spécifique des utilisateurs pour apporter des réponses plus pertinentes, en fonction des requêtes.
Ce projet s'est plus particulièrement intéressé à la prise en compte des types de requêtes, des types de documents interrogés et aux paramètres du moteur de recherche pour optimiser les réponses du système. Nous avons réalisé une étude fine de sessions de recherche en nous intéressant aux reformulations manuelles des utilisateurs pour extraire des termes pivots dans les formulations. Nous avons également montré qu'il était possible de déterminer quelles requêtes devaient être désambiguïsées avant d'être traitées par le moteur ; de la même façon d'autres requêtes gagnent à subir des traitements additionnels d'expansion par exemple et cela en s'appuyant sur différentes ressources ; ressources qui peuvent être choisies de façon contextuelle.

Exploration et apprentissage pour la contextualisation des moteurs de recherche
Une analyse manuelle des formulations de requêtes a été conduite, aidée par la mise en base de données orientée graphe des formulations des requêtes. Par ailleurs, des méthodes d'apprentissage automatique (machine à vecteur de supports) et des méthodes d'exploration de données (arbres de décision, classification, analyses factorielles) ont été utilisées afin d'apprendre les meilleurs paramètres à utiliser en fonction des requêtes ou de leur type. Nous nous sommes intéressés à la fois aux types de documents et aux types de requêtes. Les caractéristiques des requêtes et celles des documents pour leur typage correspondent à des indicateurs linguistiques et statistiques, soit que nous avons définis, soit issus de la littérature. Ils ont été extraits automatiquement des collections que nous avons utilisées par des outils que nous avons développés à base de règles pour certains, existants pour d'autres (treetagger, wordnet). Nos recherches se sont appuyées sur des collections internationales de référence (TREC, CLEF, INEX tâche *Tweet contextualization* que nous avons créée) mais également sur des collections fournies par Nomao et le CLEO.

C.2 Enjeux et problématique, état de l'art

Les systèmes de recherche d'informations (SRI) visent à restituer les informations qui répondent aux besoins d'un utilisateur exprimé sous la forme d'une requête. La recherche automatique d'informations pertinentes par rapport à une requête implique un processus en deux étapes: hors ligne, le système indexe des documents en utilisant généralement une représentation sac de de mots; en ligne, le système calcule la similitude entre la requête de l'utilisateur et les représentations de chaque document (ses termes d'indexation) pour décider des documents les plus ressemblants. Les SRI actuels, par exemple les moteurs de recherche sur le Web, sont des outils de recherche généraux mettant en oeuvre les mêmes mécanismes et les mêmes méthodes de traitement de données et de correspondance, quel que soit le contexte de la recherche : l'utilisateur, le type de besoins de l'information, ou l'utilisation de l'information.
L'hypothèse du projet CAAS est que la prise en compte de ce contexte pourrait pourtant améliorer les performances d'un SRI en expliquant certains éléments de la recherche d'information. L'aspect contextuel se réfère à la connaissance tacite ou explicite concernant les intentions des utilisateurs, l'environnement de ceux-ci et le système lui-même.

Les questions scientifiques fondamentales traitées dans le projet CAAS sont:

- **Le contrôle de la variété des contextes:** Pour traiter cette question, nous avons défini des modèles de représentation des aspects divers du contexte en RI. Une question importante concerne la variété des traitements et leur adéquation aux différents contextes connus.
- **Adaptation au contexte:** La modélisation du contexte n'est pas une fin en soi. Le système doit pouvoir s'appuyer sur cette modélisation pour décider lui-même des technologies ou des méthodes de RI les plus adéquates en fonction d'un contexte donné, c'est-à-dire qu'il doit adapter les méthodes de RI au contexte.
- **Reconnaître un contexte:** Quand un contexte est rencontré, le système doit le détecter parmi les contextes connus ou appris pour pouvoir décider quelle méthode il doit appliquer.

C.3 Approche scientifique et technique

Pour aborder ces défis, CAAS a considéré les divers aspects qui peuvent impacter le processus de RI d'abord de façon indépendantes, considérant ensuite les effets croisés. Nous nous sommes principalement concentrés sur les éléments contextuels suivant:

- les besoins des utilisateurs et les requêtes
- les documents
- les composants du système

Pour chacun d'entre eux, nous avons considéré des diverses collections que nous avons caractérisées. Nous les avons analysés en détail dans le but d'extraire des modèles et des comportements. Une fois que chaque élément contextuel a été étudié le plus indépendamment que possible des autres, nous avons considéré les effets croisés. Par exemple, nous avons montré que les caractéristiques des requêtes permettait de savoir si l'expansion de requêtes était utile ou non.

Dans CAAS, nous avons utilisé à la fois des collections de référence des programmes internationaux comme TREC (TREC7 & TREC8 adhoc, GOV2, WT10G) et des collections plus réalistes issues d'entreprises.

Partenariat

Le partenariat du projet est uniquement universitaire ; cependant, les entreprises sont également présentes notamment concernant la fourniture des logs de connexion (Nomao et Revue.org).

Plus précisément, au début du projet le consortium était composé de deux instituts en sciences informatiques, tous les deux spécialistes en RI, mais avec des compétences complémentaires. Le LIA (Laboratoire Informatique d'Avignon) travaille sur les systèmes Question/Réponse, tandis que l'IRIT (Institut de Recherche en Informatique de Toulouse) comprend des spécialistes en recherche Adhoc et détection de la nouveauté. L'IRIT travaille en relation avec l'IMT (Institut de Mathématique de Toulouse) et pour ce projet avec le groupe Statistique et Probabilité. Même si l'IMT n'apparaît pas comme un partenaire, ils travailleront dans ce projet. En effet l'IRIT et l'IMT sont des partenaires dans le Plan Pluri-formation FREMIT qui vise à développer le travail collaboratif. Le laboratoire CLLE (Connaissance, Langues, Langage, Ergonomie) est partenaire de ce projet pour leurs

compétences linguistiques. CLLE collabore depuis de nombreuses années avec l'IRIT (Mothe et Tanguy, 2005) pour la RI et le traitement automatique des langues. Un des membres du consortium ayant été promu au LSIS, ce laboratoire a rejoint le projet en 2012.

Exploration et apprentissage pour la contextualisation des moteurs de recherche

Ce projet s'est plus particulièrement intéressé à la prise en compte des types de requêtes, des types de documents interrogés et aux paramètres du moteur de recherche pour optimiser les réponses du système. Nous avons réalisé une étude fine de sessions de recherche en nous intéressant aux reformulations manuelles des utilisateurs pour extraire des termes pivots dans les formulations. Nous avons également montré qu'il était possible de déterminer quelles requêtes devaient être désambiguïsées avant d'être traitées par le moteur ; de la même façon d'autres requêtes gagnent à subir des traitements additionnel d'expansion par exemple et cela en s'appuyant sur différentes ressources ; ressources qui peuvent être choisies de façon contextuelle.

L'analyse et l'extraction de modèles sont au coeur du projet. Cependant, nous avons également développé des prototypes issus de nos recherches. Certains modules s'appuient sur les plateformes de recherche comme Terrier et Indri ainsi que sur le SGBDR PostGreSQL, de sorte qu'ils peuvent être réutilisés dans d'autres applications.

C.4 Résultats obtenus

Positionner les résultats par rapports aux livrables du projet et aux publications, brevets etc. Revisiter l'état de l'art et les enjeux à la fin du projet.

Ressources et prototypes

Durant le projet, nous avons créé de multiples ressources :

- Une collection de résultats de configurations de moteurs de recherche (80 000 configurations) sur les collections internationales TREC7 & 8 adhoc et l'outil permettant de les générer à partir de la plate-forme Terrier.
- Un outil opérationnel permettant d'extraire les références bibliographique de documents (bilbo) http://lab.hypotheses.org/955 .
- Une indexation XML combinée de larges ressources documentaires incluant le WikiPedia, 10 ans du NewYorkTimes, le ClueWeb (très large corpus extrait du Web) et le GigaWord (Deveaud et al. 2014).
- Un cadre d'évaluation et la mise en œuvre d'une campagne d'évaluation internationale sur la contextualisation de textes courts (tweets) : 3 campagnes dans le cadre d'INEX. https://inex.mmci.uni-saarland.de/tracks/qa/ (SanJuan et al, 2011) (Bellot et al., 2012) (Bellot et al. 2013)
- Une collection de log de connexion annotée en sessions et une collection de requêtes et de leur reformulation (1,5 millions de requêtes)

Ainsi que différents prototypes :

- Un prototype complet de contextualisation de requêtes, de passages courts ou de micro blogs d'après des ressources documentaires combinant méthode statistique (LDA) et TALN (graphes terminologiques). L'approche statistique permet

d'ordonnancer les ressources en fonction de la requête (Deveaud et al. 2013). L'approche TALN a servi de système de référence à QA@ INEX (SanJuan et al. 2011)
- Un outil opérationnel permettant d'extraire avec des méthodes de machine learning de nombreuses métadonnées documentaires par analyse des notes de bas de page et extraction de références bibliographique (bilbo)

Autres résultats

- Nous avons proposé un modèle qui permet de sélectionner les variables les plus importantes pour classer des objets décrits par de nombreuses variables, (Laporte et al., 2014)
- Nous avons déterminé les paramètres des requêtes et des systèmes les plus influents dans la recherche d'information et avons montré que ceux-ci pouvaient être dépendants du contexte (ici le type de requêtes) – thèse de Anthony Bigot soutenance prévue en 2014-2015, (Bigot et al., 2011).
- Nous avons montré que l'utilisation de modèles de langages spécifiques par ressources documentaire sont performants pour détecter autant que possible le contexte documentaire implicite à une requête (encyclopédique, journal, professionnel, Web, etc.) (Deveaud et al. 2012, 2013a). Nous avons ensuite montré que le LDA appliqué localement permettait de conceptualiser et préciser le contexte trouvé (Deveaud 2013b).
- Nous avons montré qu'il était possible de sélectionner les paramètres du moteur de recherche les plus adaptés aux requêtes en fonction de leurs paramètres et d'adapter ainsi les processus de RI aux contextes rencontrés – thèse de Adrian Chifu soutenance prévue fin 2014. (Chifu et Ionescu, 2012) ()
- Nous avons dégagé, au travers d'une étude approfondie d'un log de requêtes, un ensemble de comportements des utilisateurs (reformulations, parcours de documents) permettant d'identifier des besoins d'information spécifiques au contexte étudié (l'interrogation de collections de publications scientifiques)

C.5 Exploitation des résultats

Plusieurs exploitations sont réalisées, en cours ou prévues :

- Début mars 2014, OpenEdition a déployé sur la plateforme Revues.org son outil d'annotation automatique des références bibliographiques, Bilbo, dans le cadre du programme de recherche et développement « Robust and Language Independent Machine Learning Approaches for Automatic Annotation of Bibliographical References in DH Books, Articles and Blogs ». Initié en 2011 suite à l'obtention d'un Google Grant for Digital Humanities, ce programme a été mené par les équipes du LIA (université d'Avignon) puis du LSIS (Aix-Marseille université – CNRS) et du Cléo sous la direction de Patrice Bellot et Marin Dacos. Le programme bénéficie également du soutien de l'ANR CAAS et du projet Inter-Textes. Il porte sur l'ensemble des références bibliographiques présentes sur les quatre plateformes : Revues.org, Calenda, Hypothèses et OpenEdition Books. Il doit permettre de développer des fonctionnalités avancées de cross-linking (références croisées) entre les contenus d'OpenEdition et vers les contenus extérieurs.
- Depuis 2011, chaque année, nous proposons à la communauté internationale de tester leurs systèmes sur la tâche « contextualisation de tweets ». https://inex.mmci.uni-saarland.de/tracks/qa/. L'objectif de la contextualisation de textes courts (ici des

tweets, mais ce pourrait être des requêtes) est de fournir à son lecteur des informations qui rendent le message compréhensible. L'information est alors fournie sous la forme d'un résumé construit à partir de ressources existantes. Le système de contextualisation par modèle de langue et approche TALNE a servi de référence aux participants ; ce système est accessible par API.

- La collection de documents servant de source à la contextualisation est reconstruite chaque année à partir d'un dump officiel du Wikipédia antérieurs aux tweets. Nous avons développé un outil qui extrait le contenu textuel, la structure des documents ainsi que leurs annotations (liens et entités). L'ensemble est inséré dans une archive XML qui peut être indexée par Terrier, Indri ou chargée dans une base de données SQL avec ReSTaD. Cet outil fonctionne pour toute langue du wikipedia utilisant l'alphabet latin et a été repris par la communauté RI pour d'autres travaux.
  Concernant les sujets, en 2011, nous avions proposé des tweets à partir des titres d'articles du New York Times (NYT). En 2012, afin d'être plus réaliste, nous avons préféré construire une collection de tweets réels. Ainsi, nous avons collecté manuellement des tweets et choisi 63 tweets. Au-delà des 63 tweets collectés et contrôlés manuellement, nous avons constitué une collection de tweets plus importante basée sur une collecte automatique. Cet ensemble de tweets est composé à la fois des tweets contrôlés et de tweets "tout venant", en provenance toutefois des mêmes comptes prédéfinis. L'objectif était de s'assurer de la robustesse des systèmes, et d'éviter que les participants ne puissent réaliser des traitements manuels compte tenu de leur nombre. Par ailleurs, le fait d'avoir plusieurs types de tweets peut également être valorisé dans des expérimentations variées. Le même principe a été utilisé les années suivantes.
  Nous avons également proposé une mesure d'évaluation, non sensible à la taille des résumés de référence utilisés pour l'évaluation, contrairement aux mesures de la littérature.
- Nous avons développé une surcouche à Terrier permettant de lancer en parallèle plusieurs exécutions sur un ensemble de requêtes avec des paramétrages de moteur différents. Cette surcouche devra être consolidée et pourra ensuite être proposée à la communauté des utilisateurs de Terrier.
- Le système de contextualisation à partir de références documentaires sur une approche statistique est actuellement repris pour des tâches d'analyse d'e-reputation dans le cadre de la campagne RepLab par le LIA (Cossu et al) et la SCS de l'Université de Glasgow (équipe Terrier) de manière indépendante.
- ReSTaD est actuellement utilisé par le LIA pour expérimenter des modèles de RI fondés sur les logiques multi-valuées et possibilistes.

C.6 Discussion

Les objectifs initiaux du projet ont été atteints : nous avons pu analyser les effets des différents aspects (requêtes, documents, systèmes) sur la recherche et proposé des modèles qui peuvent s'adapter au contexte, soit par apprentissage de la meilleure stratégie pour des requêtes répétées, soit sur la base de caractéristiques pour sélectionner une approche de recherche adaptée. Nous avons par ailleurs développé plusieurs prototypes dont un est d'ores et déjà intégré dans un système opérationnel d'une organisation (Revue.org). Nous avons conçu et mis à disposition de la communauté internationale une collection de test qui a

été utilisée 3 ans de suite par différents groupes internationaux (INEX Tweet Contextualization).
Les résultats des travaux que nous avons menés pourront être transférés chez des partenaires industriels qui ont la problématique de fournir des outils soit à leurs clients, soit à leurs employés des réponses à des besoins d'information variés et qui nécessite des approches différentes pour répondre au besoin.
Par ailleurs, plusieurs perspectives de recherche aux travaux menés dans le projet peuevtn être envisagées :

- Caractéristiques des requêtes : les caractéristiques des requêtes que nous avons choisies pour leur typage a fait appel à des considérations statistiques et linguistique pré et post recherche. Nous pensons que d'autres caractéristiques devraient être étudiées afin d'améliorer la pertinence des résultats et le typage des requêtes. Nous pensons en particulier à l'étude de l'usage des termes, soit au sein des requêtes, soit au sein des documents,
- Caractéristiques des documents : nous pensons que plus de recherche est nécessaire pour mieux appréhender les types de documents au sens de leurs usages possibles et le lien avec leur pertinence. Par exemple, lorsqu'une requête s'intéresse à un sujet, l'utilisateur pourra être intéressé par la définition de ce sujet ou au contraire plus intéressé par un ou des documents qui discutent le sujet. Distinguer ces deux types de besoin pourrait permettre de mieux choisir les documents à restituer à l'utilisateur,
- Le contexte a été appréhendé dans ce projet sur ses facettes : requêtes, documents, paramètres du système. D'autres éléments du contexte interviennent et n'étaient pas objet d'étude dans ce projet mais mériteraient une attention particulière. L'utilisateur et son environnement sont deux éléments qui mériteraient d'être étudiés en détail. L'environnement correspond par exemple au lieu géographique dans lequel se trouve l'utilisateur, le type de matériel qu'il utilise pour interroger le système, …

C.7 Conclusions

Le projet CAAS est un projet de recherche fondamentale coordonné par Josiane Mothe, de l'institut de recherche en informatique de Toulouse – IRIT - UMR 5505 en partenariat avec le laboratoire Cognition, Langue, Langage, Ergonomie – CLLE – URM 5263 (Ludovic Tanguy), le Laboratoire Informatique d'Avignon – LIA - (Eric San-Juan) et le laboratoire des Sciences de l'Information et des Systèmes - LISI – UMR 6168 (Patrice Bellot). Le projet a débuté en décembre 2010 et s'est terminé le 15 juin 2014.
Les résultats de ce projet sont nombreux, tant au niveau scientifique avec de nombreuses publications nationales et internationale, qu'au niveau des ressources et prototypes qui ont été créées lors du projet. Certains résultats sont déjà intégrés dans des systèmes opérationnels et d'autres le seront prochainement. Par ailleurs, d'autres résultats scientifiques restent à publier.

C.8 Références

La liste complète des références se trouve à : http://www.irit.fr/CAAS/fr/publications.html

Eric Sanjuan, Véronique Moriceau, Xavier Tannier, Patrice Bellot, Josiane Mothe. Overview of the INEX 2011 Question Answering Track (QA@INEX) (regular paper). International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2011), S. Geva, J.

Kamps, and R. Schenkel (Eds.): INEX 2011, LNCS 7424, pp. 188--206. Springer, Heidelberg, 2012.

P. Bellot T. Chappell A. Doucet S. Geva S. Gurajada J. Kamps, G. Kazai M. Koolen M. Landoni M. Marx A. Mishra V. Moriceau, J. Mothe M. Preminger G. Ram´ırez M. Sanderson E. Sanjuan F. Scholer, A. Schuh X. Tannier M. Theobald M. Trappett A. Trotman Q. Wang, Report on INEX 2012. SIGIR Forum, ACM, Vol. 46 N. 3, p. 50-59, 2012.

Patrice Bellot, Antoine Doucet, Shlomo Geva, Sairam Gurajada, Jaap Kamps, Gabriella Kazai, Marijn Koolen, Arunav Mishra, Véronique Moriceau, Josiane Mothe, Michael Preminger, Eric SanJuan, Ralf Schenkel, Xavier Tannier, Martin Theobald, Matthew Trappett, Qiuyue Wang. Overview of INEX 2013, Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2013), Valence-Espagne, 23/09/2013-26/09/2013, Kent State UNiversity, p. 269-281, 2013.

Anthony Bigot, Claude Chrisment, Taoufiq Dkaki, Gilles Hubert, Josiane Mothe. Fusing Different Information Retrieval Systems According to Query-Topics. Dans / In : Information Retrieval Journal, Kluwer, Vol. 14 N. 6, p. 617-648, avril / april 2011.

Adrian Chifu, Radu Tudor Ionescu. Word Sense Disambiguation to Improve Precision for Ambiguous Queries. Dans / In : Central European Journal of Computer Science, Versita, co-éditeur Springer Verlag, Londres - GB, Vol. 2 N. 4, p. 398-411, 2012.

D Liste des livrables

Quand le projet en comporte, reproduire ici le tableau des livrables fourni au début du projet. Mentionner l'ensemble des livrables, y compris les éventuels livrables abandonnés, et ceux non prévus dans la liste initiale.

Date de livraison	**N°**	**Titre**	**Nature** (rapport, logiciel, prototype, données, …)	**Partenaires** (souligner le responsable)	**Commentaires**
Livrables prévus initialement (les livrables abandonnés sont indiqués dans la colonne commentaire					
	D1.1	Wiki			abandonné
T0+12	D.1.2 a	Rapport intermédiaire interne	Rapport	CLLE, IRIT, LIA	
T0+24	D1.2. b	Rapport intermédiaire interne	Rapport	CLLE, IRIT, LIA	
T0+6	D1.3. a	Rapport ANR intermédiaire	Rapport	CLLE, IRIT, LIA	
T0+18	D1.3 b	Rapport ANR intermédiaire	Rapport	CLLE, IRIT, LIA	
T0+42	D1.3. c	Rapport ANR final	Rapport	CLLE, IRIT, LIA	
T0+10	D2.1.	Etat de l'art	Publication	CLLE, IRIT, LIA	
T0+6	D2.2. & D2.3	Description des plateformes, collections et systèmes à utiliser,	Rapport	CLLE, IRIT, LIA	Un seul livrable combinant les deux prévus

Date de livraison	N°	Titre	Nature (rapport, logiciel, prototype, données, …)	Partenaires (souligner le responsable)	Commentaires
		formats d'échange			
/	D2.4	Consortium agreement			Abandonné.
T0+42	D2.5	Description des ressources utilisées et créées	Rapport	CLLE, IRIT, LIA	Remplace le livrable initialement prévu en T0+8 sur les données utilisées
T0+24	D3.1.	Analyse des requêtes	Rapport	CLLE, IRIT	
T0+18	D3.2.	Analyse de la difficulté des requêtes	Rapport et publication	IRIT	
T0+24	D3.3	Classifieur de requêtes	Logiciel (programmes R)	CLLE, IRIT	
T0+24	D3.4	Détecteur de difficulté de requêtes	Logiciel (programmes R)	CLLE, IRIT	
T0+12	D3.5	Workshop dans le cadre de EGC		CLLE, IRIT, LIA	
	D4.1	Analyse de la variété des documents	Indexation XML ReSTaD (surcouche PostGreSQL) et modèle de langage (surcouche Indri)	LIA	
	D4.2	Classifieur de documents	Logiciel de classification de passages (surcouche Indri) et publications	LIA	
T0+	D5.1	Analyse de la variété de système	Rapports et publications	IRIT	
	D5.2	Classifieur de systèmes	Logiciel (programmes R)	IRIT	
	D6.1	Analyse des effets croisés	Publications	CLLE, IRIT, LIA	
	D7.1	Outil implantant la recherche sélective en fonction des requêtes	Logiciel (surcouche de Terrier)	IRIT	
Mise à jour	D8.1	Site web			
Livrables non prévus initialement					
T0+6 T0+16 T0+28 T0+40		Collections « Tweet contextualization »	Ressource	IRIT, LIA, LSIS	

Date de livraison	N°	Titre	Nature (rapport, logiciel, prototype, données, …)	Partenaires (souligner le responsable)	Commentaires
T0+		Extracteur de référence – Moteur Bilbo	Logiciel	LIA	
T0+40		Session de requêtes et reformulations	Ressource	CLLE, IRIT	
T0+15		80 000 exécutions Terrier	Ressource	IRIT	

E Impact du projet

E.1 Indicateurs d'impact

Nombre de publications et de communications (à détailler en Liste des publications et communications)

		Publications multipartenaires	Publications monopartenaires
International	**Revues à comité de lecture**	3	3
	Ouvrages ou chapitres d'ouvrage		
	Communications (conférence)	4	14
France	**Revues à comité de lecture**	1	
	Ouvrages ou chapitres d'ouvrage	3	
	Communications (conférence)	12	13
Actions de diffusion	**Articles vulgarisation**		
	Conférences vulgarisation	1	
	Autres	2	

Autres valorisations scientifiques (à détailler en Liste des éléments de valorisation)

	Nombre, années et commentaires (valorisations avérées ou probables)
Brevets internationaux obtenus	
Brevet internationaux en cours d'obtention	
Brevets nationaux obtenus	
Brevet nationaux en cours d'obtention	
Licences d'exploitation (obtention / cession)	
Créations d'entreprises ou essaimage	
Nouveaux projets collaboratifs	
Colloques scientifiques	Atelier EGC 2012 Contextualisation de messages courts Labs Tweet contextualisation, INEX – CLEF 2012 ; 2013 ; 2014
Autres (préciser)	

E.2 Liste des publications et communications

- International journals

Iria da Cunha, Eric SanJuan, Juan Manuel Torres Moreno, Marina Lloberes, Irene Castellón: DiSeg 1.0: The first system for Spanish discourse segmentation. Expert Systems with Applications, Elsevier, 39(2): 1671-1678, 2012.

P. Bellot T. Chappell A. Doucet S. Geva S. Gurajada J. Kamps, G. Kazai M. Koolen M. Landoni M. Marx A. Mishra V. Moriceau, J. Mothe M. Preminger G. Ram´ırez M. Sanderson E. Sanjuan F. Scholer, A. Schuh X. Tannier M. Theobald M. Trappett A. Trotman Q. Wang, Report on INEX 2012. SIGIR Forum, ACM, Vol. 46 N. 3, p. 50-59, 2012.

Adrian Chifu, Radu Tudor Ionescu. *Word Sense Disambiguation to Improve Precision for Ambiguous Queries.* Dans / *In* : *Central European Journal of Computer Science*, Versita, co-éditeur Springer Verlag, Londres - GB, Vol. 2 N. 4, p. 398-411, 2012. ftp://ftp.irit.fr/IRIT/SIG/2012_CEJCS_CI.pdf

Patrice Bellot, Josiane Mothe, . Report on INEX 2011. SIGIR Forum, ACM, Vol. 46 N. 1, p. 33-42, 2012. http://www.sigir.org/forum/2012J/2012j_sigirforum_B_bellotInex2011Report.pdf

Eric Sanjuan, Véronique Moriceau, Xavier Tannier, Patrice Bellot, Josiane Mothe. Overview of the INEX 2011 Question Answering Track (QA@INEX) (regular paper). International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2011), S. Geva, J. Kamps, and R. Schenkel (Eds.): INEX 2011, LNCS 7424, pp. 188--206. Springer, Heidelberg, 2012. ftp://ftp.irit.fr/IRIT/SIG/inexqa2011.pdf

Anthony Bigot, Claude Chrisment, Taoufiq Dkaki, Gilles Hubert, Josiane Mothe. *Fusing Different Information Retrieval Systems According to Query-Topics.* Dans / *In* : *Information Retrieval Journal*, Springer, Vol. 14 N. 6, p. 617-648, avril / *april* 2011. ftp://ftp.irit.fr/IRIT/SIG/2011_IRJ_BCDHM.pdf

- International conferences

Patrice Bellot, Josiane Mothe, et al.. Overview of INEX 2013, Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2013), Valence-Espagne, 23/09/2013-26/09/2013, Kent State UNiversity, p. 269-281, 2013. ftp://ftp.irit.fr/IRIT/SIG/2013_CLEF_BDGGKKKMMMPSSTTTW.pdf

Romain Deveaud, Eric San-Juan, Patrice Bellot, Are Semantically Coherent Topic Models Useful for Ad Hoc Information Retrieval?, ACL, The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia (Bulgarie), 2013.

Romain Deveaud, Eric San-Juan, Patrice Bellot, Estimating Topical Context by Diverging from External Resources, ACM Press, The 36th Annual ACM SIGIR Conference SIGIR'13, Dublin (Ireland), 2013.

Romain Deveaud, Eric San-Juan, Patrice Bellot, Unsupervised Latent Concept Modeling to Identify Query Facets" , ACM Press, OAIR (Open research Areas in Information Retrieval) 10th ACM International Conference in the RIAO series, Lisboa (Portugal), 2013.

Liana Ermakova, Josiane Mothe. *IRIT at INEX 2013: Tweet Contextualization Track (regular paper).* Dans / *In* : *INitiative for the Evaluation of XML Retrieval (INEX 2013), Valencia, Spain, 23/09/2013-26/09/2013*, Polytechnic University of Valencia, (en ligne), 2013.
http://www.clef-initiative.eu/documents/71612/58a64b0a-cf0c-4751-a91f-9c8aba4312e1

Sébastien Déjean, Nicolas Faessel, Lucille Marty, Josiane Mothe, Samantha Sadala, Soda Thiam. *Analysis of Patents for Prior Art Candidate Search. International Conference on Advances in Information Mining and Management, Lisbon, Portugal, 17/11/2013-22/11/2013*, (Eds.), Xpert Publishing Service (XPS), (en ligne), 2013

Eric Sanjuan, Véronique Moriceau, Xavier Tannier, Patrice Bellot, Josiane Mothe. Overview of the INEX 2012 Tweet Contextualization Track. Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2012), Rome, Italie, 17/09/2012-20/09/2012, Springer, (en ligne), 2012.
http://www.clef-initiative.eu/documents/71612/0daaae31-ae99-4fb4-8982-b5a3a2ebbb6b

Liana Ermakova. *IRIT at INEX 2012: Tweet Contextualization (regular paper).* Dans / *In* : *Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2012), Rome, Italy, 17/09/2012-20/09/2012*, Univesity La Sapienza, 2012.
http://www.clef-initiative.eu/documents/71612/3e9ecc64-fae6-4af3-93fd-1a6a6fabb5d6

Josiane Mothe, Liana Ermakova. *IRIT at INEX: Question Answering Task (regular paper). INitiative for the Evaluation of XML Retrieval (INEX 2011), Sarrebruck, 12/12/2011-14/12/2011*, Vol. 7424, Shlomo Geva, Jaap Kamps, Ralf Schenkel (Eds.), Springer, Lecture Notes in Computer Science, p. 219-226, 2012.
ftp://ftp.irit.fr/IRIT/SIG/2012_INEX_EM.pdf

Young-Min Kim, Patrice Bellot, Élodie Faath and Marin Dacos. Machine Learning for Automatic Annotation of References in DH scholarly papers. 16th international conference on Digital Humanities, DH 2012, Hamburg, Allemagne, 2012.

Young-Min Kim, Patrice Bellot, Elodie Faath, Marin Dacos : Machine Learning for Automatic Annotation of References in DH. Digital Humanities 2012, , 2012.

Young-Min Kim, Patrice Bellot, Elodie Faath, Marin Dacos : Annotated Bibliographical Reference Corpora in Digital Humanities. Proceedings of the Eight

International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012.

Young-Min Kim, Patrice Bellot, Jade Tavernier, Élodie Faath and Marin Dacos. Evaluation of BILBO Reference Parsing in Digital Humanities via a Comparison of Different Tools. 12th ACM Symposium on Document Engineering (DocEng 2012), pp 209 – 212.

Jade Tavernier, Patrice Bellot: Flesch and Dale-Chall Readability Measures for INEX 2011 Question-Answering Track. INitiative for the Evaluation of XML Retrieval. S. Geva, J. Kamps, and R. Schenkel (Eds.): INEX 2011, LNCS 7424, pp. 235--246. Springer, Heidelberg, 2012.

Eric SanJuan: Mapping Knowledge Domains - Combining Symbolic Relations with Graph Theory. KDIR 2011: 527-536

Jonathan Compaoré, Sébastien Déjean, Adji Maïram Gueye, Josiane Mothe, Joelson Randriamparany. Mining Information Retrieval Results: Significant IR parameters (regular paper), Advances in Information Mining and Management, Barcelone, IARIA, (support électronique), 2011. ftp://ftp.irit.fr/IRIT/SIG/2011_IMMM_CGDMR.pdf

Iria da Cunha, M. Teresa Cabré, Eric SanJuan, Gerardo Sierra, Juan Manuel Torres Moreno, Jorge Vivaldi: Automatic Specialized vs. Non-specialized Sentence Differentiation. CICLing, Lecture Notes in Computer Science 6609, Springer 2011 (2) 2011: 266-276.

Young-Min Kim, Patrice Bellot, Élodie Faath, and Marin Dacos. 2011. Automatic annotation of bibliographical reference in digital humanities books, articles and blogs. In Proceedings of the CIKM 2011 BooksOnline11 Workshop, pages 41–4.

- National journal and conferences

Julie Ayter, Cecile Desclaux, Adrian Chifu, Sébastien Déjean, Josiane Mothe. *Performance Analysis of Information Retrieval Systems (regular paper).* Dans / *In* : *Spanish Conference on Information Retrieval*, *Coruna*, *19/06/2014-20/06/2014*, (Eds.), Springer-Verlag, (support électronique), 2014.

Anthony Bigot, Sébastien Déjean, Josiane Mothe. *Learning to choose: conférence nationale avec actes (regular paper).* Dans / *In* : *Spanish Conference on Information Retrieval*, *Coruna*, *19/06/2014-20/06/2014*, (Eds.), Springer-Verlag, (support électronique), 2014. (best student paper)

Adrian Chifu, Josiane Mothe. *Expansion sélective de requêtes par apprentissage. Conférence francophone en Recherche d'Information et Applications (CORIA 2014)*, *Nancy*, *France*, *19/03/2014-21/03/2014*, (Eds.), LORIA, p. 231-246, 2014.

Patrice Bellot, Véronique Moriceau, Josiane Mothe, Eric Sanjuan, Xavier Tannier. Évaluation de la contextualisation de tweets (short paper). Conférence francophone en Recherche d'Information et Applications (CORIA 2013), Neuchâtel, Suisse, 03/04/2013-05/04/2013, Association Francophone de Recherche d'Information et Applications (ARIA), 2013.
ftp://ftp.irit.fr/IRIT/SIG/2013_CORIA_BMMST.pdf

Patrice Bellot, Josiane Mothe, Eric Sanjuan, Ludovic Tanguy. Actes de conférence: Atelier Contextualisation de Messages Courts (EGC 2013), Toulouse, France, Cépaduès, Revue des Nouvelles Technologies de l'Information, 2013.
ftp://ftp.irit.fr/IRIT/SIG/2013_EGC_BMST.pdf

Liana Ermakova, Nicolas Faessel. *Création de snippets : une application de la génération automatique de résumés (regular paper).* Dans : *Atelier Contextualisation de Messages Courts (EGC 2013), Toulouse, France, 29/01/2013*, Patrice Bellot, Josiane Mothe, Éric SanJuan, Ludovic Tanguy (Eds.), Cépaduès, Revue des Nouvelles Technologies de l'Information, p. 27-36, janvier 2013.
URL : ftp://ftp.irit.fr/IRIT/SIG/2013_EGC_EF.pdf

Adrian Chifu. *Prédire la difficulté des requêtes : la combinaison de mesures statistiques et sémantiques (short paper).* Dans / *In* : *Conférence francophone en Recherche d'Information et Applications (CORIA 2013), Neuchatel, Suisse, 03/04/2013-05/04/2013*, Université de Neuchétel, p. 191-200, 2013.
URL : ftp://ftp.irit.fr/IRIT/SIG/2013_CORIA_C.pdf

Anthony Bigot. *Adapter les moteurs de recherche aux besoins en information - Prise en compte de la difficulté du besoin (regular paper). INFormatique des Organisations et Systemes d'Information et de Decision (INFORSID 2013), Paris, 29/05/2013-31/05/2013*, Université Paris 1, p. 59-74, 2013.
ftp://ftp.irit.fr/IRIT/SIG/Bigot_Anthony_INFORSID_2013.pdf

Simon Leva, Nicolas Faessel. Détection automatique des sessions de recherche par similarité des résultats provenant d'une collection de documents externe (regular paper). Dans / In : Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, Sables D'Olonne, France, 17/06/2013-21/06/2013, 2013.

Simon Leva. Les sessions de recherche comme contexte des requêtes (regular paper). Dans / In : Atelier Contextualisation de Messages Courts (EGC 2013), Toulouse, France, 29/01/2013, Patrice Bellot, Josiane Mothe, Éric SanJuan, Ludovic Tanguy (Eds.), Cépaduès, Revue des Nouvelles Technologies de l'Information, p. 1-12, janvier / january 2013.
URL : ftp://ftp.irit.fr/IRIT/SIG/2013_EGC_L.pdf

Young-Min Kim, Patrice Bellot, Elodie Faath, Marin Dacos : Automatic annotation of incomplete and scattered bibliographical references in Digital Humanities papers, CORIA 2012, pp. 329–340, Bordeaux, 21-23 mars 2012

Sébastien Déjean, Josiane Mothe, Julia Poirier, Benoît Sansas, Joelson Randriamparany. *Etude des résultats des systèmes de RI à grande échelle (regular paper).* Dans / *In* : *EGC - Atelier Extraction et Gestion Parallèles Distribuées des Connaissances, Brest, France, 25/01/2011-28/01/2011*, Dominique GAY, Hamid HACID (Eds.), Revue des Nouvelles Technologies de l'Information (RNTI), (support électronique), 2011. ftp://ftp.irit.fr/IRIT/SIG/2011_EGC_DMPSR.pdf

Fanny Lalleman , Cécile Fabre, et Johanne Heinecke,. Détecter le potentiel d'ambiguïté d'une requête – le cas des recherches portant sur l'actualité, Actes du congrès mondial de linguistique française, 2012.

Sébastien Déjean, Josiane Mothe, Julia Poirier, Benoît Sansas, Joelson Randriamparany. Etude des résultats des systèmes de RI à grande échelle (regular paper). EGC - Atelier Extraction et Gestion Parallèles Distribuées des Connaissances, Brest, Revue des Nouvelles Technologies de l'Information (RNTI), (support électronique), 2011

Romain Deveaud, Florian Boudin, Eric SanJuan, Patrice Bellot: Correction de césures et enrichissement de requêtes pour la recherche de livres. CORIA 2011: 89-96.

C. Adam, C. Fabre et L. Tanguy : *Etude des relations sémantiques dans les reformulations de requêtes sous la loupe de l'analyse distributionnelle*. Actes de l'atelier SemDis, Sables D'Olonne, 2013.

- Book chapters

Patrice Bellot. Recherche d'information contextuelle, assistée et personnalisée, ouvrage collectif, Hermès Lavoisier, Paris, ISBN 978-2-7462-2583-1, 2011.

Patrice Bellot. Difficultés de lecture, dyslexies et recherche d'information. In P. Bellot Eds. (2011), chapitre 7, p. 191 à 226, 2011.

Josiane Mothe. Recherche d'information contextuelle : le cas des requêtes. In P. Bellot Eds. (2011), chapitre 1, p. 27 à 56, 2011

- Autres actions de valorisation

Ludovic Tanguy, Josiane Mothe, Patrice Bellot, Eric San Juan. CAAS: Contextual Analysis and Adaptive Search. 2013. poster à "Les rencontres du numériques" URL : http://www.irit.fr/publis/SIG/CAAS_2013_LM.pdf

- PhD thesis

Romain Deveaud, Vers une représentation du contexte thématique en Recherche d'Information, Thèse de l'université d'Avignon et des pays du Vaucluse, soutenue en novembre 2013.

Thèses à soutenir fin 2014:

Adrian Chifu, Université de Toulouse
Anthony Bigot, Université de Toulouse

E.3 Liste des éléments de valorisation

- Une collection de résultats de configurations de moteurs de recherche (80 000 configurations) sur les collections internationales TREC7 & 8 adhoc et l'outil permettant de les générer à partir de la plate-forme Terrier.
- Un outil opérationnel permettant d'extraire les références bibliographique de documents (bilbo) http://lab.hypotheses.org/955 .
- Une indexation XML combinée de larges ressources documentaires incluant le WikiPedia, 10 ans du NewYorkTimes, le ClueWeb (très large corpus extrait du Web) et le GigaWord (Deveaud et al. 2014).
- Un cadre d'évaluation et la mise en œuvre d'une campagne d'évaluation internationale sur la contextualisation de textes courts (tweets) : 3 campagnes dans le cadre d'INEX, intégrée depuis 2012 à la conférence CLEF. https://inex.mmci.uni-saarland.de/tracks/qa/ (SanJuan et al, 2011) (Bellot et al., 2012) (Bellot et al. 2013)
- Une collection de log de connexion annotée en sessions et une collection de requêtes et de leur reformulation (1,5 millions de requêtes)
- Un prototype complet de contextualisation de requêtes, de passages courts ou de micro blogs d'après des ressources documentaires combinant méthode statistique (LDA) et TALN (graphes terminologiques). L'approche statistique permet d'ordonnancer les ressources en fonction de la requête (Deveaud et al. 2013). L'approche TALN a servi de système de référence à QA@ INEX (SanJuan et al. 2011)
- Un outil opérationnel permettant d'extraire avec des méthodes de machine learning de nombreuses métadonnées documentaires par analyse des notes de bas de page et extraction de références bibliographique (bilbo).
- L'organisation d'un atelier sur la contextualisation des messages courts (CMC) dans le cadre de la conférence EGC (29 janvier 2012, 6 soumissions, 15 participants.

E.4 Bilan et suivi des personnels recrutés en CDD (hors stagiaires)

Ce tableau dresse le bilan du projet en termes de recrutement de personnels non permanents sur CDD ou assimilé. Renseigner une ligne par personne embauchée sur le projet quand l'embauche a été financée partiellement ou en totalité par l'aide de l'ANR et quand la contribution au projet a été d'une durée au moins égale à 3 mois, tous contrats confondus, l'aide de l'ANR pouvant ne représenter qu'une partie de la rémunération de la personne sur la durée de sa participation au projet.
Les stagiaires bénéficiant d'une convention de stage avec un établissement d'enseignement ne doivent pas être mentionnés.

Les données recueillies pourront faire l'objet d'une demande de mise à jour par l'ANR jusqu'à 5 ans après la fin du projet.

Identification			Avant le recrutement sur le projet	Recrutement sur le projet	Après le projet										
Nom et prénom	Sexe H/F	Adresse email (1)	Date des dernières nouvelles	Dernier diplôme obtenu au moment du recrutement	Lieu d'études (France, UE, hors UE)	Expérience prof. Antérieure, y compris post-docs (ans)	Partenaire ayant embauché la personne	Poste dans le projet (2)	Durée missions (mois) (3)	Date de fin de mission sur le projet	Devenir professionnel (4)	Type d'employeur (5)	Type d'emploi (6)	Lien au projet ANR (7)	Valorisation expérience (8)
Anthony Bigot	M	Anthony.bigot@irit.fr	CDD en cours	Master 2	France	aucune	IRIT	Doctorant	3 ans	14/06/2014					
Youngmin Kim	F	youngminn.kim@gmail.com		Doctorat	France et hors UE	Aucune	LIA	Post-Doctorant							
Faessel Nicolas	M	Nicolas.faessel@irit.fr		Doctorat	France	1 ans (ATER)	CLLE	Ingénieur	1 an	31/10/2013	*recherche d'emploi*				
Tavernier Jade	F	jade.tavernier@etd.univ-avignon.fr		Master 1	France	Aucune	LIA	Ingenieur en alternanace Master Informatique 2							
Yoann Moreau	M	moreau.yo@gmail.com		Master 2	France	Ingénieur recherche	LIA	Ingénieur							
Chifu Adrian	M	Adrian.chifu@irit.fr	CDD en cours	Master 2	Rouma	Doctorant	IRIT	Doctorant (non financé	3 ans	14/06/2014					

					nie			sur CAAS, allocation Université)							
Romain Deveaud	M	romain.deveaud@etd.univ-avignon.fr		Master 2	France	Ingénieur (master en alternance)	LIA	Doctorant (non financé sur CAAS, allocation MESR)							
Simon Leva	M	simon.leva@univ-tlse2.f	thèse en cours	Master 2	France	Aucune	CLLE	Doctorant (non financé sur CAAS)	3 ans	15/06/14					
Clémentine Adam	F	adam@univ-tlse2.fr	Décembre 2013	Doctorat	France	Aucune	CLLE	Post-Doctorat	1 an	30/11/13					
Fanny Lalleman	F	lalleman@univ-tlse2.fr	Juin 2014	Doctorat	France	Aucune	CLLE	Chercheur	4 mois	15/06/14					

Aide pour le remplissage

(1) Adresse email : indiquer une adresse email la plus pérenne possible
(2) Poste dans le projet : post-doc, doctorant, ingénieur ou niveau ingénieur, technicien, vacataire, autre (préciser)
(3) Durée missions : indiquer en mois la durée totale des missions (y compris celles non financées par l'ANR) effectuées sur le projet
(4) Devenir professionnel : CDI, CDD, chef d'entreprise, encore sur le projet, post-doc France, post-doc étranger, étudiant, recherche d'emploi, sans nouvelles
(5) Type d'employeur : enseignement et recherche publique, EPIC de recherche, grande entreprise, PME/TPE, création d'entreprise, autre public, autre privé, libéral, autre (préciser)
(6) Type d'emploi : ingénieur, chercheur, enseignant-chercheur, cadre, technicien, autre (préciser)
(7) Lien au projet ANR : préciser si l'employeur est ou non un partenaire du projet
(8) Valorisation expérience : préciser si le poste occupé valorise l'expérience acquise pendant le projet.

Les informations personnelles recueillies feront l'objet d'un traitement de données informatisées pour les seuls besoins de l'étude anonymisée sur le devenir professionnel des personnes recrutées sur les projets ANR. Elles ne feront l'objet d'aucune cession et seront conservées par l'ANR pendant une durée maximale de 5 ans après la fin du projet concerné. Conformément à la loi n° 78-17 du 6 janvier 1978 modifiée, relative à l'Informatique, aux Fichiers et aux Libertés, les personnes concernées disposent d'un droit d'accès, de rectification et de suppression des données personnelles les concernant. Les personnes concernées seront informées directement de ce droit lorsque leurs coordonnées sont renseignées. Elles peuvent exercer ce droit en s'adressant l'ANR (http://www.agence-nationale-recherche.fr/Contact).