Projet d'action spécifique : RIP-WEB

"Recherche d'Informations Précises sur le WEB"

Demandeurs :

Brigitte Grau, Groupe LIR, LIMSI

Jean-Pierre Chevallet, groupe MRIM, CLIPS-IMAG

I Positionnement

La production de documents augmente, principalement dans le domaine du document numérique. Plus que jamais il devient indispensable de trouver une aide efficace dans l'assistance à la recherche d'information, particulièrement sur le large corpus hétérogène qu'est le WEB. Les prochains systèmes de recherche d'information se doivent donc de répondre à des besoins précis en matière d'information afin de satisfaire au mieux les utilisateurs.

Lorsque l'intérêt de l'utilisateur porte sur des thèmes particuliers ou sur le suivi de l'évolution d'un domaine, il est pertinent d'utiliser des systèmes de filtrage et de résumé. Lorsque ce même utilisateur recherche une donnée factuelle, l'information pertinente ne peut être apportée que par des systèmes dédiés à ce type de tâche. En effet, face à une question telle que « Quelle est la voiture la plus chère du monde ? », les moteurs de recherche traditionnels renvoient tous les documents jugés pertinents par rapport à la question, en sélectionnant éventuellement un passage, et c'est à l'utilisateur que revient la tâche d'explorer ces documents afin de trouver la réponse à sa question : répondre à des questions précises requiert une analyse plus en profondeur des documents sélectionnés afin d'en extraire l'information pertinente.

La problématique des systèmes de question/réponse se situe ainsi à l'intersection de plusieurs domaines, dont notamment la recherche d'information et le traitement de la langue naturelle. La recherche de documents pertinents est donc enrichie par l'intégration de modules de TAL s'appliquant à large échelle, i.e. quel que soit le domaine abordé, et possédant une grande couverture linguistique. Cette intégration permet la sélection de passages pertinents reposant sur des critères numériques et aussi sur l'exploitation de traits linguistiques, qu'ils soient de nature syntaxique ou sémantique. L'apprentissage automatique trouve alors aussi naturellement sa place dans cette problématique, plus précisément l'apprentissage de critères de sélection et de classification d'extraits.

Nous nous intéressons spécifiquement au Web pour ses caractéristiques propres. En effet, par sa diversité, le Web pose des problèmes différents de ceux relatifs aux collections plus homogènes. Outre l'accessibilité, la diversité et l'hétérogénéité du Web, aussi bien dans la forme (la structure) que dans le contenu des documents, amènent à aborder le problème de la nature même des documents, qui peuvent être un site, un ensemble de pages ou un regroupement de parties de pages, et qui influent sur le type de question posée et de réponse attendue. Ces caractéristiques amènent aussi à traiter le problème de la validité des réponses trouvées ainsi que celui du multilinguisme.

Finalement, le contexte dans lequel s'effectue la recherche, et plus précisément le but poursuivit pour l'exploitation des informations trouvées, ainsi que les connaissances de l'utilisateur sur le domaine interrogé, et également sur le Système de Recherche d'Informations (SRI) qu'il met en œuvre, influencent sa satisfaction à la vue de la réponse du système. La tâche de recherche d'information est essentiellement humaine, et le SRI ne peut être que le médiateur informatisé qui va guider l'utilisateur de son besoin à l'information exprimée dans les documents. Le jugement de pertinence de cet utilisateur ne peut alors pas se détacher du contexte d'utilisation d'un tel outil.

II L'évaluation des SRI

L'évaluation des SRI a toujours constitué une activité importante dans la mise au point de ces systèmes. Malheureusement, la plupart des évaluations ne mesurent que le comportement du moteur de recherche, en dehors de tout utilisateur : donc en dehors de tout contexte. Des tentatives dans le cadre de la conférence TREC ont tout de même été envisagé et mise en place. Il ne faut pas pour autant négliger l'importance d'une campagne d'évaluation. En effet, c'est quasiment le seul moyen d'avoir une évaluation comparative des modèles, des technologies et des méthodes mise en oeuvre pour solutionner le problème de la RI en général. De plus, une campagne d’évaluation offre la possibilité de faire collaborer des chercheurs de domaines différents, et de confronter les différentes approches sur un même problème. Alors qu’à l’heure actuelle, les moteurs de recherche documentaire ont tendance à stagner, des avancées en question-réponses profiteront largement à cette activité.

La campagne proposée dans la réponse à l’appel d’offre de Technolangue, en offrant un cadre d'évaluation à des systèmes complets de question-réponse, offre aussi une oportunité d'évaluer l'apport de modules de différentes natures (moteur de recherche, modules de TAL, etc.) par leur impact sur la tâche complète. Il y a actuellement 11 participants à cette campagne, 6 laboratoires de recherche (dont 2 laboratoires étrangers francophones) et 5 entreprises. D’autres laboratoires se sont montrés intéressés par une telle tâche, mais ne sont pas prêts à participer de suite à une campagne d’évaluation. Il semble donc opportun d'élargir la communauté touchée par cette campagne dans l'objectif de définir des axes prioritaires de travail et d'étude sur le thème de la recherche d'informations précises sur le WEB.

Dans cette action spécifique, nous n'envisageons donc pas de proposer une campagne d'évaluation concurrente. Au contraire, notre objectif est de se faire rencontrer les acteurs de la recherche d'information, des traitements des langues et de l'apprentissage, en s'appuyant éventuellement sur les campagnes d'évaluation existantes (Technolangue, TREC).

III Proposition

L'objectif de cette action spécifique est d'étudier les grands axes importants qui se croisent dans les domaines de la recherche d'information, du traitement de la langue et de l'apprentissage dans le cadre des corpus textuels du WEB. Dans ce croisement se trouve un nouveau type de système, plus apte à extraire de l'information précise d'une masse gigantesque et informe d'informations de toute nature, toute langue et de toute origine. Nous parlerons alors de "Recherche d'Informations Précises sur le WEB".

Le thème de cette AS reprend à son compte la thématique des systèmes de questions-réponses et celle des systèmes de recherche d'informations dans des documents structurés orientés vers la précision des réponses. En effet, en fonction du besoin de l'utilisateur, qui peut être identifié en fonction du contexte de l'interrogation, il est opportun de déterminer ce qui constitue la réponse attendue de manière à fournir la plus petite parcelle de documents qui répond à la question de l'utilisateur.

Les communautés du domaine du traitement de la langue et celle du domaine de la recherche d'information ont travaillé en parallèle sur un problème finalement commun : la recherche dans des textes de la meilleure réponse à une requête issue du besoin d'information d'un utilisateur. Cette action spécifique sera alors une première étape pour fusionner les synergies positives existantes, par l'appropriation d'un vocabulaire commun, le groupement de compétences et la comparaison de méthodologies différentes pour aborder le problème.

1 Objectifs

Cette proposition d'action spécifique a donc pour objectif l'étude des verrous principaux pour aboutir à un système d'interrogation réaliste sur le WEB, orienté vers la précision des réponses. Pour cela nous proposons l'organisation de 7 journées thématiques qui auront pour but de tracer les limites du problème de la résolution de questions précises dans un environnement aussi vaste qu'est le WEB.

Nous proposons une progression en 7 axes qui correspondront à au moins 7 thèmes de réunions. Pour chaque axe, nous produirons un résumé des problèmes analysés.

> **1) Requête, question, document, réponse** : l'objet de cet axe est de préciser les notions fondamentales à la jonction du domaine de la recherche d'information et des systèmes de questions-réponses. Il s'agit donc de définir dans le contexte particulier d'une recherche précise d'information, ce que nous entendons par *requête* de l'utilisateur, par *document* source d'information et par la *réponse* que le système doit fournir. Cet axe doit également étudier la notion de *contexte* de recherche. De manière concrète, l'étude de ce contexte peut nous amener à définir une typologie des requêtes. par exemple, recherche de l'information technique pour résoudre un problème précis en situation de travail, est très différent d'une recherche d'un lieu de vacances. La typologie des requêtes doit nous conduire à une typologie des documents ou de sites WEB, dans lequel se trouve l'information recherchée. Dans l'exemple précédent, connaître le type de document (manuel technique, mails de réponse à une question, site commercial), permet de mieux cibler la réponse en fonction du contexte de la recherche.
>
> **2) Interaction RI et TAL** : cet axe a pour objectif l'étude de l'articulation de techniques classiques de recherche d'information avec les techniques Traitement de la Langue pour l'extraction d'informations précises lors de la résolution d'une requête de type question/réponse. Cet axe sera l'occasion d'étudier les aspects spécifiques en traitement automatique de la langue pouvant être évalués via un système de question-réponse. Une telle réflexion permettra d’élargir les recherches actuelles et favorisera la coopération d’équipes autour d’une tâche commune, chacune pouvant conserver sa spécialité. Nous étudierons en particulier les aspects relatifs à l'analyse d'une question, l'usage de techniques RI pour la focalisation et le filtrage des documents, l'usage de TAL pour l'extraction d'informations.

3) Place de l'apprentissage : l'apprentissage a déjà été mis en œuvre avec succès dans les tâches recherche d'informations plus classiques ou dans les tâches de filtrage et classification de documents. La question qui est posée concerne la place et le rôle que doit jouer l'apprentissage dans le cadre d'un système de recherche d'informations précises dans un univers aussi vaste que le WEB.

4) Spécificité des informations sur le WEB : hormis la spécificité de travailler avec une source d'information extrêmement vaste, il est nécessaire d'étudier d'autres traits spécifiques du WEB qui doivent être pris en compte de manière explicite dans la tâche d'extraction d'informations précises. Nous pouvons noter en exemple, le fait que l'information qui s'y trouve disponible, n'est pas pour autant valide. De même, nous pouvons trouver des informations imprécises voire contradictoires. Ce problème peut être abordé sous l'aspect du typage des pages et des liens du WEB, et donc aussi sous l'aspect de la structure des documents présents sur le WEB.

5) Extraction multilingue d'informations : le WEB est de fait une source multilingue d'informations hétérogènes. Les outils de traitement de la langue à mettre en œuvre dans ce cadre doivent donc tenir compte du multilinguisme. Il s'agira ici de mettre en évidence les problèmes que le multilinguisme pose dans un système de questions-réponses. Cet axe doit alors recenser les différentes voies possibles à court et long terme, qui mènent à un WEB multilingue ou bien un WEB interlingue par langage pivot.

6) Mise en forme et présentation de l'information : que ce soit pour la sélection de parties de documents en réponses ou pour la formulation d'une réponse à une question précise, la réponse du système doit être mise en forme et doit être présentée à l'utilisateur. Cette mise en forme doit tenir compte du contexte dans lequel la réponse a été trouvée : le contexte (ou type) des documents, le contexte du site WEB ou se trouve la réponse, et le contexte dans lequel la question a été posée. La mise en forme peut impliquer une fusion de plusieurs sources d'informations pour la construction automatique d'un résumé qui constitue la réponse finale.

7) Architecture distribuée d'un Système de Recherche d'Informations Précises : ce dernier axe a pour objectif l'étude préliminaire de l'architecture d'un SRIP pour l'expérimentation en vraie grandeur d'un système d'interrogation du WEB. Cette étude est le point de départ pour la définition d'une équipe projet chargée de la mise en place et en service d'une plate-forme distribuée expérimentale dédiée à la recherche d'informations précises sur le WEB.

2 Résultats attendus

Le travail sera réalisé sur une année. Les résultats des différents axes de travail seront consignés dans un document collectif final qui sera présenté à la fin de l'action lors d'une journée de présentation ouverte.

Un autre résultat attendu de cette action concerne l'élargissement de la communauté actuelle à d’autres équipes en TAL et en recherche d’information, mais aussi à des équipes en apprentissage pour lesquelles la réalisation d’un système complet représente un travail supplémentaire trop important, en dehors de leurs spécialités qui plus est, mais qui pourraient collaborer avec les premières citées.

Il s'agit aussi d'approfondir la réflexion sur les systèmes de questions réponses et leur évaluation et rédiger un « roadmap » à 5 ans. L'évaluation Technolangue, si elle a lieu, est prévue pour fin 2003. Il serait opportun de ne pas en rester là, la richesse de cette tâche permettant de renouveler l'opération, en définissant des évaluations supplémentaires incitant à étudier des problèmes plus complexes tels que l'inférence, la génération de réponse à partir de plusieurs documents, la mise en œuvre d'un dialogue, etc.

La fin de cette action spécifique peut se conclure sur la définition d'une équipe projet chargée de la mise en place effective des briques de bases de traitement de la langue et de traitement statistique de recherche d'information dans le but de mutualiser des savoirs et savoir-faire des différentes équipes participantes pour constituer une plate-forme d'expérimentation et de démonstration d'un système de recherche d'informations précises sur le WEB. Par ce regroupement, le travail des équipes sera facilité, et les non-spécialistes de l'un des domaines auront ainsi accès aux outils nécessaires à la réalisation d'un système complet et à l'évaluation de leurs travaux via la tâche de question-réponse. Ce travail important de mutualisation devra se synchroniser avec les projets similaires dans les autres RTP.

3 Positionnement de l'AS par rapport aux projets existants :

- *Technolangue* : le projet EqueR est une campagne d'évaluation de systèmes de question-réponse sur le français, dont l'organisateur est la société Elda. Les participants sont subventionnés pour participer au comité scientifique dont le rôle est de préciser les modalités de la campagne d'évaluation, la proposition se fondant sur une évaluation du type de celle effectuée dans le cadre de Trec : un ensemble de questions factuelles, indépendantes, dont il faut fournir la réponse. Il s'agira de préciser le type de question qui seront à traiter (questions tout domaine et sur des domaines restreints (médecine, commerce, etc.), l'organisation des types d'évaluation (une tâche générale plus des tâches spécifiques), les modalités pour fournir les réponses (1 ou plusieurs, seulement la réponse ou une phrase, etc.), et enfin les mesures appliquées (en repartant des mesures définies par le NIST dans le cadre de Trec). Les participants reçoivent aussi une subvention pour adapter leur système aux modalités de l'évaluation.

Ce financement est clairement séparé de l'AS, car il est destiné à des actions qui sont spécifiques à ce projet.

- *Société de l'information* : la proposition qui a été acceptée consiste à étudier des types d'évaluation différents afin d'aller au-delà des évaluations actuelles qui sont fondées sur un jugement binaire hors du contexte du besoin de l'utilisateur. En effet, en l'absence d'informations claires sur l'origine du besoin d'information, sur le contexte d'utilisation de l'information qui sera trouvée par le système, les juges qui choisissent les bonnes réponses de la collection de test, sont seuls pour déterminer le niveau de réponse accepté. Notre objectif est d'une part de tenir compte de la complexité et de la diversité des questions : nous ne nous limitons pas aux questions factuelles hors de tout contexte d'interrogation. Notre objectif est d'autre part de tenir compte de la complexité des systèmes de question-réponse. En effet, cette problématique fait intervenir de nombreux modules de traitement de textes en langue naturelle. Certains sont fondés sur des approches du domaine du Traitement Automatique des Langues (TAL) : ils sont alors capables de structurer en profondeur des énoncés textuels, mais sont limités dans les documents que l'on peut traiter tant sur le plan de la quantité, que sur la diversité des domaines. En effet, certains traitements nécessitent des ressources linguistiques propre à un domaine d'application.

> D'autres modules sont fondés sur une approche de type Recherche d'Information où les traitement linguistiques sont plus légers (ou inexistant) ce qui permet un traitement plus rapide et plus large thématiquement. En contre partie, ces modules n'ont qu'une faible précision dans la localisation de l'information. Finalement les modules capable d'apprentissage nécessitent une phase d'entraînement et viennent compléter les deux autres modules pour former un système complet de recherche d'informations capable de répondre avec précision à des questions précises. Le but de cette proposition est alors de réfléchir à une manière d'évaluer les différentes approches combinaison de ces trois types de modules pour mesurer leurs apports mutuels. C'est le système global ainsi constitué qui forme un Système de Question-Réponse qui sera évalué.

L'action spécifique présentée ici est un bon complément aux questions posées dans cette proposition. Le point 1 de la proposition RIP-WEB (Requête, question, document, réponse) est complémentaire car nous mettons en place un vocabulaire commun pour décrire le problème qui nous intéresse : la recherche d'informations précises sur le WEB. Dans les points suivants (2 et 3) nous nous posons les questions concernant la place et le rôle des différents types de modules (TAL,RI et apprentissage) dans une chaîne de traitement de documents. C'est précisément sur ces deux points que se réalise l'intersection entre l'AS RIP-WEB et la proposition de la Société de l'information. Cette proposition étudie en plus les problèmes liés à l'évaluation, et à la prise en compte du contexte ou les requêtes sont posés.

De nouvelles dimensions étudiées dans RIP-WEB viennent compléter cette initiative : elles concernent l'influence du WEB comme source d'information sur le système question/réponse à construire. Cette influence porte sur la quantité d'informations disponible, mais surtout sur le problème de la diversité, et celui de la fiabilité des réponses trouvées ainsi que des contradictions possibles. Cela fait l'objet du point 4 (Spécificité des informations sur le WEB)

La notion du multilinguisme est également nouvelle : c'est le thème de la partie 5 (extraction multilingue d'informations). Cet aspect est très important dans le contexte du WEB.

Finalement, la forme de présentation des résultats, ainsi que leur possible fusion, est également un axe original de cet action et sera étudiée dans le point 6 (Mise en forme et présentation de l'information). La notion de l'architecture (point 7) sera abordé mais, il dépasse un peu le cadre d'une telle action et mérite un développement à part pour l'étude d'une plate-forme logiciel, comme par exemple une équipe projet.

4 Participants :

La liste suivante mentionne les organismes et personnes ayant déclaré leur intérêt pour cette proposition.

Laboratoires associés CNRS :

Groupe LIR, LIMSI Orsay, Brigitte Grau, Patrick Paroubek, Gabriel Illouz, Anne Vilnat
Groupe MRIM, CLIPS-IMAG Grenoble, Jean-Pierre Chevallet, Christophe Brouard.
Groupe GETA, CLIPS-IMAG Grenoble, Gilles Serasset.
Groupe Connex, LIP6 Paris VI, Patrick Gallinari, Massih-Reza Amini.
Groupe SIG, IRIT Toulouse III, Josiane Mothe, Mohand Boughanem.
Laboratoire LaLICC, Paris-Sorbonne, Jean-Luc Minel
Laboratoire LATTICE, ENS Paris, Bernard Victorri.

Laboratoire LIPN, Villetaneuse, Adeline Nazarenko
Laboratoire ERSS, Toulouse, Didier Bourigault
Equipe I&A, LRI, Orsay, Antoine Cornuejols, Michele Sebag, Yves Kodratoff
Laboratoire LIA, Avignon, Patrice Bellot

Laboratoires, école ou sociétés :

Equipe Grappa, univ Lille3, Rémi Gilleron
Laboratoire INaLCO/CRIM, Thierry Poibeau
Laboratoire LIC2M, CEA Fontenay aux Roses, Christian Fluhr, Olivier Ferret, Gael de Chalendar.
Laboratoire Gresec - Equipe Cristal, Grenoble , Laurence BALICCO
Laboratoire STIM, AP-HP Paris, Pierre Zweigenbaum.
Laboratoire RALI, Montréal, Guy Lapalme
Laboratoire PSI, Rouen, Thierry Paquet
France Telécom recherche, Lannion, Olivier Collin
ENSM Saint-Etienne, Michel Beigbeder
Société Sinequa, Ivry suerSeine, Claude de Loupy
Société iSmart, Aix en Provence, Michel Benoit
Société Synapse Developpement, Toulouse, Patrick Seguela

5 Soutient demandé

Pour l'AS RIP-WEB, le budget demandé est de 30K euros. Il y a 27 personnes qui participent au projet, dans 8 laboratoires affiliés au CNRS, 4 autres laboratoires, une école et 3 sociétés. Le budget total prévu pour l'action conjointe RIP-WEB et SI est donc de 15 + 30 = 45Keuros en globalisant les 2 projets. Nous proposons globalement la répartition suivante :

Missions par laboratoires	*2000*
Nombre de laboratoires	*13*
Total mission RIP-WEB	**26000**
Missions et projets spécifique à SI	**5000**
Vacation site WEB	**2000**
Achat Ressources (CORPUS TREC WT10g * 3)	**3000**
Colloque Ouvert	**3000**
Vacation pour constitution de ressources	**6000**
TOTAL	**45000**

Le coût des déplacements correspond à une estimation basée sur une moyenne de 180€ par missions pour les 7 réunions de prévues dans les objectifs. Compte tenu du nombre variable de personnes présentes par laboratoire, nous avons estimé à 2000 € l'enveloppe mission par laboratoire. Le total des missions est donc de 26000 € pour tout le projet.

Le projet financé par SI sera utilisé pour des missions complémentaires et des projets spécifiques qui seront définis et acceptés par l'ensemble de la communauté formée. Nous envisageons la constitution de ressources pour cette action. Nous estimons 5000€ le coût propre à ces missions.

Un site WEB sera constitué et maintenu pour cette action. Ce site a pour objectif la mise à disposition des informations et des rapports de réunions produit, ainsi que la mise en accès des ressources produite par et pour le groupe. Ce coût est estimé à 2000€.

La collection qui est actuellement la référence pour l'évaluation de recherche d'informations sur le WEB est la collection WT10g qu'il faut acquérir pour 1000$. Nous estimons que 3 groupes au plus voudront travailler avec cette collection, pour un coût de 3000€.

Nous envisageons d'organiser une journée de présentation ouverte à d'autre participant que le groupe formé à la fin de cette action spécifique. (Estimation 3000€).

Finalement, pour constituer des ressources linguistiques particulière à ce projet, nous avons décidé de conserver 6000€ sur le budget global pour des vacations liées à cette tâche. Rappelons qu'une concrétisation de cette action spécifique doit être l'étude préliminaire de l'architecture d'un SRIP pour l'expérimentation en vraie grandeur d'un système d'interrogation du WEB.

6 Répartition en laboratoires

Dans le groupe, 4 laboratoires et une école n'appartiennent pas au CNRS. Les laboratoires CLIPS et LIMSI seront chargés de gérer les missions pour ces laboratoires. Nous proposons alors la répartition suivante sur les 8 laboratoires du CNRS qui participent à cette action. Le LIMSI apparaît deux fois car il gère la partie du financement coté SHS pour le projet SI.

Répartition laboratoires		
LIMSI	SI, misions, vacation, ressources	15000
LIP6	mission	2000
IRIT	mission	2000
ERSS	mission	2000
LRI	mission	2000
LIA	mission	2000
LIMSI	mission, vacation	8000
CLIPS	mission, vacation, site WEB, ressources	12000
TOTAL		45000