

Exploitation du contexte utilisateur et de la structure des documents pour une recherche d'information ciblée. Application à un domaine de spécialité et à la recherche d'information dans un contexte mobile.

Directeur de thèse : Patrice Bellot (MCF HDR, 50%); **Co-encadrant :** Eric SanJuan (MCF, 50%) ;

Mots clefs: Recherche d'Information contextuelle, Système de Question Réponse, Interface Homme Machine, Traitement automatique de la langue naturelle écrite, Indexation, Classification Automatique.

Les moteurs de recherche d'information de l'Internet emploient tous plus ou moins les mêmes approches pour répondre aux requêtes des utilisateurs. Ils partent notamment de l'hypothèse que l'utilisateur souhaite obtenir rapidement une liste de réponses quitte à passer du temps ensuite à rechercher manuellement l'information désirée au sein de celles-ci. L'amélioration des méthodes de traitement automatique des langues (TAL) permet désormais d'envisager des systèmes plus puissants qui non seulement vont *fouiller* les pages web liées à la requête mais vont également pouvoir tenir compte du *contexte* de cette dernière. La nature de l'information recherchée diffère selon que la requête est effectuée dans un *contexte* de veille technologique, d'analyse d'opinion, de recherche d'information factuelle précise...

Le sujet de thèse que nous proposons s'inscrit dans le domaine de la *recherche d'information contextuelle*. Il consiste d'une part à analyser le contexte de recherche et d'autre part à proposer des méthodes d'accès à l'information qui en tiendront compte. La problématique étant très large, nous nous concentrerons sur deux contextes pour des requêtes liées à des domaines (de spécialité ou non) précis. Le premier contexte sera centré sur la recherche d'informations à caractère historique: les notions de chronologie, de points de vue multiples et de « crédibilité » seront prépondérantes. Le deuxième contexte concernera la recherche d'information en environnement mobile et nous traiterons plus particulièrement des requêtes dans lesquelles l'analyse d'opinion et les informations précises (quantités, localisations, caractéristiques...) jouent un rôle premier — par exemple : *quelle est la pièce de théâtre la plus appréciée* (implicitement : à l'endroit où je me trouve et actuellement) —. Ces deux contextes s'inscrivent dans la lignée des objets d'étude des différents projets que nous avons soumis à l'ANR en 2010 avec, entre autres, les laboratoires « Culture et Communication » et « Biens, Normes, Contrats » de l'UAPV.

Le principal objectif scientifique de cette thèse est d'étudier les méthodes complémentaires à l'utilisation des historiques pour la détection et la prise en compte des contextes de recherche. Il a été démontré que la reformulation et l'expansion des requêtes fondées sur une analyse des historiques (*logs*) des moteurs de recherche du web permet d'améliorer la qualité de la recherche d'information (Uichin L., Zhenyu L., Junghoo C.2005). Lorsqu'ils sont de très grande dimension, ces historiques permettent une modélisation probabiliste efficace du comportement de l'utilisateur. Les requêtes sont classées selon les documents communs les plus souvent visualisés et sont alors reformulées sur la base de similarités et leurs résultats réordonnés sur la base de classes de documents. Cependant ces historiques ne sont pas publics. Les moteurs de recherche, tels ceux de Google et de Microsoft, qui disposent des plus larges historiques de requêtes, ont un avantage majeur vis à vis des autres acteurs du domaine, avantage qui ne cesse de s'amplifier: les bonnes performances pouvant attirer plus d'utilisateurs qui viennent à leur tour enrichir ce modèle probabiliste. Différentes alternatives ont été proposées. La première à avoir montré de bonnes performances lors des campagnes internationales d'évaluation TREC (Dang, V., Croft, W. B. 2010) repose sur l'utilisation des liens hypertextes. Par ailleurs, de multiples expérimentations (SanJuan E., Ibekwe F. 2010) montrent qu'un léger processus interactif fondé sur des approches robustes de traitement automatiques des langues naturelles permet d'obtenir des reformulations de requêtes équivalentes. Ces deux approches présentent l'avantage de pouvoir être adaptées à la recherche d'information de type questions-réponses où l'on cible des passages de documents au lieu de documents entiers ainsi qu'au cas où la réponse doit être extraite et recomposée à partir de documents non ou peu structurés (questions dites complexes).

Au-delà de l'expertise du LIA en recherche d'information et en traitement automatique des langues naturelles, le travail de recherche passera par exemple par le croisement d'une analyse (classification) de la requête et des documents associés afin de détecter des classes stylistiques, d'identifier des niveaux de spécialisation et d'expertise mais aussi par l'emploi de méthodes robustes de fouille de texte applicables à des contenus variés (détection d'entités représentatives de classes lexicales non prédéfinies, extraction de valeurs numériques associées à leur signification). Les méthodes numériques choisies seront associées à des approches interactives ainsi qu'à des évaluations grandeur nature: campagnes internationales et définition de tâches précises en partenariat avec d'autres laboratoires de l'UAPV selon les contextes cités plus haut.

Dang, V, Croft, W. B. , « Query Reformulation Using Anchor Text », to appear in the Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM) 2010.

Uichin L., Zhenyu L., Junghoo C., « Automatic identification of user goals in Web search », Proceedings of the 14th international conference on World Wide Web 2005, pp: 391 – 400.

SanJuan E., Ibekwe-Sanjuan F.: Multi Word Term Queries for Focused Information Retrieval. CILing 2010: 590-601