ANNEXE III.1

Quelques notes à propos de TREC

Table des matières de l'Annexe III.1

1. Cadre

TREC (*Text Retrieval Conference*) est une rencontre scientifique de grande envergure autour d'un large banc d'essai des systèmes de recherche documentaire automatisée. Elle se tient chaque année (depuis novembre 1992) au *National Institute of Standards and Technology* (NIST). Trois caractéristiques concourent à son attrait :

1) les acteurs majeurs du domaine sont représentés, aussi bien les laboratoires de recherche que l'industrie. C'est une occasion propice aux échanges entre les deux communautés, tant sur le plan des méthodes que sur les problèmes « réels » (rencontrés dans l'utilisation concrète des systèmes). Pour les produits commercialisés, la Conférence donne la possibilité de ne pas publier les algorithmes.

	TREC-1	TREC-2	TREC-3	TREC-4
Date	novembre 1992	août 1993	1994	novembre 1995
Nombre d'équipes participant	22	31	33	36

2) le protocole d'expérimentation des systèmes est centré sur la tâche à accomplir (recherche, ciblage), de telle sorte que puissent être confrontées des méthodes très différentes. Comme on se donne des modes d'évaluation des systèmes, divers types de résultats peuvent être visés. En se situant sur le plan d'une compétition des équipes et des produits, c'est une comparaison pour reconnaître les meilleures approches et les meilleures réalisations (évaluation entre les systèmes) ; en revanche, sur un plan plus expérimental et méthodologique, c'est une jauge, une mesure à la disposition des équipes pour apprécier l'effet d'un paramètre de leur système, pour estimer la qualité des résultats en se plaçant dans tel ou tel cas de figure (évaluation interne à un système).
3) l'expérience est prévue avec un grand volume de données (plusieurs Giga-octets de textes) : il s'agit donc de résoudre les problèmes d'échelle, susceptibles de jouer sur les structures de données, les procédures de caractérisation des textes (indexation), le stockage (compression).

2. Description de la tâche

Le principe général est que l'on dispose d'une collection de requêtes (ou plus exactement d'expressions de besoins d'information, sans préjuger de la forme que peut prendre la requête effective devant sélectionner les documents), d'une collection de documents, et d'un ensemble complet de valeurs de pertinence : toute association requête-document a été jugée soit satisfaisante, soit invalide (selon l'appréciation d'un arbitre).

Sur cette base, deux cas de figure sont envisagés :

1) recherche *adhoc* (sic) : de nouvelles requêtes sont soumises, le fonds documentaire restant stable. La situation est à rapprocher du fonctionnement d'une bibliothèque.
2) *routing* : un ensemble de requêtes sont fixées, un flux de documents leur est confronté. La situation serait celle de profils qui filtrent ou sélectionnent des documents au fur et à mesure qu'ils se présentent.

L'évaluation porte sur les documents sélectionnés par chaque requête, classés par ordre de pertinence. L'ergonomie de l'interface ou les temps de réponse n'entrent pas en ligne de compte : l'optique est plutôt celle d'un traitement *batch*.

Le volume de données et la disponibilité d'une base d'apprentissage se prêtent à des approches automatiques ou semi-automatiques. Le cas échéant, les interventions manuelles humaines portent sur la formulation et la structuration des requêtes (écriture d'une équation de recherche booléenne, introduction d'opérateurs de proximité), et/ou sur la modification et l'affinement de requêtes construites automatiquement ou non (ajout ou élimination de termes, ajustement de pondérations).

3. Résultats

Trois conclusions générales ressortent de l'expérience :

- il n'y a pas une méthode supérieure à toutes les autres. Pour chaque tâche (*adhoc* et *routing*) les systèmes se révélant les meilleurs reflètent une grande variété d'approches mais ne montrent aucune différence significative pour le résultat global. Leurs niveaux de qualité (fonction de la précision et du rappel) sont analogues et les systèmes retrouvent en bonne partie les mêmes documents en réponse aux mêmes requêtes.
- en revanche, le niveau de qualité globale recouvre des résultats très inégaux : (i) d'une requête à l'autre pour un même système (requêtes « difficiles »), (ii) d'un système à l'autre pour une même requête, (iii) d'une famille de documents à une autre (en particulier, les documents longs apparaissent pénalisés par les méthodes[1]).
- enfin, dans le cadre du protocole de TREC, les approches automatiques réussissent généralement mieux (ou aussi bien) que les approches manuelles. Les atouts de l'automatisation sont notamment de tirer grand parti de la base d'apprentissage (particulièrement pour le routing), et de s'adapter parfaitement au contexte des données (le vocabulaire présent dans les textes et sa répartition dans la base). Les modifications manuelles, elles, peuvent quelquefois apporter un gain de précision notable, par l'élimination précise de termes facteurs de bruit.

4. Analyse des conditions de l'expérience

a) Les documents

Ils sont représentés par du texte intégral (pas de mots-clé : on n'a à sa disposition que des documents primaires). Il semble que certains soient (faiblement) structurés : présence d'un titre, indication des paragraphes.

Les documents sont sous forme électronique propre : ils n'ont pas été obtenus par reconversion automatique d'un fonds sous forme papier, du moins ils ne présentent pas les erreurs liées à une reconnaissance optique de caractères.

Le corpus a été rassemblé avec un souci de représentativité de la variété des documents rencontrés dans la réalité. Quatre dimensions de variation sont affichées, mais certaines appellent des commentaires :

1) *longueur* : la très grande majorité des documents (plus de 99 % d'entre eux) comptent de l'ordre de 300 mots ou moins : c'est relativement court. Les quelques documents plus longs sont des brevets d'environ 3 000 mots.
2) *genre* : une petite dizaine de sources sont distinguées ; mais une bonne moitié d'entre elles fournissent des articles de presse. Les autres genres concernés sont des résumés courts de publications, et (marginalement) une collection de documents légaux, des brevets, ainsi que (dernièrement) des documents d'informatique mis sur Internet.
3) *domaine*
4) *date* : le plus anciens datent de 1987. Si variation il y a, il faudrait qu'elle soit sensible à l'échelle de la décennie.

b) Les requêtes

On distingue bien d'une part l'expression des besoins d'un utilisateur, et d'autre part les données qui seront finalement utilisées par un système pour déterminer, par un calcul, les documents concernés. Cette distinction est capitale ici, car elle conditionne la possibilité de tester les méthodes les plus variées.

Thème de recherche –[traduction]→ Requête (dans un format qui se prête au calcul) –[apprentissage si routing]→ Requête affinée

On continuera cependant ici à parler de « requête » pour désigner les thèmes de recherche, pour ne pas alourdir l'exposé.

[1] Nous aurons l'occasion de revenir à cette question des documents longs, *i.e.* de plus de quelques pages.

Les requêtes pour la recherche *adhoc* sont reprises l'année suivante pour le *routing*. L'opposition entre un besoin ponctuel et une veille documentaire à plus long terme (avec une formulation potentiellement plus fouillée) n'apparaît pas au niveau du texte des requêtes initiales.

Au fil des sessions, la forme des requêtes a notablement évolué. La première forme était très structurée. Elle comportait quatre champs : un titre nommant le thème, une description énonçant complètement l'objet de la recherche, un développement explicitant des critères de validité des rapprochements, des mots-clés fournissant le contexte terminologique et les concepts concernés. Cette forme apportait donc une information aussi complète et détaillée que possible, y compris des connaissances avancées sur le domaine grâce aux mots-clés. Cette forme s'est allégée progressivement, dans le but de se placer dans une position plus réaliste. C'est d'abord les mots-clés qui disparaissent, car ils explicitent artificiellement les connaissances du domaine. Puis ne sont conservés qu'un titre et une description très brève (d'une phrase ou deux), plus représentatifs des demandes réellement soumises aux systèmes documentaires en service courant.

La création des requêtes est plus ou moins calibrée et ajustée aux documents :

- les critères de validité que comprenaient les premières requêtes ont été inspirés par un premier balayage des documents de la base
- les requêtes qui apparaissent trop générales sont exclues
- les requêtes perçues comme ambiguës sont également rejetées
- les profils sélectionnés pour TREC-4 sont choisis pour leur propension à retrouver des documents d'une certaine collection (des documents longs en l'occurrence).

c) Pertinence

La pertinence d'un document pour une requête est codée par une valeur binaire : soit le document est pertinent pour la requête, soit il ne l'est pas. Pour une requête et un ensemble de documents proposés en résultat, on est donc en mesure de calculer les indicateurs classiques :

recall : nombre de documents pertinents proposés en réponse,
rapporté au nombre total de documents pertinents dans la collection où s'effectue la recherche ;

precision : nombre de documents pertinents proposés en réponse,
rapporté au nombre total de documents proposés en réponse ;

fallout : nombre de documents non pertinents proposés en réponse,
rapporté au nombre total de documents non pertinents dans la collection où s'effectue la recherche.

Chaque système indique pour chaque requête 1 000 documents, classés par ordre de pertinence. On examine alors comment varient les indicateurs ci-dessus quand considère les 1, 2,..., n,..., 1 000 premiers documents. La qualité des résultats d'un système est alors représentée par la courbe du *recall* en fonction du *fallout* (point de vue système), ou la courbe du *recall* en fonction de la *precision* (point de vue utilisateur).

L'ensemble total des documents étant trop grand pour pouvoir évaluer (à l'avance) la pertinence de chacun vis-à-vis de chaque requête, on fait l'hypothèse que tous les documents pertinents ont été retrouvés, et que chacun a été classé dans les 100 premiers par au moins un système. Le dépouillement manuel et l'affectation des valeurs de pertinence se fait donc par une personne (une seule pour une bonne consistance des jugements) sur l'union des ensembles des 100 premiers documents retournés par chaque système. Deux observations viennent cautionner cette approche : d'une part, le nombre de documents pertinents (environ 200) est très inférieur au nombre de documents examinés (environ 1 200) : d'autre part, le fait de dépouiller davantage de documents par système apporte relativement peu de documents pertinents supplémentaires.

Le but consiste donc à maximiser la satisfaction de l'utilisateur, ceci étant équivalent à obtenir la meilleure adéquation possible entre les appariements requête-document calculés et les appariements requête-document de référence. Autrement dit, on dispose d'une fonction injective « bonne réponse » qui à tout couple requête-document associe soit la valeur « pertinent », soit la valeur « non pertinent » (la pertinence est donc anonyme, atemporelle, absolue, univoque...), et tout le problème se ramène à trouver un modèle, une expression mathématique, de cette fonction. L'évaluation pratiquée se situe de fait au niveau des résultats (empirisme : on produit le comportement voulu) et non au niveau de l'analyse des mécanismes en jeu dans la recherche documentaire (linguistique, cognition) : les

formules de calcul ne sont pas construites pour transcrire des phénomènes ou des processus ; elles sont testées, ajustées, et les décisions sont expliquées et arbitrées par le principe d'empirisme (« ça marche »).

5. Variantes des tâches

	Variation sur les données	Variation sur la nature des résultats
Recherche adhoc	- données bruitées issues d'OCR (étude de la robustesse) - autre langue (étude de la portabilité) - prise en compte des sous-collections (diversité des formats et des genres)	- recherche interactive (étude du comportement de l'utilisateur ; conception d'interfaces imagées de type espace des mots / espace des documents)
Routing		- filtrage, à savoir pas d'ordre de pertinence sur les documents, mais 3 valeurs de seuil.

6. Eléments qui peuvent être exploités dans chacune des tâches : dissymétrie

	Recherche adhoc	Routing
Référentiel (servant par exemple à repérer des mots discriminants)	L'ensemble des documents, formant la base de recherche.	Eventuellement les documents de la base d'apprentissage. Les nouveaux documents doivent être considérés indépendamment les uns des autres. Quant aux rapports entre les profils-requêtes, ils n'entrent pas dans la problématique posée par TREC.
Base de travail (stable)	Des centaines de milliers de documents de quelques centaines de mots.	Une cinquantaine de profils d'une centaine de mots ou moins.
Rôle de la formulation de la requête	La formulation de la requête joue un rôle majeur : en effet, la requête est la seule information disponible sur les documents qu'il faut sélectionner.	La formulation de la requête joue un rôle variable, éventuellement nul. En effet, l'ensemble des documents pertinents donnés en apprentissage indique aussi ce qui est recherché, et la requête peut être élaborée à partir de ces documents.
Relevance feedback	Le relevance-feedback peut intervenir *a posteriori*, en faisant un traitement en 2 passes et en utilisant un premier ensemble de documents retrouvés pour ajuster la requête.	Le relevance-feedback peut intervenir *a priori*, au moment de l'apprentissage (la requête est ajustée en fonction des documents pertinents de l'ensemble d'apprentissage). En revanche, le règlement de TREC interdit une retouche des requêtes en fonction des documents de test.

7. Données de TREC

TREC-2	Requêtes 1-50	Requêtes 51-100 (149 mots)	Requêtes 101-150 (178 mots)
Documents du disque 1	Apprentissage	Apprentissage	Adhoc
Documents du disque 2	Apprentissage	Apprentissage	Adhoc
Documents du disque 3		Routing	

TREC-3	Requêtes 1-50	Requêtes 51-100 (149 mots)	Requêtes 101-150 (178 mots)	Requêtes 151-200 (119 mots)
Documents du disque 1			Apprentissage	Adhoc
Documents du disque 2			Apprentissage	Adhoc
Documents du disque 3			Routing	

TREC-4	Requêtes 1-50	Requêtes 51-100 (149 mots)	Requêtes 101-150 (178 mots)	Requêtes 151-200 (119 mots)	Requêtes 201-250 (16 mots)
Documents du disque 1	Apprentissage	Apprentissage	Apprentissage	Apprentissage	
Documents du disque 2	Apprentissage	Apprentissage	Apprentissage	Apprentissage	Adhoc
¾ des doc. du disque 3	Apprentissage	Apprentissage	Apprentissage	Apprentissage	Adhoc
¼ doc. d.3 + ½ disque nouveaux	R.	R.	R.	R.	