Analyses statistiques pour l'évaluation des systèmes de recherche d'information

S. Déjean, J. Mothe

www.math.univ-toulouse.fr/~sdejean

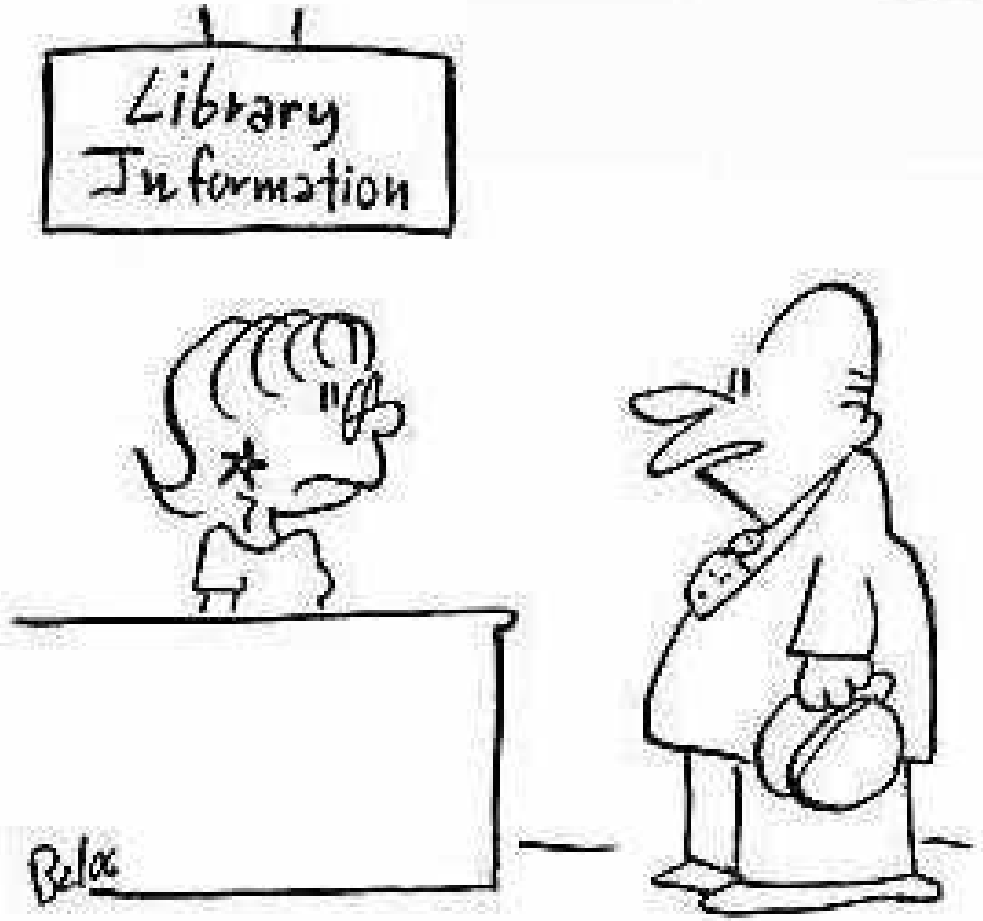
INSTITUT de MATHEMATIQUES de TOULOUSE

Institut de Recherche en Informatique de Toulouse

Recherche d'information

http://boston.lti.cs.cmu.edu/classes/11-744/

Recherche d'information

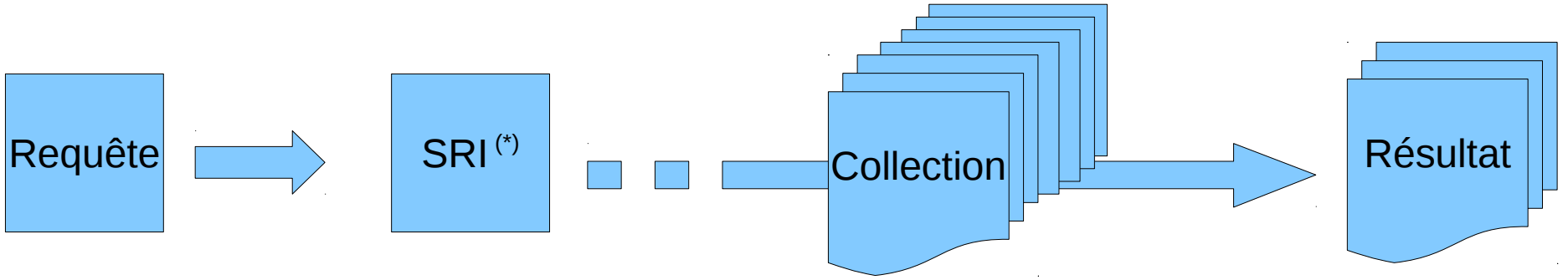


"I CAN'T FIND THE BOOKS ON INFORMATION RETRIEVAL."

https://oncealibrarian.wordpress.com/2012/03/12/62/

Recherche d'information

Recherche d'Information :
domaine de la recherche qui s'intéresse à la structure, à l'analyse, à l'organisation, au stockage, à la recherche et à la découverte de l'information (G. Salton).



Quoi ? Comment ? Où ? Comment évaluer le résultat ?

(*) SRI : Système de Recherche d'Information

Partie I
Mesures de performance

A. Baccini, S. Déjean, L. Lafage, J. Mothe (2011). **How many performance measures to evaluate Information Retrieval Systems?** *Knowledge and Information System*, DOI 10.1007/s10115-011-0391-7

TREC : Text REtrieval Campaign

- Tâches : adhoc (évaluation de performances des systèmes), *filtering track* (filtrage d'informations), *cross-language track* (RI multilingue), *web track, question answering*...
- Requête(*) : titre, description, partie narrative
- Les « bons » documents sont connus a priori
- *Trec_eval* : outil de calcul des performances des différents systèmes (« pertinence » des documents retournés)

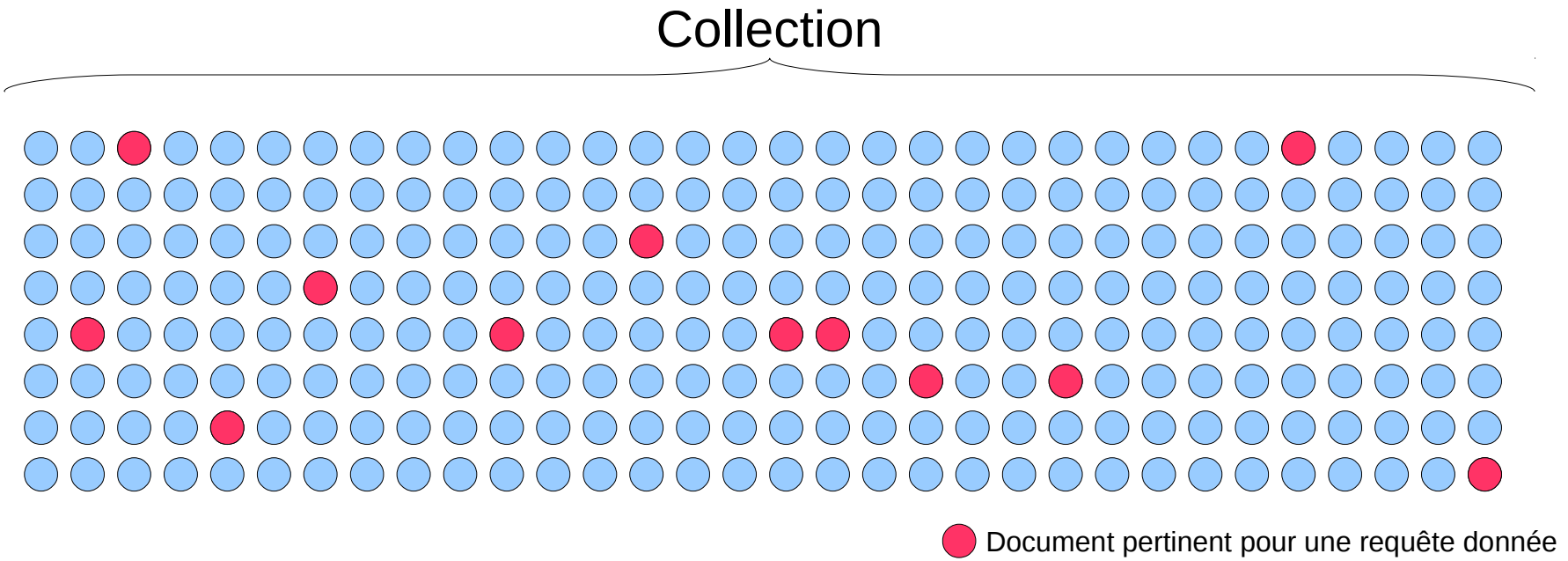
<title> **Topic:** Corporate Pension Plans/Funds ***(*) Exemple de requête***

<desc> **Description:**
Document will report on problems associated with pension plans/funds such as fraud, skimming, tapping or raiding.

<narr> **Narrative:**
A relevant document will report on problems associated with pension plans/funds and also U.S. Government regulatory controls on pension plans. Examples of problems that are considered relevant are fraud, skimming, tapping or raiding.

Mesures de performance : rappel et précision

Rappel : nombre de documents pertinents retournés / nombre total de documents pertinents
Précision : nombre de documents pertinents retournés / nombre de documents retournés



Résultat du SRI

- 5 premiers résultats :  Rappel = 3/12, Précision = 3/5
- 10 premiers résultats :  Rappel = 5/12, Précision = 5/10
- ...

Trec_eval : 135 mesures de performance

`http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/trec_eval_video/README`

trec_eval is the standard tool used by the TREC community for evaluating an ad hoc retrieval run, given the results file and a standard set of judged results

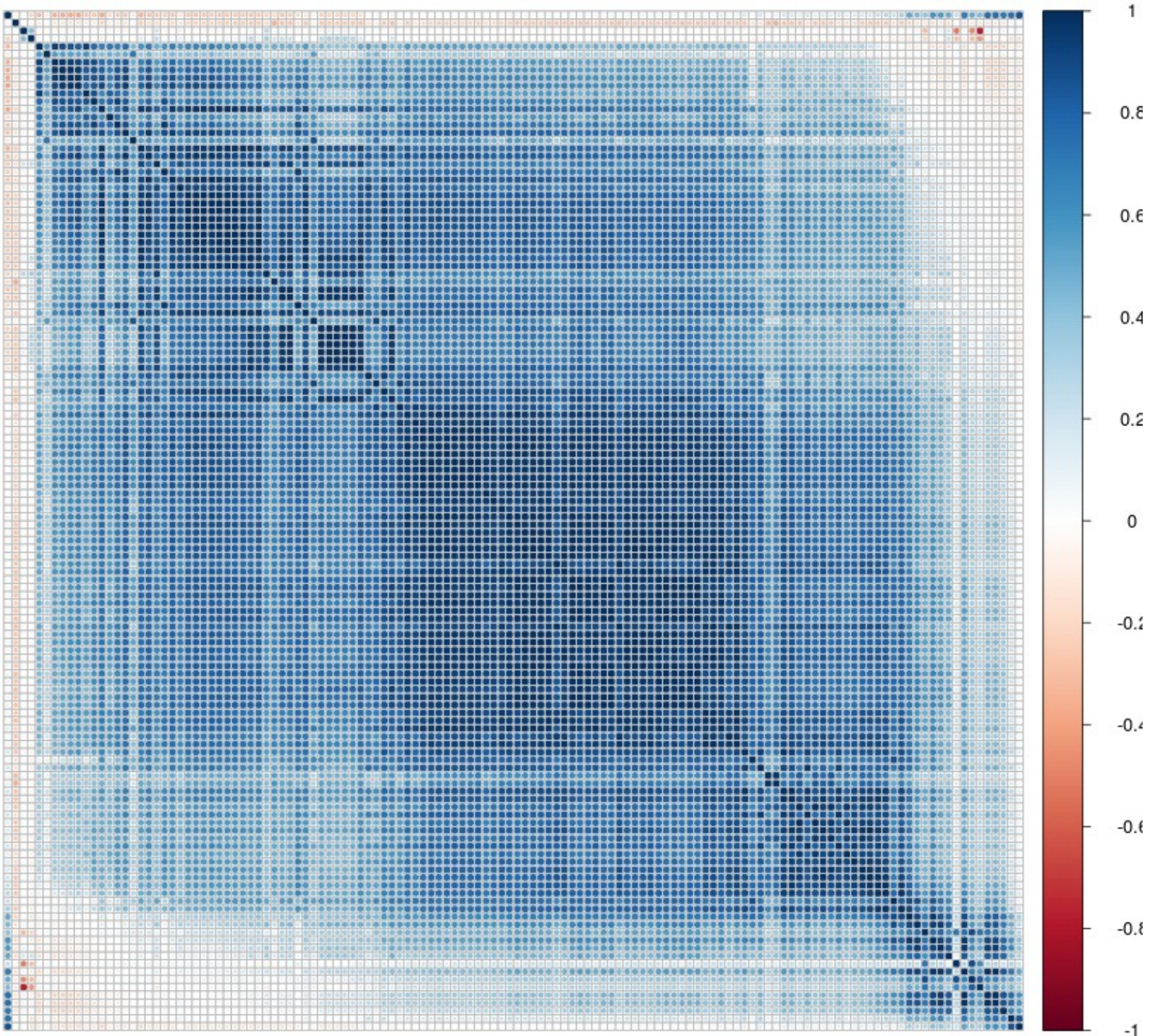
map	*Mean Average Precision (MAP)*
gm_ap	*Average Precision. Geometric Mean, q_score=log(MAX(map,.00001))*
R-prec	*R-Precision (Precision after R (= num-rel for topic) documents retrieved)*
bpref	*Binary Preference, top R judged nonrel*
recip_rank	*Reciprical rank of top relevant document*
ircl_prn.0.00	*Interpolated Recall - Precision Averages at 0.00 recall*
[…]	
ircl_prn.1.00	*Interpolated Recall - Precision Averages at 1.00 recall*
P5	*Precision after 5 docs retrieved*
[…]	
P1000	*Precision after 1000 docs retrieved*
exact_prec	*Exact Precision over retrieved set*
exact_recall	*Exact Recall over retrieved set*
3-pt_avg	*Average over 3 points of recall-precision graph*
avg_doc_prec	*Rel doc precision averaged over all relevant docs (NOT over topics)*
exact_relative_prec	*Exact relative precision*
avg_relative_prec	*Average relative precision*
exact_unranked_avg_prec	*Exact Unranked Average Precision*
map_at_R	*Average Precision over first R docs retrieved*
int_map	*Interpolated Mean Average Precision*
exact_int_R_rcl_prec	*Exact R-based-interpolated-Precision*
int_map_at_R	*Average Interpolated Precision for first R docs retrieved*
bpref_topnonrel	*Binary Preference, top 100 judged nonrel*
bpref_top5Rnonrel	*Binary Preference, top 5R judged nonrel*
bpref_top25p2Rnonrel	*Binary Preference, top 25 + 2*R judged nonrel*
bpref_retall	*Binary Preference, Only retrieved judged rel and nonrel*
bpref_5	*Binary Preference, top 5 rel, top 5 nonrel*
bpref_10	*Binary Preference, top 10 rel, top 10 nonrel*
bpref_num_all	*Binary Preference, Number not retrieved before (all judged)*
bpref_num_ret	*Binary Preference, Number retrieved after*
recall5	*Recall after 5 docs retrieved*
[...]	

- Un extrait des 135 mesures fournies par *trec_eval*
- La plupart sont des combinaisons de Rappel et Précision

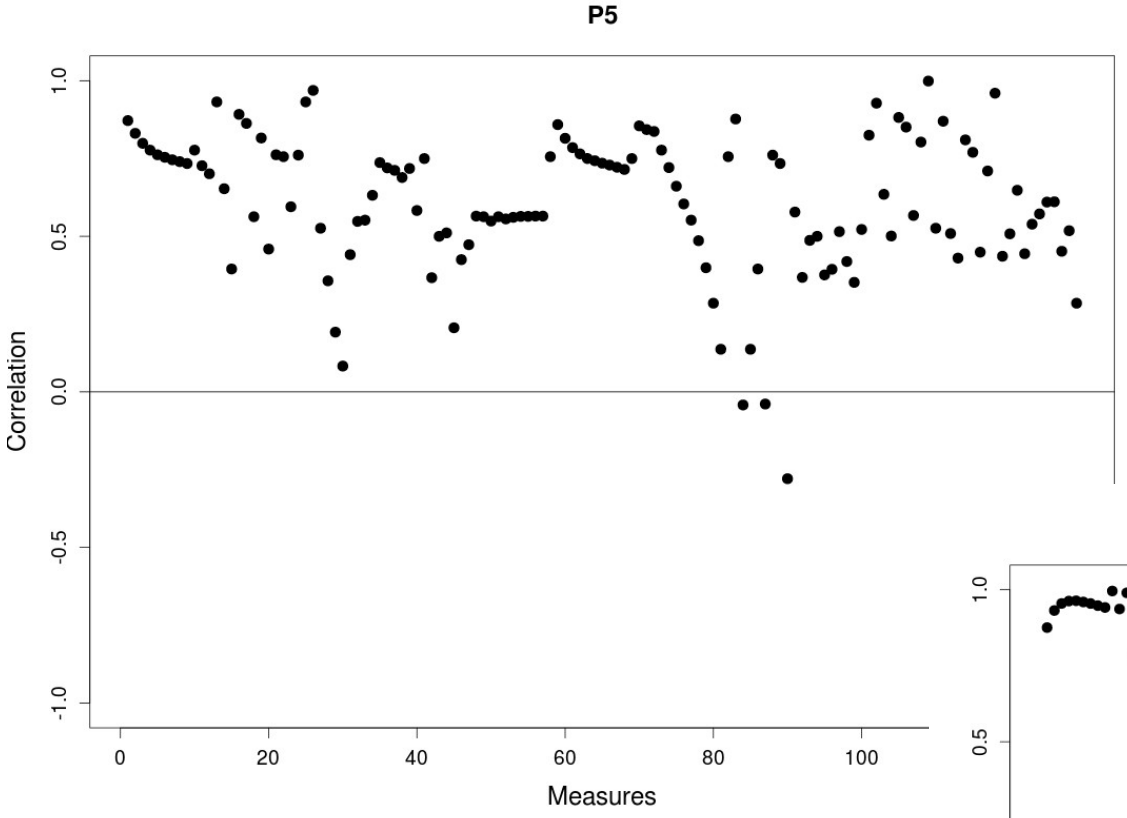
Redondance des mesures

- L'interprétation globale des 135 mesures est délicate et pas forcément pertinente
- Certaines mesures sont, par nature, redondantes (corrélées) : P5, P10, P15...
- La caractérisation d'un SRI doit pouvoir se faire à partir d'une nombre restreint de mesures
- Étude réalisée sur des données issues de 7 campagnes TREC : 23 518 runs (requête $\times$ système) pour lesquels on dispose des 135 mesures

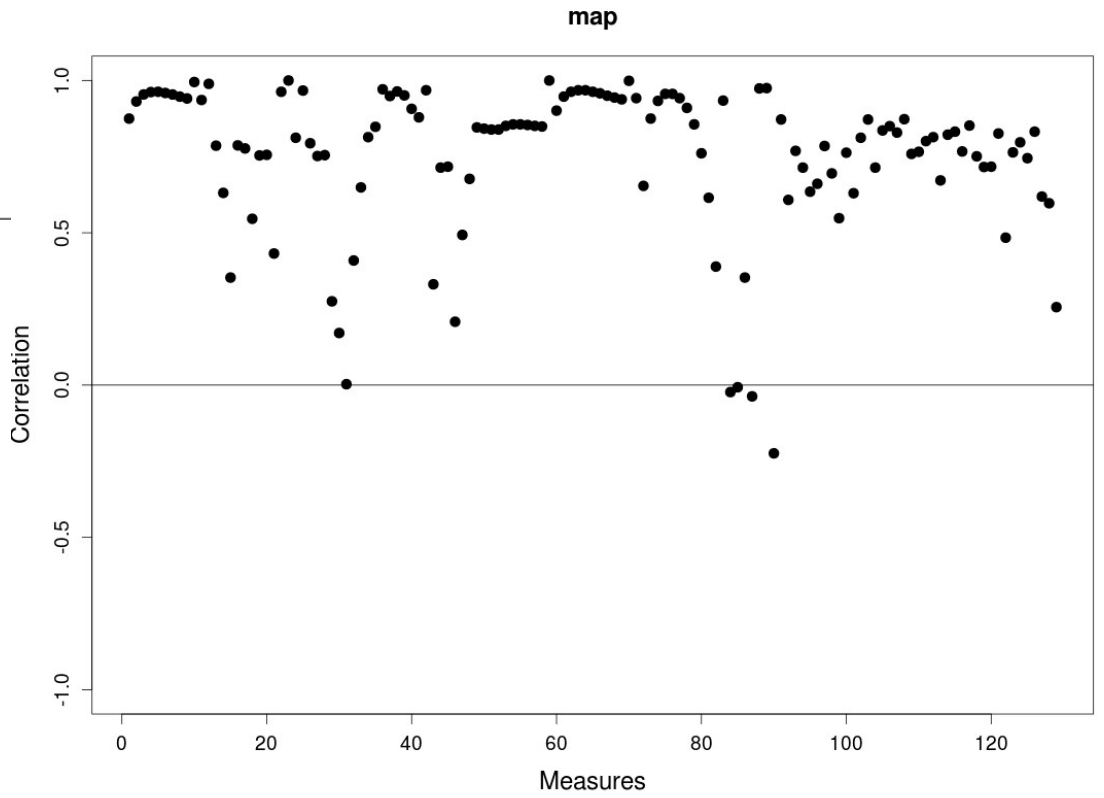
Matrice de corrélation



Corrélations (suite)



Représentation des corrélations entre 2 indicateurs particuliers (P5 et MAP) et l'ensemble des autres indicateurs.



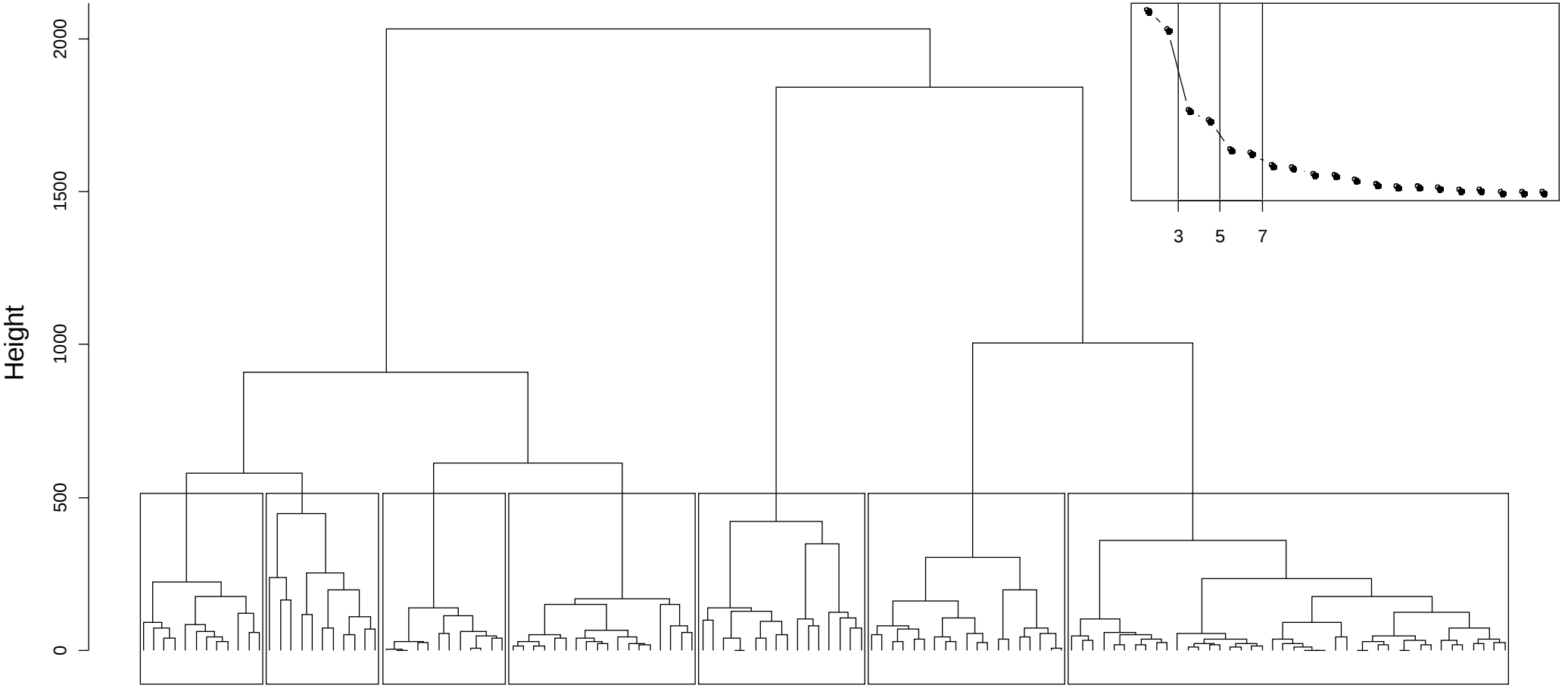
Analyse multivariée

- **Objectifs** :
  - Identifier des groupes de mesures homogènes
  - Définir un représentant par groupe
- **Méthodes** :
  - Classification hiérarchique + k-means
  - Heuristique pour le choix de représentant
- **Validation** :
  - Comparaison de classements élaborés soit avec l'ensemble complet des mesures, soit avec un ensemble restreint

Analyse multivariée

- **Objectifs** :
  - Identifier des groupes de mesures homogènes
  - Définir un représentant par groupe
- **Méthodes** :
  - Classification hiérarchique + k-means
  - Heuristique pour le choix de représentant
- **Validation** :
  - Comparaison de classements élaborés soit avec l'ensemble complet des mesures, soit avec un ensemble restreint

Classification des mesures de performance



Classification hiérarchique (distance euclidienne + critère de Ward) + k-means
→ 7 groupes homogènes (dont un contenant des mesures non informatives : nombre de documents retournés par exemple)

Interprétation des groupes

Cluster 1 (23 measures)

relative unranked avg prec30 - relative unranked avg prec20 - relative prec30 - map at R - relative unranked avg prec15 - relative prec20 - P30 - relative prec15 - int 0.20R.prec - relative unranked avg prec10 - X0.20R.prec - ircl prn.0.10 - P20 - P15 - relative prec10 - bpref 10 - P10 - relative unranked avg prec5 - relative prec5 - P5 - bpref 5 - recip rank – ircl prn.0.00

Précision avec un faible nombre de documents retournés

Cluster 2 (16 measures)

P100 - P200 - unranked avg prec500 - unranked avg prec1000 - bpref num ret – P500 - bpref num all - P1000 - num rel ret - exact unranked avg prec - num rel - exact prec - bpref num correct - utility 1.0 .1.0 0.0 0.0 - exact relative unranked avg prec - bpref num possible

Précision avec un nombre élevé de documents retournés

Cluster 3 (12 measures)

bpref top10Rnonrel - bpref retnonrel - relative unranked avg prec500 - avg relative prec - recall500 - relative prec500 - bpref allnonrel - relative unranked avg prec1000 - exact recall - recall1000 - relative prec1000 - exact relative prec

Rappel avec un nombre élevé de documents retournés

Cluster 4 (45 measures)

X1.20R.prec - ircl prn.0.30 - X1.40R.prec - int map - X1.00R.prec - R.prec - int 1.20R.prec - exact int R rcl prec - int 1.00R.prec infAP - avg doc prec - map - X11.pt avg - X1.60R.prec - int 0.80R.prec - int 1.40R.prec - X0.80R.prec - old bpref top10pRnonrel - ircl prn.0.40 - X1.80R.prec - int 1.60R.prec - X0.pt avg bpref - X2.00R.prec - bpref top25p2Rnonrel - old bpref – bpref top10pRnonrel - int 1.80R.prec - int 0.60R.prec - int 2.00R.prec - bpref top25pRnonrel - X0.60R.prec - bpref top50pRnonrel - bpref top5Rnonrel - ircl prn.0.20 - ircl prn.0.50 - int 0.40R.prec - X0.40R.prec - int map at R - ircl prn.0.60 - unranked avg prec30 - ircl prn.0.70 - ircl prn.0.80 - unranked avg prec200 - unranked avg prec100

Mesures « moyennes »

Cluster 5 (18 measures)

bpref topnonrel - fallout recall 42 - fallout recall 28 - fallout recall 56 - rcl at 142 nonrel - fallout recall 71 - fallout recall 85 - relative unranked avg prec100 - fallout recall 99 - fallout recall 113 - relative prec100 - fallout recall 127 - relative unranked avg prec200 - fallout recall 142 - recall100 - relative prec200 - recall200 - bpref retall

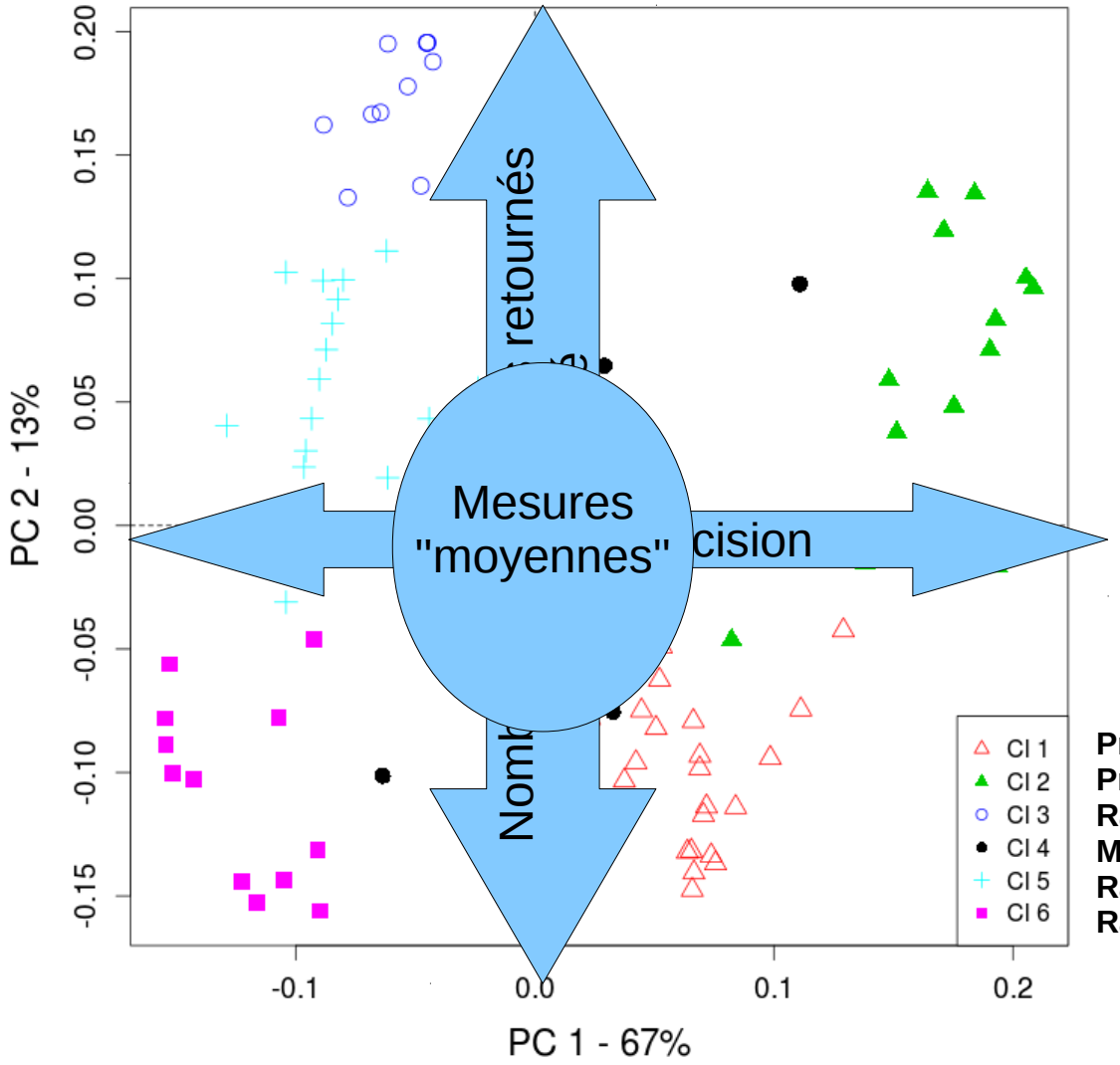
Rappel (???)

Cluster 6 (13 measures)

fallout recall 14 - unranked avg prec20 - unranked avg prec15 - ircl prn.0.90 - fallout recall 0 - unranked avg prec10 - recall30 - ircl prn.1.00 - recall20 - recall15 - unranked avg prec5 - recall10 - recall5

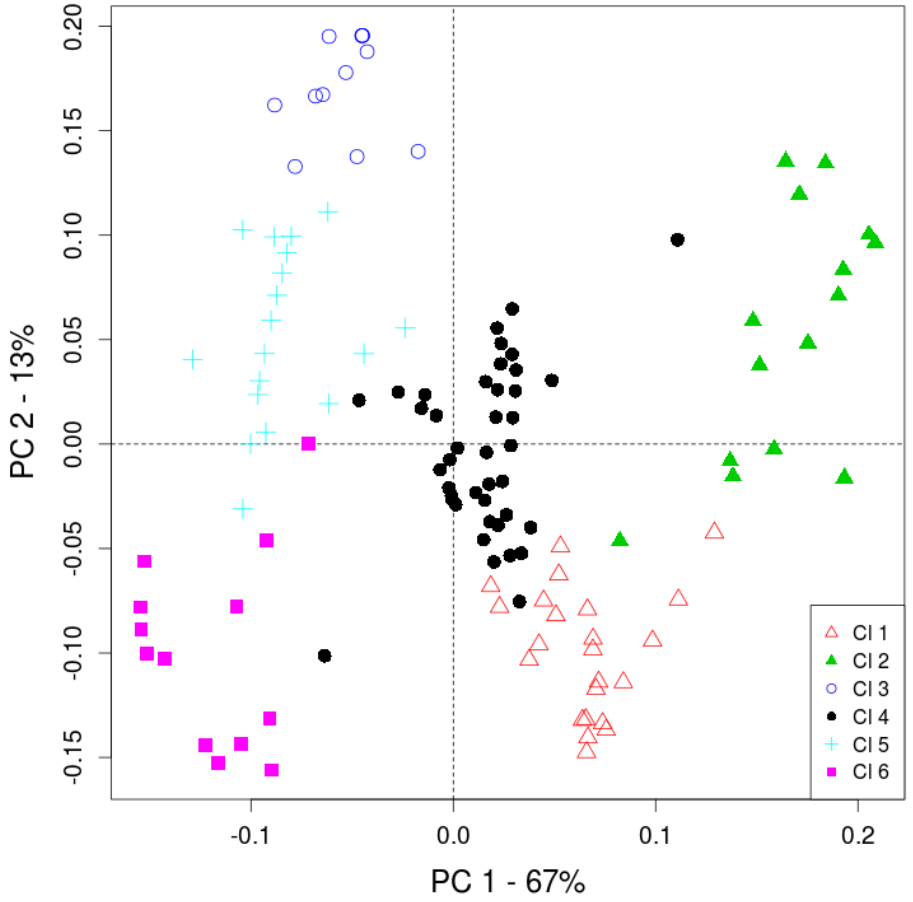
Rappel avec un faible nombre de documents retournés

Analyse en Composantes Principales



Précision avec un faible nombre de documents
Précision avec un nombre élevé de documents
Rappel avec un nombre élevé de documents
Mesures « moyennes »
Rappel (???)
Rappel avec un faible nombre de documents

Choix d'un représentant



Analyse en Composantes Principales – Représentation des mesures

- Les groupes apparaissent homogènes et relativement bien distincts les uns des autres

- Proposition : le représentant d'un groupe doit occuper une position « centrale » dans son groupe

- Heuristique :
  1) Calculer le barycentre de chaque groupe
  2) Calculer la distance de chaque mesure du groupe au barycentre
  3) Prendre comme représentant la mesure la plus proche du barycentre

- Liste proposée :

`P30 – P100 - Exact recall – MAP - bpref retall - recall30`

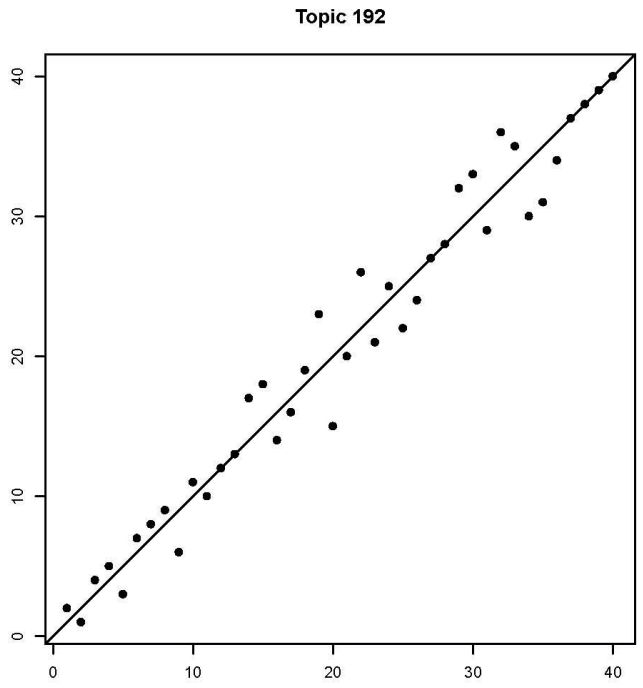
Vers un classement des systèmes

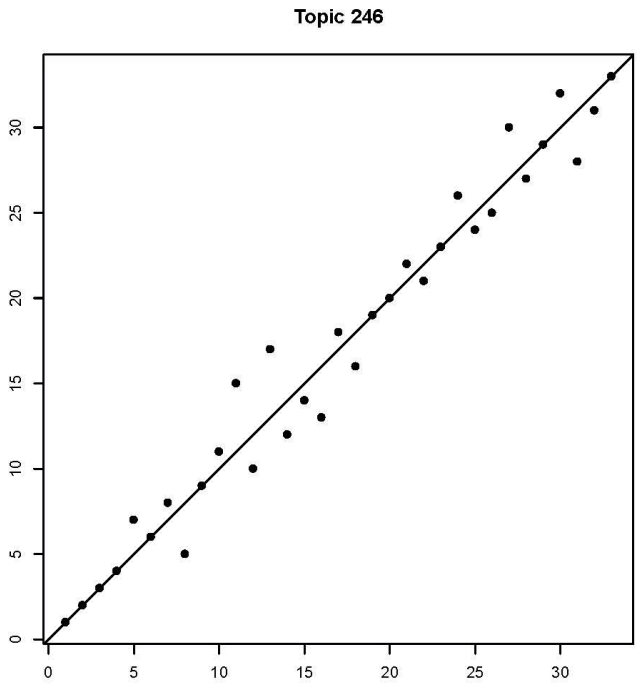
$$\begin{pmatrix} X_{11} & \dots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix} \qquad \underbrace{\begin{matrix} \overline{X}_1 \\ \vdots \\ \overline{X}_n \end{matrix}}_{\text{score moyen}} \qquad \underbrace{\begin{matrix} R_1 \\ \vdots \\ R_n \end{matrix}}_{\text{rang basé sur le score moyen}}$$

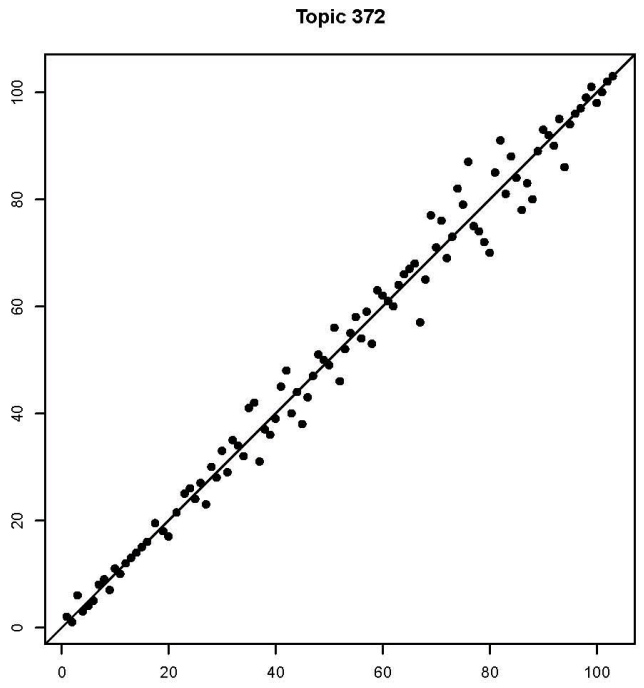
X_{ij} : score obtenu par le système i pour l'indice de performance j.
i = 1 à n, nombre de systèmes
j = 1 à 135 (version complète)
j = 1 à 6 (version réduite)

Classement de systèmes

Illustration : pour 3 requêtes particulières (#192 – TREC3, #246 – TREC4, #372 – TREC7), chaque point a pour coordonnées son rang selon le classement obtenu à partir de l'ensemble complet (horizontal) ou de l'ensemble réduit de mesures (vertical).







→ Le classement des systèmes reste globalement le même.

Conclusion Partie 1

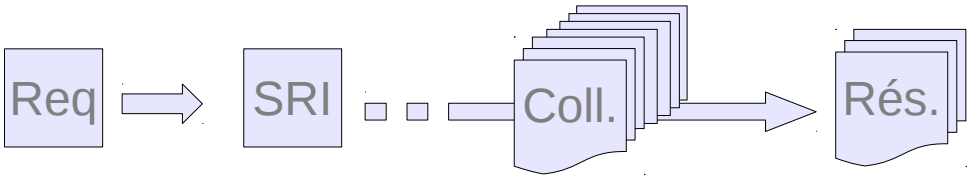
Typologie des mesures de performances : 6 grandes catégories de mesures

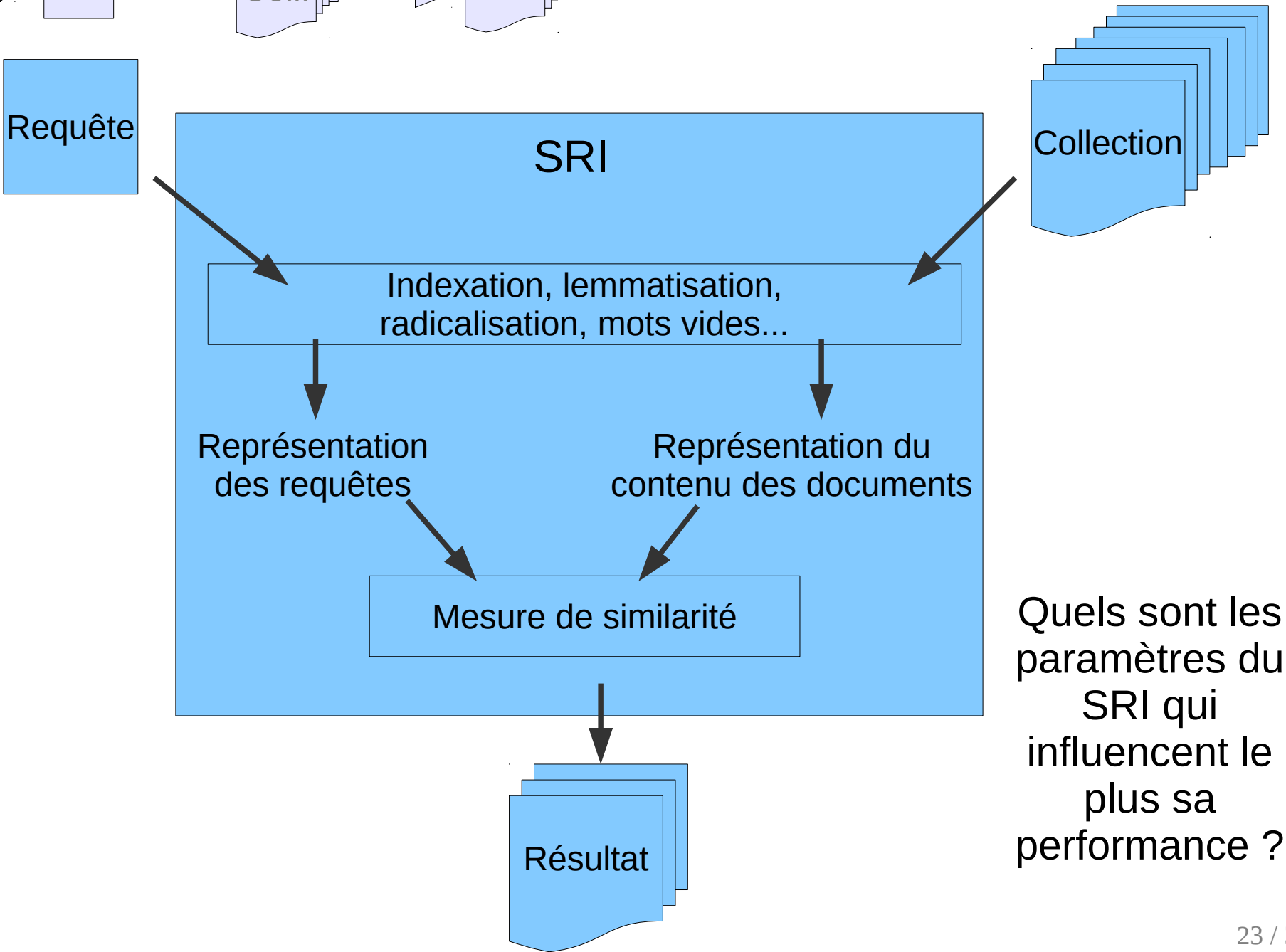
- Caractériser un SRI par l'évaluation de 6 mesures (une par groupe)
- Faciliter la comparaison de systèmes
- Cibler les mesures pertinentes (restreintes à 1 ou 2 groupes) à évaluer en vue d'une optimisation particulière d'un SRI
- Positionnement d'une nouvelle mesure vis-à-vis de cette typologie

Partie II
Paramètres d'un SRI

J. Compaoré, S. Déjean, A.M. Gueye, J. Mothe, J. Randriamparany. **Mining Information Retrieval Results: Significant IR parameters**. Dans / In : *Advances in Information Mining and Management*, Barcelone, 23/10/2011-28/10/2011, IARIA

Paramètres d'un SRI





Quels sont les paramètres du SRI qui influencent le plus sa performance ?

La plateforme Terrier

Terrier is a highly flexible, efficient, and effective open source search engine, readily deployable on large-scale collections of documents.

http://terrier.org/

Terrier implements state-of-the-art indexing and retrieval functionalities, and provides an ideal platform for the rapid development and evaluation of large-scale retrieval applications.

Terrier is open source, and is a comprehensive, flexible and transparent platform for research and experimentation in text retrieval. Research can easily be carried out on standard TREC and CLEF test collections.

Quelques paramètres

Parameters	**Meaning**	**Values**
Top	Numéro de la requête (*topic*)	`351, …, 400`
Field	Champ de la requête	`T, T+D, T+D+N`
Bloc	Taille du bloc d'indexation	`0, 1, 5, 10`
Idf	Fréquence inverse de document	`FALSE, TRUE`
Model	Modèle de recherche (*retrieval model*)	`BB2c1, BM25b0.5, DFRBM25c1.0, IFB2c1.0, InexpB2c1.0, InexpC2c1.0, InL2c1.0, PL2c1.0, TFIDF`
Ref	Reformulation de requête	`None, Bo1bfree, Bo2bfree, KLbfree`
DocNb	Nombre de documents pour la reformulation	`0, 3, 5, 10, 50, 100, 200`

Jeu de données

• 98 650 lignes : 1 ligne = 1 *run*, une requête (*topic*) traitée en utilisant une configuration de 6 paramètres
• 8 colonnes : 1 requête + 6 paramètres + 1 mesure de performance (map)

#	Top	Field	Bloc	Idf	Model	Ref	DocNb	map
1	351	T	1	false	BB2c1.0	Bo1bfree	3	0.6134
2	352	T	1	false	BB2c1.0	Bo1bfree	3	0.3412
3	353	T	1	false	BB2c1.0	Bo1bfree	3	0.3479
4	354	T	1	false	BB2c1.0	Bo1bfree	3	0.0662
5	355	T	1	false	BB2c1.0	Bo1bfree	3	0.2794
6	356	T	1	false	BB2c1.0	Bo1bfree	3	0.0460
…								
98645	445	T	0	true	TFIDF	NONE	1	0.1514
98646	446	T	0	true	TFIDF	NONE	1	0.2234
98647	447	T	0	true	TFIDF	NONE	1	0.1121
98648	448	T	0	true	TFIDF	NONE	1	0.0114
98649	449	T	0	true	TFIDF	NONE	1	0.0714
98650	450	T	0	true	TFIDF	NONE	1	0.3226

Requêtes plus ou moins faciles

• La plupart du temps, les résultats sont moyennées sur un ensemble de requêtes alors que les résultats d'un système sont dépendants de la requête.

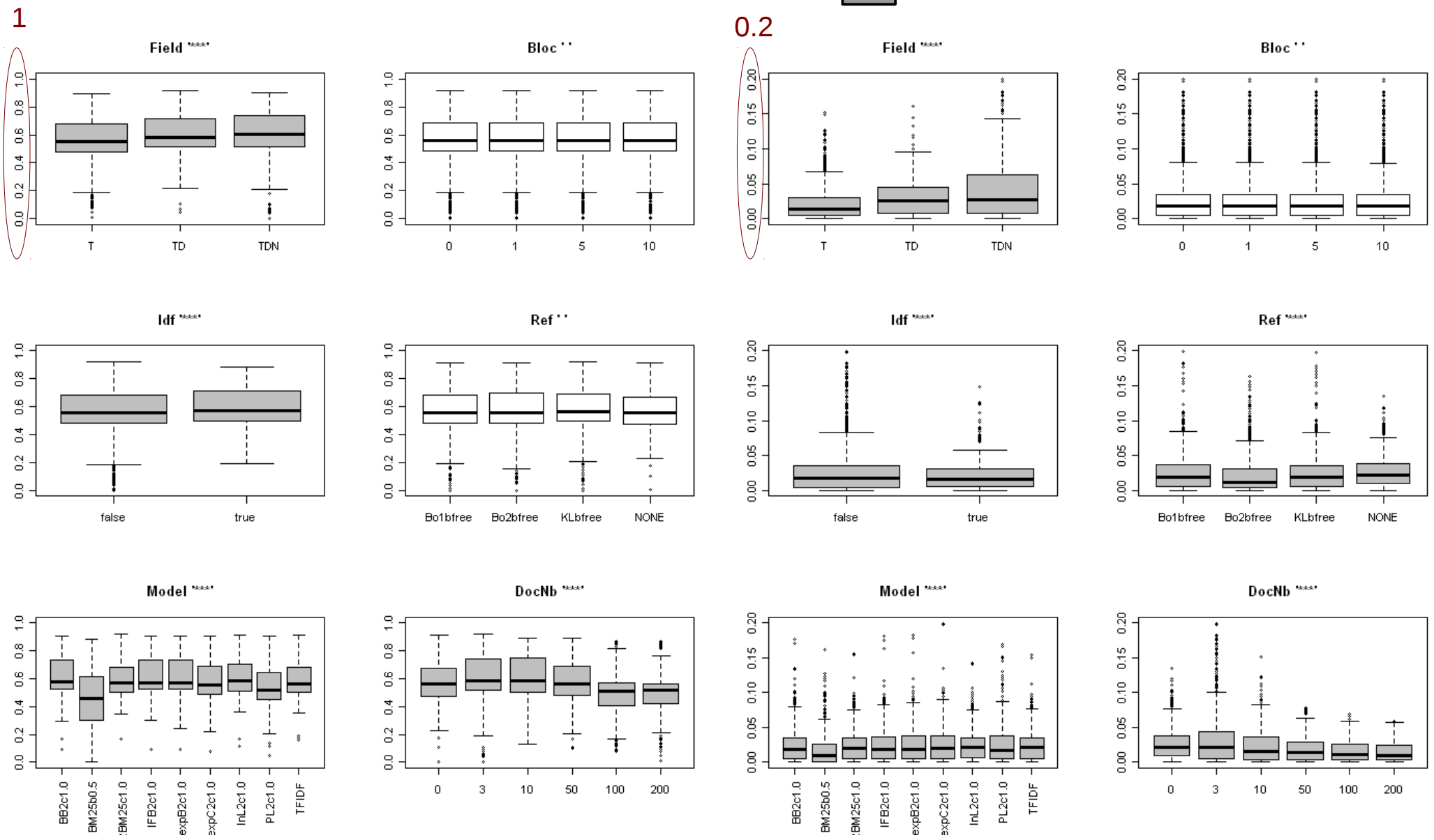
• Étude de l'impact des paramètres en fonction de la difficulté de la requête :

- Requêtes « faciles »: requêtes pour lesquelles, l'AP moyenne sur les systèmes est la plus élevée ; dans notre cas, **13** requêtes avec une AP moyenne > 0.45)

- Requêtes « difficiles »: requêtes pour lesquelles, l'AP moyenne sur les systèmes est la plus faible ; dans notre cas, **19** requêtes avec une AP moyenne < 0.045)

Analyses univariées

Effect significatif (ANOVA 1 facteur)



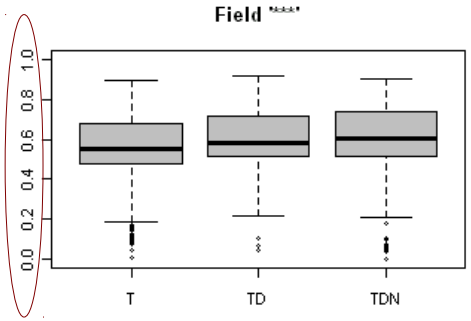
Requêtes faciles

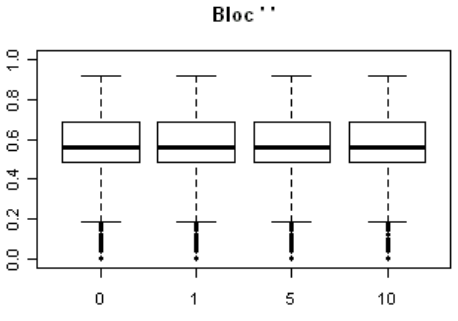
Requêtes difficiles

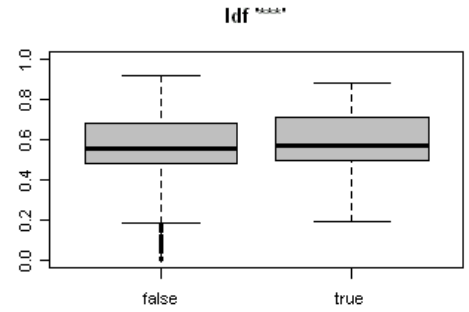
Analyses univariées

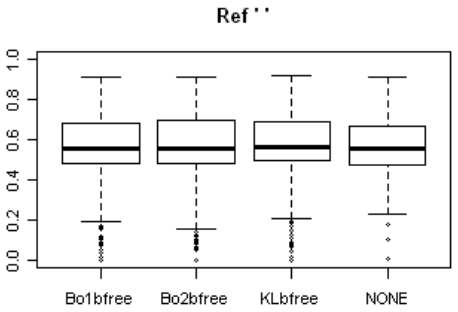
Effect significatif (ANOVA 1 facteur)

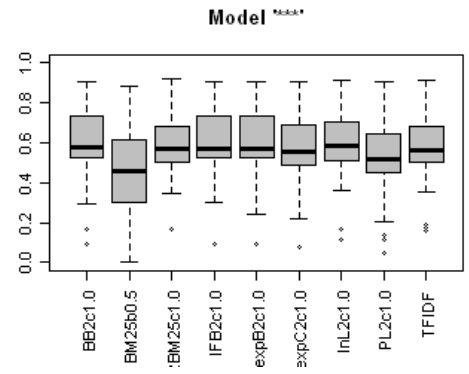
1

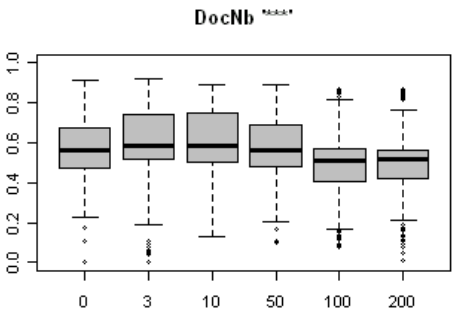






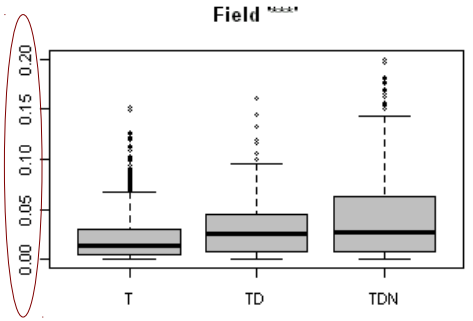


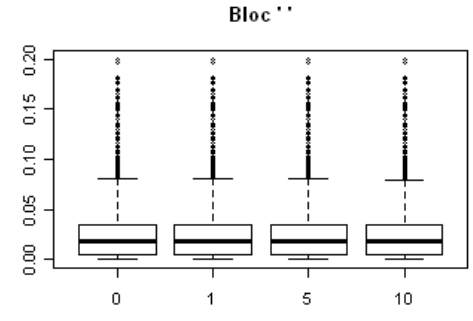


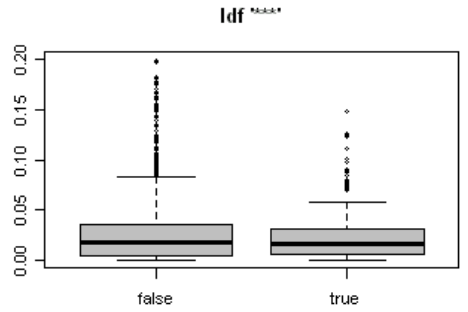


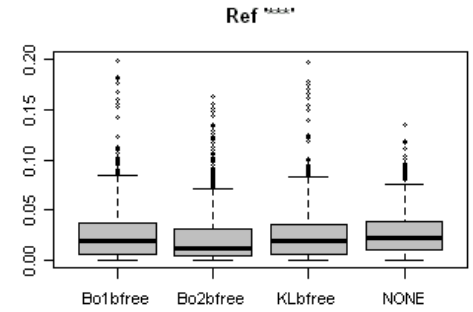
Requêtes faciles

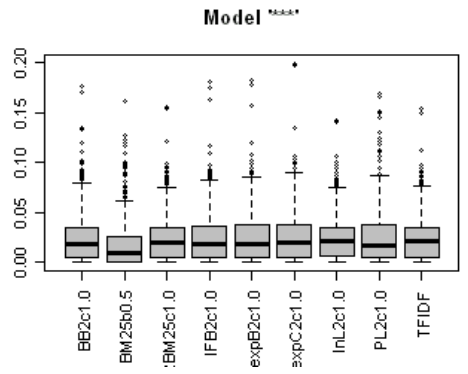
0.2

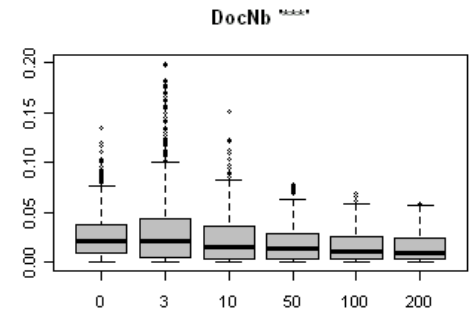












Requêtes difficiles

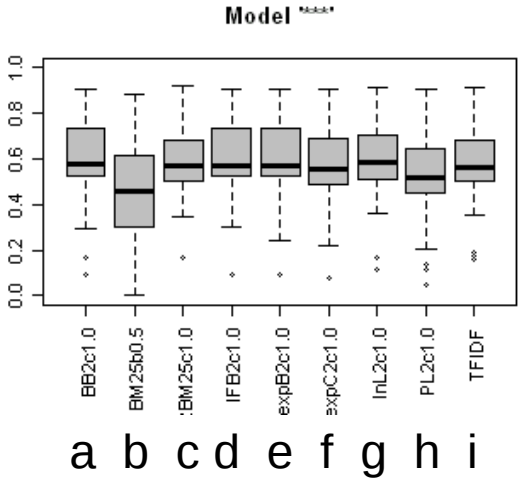
Analyse multivariée

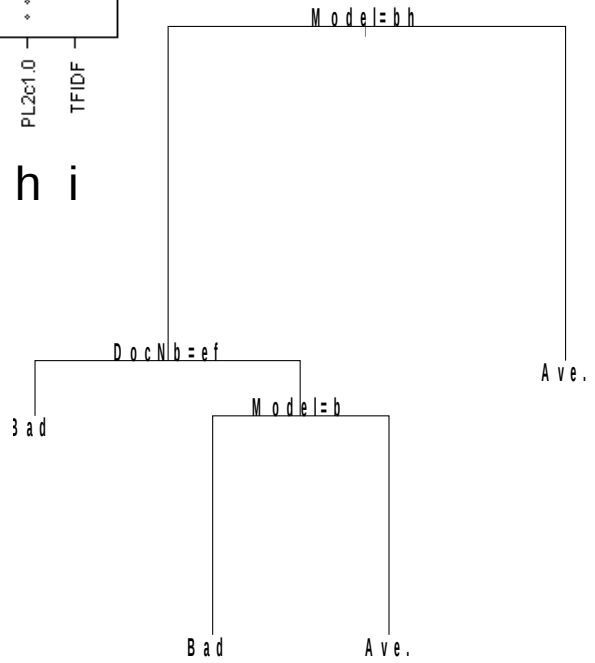
• Arbre de décision (***C****lassification* ***A****nd* ***R****egression* ***T****rees, CART*)

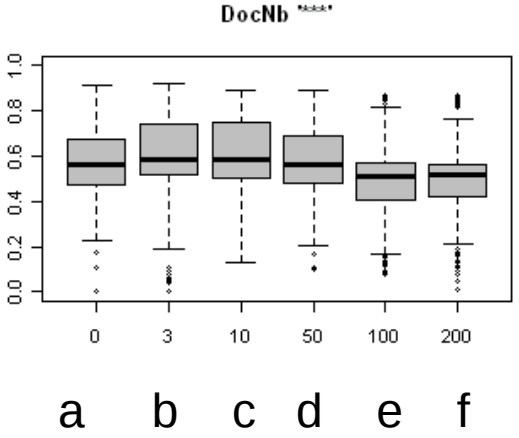
• Construction d'un arbre de décision par découpages successifs de variables explicatives selon les valeurs d'une variable à expliquer (quantitative pour la régression ou qualitative pour la classification).

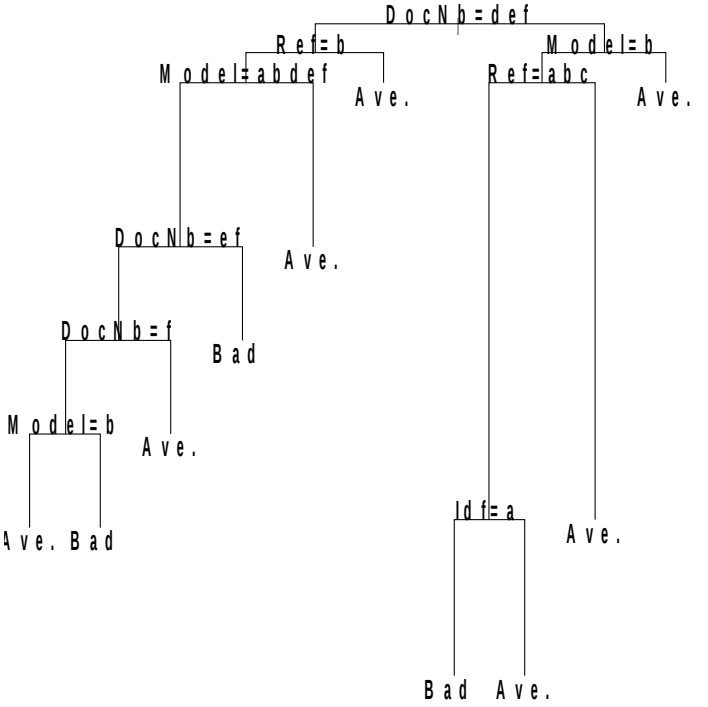
• Nous avons opté pour le cadre de la classification. Les valeurs de MAP ont été converties en classes :
  • *Bad*: MAP inférieure au premier quartile
  • *Average*: MAP comprise entre le premier et le troisième quartile
  • *Good*: MAP supérieure au troisième quartile

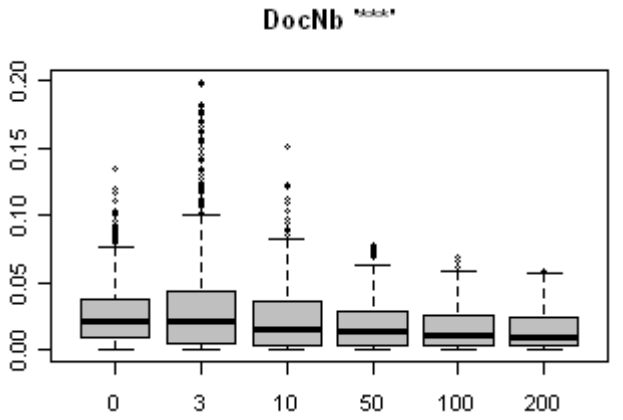
Résultats de CART







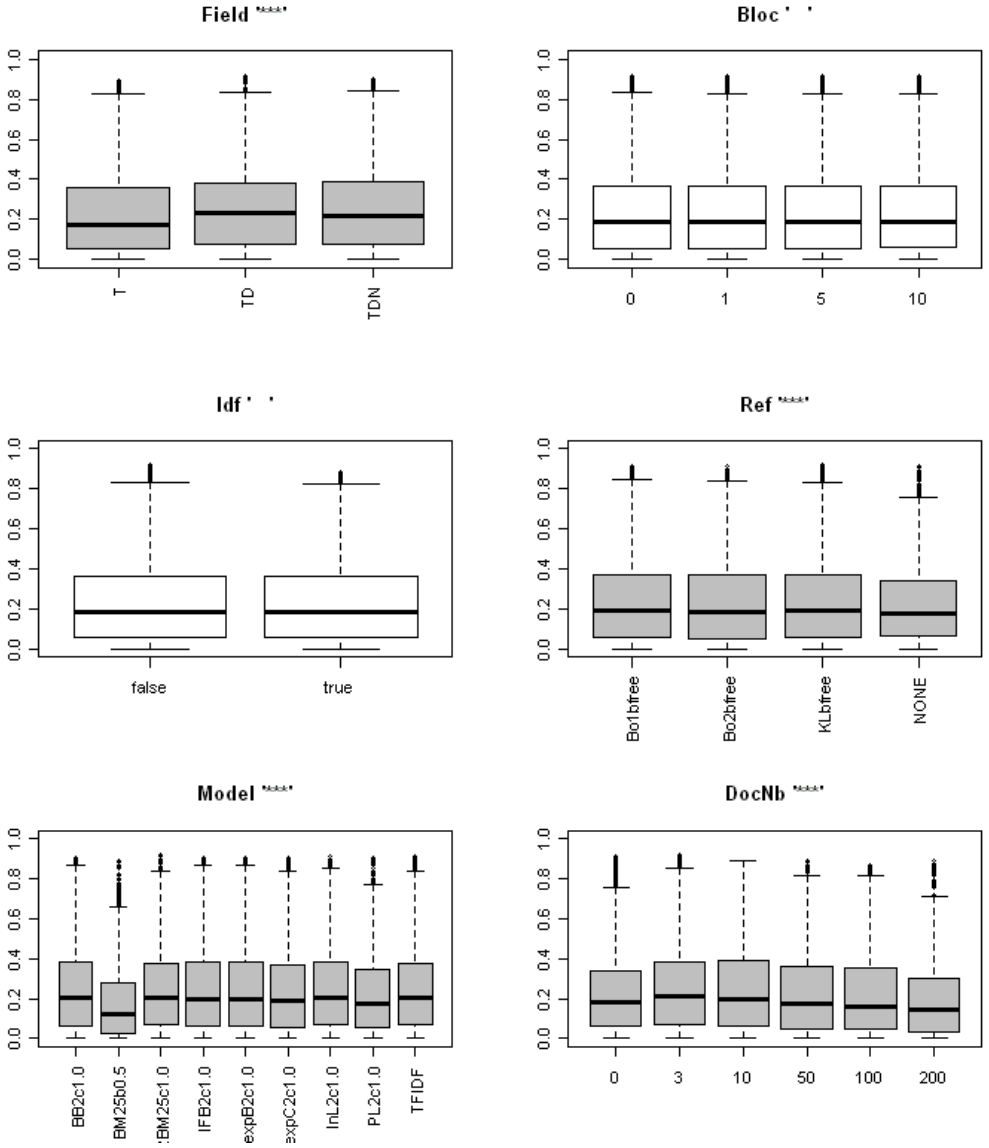


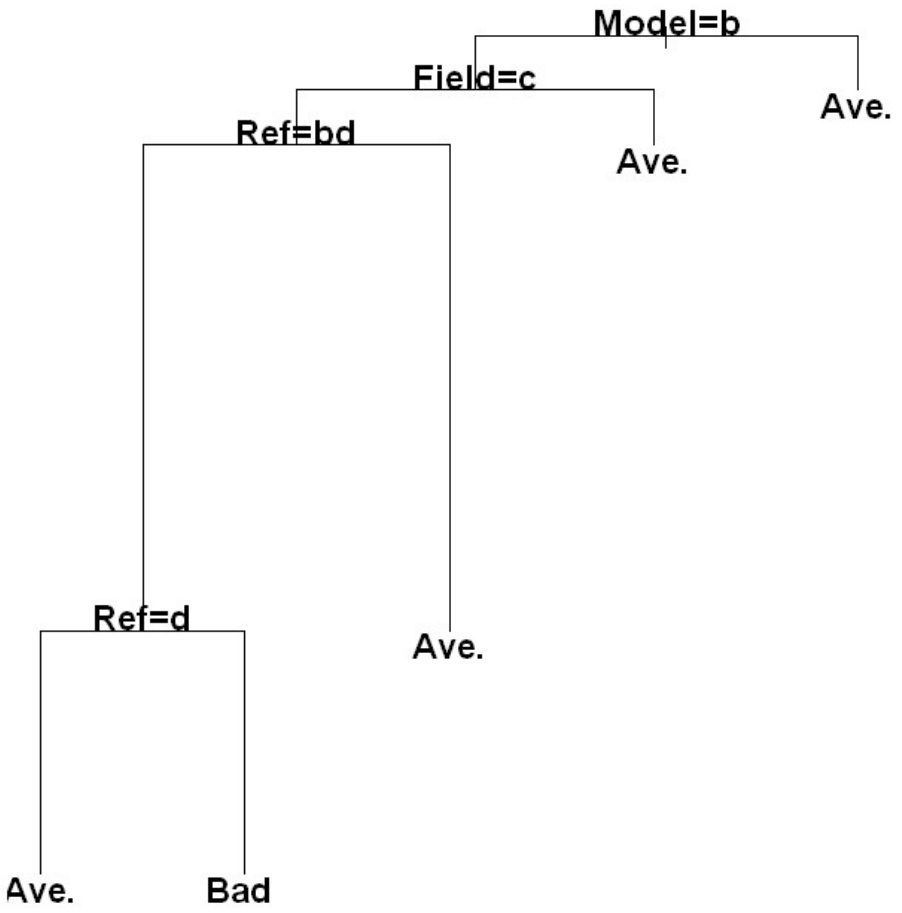


Requêtes faciles

Requêtes difficiles

Sur l'ensemble des requêtes





Conclusion Partie II

• Certains paramètres produisent des changements significatifs dans les performances d'un SRI.

• Ces changements dépendent de la difficulté des requêtes.

• Ces travaux fournissent des pistes concernant le réglage des paramètres susceptibles d'améliorer la performance d'un SRI.

• Résultats à étendre à d'autres mesures de performance.

• Besoin d'un procédure plus systématique et *non auto-référente* pour définir les groupes de requêtes.

Suites

Besoin d'un procédure plus systématique et non auto-référente pour définir les groupes de requête

A. CHIFU, L. LAPORTE, J. MOTHE
La prédiction efficace de la difficulté des requêtes : une tâche impossible ?
CORIA 2015 • Paris • 18-20 Mars

Prédicteurs pré-recherche

• L’ambiguïté des termes. Nombre de sens deWordNet (WNS) : prédicteur linguistique de pré-recherche, mesure de l’ambiguïté

• La discrimination des termes. Fréquence Inverse (idf) : prédicteur statistique de pré-recherche, mesure la rareté/popularité d’un terme

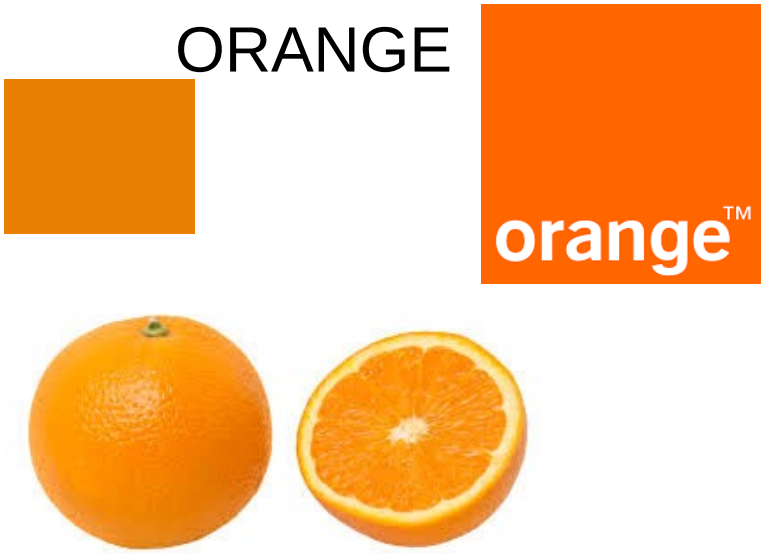
Prédicteurs post-recherche (?!)

• L’homogénéité des listes de documents. Ecart-type (STD) : prédicteur post-recherche statistique,mesure le degré de variation de la liste des scores

• La divergence des listes. Retour sur la requête (QF) : un prédicteur post-recherche, calcule le chevauchement entre deux listes de documents retrouvés

Ambiguïté des termes

ORANGE



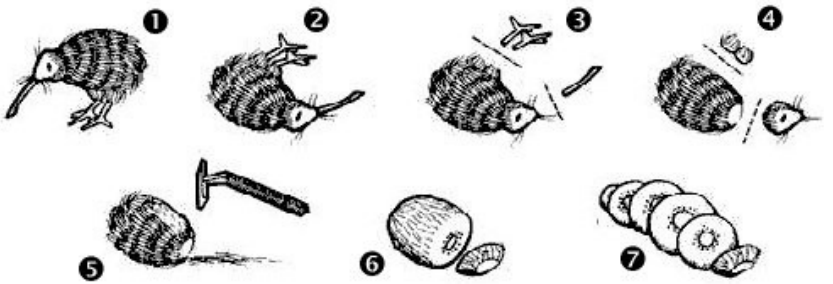
KIWI

TERRIER

Terrier

http://terrier.org/

How to prepare a kiwi



Analyses statistiques pour l'évaluation des systèmes de recherche d'information

S. Déjean, J. Mothe

www.math.univ-toulouse.fr/~sdejean

INSTITUT de MATHEMATIQUES de TOULOUSE

Institut de Recherche en Informatique de Toulouse