

CONTENUS NUMÉRIQUES

CAAS

Contextual Analysis and Adaptive Search

Contint 2010



CLLE – Cognition, Langue, Langage Ergonomie, Université de Toulouse (L. Tanguy)
Institut de Recherche en Informatique de Toulouse, Université de Toulouse (J. Mothe)
Laboratoire d'informatique Avignon, Université d'Avignon et des pays de Vaucluse (E. San Juan)
LSIS / Polytech Marseille - Aix-Marseille Université (P. Bellot)

OBJECTIFS DU PROJET

Les systèmes de recherche d'information (SRI) ont pour objet de retrouver l'information qui répond au besoin d'information exprimé par l'utilisateur via une requête. Les SRI actuels comme les moteurs du web se comportent de la même façon quelque soit le contexte de recherche, l'utilisateur, le type du besoin d'information et d'usage de l'information.

L'hypothèse du projet CAAS est que la prise en compte du contexte pourrait améliorer les performances du système. Le contexte fait référence ici à des connaissances implicites ou explicites concernant les intentions de l'utilisateur, son environnement et le système lui-même.

Les verrous scientifiques que le projet vise à lever sont :

- . La définition du contexte en RI
- . L'apprentissage des contextes
- . La reconnaissance d'un contexte lorsqu'il se présente

MÉTHODOLOGIE ET RÉSULTATS

CAAS considère les divers aspects qui peuvent influencer les résultats d'une recherche d'information, d'abord de façon la plus indépendante possible, puis en considérant les effets croisés. Nous étudions ainsi les éléments contextuels suivants : les attentes des utilisateurs via leurs requêtes et actions, les documents et les composants des SRI.

Nous considérons à la fois des collections internationales largement utilisées par la communauté RI et des collections issues d'entreprises. Notre objectif est aussi de développer des modules à partir des résultats qui pourront être intégrés dans des plateformes de RI pour permettre leur réutilisation comme composants d'un système de RI.

Analyse des requêtes et log de connexion.

Certaines requêtes sont plus faciles que d'autres à traiter par les SRI. La difficulté peut être liée à de nombreux paramètres liés à la requête et/ou aux documents. Nous avons proposé une combinaison efficace de prédicteurs et travaillons sur de nouveaux prédicteurs à base d'indicateurs linguistiques.

Une requête est souvent un des éléments d'une session plus longue correspondant à la recherche pour satisfaire un besoin d'information. Nous avons proposé une méthode pour détecter les sessions de recherche. Nous avons également mis en place deux collections de référence à partir de données issues de deux moteurs en ligne, une pour évaluer l'extraction correcte des sessions (Revue.org), l'autre pour évaluer la pertinence des documents sur la base des clics utilisateurs (Nomao).

Analyse des documents.

Nous avons mis au point une nouvelle méthode afin d'extraire différentes formes de citation et de référence dans les documents (Revue.org). Notre hypothèse est que ces éléments sont des indicateurs sur la variété des documents. Nous travaillons également sur d'autres indicateurs, en particulier linguistiques, pour créer des groupes de documents.

Analyse des composants du processus de RI

Sur des collections internationales de référence, nous avons montré qu'il était possible d'apprendre pour une requête donnée quelle était la combinaison de module la plus pertinente à utiliser pour répondre au mieux au besoin d'information. Nous avons également étudié quels étaient les paramètres les plus influents lors d'une recherche.

Contextualisation de messages courts

Nous avons organisé un atelier dans le cadre de la conférence EGC 2013 sur ce thème. Par ailleurs, nous sommes co-organisateur d'une tâche « Tweet contextualization » à la campagne INEX (labs de CLEF).

CONCLUSION ET PERSPECTIVES

L'étude du contexte à travers les différents éléments qu'il comporte a conduit à des résultats qui nous permettent d'envisager maintenant leur combinaison.

Nous travaillons actuellement à la mise au point d'un modèle de recherche sélectif : à partir de l'indice de difficulté prédit d'une requête, le système choisit lui-même les traitements les plus adaptés. Nous visons l'intégration de ce modèle sélectif à la plateforme Terrier qui permettra son utilisation en particulier dans les campagnes d'évaluation internationales.

Parallèlement, nous mettons au point une méthode qui, en fonction du type de requête, restituera des documents du type le plus adapté en se basant sur des classifications préalables de documents, non pas par rapport à leur contenu thématique, mais par rapport à des critères d'usage. Ce travail s'appuie sur les données issues de la plateforme Revue.org.

CONTACT :

caas@irit.fr



LES RENCONTRES DU NUMÉRIQUE

17 et 18 avril 2013