RI géolocalisée et contextualisation

Léa Laporte

Université de Toulouse, UT3, IRIT-IMT
118 route de Narbonne
31062 Toulouse cedex 9
lea.laporte@irit.fr

Nomao
1 avenue Jean Rieux
31500 Toulouse
lea.laporte@wikiogroup.com

Learning-to-rank et contextualisation

Un grand nombre de moteurs de recherche d'informations géoréférencées sont apparus au cours des dernières années (Yelp, AroundMe, Nomao). Ils proposent aux internautes et aux utilisateurs de smartphones des services de recommandation personnalisée de lieux. Ceux-ci sont sélectionnés et ordonnés suivant leur proximité géographique, leur e-réputation et leur adéquation aux goûts et aux réseaux sociaux de l'utilisateur.

Dans ce cadre, nos travaux portent sur la problématique de l'optimisation automatique des résultats (*learning-to-rank*) et y intègrent un challenge supplémentaire : la contextualisation. Il s'agit d'adapter le système à l'ensemble des éléments caractérisant les interactions entre l'utilisateur, son environnement et le système de recherche d'information. L'objectif principal peut se résumer de la façon suivante : comment obtenir de façon automatique un classement optimal des documents (des lieux) les plus pertinents vis-à-vis d'une requête et d'un contexte ?

En *learning-to-rank*, il s'agit de considérer une collection de couples (requête, document) pour lesquels la pertinence est connue. Une fonction d'ordonnancement est apprise à l'aide d'un algorithme d'apprentissage sur ce jeu de données. Cette fonction est ensuite utilisée pour prédire la pertinence de nouvelles paires (requête, document) et fournir un classement.

Beaucoup d'algorithmes ont ainsi été développés au cours de la dernière décennie. Ils sont basés sur des approches distinctes, qui diffèrent sur deux points : la façon de considérer les documents en entrée du système d'apprentissage (document seul, paire de documents, ou liste de documents) et la modélisation du problème d'ordonnancement (régression, classification, optimisation d'une fonction de perte). Tie-Yan Liu en dresse un panorama dans son article de synthèse (Liu, 2009).

La performance des algorithmes est évaluée sur des collections dans lesquelles la pertinence d'un document vis-à-vis de la requête est donnée par un expert

(méthodologie Cranfield). Ceci pose un problème dans le cadre de la personnalisation, car les besoins et les jugements des utilisateurs ne sont pas connus. Des approches utilisant les logs de requêtes pour construire la base d'apprentissage ont ainsi émergé. Les jugements sont connus au travers des clics, qui sont traduits comme une préférence pour un document par rapport à un autre (Joachims, 2002).

Nos travaux portent sur la contextualisation en recherche d'information, via les algorithmes d'ordonnancement. Nous nous intéressons à trois composantes du contexte : le support d'accès à l'information, les catégories de requêtes et l'utilisateur. Nous utilisons une approche basée sur les logs, adaptée à ce cadre. Le paragraphe suivant présente les avancées et les perspectives des travaux.

Premiers résultats et perspectives

L'analyse des données de notre partenaire industriel a permis de mettre en évidence des adaptations possibles suivant la catégorie associée à la requête (prise en compte plus ou moins forte de critères de popularité suivant la catégorie) et le type d'usage. L'étude avait en effet révélé des usages différents suivant le support utilisé (web ou smartphone). Ces résultats constituent une base préliminaire pour l'adaptation des fonctions d'ordonnancement. D'autres études doivent être menées sur les logs afin de déterminer les pistes pour l'adaptation du système, comme l'existence de comportements types des utilisateurs ou des catégories de requêtes, et d'étudier les difficultés d'interprétation des clics en tant que critères de satisfaction utilisateur (Radlinski *et al*, 2008). Un outil d'analyse et d'extraction des données à partir des logs est en cours d'implémentation. L'intégration de ces éléments au sein d'un algorithme d'ordonnancement, l'implémentation et l'évaluation d'un tel algorithme auront lieu ultérieurement.

Ces travaux sont co-dirigés par Josiane Mothe à l'IRIT et Sébastien Déjean à l'IMT, et co-encadrés par Laurent Candillier (Nomao). Ils sont financés par la région Midi-Pyrénées, et NOMAO, moteur de recherche géolocalisé et personnalisé.

Bibliographie

Joachims T., « Optimizing search engines using clickthrough data », *Conference on Knowledge Discovery and Datamining*, 2002, p. 133-142.

Liu T.-Y., « Learning to rank for Information Retrieval », *Foundations and Trends in Information Retrieval*, vol. 3, n°3, 2009, p. 225-331.

Radlinski F., Kurup M., Joachims T., « How does clickthrough data reflect retrieval quality ? », *Conference on Information and Knowledge Management (CIKM)*, 2008, p. 43-52.