Mohand Boughanem, Professeur.
CNRS-IRIT
Université Paul Sabatier
118, Route de narbonne 31062 Toulouse Cedex 04
bougha@irit.fr

Outils de validation en recherche d'information : la campagne d'évaluation TREC

La démarche de validation en RI (Recherche d'Information) se base sur l'évaluation expérimentale des performances du modèle ou du système proposé. Cette évaluation peut porter sur plusieurs critères : le temps de réponse, la pertinence, la qualité et la présentation des résultats, etc. Le critère le plus important est celui qui mesure la capacité du système à satisfaire le besoin en information de l'utilisateur, c'est à dire la pertinence. Deux facteurs permettent d'évaluer ce critère. Le premier est le rappel, il mesure la capacité du système à sélectionner tous les documents pertinents. Le second est la précision, il mesure la capacité du système à rejeter tous les documents non pertinents.

Pour pouvoir effectuer ce type d'évaluation, la communauté en recherche d'information s'est dotée de collections de tests. Une collection de tests est composée d'un ensemble de documents, un ensemble de requêtes et un ensemble de jugements de pertinence. Un jugement de pertinence est une liste de documents « censés » être pertinents pour une requête donnée. La problématique réside dans la construction d'une « bonne » collection de test. Plusieurs initiatives ont été lancées, la première d'entre elles date de 1968 avec la constitution de la collection de test Cranfield [Claverdon 1968].

L'initiative la plus importante actuellement est sans conteste TREC (Text REtrieval Conference, http://trec.nist.gov) [Harman 1992]. TREC est plus qu'une collection de test, c'est un programme d'évaluation de systèmes de recherche d'information, initié par le NIST aux USA. Il fournit une plate-forme comportant des collections de tests, des tâches spécifiques et des protocoles d'évaluation pour chaque tâche, pour l'évaluation et la comparaison d'expérimentations sur des collections volumineuses de textes. TREC a débuté en 1992, il comptait à cette époque 25 groupes. Aujourd'hui, plus 80 groupes ont envoyé des résultats de tests effectués sur les sept tâches pré-définies pour la campagne de 2002.

D'autres initiatives ont vu le jour récemment en Europe et au Japon. On y trouve par exemple :

- en Europe les campagnes annuelles CLEF (Cross Language Evaluation Forum) [Peters 2000] lancé en 2000 pour l'évaluation de systèmes de recherche d'information multilingue et INEX lancé en 2002 (http://www.is.informatik.uni-duisburg.de/projects/inex03/) pour l'évaluation de systèmes de recherche d'information sur des documents XML.
- Amaryllis, une version française de TREC de 1996 à 1999. Amaryllis a intégré la campagne CLEF en 2002. Il s'occupe de la tâche évaluation monolingue de collections de textes Français.
- le projet annuel NTCIR (NII-NACSIS Test Collection for IR Systems).

L'évaluation des performances de SRI basée sur les collections de tests actuelles ne fait l'unanimité au sein de la communauté RI. Une des critiques récurrente réside dans le fait que les collections de tests présentent la pertinence vis-à-vis d'un besoin comme une décision binaire « objective », or il est reconnu que la pertinence est plutôt une notion subjective

dépendant complètement de l'utilisateur. De plus, certains travaux [Harter 1996] ont montré que les jugements de pertinence pour un même besoin, diffèrent selon les assesseurs et l'instant du jugement.

L'objectif de cette présentation est de décrire les outils dont on dispose aujourd'hui pour l'évaluation d'expérimentations (de laboratoire) en recherche d'information. Cette présentation sera principalement axée sur TREC, pour mieux montrer les différentes phases allant de la construction de collections, l'envoi des résultats d'expérimentations par les participants jusqu'à la publication des évaluations en passant par la phase de jugement des pertinences.

Bibliographie

[Cleverdon 1968] Claverdon C. W., Mills J., & Keen E. M. *Factors determining the performance of indexing systems*. Two volumes. Cranfield, England

[Harman 1992] Harman D. *Overview of the First Text REtrieval Conference (TREC-1).* NIST Special Publication 500-207. Proceedings of the First Text REtrieval Conference pp 1-20.

[Harter 1996] Harter S. P. *Variations in relevance assessments and the measurement on retrieval experiments*. Journal of the American Society for Information Science, 29 (4), 411-414.

[Peters 2000] Peters C. *Cross-Language Information Retrieval and Evaluation.* Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000, Lecture Notes in Computer Science 2069 Springer 2001, ISBN 3-540-42446-6