Etude des résultats des systèmes de RI à grande échelle
Sélection des mesures de performance pour l'évaluation

Sébastien Déjean*, Josiane Mothe**, Julia Poirier***,
Benoît Sansas***, Joelson Randriamparany **,

*IMT, UMR, Université de Toulouse
**IRIT, URM5505, Université de Toulouse
***INSA, Université de Toulouse

Résumé. Cet article présente un usage des méthodes d'analyse de données dans le domaine de la recherche d'information. Plus spécifiquement, nous analysons l'ensemble des données issues de la campagne d'évaluation TREC et montrons que les mesures utilisées pour évaluer les systèmes peuvent se réduire à 7 au lieu des 128 qui sont intégrées dans le programme d'évaluation trec_eval.

1 Introduction

Les moteurs de recherche d'information (RI) visent à retrouver l'information pertinente par rapport à un besoin formulé sous forme de requête d'un utilisateur. Pour permettre cette RI, les systèmes se basent sur plusieurs principes fondamentaux dont l'indexation, qui vise à représenter chaque document à l'aide de termes d'indexation et l'appariement de la requête avec chacun des documents. Ainsi, les différents moteurs de recherche diffèrent par les mécanismes mettant en œuvre l'indexation d'une part et l'appariement d'autre part.

Les performances d'une chaine de traitement sont évaluées lors de campagnes internationales comme Text Retrieval Conference (TREC). Pour une tâche donnée, TREC propose l'ensemble des ressources nécessaires aux participants pour tester leurs systèmes. Ainsi, par exemple, la tâche TREC *adhoc* vise à évaluer les moteurs qui restituent une liste de documents à partir d'une requête. TREC fourni pour cela un ensemble de documents et un ensemble de besoins d'information. Les participants fournissent en retour la liste des documents que leurs systèmes retrouvent. L'évaluation est ensuite réalisée en considérant les réponses attendues (données par des acteurs humains) et des mesures de performance.

Actuellement, plusieurs centaines de requêtes de test sont disponibles et plusieurs dizaines de résultats provenant de variantes de chaines de traitement. Cette masse d'information n'est que très peu exploitée dans le domaine de la recherche d'information. Cet article vise à proposer quelques pistes d'analyse qui pourraient ouvrir de nouvelles voies dans le domaine de la recherche d'information. Les outils statistiques nous ont paru pertinents pour une analyse poussée de ces masses de données.

La suite de l'article est organisée comme suit : dans la section 2, nous présentons quelques travaux reliés. Dans la section 3, nous présentons les données que nous avons analysées. Dans la section 4 nous montrons qu'il est possible, grâce à l'analyse de données, de regrouper les mesures d'évaluation selon des classes homogènes. Dans la section 5, nous montrons que le classement d'un ensemble de systèmes est stable lorsque nous considérons 6 mesures au lieu des 128 initiales, que l'on se base sur le score ou sur le rang moyen des systèmes. Nous concluons ces travaux et présentons plusieurs pistes pour les prolonger.

2 Travaux reliés

Peu de travaux se sont intéressés à l'analyse des résultats issus des campagnes d'évaluation.

Banks *et al.* (1999) décrivent différentes analyses sur différentes données de TREC. En particulier, ils considèrent une matrice dans laquelle les lignes et les colonnes représentent les systèmes et les besoins d'information ; les cellules correspondent à la *précision moyenne* (moyenne des précisions à chaque fois qu'un document pertinent est retrouvé). Cette matrice est alors utilisée pour étudier les groupes qui pourraient en être extraits. Pour cela, les auteurs utilisent une classification hiérarchique par simple liage. Cette méthode combine les groupes qui minimisent la distance entre les éléments les plus proches. Les figures présentées dans l'article cachent la distance entre les groupes détectés ; de plus l'article ne discute pas l'arbre obtenu ni les groupes qui auraient été obtenus en coupant l'arbre à différents niveaux. Les auteurs concluent que l'analyse n'apporte pas d'information vraiment utilisable.

Une analyse différente est présentée par Mizzaro et Robertson (2007). L'objectif de leur étude est d'identifier un petit ensemble de requêtes qui pourrait être utilisé pour distinguer les systèmes efficaces des systèmes qui ne le sont pas. Les auteurs utilisent les mesures de performance sur les paires systèmes/requêtes. Les auteurs concluent que certains systèmes sont meilleurs pour distinguer les requêtes faciles de celles qui sont difficiles. Ils concluent également que les systèmes peu performants le sont quelque soit la requête alors que de bons systèmes peuvent échouer sur certaines requêtes. Enfin, les auteurs montrent que les requêtes les plus faciles sont les plus aptes à distinguer les systèmes par rapport à leurs performances.

3 Données et objectifs d'analyse

3.1 Matrice de données à analyser

Les données que nous utilisons sont obtenues à partir des résultats des participants à une tâche d'évaluation. Plus précisément, les données initiales sont les listes de documents retrouvées par chaque système participant à une tâche. Nous nous sommes basés sur la tâche *adhoc* de TREC. Cette tâche se compose chaque année de 50 requêtes à traiter ; le nombre de systèmes et les systèmes eux-mêmes varient d'une année sur l'autre. Par exemple, en 1999, 129 systèmes ont traité les requêtes. A partir de ces éléments, pour cette seule année, nous pouvons considérer un ensemble de 6450 individus à analyser (un individu pour chaque système et pour chaque requête). Pour chaque individu, nous calculons un ensemble de valeurs correspondant à des mesures de performance, ce seront les variables. Elles sont calculées par l'outil trec_eval (Buckley, 1991) et sont au nombre de 128. La majorité de ces mesures ont des valeurs ou scores compris entre 0 et 1.

Un extrait de la matrice qui nous sert de base dans l'analyse est présenté à la figure 1.

3.2 Objectif d'analyse

Nous nous sommes fixés deux objectifs pour l'analyse :

- Compte tenu du nombre important de mesures permettant d'évaluer la performance des systèmes, nous souhaitons vérifier le niveau de redondance et la complémentarité

de celles-ci. En effet, lorsqu'un nouveau système de RI est mis au point, la mesure de ses performances est une étape cruciale. Comparer le nouveau système aux autres systèmes existants sur 128 mesures s'avère fastidieux. A l'opposer, se limiter à la mesure *Mean Average Precision* (moyenne sur un ensemble de besoins d'information des *précisions moyennes*) risque de cacher une partie des performances. L'analyse des données pourrait permettre de déterminer un nombre optimal de mesures à considérer lors d'une évaluation, voire à les identifier (Baccini *et al.*, 2010).

- La mise au point de nouveaux modèles de RI ou l'adaptation de modèles existants visent à obtenir un « meilleur » système ; il s'agit donc de pouvoir comparer des systèmes entre eux. Nous étudions l'impact de l'utilisation d'un nombre réduit de mesures de performance sur les rangs obtenus par les systèmes.

campagnes	taches	systemes	requetes	0.20R-prec	0.40R-prec	0.60R-prec
TREC1999	adhoc	1	401	0.016700	0.008300	0.011100
TREC1999	adhoc	1	402	0.000000	0.000000	0.000000
TREC1999	adhoc	1	403	0.000000	0.000000	0.000000
TREC1999	adhoc	1	404	0.000000	0.000000	0.000000
[...]						
TREC1999	adhoc	weaver2	447	0.750000	0.571400	0.600000
TREC1999	adhoc	weaver2	448	0.000000	0.000000	0.035700
TREC1999	adhoc	weaver2	449	0.000000	0.000000	0.000000
TREC1999	adhoc	weaver2	450	0.762700	0.686400	0.596600

Figure 1 : Extrait de la matrice analysée

4 Redondance et complémentarité des mesures

4.1 Données et méthode d'analyse

Pour étudier la redondance dans les mesures de performance des systèmes de RI, nous nous sommes appuyés sur l'ensemble des données collectées pour la tâche adhoc de TREC (TREC-2 à TREC-8). Nous avons éliminé la première année qui a servi de mise au point de la définition de la tâche (TREC-1). Au total, nous obtenons une matrice composée de 23 518 individus. Les variables sont celles présentées dans la section 3 (mesures de performances obtenues par trec_eval).

D'un point de vue outils mathématiques, pour analyser ces données, nous nous sommes appuyés sur une **classification ascendante hiérarchique**. En effet, notre objectif est de regrouper les mesures de performance en groupes le plus homogènes possibles. N'ayant pas d'idée *a priori* sur le nombre de classes à obtenir, une telle classification, nous permet de choisir à postériori un ensemble de classes. Seber (1984) préconise, lorsque cela est possible de stabiliser ce type de classification en la faisant suivre d'une classification supervisée du type k-means. Lors de la classification hiérarchique, nous avons utilisé la mesure de Ward qui est la plus utilisée. Elle consiste à fusionner à chaque étape les groupes qui minimisent l'augmentation du total de la somme des carrés des distances dans les groupes.

4.2 Résultats d'analyse

4.2.1 Classes de mesures de performances des systèmes de RI

La figure 2 présente le dendrogramme résultant de la classification hiérarchique des mesures de performance des systèmes de RI. Le graphique en haut à droite de la figure représente la distance entre les nœuds. Ce graphique suggère une coupe pertinente en considérant 3 groupes ; une autre en considérant 5 groupes et une dernière en considérant 7 groupes (*cf.* la pente sur le graphique cité plus haut).

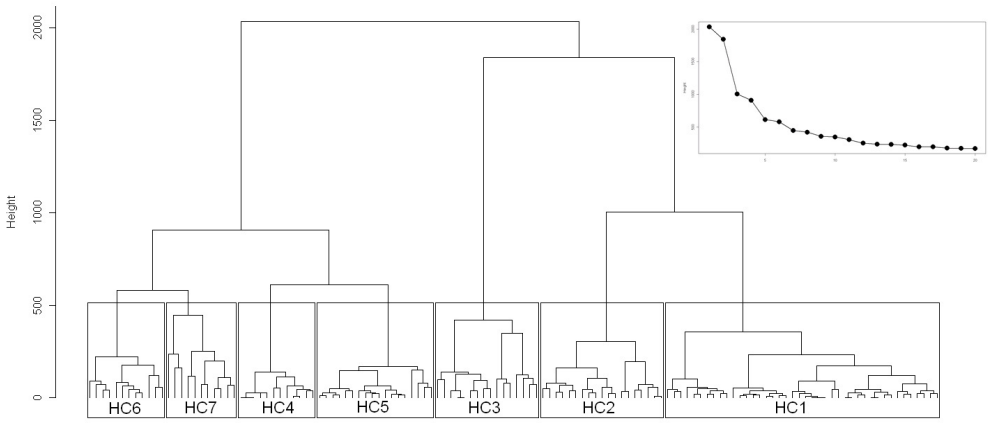


Figure 2. Dendrogramme représentant la classification hiérarchique des mesures de performance

L'application d'une classification k-means à partir des groupes ainsi obtenus à montré la bonne stabilité des groupes puisque seulement 13 mesures ont changé de groupe après son application. La classification supervisée a isolé un groupe de 3 mesures ne correspondant pas à des scores (*nombre de documents retrouvés*, *rang du premier document pertinent retrouvé* et *nombre de documents jugés non pertinents retrouvés).*

Au final, le nombre de groupes (7) correspond donc au nombre minimal de mesures permettant de couvrir les différents aspects mesurés par l'ensemble complet des mesures.

4.2.2 Homogénéité des groupes et représentant de groupe

Une fois les groupes déterminés, pour permettre leur utilisation, nous avons sélectionné un représentant de groupe : pour un groupe donné, il s'agit de la mesure qui devrait être étudiée pour évaluer le système de RI. Selon les résultats de l'analyse, utiliser une autre mesure du même groupe conduirait aux mêmes conclusions lors de la comparaison du système évalué avec d'autres. En effet, les mesures issues d'un même groupe sont trouvées comme redondantes.

Avant de choisir le représentant de chaque groupe, nous nous sommes assurés de leur homogénéité. Compte tenu de la forte homogénéité des groupes obtenus, n'importe quelle mesure pourrait servir de représentant de groupe. Cependant, nous avons arbitré en choisissant, soit la mesure du groupe la plus communément utilisée dans la littérature, soit la mesure la plus proche du centroïde. Le détail des groupes obtenus n'est pas présenté dans cet article.

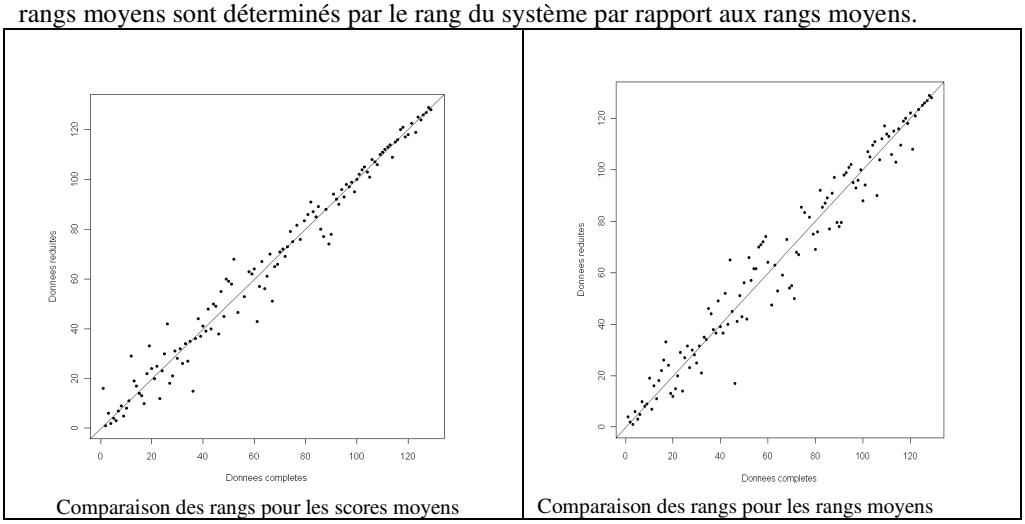
Les mesures représentatives des groupes ainsi que les caractéristiques des groupes sont les suivantes :

- MAP (Mean Average Precision) est un bon représentant du premier groupe. Ce groupe contient également les mesures permettant de réaliser les courbes de rappel/précision. La mesure MAP est connue pour permettre une représentation globale des performances.
- P10 (Précision lorsque 10 documents sont retrouvés) représente bien le groupe qui associe dans le résultat de l'analyse les mesures de hautes précisions.
- P100 représente bien le groupe qui associe les mesures de précision lorsque de grands ensembles de documents retrouvés sont considérés.
- Exact recall (Rappel exact, non interpolé) peut représenter le groupe correspondant aux mesures orientées rappel et qui ont été regroupées par l'analyse.
- rank first rel est le représentant des 3 mesures regroupées qui ne sont pas des scores.
- Recall 30 (rappel lorsque 30 documents sont retrouvés) et bpref topnonrel sont deux mesures que nous avons choisies pour représenter les deux groupes restants.

5 Classement des systèmes : effet de score et de rang

Généralement, la mise au point d'un système de RI implique la comparaison avec d'autres et donc un classement du système. Le principal objectif de l'analyse présentée dans cette section est de comparer les méthodes de classement des systèmes lorsque l'ensemble des mesures sont utilisées et lorsque l'ensemble réduit de mesures est choisi (Poirier et Sansas, 2009).

Nous définissons le **score moyen d'un système** par la moyenne des valeurs sur l'ensemble des mesures de performance pour chaque système. Le rang d'un système est déterminé par rapport aux scores moyens après avoir ordonné les systèmes par ordre croissant des scores moyens. Le **rang moyen d'un système** sur l'ensemble des mesures correspond à la moyenne des rangs obtenu pour chaque mesure. Les systèmes qui ont les meilleurs rangs moyens sont déterminés par le rang du système par rapport aux rangs moyens.



Comparaison des rangs pour les scores moyens | Comparaison des rangs pour les rangs moyens

Figure 3 – Comparaison des rangs obtenus pour les données complètes et réduites (req. 425)

La figure 3 montre les différences dans le classement obtenu lorsque l'ensemble des mesures de performance sont considérées (axe des X) et lorsque l'ensemble réduit de mesure est utilisé (axe des Y) pour la requête 425. La partie de droite considère les scores moyens alors que celle de gauche considère les rangs moyens. Quelle que soit la méthode, on constate que les résultats sont stables. Cependant, le classement est plus homogène en utilisant la méthode des scores moyens. Ceci est dû à un nombre d'ex-aequo beaucoup plus important en utilisant la méthode du rang des rangs moyens qu'en utilisant la méthode du rang des scores moyens. Le test de corrélation de Kendall indique une corrélation de 0,916 pour les rangs des scores moyens et de 0,878 pour les rangs des rangs moyens.

6 Conclusions et perspectives

Dans cet article, nous nous sommes intéressés à l'analyse de données issues du domaine de la RI. Grâce aux méthodes d'analyse de données, nous avons pu réduire le nombre de mesures de performance à utiliser pour comparer deux systèmes. Nous avons montré que 7 mesures de score étaient suffisantes pour représenter une plus large gamme de mesures.

Le prolongement de ces travaux vise à utiliser les méthodes d'analyse de données pour l'étude fine des impacts des différents modules de RI et des caractéristiques des requêtes. En effet, dans cette présente étude, nous avons considéré les systèmes de RI comme des boites noires et n'avons pas considéré leurs caractéristiques. De plus, nous avons considéré de la même façon tous les besoins d'information. Nous souhaitons dans le futur réaliser une étude plus fine qui prendrait en compte ces aspects. Ces travaux s'inscrirons dans le cadre du projet ANR CAAS (Analyse Contextuelle et Recherche d'information Adaptative) dans lequel deux partenaires industriels du domaine de la RI collaborent.

Références

Baccini A., Déjean S., Mothe J., (2010). Analyse des critères d'évaluation des systèmes de recherche d'information. Technique et Science Informatiques, 29(3) :289-308.

Banks D., Over P., Zhang N.-F., (1999). Blind Men and Elephants: Six Approaches to TREC data, Information Retrieval, 1(1-2), 7-34.

Buckley, C. (1991). Trec eval,available at http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/trec_eval_video/README

Poirier J. and Sansas B. (2009). Comparaison des classements de systèmes de recherche d'information en fonction des mesures de performances utilisées. Rapport Interne IRIT/RR–2009-31–FR, IRIT.

Seber, G.A.F. (1984). Multivariate Observations, Wiley.

Summary

This paper presents the use of data analysis methods for information retrieval. More specifically, we analyze the set of the data resulting from TREC survey and show that performance measures used to evaluate systems can be reduced to 7 instead of the 128 which are integrated in the trec_eval evaluation program.