Recherche d'information – analyse des résultats de différents systèmes réalisant la même tâche

Claude Chrisment, Taoufiq Dkaki**, Josiane Mothe *,***, Sandra Poulain **, Ludovic Tanguy *****

(*) ERT34, Institut Universitaire de Formation des Maîtres, 56 Avenue de l'URSS, 31400 Toulouse, France
(**) Institut de Recherche en Informatique de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 04, France
(***) ERSS
{mothe}@irit.fr tél: 05 61 55 64 44

RESUME: Les systèmes de recherché d'information visent à optimiser les résultats qu'ils fournissent en réponse à une requête de l'utilisateur. Les performances de ces systèmes sont généralement mesurées par rapport à des collections de test communes, comme les collections de TREC (Text REtrieval Conférence). Cette évaluation est réalisée de façon globale, en calculant une moyenne des résultats sur un ensemble de cinquante requêtes. Ainsi, la valeur ajoutée des différentes techniques utilisées par tel ou tel système n'apparaissent pas clairement. Cet article vise à étudier plus finement les résultats obtenus dans une tâche de recherche d'information et répond aux questions suivantes : les requêtes peuvent-elles être classifiées? Y a-t-il une corrélation entre les performances des systèmes sur les différentes requêtes ? La tâche choisie est celle de recherche de passages pertinents et/ou nouveaux. Nous montrons que les variations dans les résultats sont plus corrélées aux outils qu'à leurs différentes versions.

ABSTRACT : Information retrieval systems aim at answering users' needs. Information Retrieval System performances are evaluated using benchmark collections such as TREC (TExt Retrieval Conference) collections. Evaluation is generally based on global evaluation, computing average results over a set of fifty queries. Doing so, the added value of the different techniques used is not easy to show. This paper aims at studying in more details results obtained in a IR task that answers the following questions: can queries be classified? Is there correlation between system performances and queries? The IR task we chose is passage retrieval and novelty detection. We show that variations in results more related to systems than to system versions.

MOTS-CLES: Recherche d'information, recherche de passages, détection de la nouveauté, analyse de résultats, typologie de requêtes, analyse factorielle
KEY WORDS : Information retrieval, passage retrieval, novelty detection, analysis of IR results, typology of queries, factorial analysis

SERVICE EDITORIAL HERMES – LAVOISIER
14 RUE DE PROVIGNY – 94236 CACHAN
Tel : 01-47-40-67-00 Télécopie : 01-47-40-67-02
E-mail : revues@lavoisier.fr
Serveur Web : http//www.hermes-science.com

1. Introduction

Les travaux en Recherche d'Information (RI) visent à proposer des modèles et des méthodes permettant de répondre le plus justement possible aux attentes des utilisateurs par rapport aux besoins qu'ils expriment au travers d'une requête. Il s'agit pour un système de RI de limiter le bruit et le silence documentaire, c'est-à-dire de ne pas restituer des documents qui ne répondent pas au besoin et en même temps restituer le maximum de documents pertinents. Les moteurs génériques, comme ceux qui permettent d'accéder à des documents du Web, n'intègrent pas des mécanismes qui s'adapteraient à l'usage que l'utilisateur souhaite faire de l'information retrouvée. Pourtant, la satisfaction de l'utilisateur par rapport aux réponses d'un système dépend de l'objectif de l'utilisateur : veut-il vérifier une hypothèse ? Connaître la réponse à une question précise ? ou veut-il réaliser une étude par rapport à un domaine ? En fonction de ses objectifs, l'utilisateur souhaitera un plus ou moins grand nombre de documents en réponse à sa requête. Un seul document peut servir pour répondre à une question précise ; en revanche pour une étude, un plus grand nombre de documents sera attendu. La redondance dans les réponses peut correspondre à un besoin ou au contraire à un bruit, en fonction du contexte de la recherche. Enfin, un document complet n'est peut être pas utile pour l'utilisateur et celui-ci préfèrera parfois obtenir seulement les passages de documents qui répondent à son besoin.

C'est dans ce contexte de systèmes adaptatifs que nous proposons dans cet article les résultats d'une étude sur le problème de la recherche de passages pertinents dans des documents et de la détection de la nouveauté. Ces tâches de RI ne sont pas nouvelles, cependant, leur évaluation reste difficile. Trois programmes d'évaluation internationaux s'intéressent à ces problématiques. Le programme INEX (INEX Initiative [1]) évalue les systèmes par rapport à leur capacité à détecter les composants XML pertinents pour un besoin d'information portant soit sur le contenu des éléments, soit sur une combinaison contenu / structure. Le programme TDT (Topic Detection and Tracking[2]) évalue les systèmes par rapport à leur capacité à détecter un nouvel événement (sur la base d'un flux d'information télévisée retranscrit) puis à suivre cet événement. Enfin, le programme TREC (Text Retrieval Conference[3]) dans le cadre de la tâche nouveauté (Novelty track) introduite en 2002 évalue les systèmes par rapport à leur capacité d'une part à détecter les passages (phrases) des documents qui sont pertinents par rapport à une

[1] inex.is.informatik.uni-duisburg.de

[2] www.nist.gov/speech/tests/tdt/

[3] trec.nist.gov

requête, d'autre part à déterminer parmi ces passages lesquels apportent de la nouveauté.

L'existence de tels programmes d'évaluation est intéressante à plus d'un titre :

- pour un chercheur, cela rend l'évaluation d'une nouvelle technique moins difficile et moins coûteuse. En effet la constitution d'une collection de test est assez lourde puisque, outre le choix des documents et des requêtes de test, elle doit faire intervenir des jugements humains quant à la pertinence des documents restitués par le SRI et quant aux documents qui sont effectivement pertinents et qui auraient donc dû être restitués pour chaque requête,
- les résultats obtenus par une technique peuvent être directement comparés à ceux déjà obtenus par d'autres systèmes, puisque les collections sont communes et que les publications scientifiques font état des résultats obtenus,
- des avancées importantes ont pu être obtenues au travers de ces programmes d'évaluation. Par exemple, l'utilisation maintenant très commune de la reformulation automatique de requêtes par la réinjection de pertinence aveugle (Buckley et al., 1992) a été introduite dans TREC (trec.nist.gov).

Ainsi, l'importance de l'existence de tels programmes n'est pas remise en cause par la communauté scientifique, pourtant un certain nombre de questions se posent :

- les tâches définies ne sont pas toujours directement calquées sur une utilisation " réelle " d'un SRI. Des efforts sont réalisés dans ce sens, mais les participants restent confrontés à la difficulté de la définition de tâches les plus réalistes possibles, mais qui prennent également en compte le type d'évolution des systèmes et de leurs fonctionnalités,
- l'évaluation porte sur des critères restreints comme le rappel et la précision. Ces mesures offrent l'avantage d'être quantifiables mais ne permettent pas de mesurer pleinement la satisfaction de l'utilisateur. Récemment par exemple, de nouvelles tâches ont été introduites pour mesurer l'interactivité utilisateur/système (TREC7).
- les évaluateurs qui analysent les réponses introduisent une subjectivité qui n'est pas prise en compte. Jusqu'ici, cette subjectivité est considérée comme étant gommée par le fait que l'évaluation est réalisée sur un grand nombre de requêtes, chacune évaluée par différents juges (Buckley et Voorhees, 2000),
- l'évaluation est réalisée de façon globale, en calculant des moyennes de performance sur un ensemble de requêtes. Cette analyse globale ne permet pas la compréhension fine des mécanismes de RI et cache les disparités dans les résultats obtenus. Certains travaux s'intéressent maintenant à une analyse plus fine des résultats (Buckley et Harman, 2004) pour mieux

comprendre le fonctionnement des différents mécanismes proposés. Il s'agit également de l'objectif de cet article.

Dans cet article, nous rapportons les résultats de l'analyse des résultats obtenus par les différents participants à la tâche *'Nouveauté'* du programme TREC. L'objectif de cette analyse est multiple : il s'agit d'abord de proposer une méthode d'analyse des résultats applicable dans différents contextes. Il s'agit ensuite de mesurer la difficulté de la tâche de recherche d'information choisie et de déduire des tendances pour le développement de nouveaux mécanismes de détection de nouveauté.

Cet article est organisé comme suit. Dans la section 2, nous présentons le cadre d'évaluation de la tâche nouveauté de TREC. Nous présentons dans la section 3 une revue des mécanismes proposés dans la tâche TREC. La section 4 est dédiée à la présentation des éléments nécessaires à la compréhension de l'analyse des résultats. Nous discutons ces résultats dans les sections 5 et 6 afin d'essayer de dégager des éléments utiles pour la construction de nouvelles méthodes de recherche de passages et de détection de la nouveauté.

2. Tâche *nouveauté* de TREC

2.1. Recherche des phrases nouvelles

L'étude présentée dans cet article s'intéresse à la recherche de passages de documents pertinents et à la détection de passages de documents potentiellement nouveaux pour l'utilisateur (par rapport aux documents qu'il a déjà lu). Cette étude se base sur la tâche « nouveauté » telle que définie dans TREC (Harman, 2002). Cette tâche comprend deux sous-tâches (cf. Figure 1) :

- La sélection des phrases pertinentes à partir de documents connus comme étant pertinents,
- La sélection des phrases apportant des éléments d'information nouveaux; il s'agit d'un sous-ensemble des phrases pertinentes.

Le fait de ne considérer que des documents pertinents lors de la sélection des phrases pertinentes ou nouvelles peut être vu comme une contrainte forte. Ce choix s'explique d'une part par le souhait de valider les techniques de détection de nouveauté seulement (induisant le découpage de la tâche) et d'autre part par le fait que d'autres tâches du programme s'intéressent à valider et à évaluer les mécanismes de recherche d'information au niveau du document. Le choix du niveau de découpage des documents (un passage est une phrase) s'explique :

- d'une part par le fait que d'autres programmes s'intéressent à d'autres niveaux de découpages. Par exemple INEX s'intéresse aux composants XML,
- d'autre part par le fait que la sélection de phrases correspond à une phase préliminaire essentielle de d'autres tâches de RI comme la recherche de faits (question/réponse) ou le résumé automatique.

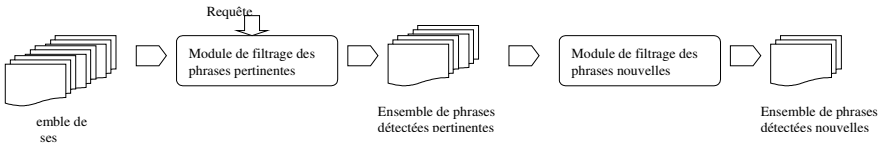


Figure 1: *Phases de détection de la nouveauté*

2.2 Caractéristiques de la collection de test

	NIST-2002
Nombre de requêtes	49
Nombre moyen de documents pertinents par requête	22,3
Nombre moyen de phrases issues des documents par requête	1321
Nombre moyen de phrases pertinentes par requête	27,9
% moyen de phrases pertinentes	2,1
Nombre moyen de phrases nouvelles par requête	25,3
% moyen de phrases nouvelles (par rapport à celles qui sont pertinentes)	90,9

Tableau 1 : *Caractéristiques de la collection de test de TREC*

En 2002, TREC a choisi de sélectionner 49 requêtes issues des requêtes 300-450 des collections TREC. Le NIST (National Institute of Standards and Technology) a sélectionné les documents effectivement pertinents pour chacune des requêtes, avec un maximum de 25 documents par requête et les a fournis aux participants. Dans une seconde étape, des évaluateurs humains ont indiqué quelles phrases de ces documents étaient effectivement pertinentes et lesquelles apportaient des éléments nouveaux. Les caractéristiques de cette collection sont fournies dans le tableau 1.

2.3 Critères d'évaluation

Les critères d'évaluation sont ceux définis par TREC et sont directement issus des critères communément utilisés pour évaluer les systèmes de recherche d'information : les taux de rappel et de précision. Ces deux taux évoluant en sens inverse, une mesure globale, la mesure F combinant rappel et précision permet une comparaison rapide des résultats obtenus par différents systèmes. Cette mesure fait jouer un rôle symétrique au rappel et à la précision, sans privilégier l'un ou l'autre de ces critères.

Ces mesures sont formulées pour la recherche de passages comme suit :

$$R = Rappel = \frac{Nombre\ de\ phrases\ pertinentes\ et\ sélectionnées}{Nombre\ de\ phrases\ jugées\ pertinentes}$$

$$P = \Pr écision = \frac{Nombre\ de\ phrases\ pertinentes\ et\ sélectionnées}{Nombre\ de\ phrases\ sélectionnées}$$

$$mesureF = \frac{2 \cdot R \cdot P}{R + P}$$

Ces mesures appliquées à la détection de la nouveauté sont définies de la façon suivante :

$$R = Rappel = \frac{Nombre\ de\ phrases\ nouvelles\ et\ sélectionnées}{Nombre\ de\ phrases\ jugées\ nouvelles}$$

$$P = \Pr écision = \frac{Nombre\ de\ phrases\ nouvelles\ et\ sélectionnées}{Nombre\ de\ phrases\ sélectionnées}$$

$$mesureF = \frac{2 \cdot R \cdot P}{R + P}$$

Lorsque l'évaluation prend en compte un ensemble de requêtes, la moyenne des résultats permet de mesurer les performances. La moyenne peut également être calculée par rapport à l'ensemble des systèmes pour une requête donnée.

2.4 Participants

13 groupes ont participé à TREC 2002, correspondant à un total de 43 systèmes ou ensembles de résultats. En pratique, un groupe utilise généralement un seul outil pour lequel il teste différents paramètres ; nous appellerons donc 'système' ou 'version de système' un outil et les paramètres associés.

3. Travaux du domaine

Les méthodes de détection de la nouveauté en recherche d'information ont pour objectif de fournir à l'utilisateur une aide lors de la prise en compte des résultats d'un système. Il s'agit plus spécifiquement de décider si une information est redondante ou si au contraire, elle apporte de la nouveauté par rapport aux informations que l'utilisateur a déjà vues.

La restitution de passages plutôt que de documents entiers est apparue avec l'essor du langage SGML qui permettait de définir la notion de passage en s'appuyant sur la structure explicite du document. Dans (Wilkinson, 1994), (Corral, 1995) les unités documentaires (parties de documents) sont extraites en s'appuyant sur la DTD (Document Type Definition). Ces unités correspondent alors aux unités manipulées par le système à chaque étape : indexation, recherche et restitution. D'autres travaux se sont intéressés à la combinaison de ressemblances dites locales, c'est à dire calculées au niveau des passages, et de ressemblances globales, c'est à dire calculées au niveau des documents. (Salton, 1994) propose une recherche en deux étapes : les documents entiers sont indexés ; lors d'une recherche, les documents supposés pertinents sont alors indexés par passage et un nouveau calcul de ressemblance avec la requête permet de sélectionner les meilleurs passages. Ces travaux ont montré que le découpage optimal était le paragraphe. D'autres types de passages ont été définis dans le cadre de documents non structurés comme les fenêtres de taille fixe (Stanfill, 1992) ou les phrases (Harman, 2002).

TREC 2002 (Harman, 2002) a défini la tâche de recherche de passages au niveau de la phrase. Cette recherche de phrases pertinentes est réalisée à partir de documents jugés pertinents par des évaluateurs. Le problème se trouve certes simplifié (il n'y a plus le risque de sélectionner une phrase issue d'un document non pertinent) mais ce choix permet de se focaliser sur un aspect précis de la recherche de passages.

La majorité des systèmes utilisés dans TREC ont considéré les phrases comme des documents et appliquent simplement les techniques existantes au niveau des phrases plutôt qu'au niveau des documents. Ainsi les phrases, considérées de façon individuelle, ont d'abord été indexées avant de calculer la ressemblance de ces phrases avec la requête. Bien que les requêtes TREC comprennent une partie description qui donne des informations sur les documents qui seront considérés comme pertinents et sur ceux qui seront considérés comme non pertinents, les participants à TREC considèrent cette description de façon globale et perdent donc des informations importantes. (Allan, 2003) a utilisé le modèle vectoriel. (Collins, 2002) a également considéré une mesure cosinus avec une pondération de type tf.idf et a complété la recherche en appliquant les techniques de réinjection de pertinence. Ils ont également étudié différents types de classificateur en se basant sur des caractéristiques lexicales et sémantiques issues de l'analyse des textes. (Schiffman, 2002) a choisi d'étendre la requête en ajoutant à la requête initiale des termes

sémantiquement équivalents et des termes fortement co-occurrents avec les termes de la requête. (Zhang, 2002, 2003) combine la réinjection de pertinence aveugle avec une classification des phrases utilisant un algorithme SVM (Support Vector Machine). (Dkaki, 2002) a amené une nouvelle approche en caractérisant les termes d'indexation selon trois classes : très pertinents, pertinents, non pertinents.

Concernant la sous tâche de détection des phrases nouvelles, différentes approches ont été proposées dans TREC. Toutes se basent sur la représentation des requêtes et des phrases des documents sous forme d'ensemble de termes. Ces ensembles de termes sont comparés pour déterminer le caractère redondant ou nouveau d'une phrase. (Allan et al., 2003) représente les textes par le modèle de langage (Ponte et Croft, 1998) en lissant les représentations en fonction de la longueur des phrases. (Kazawa et al., 2002) sélectionne les phrases nouvelles parmi les phrases pertinentes en se basant sur la mesure de la pertinence marginale maximum (MMR) (Carbonell, Goldstein, 1998). Dans (Kwok, 2002), les phrases sont d'abord étendues par des termes synonymes de ceux utilisés dans les phrases. Ces termes synonymes sont issus de WordNet[4]. Un calcul de similarité entre la phrase en cours de traitement et les phrases déjà traitées permet de décider du caractère de nouveauté de la phrase. Dans (Dkaki et Mothe, 2002) la détection de la nouveauté est basée sur une fonction de décision calculée en combinant la similarité de la phrase considérée avec chacune des phrases déjà vues par l'utilisateur et avec une phrase abstraite correspondant à l'union des phrases déjà traitées. Plutôt que de considérer la ressemblance entre la phrase en cours de traitement et les phrases précédemment traitées, (Zhang et al., 2002) base la sélection des phrases nouvelles sur une mesure de recouvrement (% de termes identiques à la phrase précédente).

Ces méthodes ont été évaluées dans le cadre bien défini de TREC. Ce cadre comporte un certain nombre d'avantages comme le fait d'évaluer sur la base de critères communs et de collections communes. Il faut cependant noter que l'évaluation est réalisée de façon globale, c'est à dire en calculant des moyennes de performance sur un ensemble de requêtes. Cette caractéristique limite l'analyse de la compréhension fine des mécanismes mis en oeuvre et cache les disparités des résultats obtenus. Peu de travaux s'intéressent à une analyse plus fine. Les principales analyses détaillées concernent les algorithmes de lemmatisation en anglais (Hull et Grefenstette, 1996) et l'expansion de requêtes. (Sormunen, 2002), (Nelson, 1995) visent généralement à encourager l'examen des comportements de systèmes simples de RI, en cherchant notamment à repérer les requêtes "difficiles". Dans le domaine de la caractérisation des requêtes en RI, peu de travaux proposent une approche au-delà de caractéristiques primitives, comme le nombre de mots que contient une requête et, dans le cas de moteurs de recherche les autorisant, l'emploi de connecteurs booléens (AND / OR / NOT). Dans (Buckley et Harman, 2004), les auteurs indiquent que la compréhension de la variabilité des résultats (le système 1 donne de bons résultats par rapport à la requête 1 mais de mauvais sur la requête 2

[4] www.cogsci.princeton.edu/~wn/

alors que le système 2 permet d'obtenir des résultats inverses) est difficile car elle est due à trois types de facteurs : la formulation du besoin, la relation entre la requête et le contenu de la collection, et les caractéristiques du système utilisé. Le workshop RIA (Harman et Buckley, 2004) vise à analyser les facteurs de variabilité des systèmes et des expressions des besoins. Ils se sont intéressés au mécanisme d'expansion de requêtes par les techniques de réinjection de pertinence (Harman, 1992). Sept systèmes ont été étudiés.

4. Outils mathématiques utilisés pour l'analyse des données

4.1 L'analyse en composante principale

L'analyse en Composante Principale (ACP) (Benzécri, 1973), (Lebart et al., 1995) permet d'obtenir une représentation du nuage de points extrait d'un tableau de données, dans un espace de dimension réduite de telle manière que l'information expliquée dans cet espace soit la plus grande possible en termes d'inertie. Cette méthode est basée sur la recherche des axes principaux d'un nuage de points.

De façon générale, le tableau de départ est une matrice de dimension n x p où n est le nombre d'individus ou d'unités d'observations et p le nombre de variables ou de critères observés. Les données sont relatives à des variables quantitatives, continues, homogènes ou non. A partir de cette matrice, une représentation graphique des individus peut être réalisée en choisissant les variables comme axes. Une telle représentation permet une interprétation simple des corrélations visualisées, mais elle n'est cependant pas satisfaisante : elle ne prend pas en compte les dépendances entre les variables, la visualisation des corrélations est partielle et aléatoire, l'hétérogénéité des données n'est pas prise en compte. L'ACP permet de répondre à ces problèmes. Il s'agit de déterminer les axes principaux c'est à dire les axes minimisant l'inertie du nuage de points. Ces nouveaux axes correspondent à des combinaisons des variables et sont déterminés par la recherche des valeurs propres de la matrice de variance-covariance. La représentation graphique de la matrice résultat permet d'observer la forme du nuage de points et d'apprécier les distances entre deux points et donc entre les deux concepts qu'ils représentent. Une représentation selon les premiers axes principaux suffit à visualiser la majorité de l'information.

Plus précisément, soient :

$X = \{Xij,\ i \in I, j \in J\}$ un tableau de données

p le nombre de variables et *n* le nombre d'individus

• Un pré-traitement de la matrice X est généralement nécessaire afin de centrer les données par rapport aux axes de représentation (le centre de gravité du nuage correspond alors au centre du repère de représentation) :

$X = \{Xij,\ i \in I, j \in J\}$

est transformé en

$$X' = \left(X'_{ij} = \frac{X_{ij} - \bar{X}_j}{\sqrt{n}} ; i \in I, j \in J \right)$$

où $\bar{X}_j$ est la moyenne arithmétique de X_j

• Les données peuvent être hétérogènes quant à leur dispersion et à leur nature. Pour pouvoir néanmoins les comparer, il faut alors les transformer de sorte que les variables soient de variance unité et de moyenne nulle.

$$X' = \left\{ X_{ij}', i \in I, j \in J \right\} \text{ (précédente)}$$

est transformé en

$$X'' = \left(X''_{ij} = \frac{X'^{ij}}{\sigma_j}; i \in I \right.$$

où σ_j désigne l'écart type de la variable x_j

• Les axes sont ensuite donnés par les valeurs propres de la matrice de variance covariance :

$$\left(\right. \qquad \qquad _{jk} = \sum_{i=1}^{n} \frac{X''^{ij} \cdot X''^{i}}{n}$$

• Les axes factoriels ainsi obtenus permettent d'obtenir la meilleure représentation visuelle des distances entre les individus et entre les variables. Il faut noter que la valeur propre correspond à l'inertie d'un facteur, c'est-à-dire à la variance des coordonnées des individus sur l'axe associé. La visualisation est donc généralement basée sur les axes de plus forte inertie, ceux qui expliquent le plus les données (qui permettent une réduction de l'espace de représentation en maximisant l'information représentée).

•*4.2 La classification ascendante hiérarchique (CAH)*

Les classes sont obtenues en agrégeant les deux éléments (individus) les plus proches. L'agrégation est répétée en rapprochant les deux éléments (individus ou classes déjà obtenues) les plus proches. Le processus est réitéré jusqu'à l'obtention d'une classe unique. Cette classification génère un arbre que l'on peut couper à différents niveaux pour obtenir des classes plus ou moins grandes ou plus ou moins nombreuses (dendrogramme).

La distance interclasse peut être définie par :

$$dis\tan ce(A,B)=\frac{1}{Card(A)\cdot Card(B)}\cdot\sum_{a\in A,b\in B} dis\tan c$$

où A et B sont deux classes, et où distance(a,b) peut être définie comme la distance euclidienne entre les éléments a et b.

D'autres formules peuvent être utilisées pour calculer la distance entre deux éléments comme la distance inférieure (qui favorise la création de classes de faible inertie) ou la distance supérieure (qui favorise la création de classes d'inertie plus importante).

5. Analyse des résultats soumis à TREC

5.1 Motivations

Comme nous l'avons indiqué précédemment, une des caractéristiques des programmes d'évaluation est que les performances d'un système sont calculées de façon globale, en calculant des mesures moyennes sur un ensemble de requêtes (rappel, précision et mesure F moyens).

Par exemple, le tableau 2 correspond au type de résultats qui sont publiés dans les actes de la conférence TREC pour la tâche *Nouveauté* et la sous tâche 1 - *Novelty track, task 1* (Harman, 2002).

Systèmes	Rappel	Précision	R*P
Dubrun	0.49	0.15	0.19
Thunv1	0.34	0.23	0.235
Thunv3	0.41	0.20	0.235
Pircs2N01	0.49	0.16	0.209
Nttcslabnvr2	0.60	0.10	0.166

Tableau 2 : *Rappel et précision moyens pour quelques systèmes*

Ce tableau 1 indique que les systèmes Dubrun et Pircs2N01 obtiennent des résultats comparables, que l'on considère le rappel ou la précision. D'autre part, le système Nttcslabnvr2 est assez performant si l'on considère le rappel, mais peu performant si l'on considère la précision. Enfin, les systèmes Thumv1 et Thumv3, qui correspondent à différentes versions de l'outil Thunv sont les plus performants

(R*P est de 0.235), avec un léger avantage en termes de rappel pour Thunv3 et un léger avantage en termes de précision pour Thunv1.

Le tableau 3 donne quelques résultats détaillés pour ces mêmes systèmes. Seul le rappel est considéré. La ligne 'meilleur rappel' indique la valeur de rappel obtenu par le meilleur système (le meilleur système peut être différent pour chacune des requêtes).

Système / Requête	305	369	377	427
Dubrun	0.13	0.25	0.33	0.36
Thunv1	0.20	0.17	0	0.27
Thunv3	0.47	0.17	0	0.27
Pircs2N01	0.13	0.58	0.33	0.27
Nttcslabnvr2	0.60	0.58	0	0.45
Best recall	0.60	0.58	0.67	0.82

Tableau 3 : *Rappel pour certaines requêtes pour les systèmes du tableau 2*

Les résultats détaillés du tableau 3 montrent que Dubrun et Pircs2N01 obtiennent des résultats bien différents pour les besoins d'information 369 et 427. Pour le besoin 369, Pircs2N01 obtient un rappel deux fois plus important que Dubrun et pour le besoin 427, c'est Dubrun qui est meilleur que Pircs2N01. Pourtant, globalement (tableau 2), leurs performances sont comparables. Nttcslabnvr2 est inefficace pour le besoin d'information 377 (rappel nul) alors que globalement, il obtient le meilleur rappel. Thunv1 et Thunv3 obtiennent les mêmes résultats pour différentes requêtes. Pourtant globalement, ils se distinguent, l'un favorise le rappel alors que l'autre favorise la précision.

Sur ce petit exemple, nous pouvons donc observer que la tendance générale, représentée par les moyennes des résultats par rapport à l'ensemble des besoins d'information peut être contredite lorsque l'on considère simplement quelques besoins d'information. Cette variabilité dans les résultats, qui peut être relativement significative doit mieux être prise en compte si l'on veut pouvoir mieux comparer les résultats entre eux et arriver à proposer des systèmes adaptatifs, c'est-à-dire qui s'adaptent au contexte de chaque besoin d'information.

Notre objectif est donc de proposer une démarche d'analyse des résultats de campagne d'évaluation pour permettre une meilleure prise en compte et exploitation de ces résultats.

5.2 Les résultats obtenus par les participants: étude des taux de rappel et précision

Notre première analyse concerne les résultats obtenus par les différents participants en termes de précision et de rappel dans la tâche globale, c'est-à-dire la détection des phrases nouvelles parmi celles qu'a sélectionné le système.

Le taux de rappel moyen (sur l'ensemble des requêtes pour un système) varie de 0,04 à 0,49. Le meilleur système permet donc de détecter la moitié environ des phrases réellement nouvelles. Le rappel moyen (sur l'ensemble des résultats envoyés par les participants) est de 0,24.

Concernant le taux de précision, il varie de 0,05 à 0,23. Le taux de précision moyen en considérant l'ensemble des systèmes est de 0,12. Ainsi, au mieux, un peu moins d'un quart des phrases détectées comme nouvelles le sont réellement. Le système ayant obtenu le meilleur rappel (0,49) a obtenu 0,09 de précision ; alors que le système qui a obtenu la meilleure précision (0,23) a obtenu 0,29 de rappel.

Les taux de rappel et de précision variant en sens inverse, il est important de s'intéresser à la mesure globale (mesure F). Celle-ci varie de 0,039 à 0,216 sur l'ensemble des données envoyées par les participants, pour une valeur moyenne sur l'ensemble des systèmes de 0,134. Le système ayant obtenu la meilleure valeur de mesure F a obtenu 0,3 de rappel et 0,22 de précision.

Les valeurs de ces mesures reflètent d'abord la difficulté à détecter la nouveauté. En moyenne, les systèmes détectent 1/4 des phrases nouvelles mais incluent également beaucoup de bruit dans leur réponse (9/10).

5.3 Requêtes difficiles?

Cette section s'intéresse à étudier l'ensemble des systèmes et à répondre à la question: les systèmes butent-ils tous sur les mêmes requêtes? Et ces requêtes peuvent-elles être caractérisées?

La figure 2a) indique pour chacune des requêtes le nombre de systèmes pour lesquels la mesure F est nulle (43 systèmes ont participés). Par exemple, pour la première requête (305), 3 systèmes indiquent une valeur nulle de la mesure F. Le graphique figure 2b) indique la valeur moyenne de la mesure F. Par exemple, pour la première requête (305), la mesure F moyenne sur l'ensemble des systèmes est de 0,1.

Un certain nombre de requêtes peuvent être considérées comme difficiles.

Les trois requêtes pour lesquelles une grande proportion de systèmes ont échoué (15 systèmes ou plus sur 43) sont celles qui ont les plus faibles taux de précision (leur taux de rappel est également très faible, mais pas nécessairement parmi les plus faibles). De la même façon, les requêtes pour lesquelles les systèmes ont obtenu en moyenne les meilleures valeurs de mesure F offrent les meilleurs taux de

précision (mais pas forcément de rappel). Ce résultat montre qu'il serait plus facile de privilégier la précision, c'est à dire limiter le bruit dans les réponses.

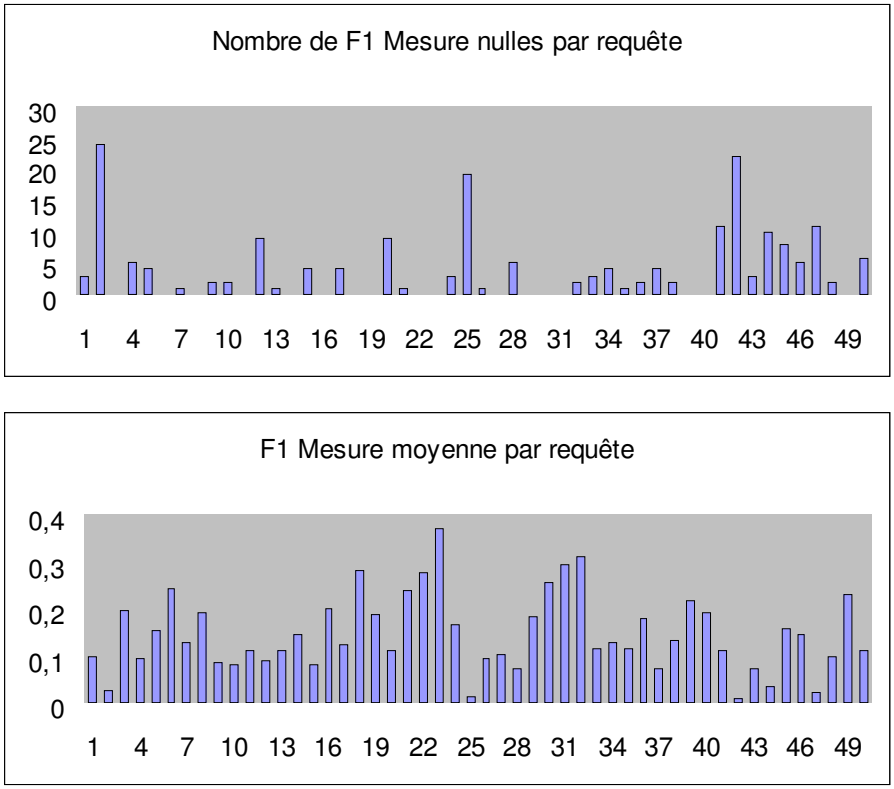


Figure 2: *Nombre de systèmes pour lesquels la mesure F est nulle et mesure F moyenne pour chacune des requêtes. (la correspondance entre le numéro de requête et sa référence est donnée dans le tableau 2).*

Parmi les requêtes pour lesquelles la mesure F moyenne est la plus faible figure la majorité des requêtes pour lesquelles moins de 1% des phrases parmi l'ensemble des phrases des documents pertinents étaient elles-mêmes pertinentes. Plus précisément, parmi les 49 requêtes, 6 d'entre elles ont moins de 1% de phrases pertinentes (cf Tableau 2, requêtes 312, 316, 351, 377, 427 et 432), 5 figurent parmi les 10 moins bons résultats moyens.

Rang/Besoin	Pert	%Total	Nouv	%Pert	Besoin	Pert	%Total	Nouv	%Pert
1 / 305	15	2.01	15	100	2 / 312	5	0.87	5	100
3 / 314	25	2.35	25	100	4 / 315	18	3.08	11	61.11
5 / 316	22	0.99	18	81.82	6 / 317	23	4.19	23	100
7 / 322	34	4.57	34	100	8 / 323	65	4.57	60	92.31
9 / 325	21	1.25	21	100	10/326	10	1	8	80
11/330	29	3.49	27	93.1	12/339	12	1.6	11	91.67
13/342	17	2.17	17	100	14/345	47	5.19	47	100
15/351	6	0.75	5	83.33	16/355	103	3.94	78	75.73
17/356	10	1.2	9	90	18/358	40	4.8	37	92.5
19/362	47	5.15	46	97.87	20/363	11	2.08	10	90.91
21/364	42	3.5	42	100	22/365	34	2.8	34	100
23/368	71	4.63	66	92.96	24/369	13	1.79	12	92.31
25/377	3	0.19	3	100	26/381	19	1.38	19	100
27/382	41	1.83	24	58.53	28/384	23	1.41	23	100
29/386	43	4.3	41	95.35	30/388	56	4.57	56	100
31/394	21	2.45	21	100	32/397	29	6.18	28	96.5
33/405	40	3.75	37	92.5	34/406	10	1.79	10	100
35/407	32	2.46	29	90.62	36409	17	3.26	12	70.59
37/410	17	1.92	15	88.24	38/411	21	1.86	19	90.48
39/414	29	3.62	25	86.21	40/416	36	1.96	30	83.33
41/419	50	3.32	36	72	42/420	18	1.4	18	100
43/427	14	0.31	11	78.57	44/432	9	0.96	8	88.89
45/433	11	1.72	7	63.64	46/440	19	1.35	19	100
47/445	10	1.07	5	50	48/448	20	2.25	20	100
49/449	57	3.65	57	100					

Tableau 2: *Caractéristiques des requêtes (nombre de phrases pertinentes et nouvelles)*

5.4 La non détection de la nouveauté

La tâche de détection de la nouveauté telle que proposée dans le programme TREC comprend en réalité deux sous-tâches : la sélection des phrases pertinentes puis la sélection des phrases nouvelles parmi les phrases retenues. Pour mesurer la pertinence des mécanismes de sélection des phrases nouvelles, nous avons calculé

les résultats qu'auraient obtenus les systèmes s'ils avaient choisi de considérer toutes les phrases décidées pertinentes comme étant nouvelles. En d'autres termes, si ces systèmes n'avaient pas appliqué de module de filtrage de la redondance (cf Figure 3).

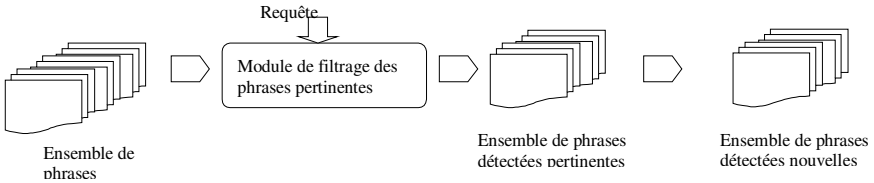
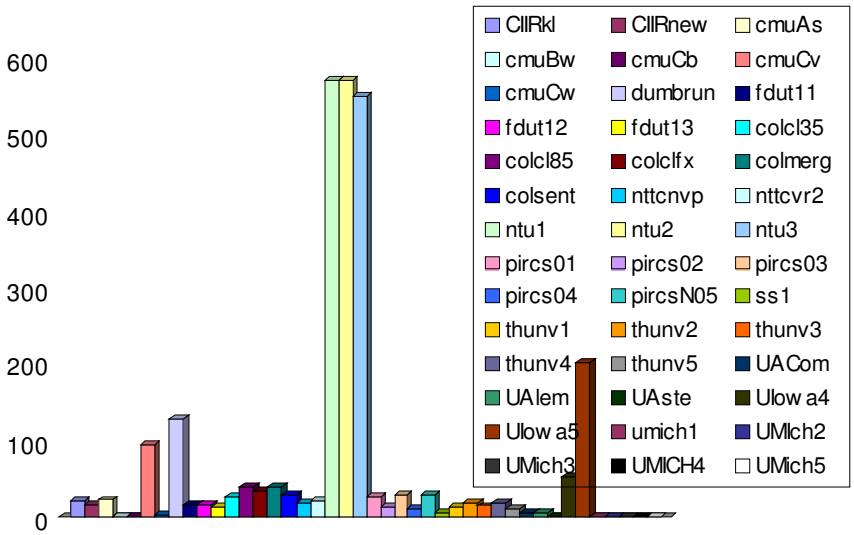


Figure 3: *Phases de sélection des phrases nouvelles modifiées*

Si l'on considère que chaque système restitue toutes les phrases détectées pertinentes en tant que phrases nouvelles, le rappel moyen sur l'ensemble des systèmes est de 0,34 (contre 0,24). La précision moyenne est elle de 0,119 (contre 0,134). Compte tenu de l'expérimentation, le rappel ne pouvait qu'augmenter (puisque aucune phrase n'est écartée). On note que la précision diminue, mais globalement, la mesure F augmente. Globalement donc, les mécanismes d'élimination de la redondance sont trop stricts et éliminent beaucoup trop de phrases qui sont effectivement nouvelles et maintiennent un bruit trop important.



La légende de cette figure permet de faire le lien entre % d'amélioration et le système correspondant. La légende se lit de gauche à droite et de haut en bas.

Figure 4 : *Amélioration de la mesure F obtenue en considérant toutes les phrases supposées pertinentes comme nouvelles.*

Les plus fortes améliorations (de l'ordre de 500%) sont obtenues par des systèmes avec de très faibles résultats (cf. Figure 4). Il s'agit en réalité d'un même outil avec trois paramétrages différents. Ces systèmes (ntu1, 2 et 3) obtenaient initialement une précision de 0,02 (contre 0,11 en moyenne sur l'ensemble des systèmes, un rappel de 0,40 à 0,47 (contre 0,33) et ainsi une mesure F de 0,06 à 0,07 (contre 0,14 en moyenne).

Il faut toutefois noter que le meilleur système (versions de Thunv) obtient une amélioration de plus de 13% de la mesure F. Aucun système ne détériore globalement les résultats obtenus.

6 Etude des corrélations entre résultats

L'étude des corrélations entre résultats obtenus est particulièrement délicate compte tenu du nombre de paramètres mis en jeu. Dans cette étude, nous avons choisi de nous focaliser indépendamment sur les mesures taux de rappel et taux de précision. En effet, la mesure F est une mesure globale qui, si elle permet de comparer des systèmes entre eux, gomme le fait qu'un système privilégie le rappel ou la précision. Pourtant, il s'agit d'une information importante, puisque, selon la tâche de l'utilisateur, celui-ci sera plutôt intéressé par un fort taux de rappel (exhaustivité de la réponse) ou par un fort taux de précision.

6.1 Méthodologie

L'analyse porte d'abord sur la classification des requêtes. Nous essayons de regrouper les requêtes en fonction des performances obtenues par tel ou tel système. L'objectif à plus long terme de l'étude est de répondre à la question: y a t-il des systèmes plus adaptés à un type donné de requête? Cette analyse est d'abord effectuée en considérant le taux de rappel, puis le taux de précision. Nous utilisons pour cela les méthodes de classification hiérarchique et l'analyse en correspondance principale réduite. Dans un deuxième temps, nous nous intéressons aux systèmes.

6.2 Classification des requêtes

Une première classification des requêtes est obtenue en considérant les requêtes comme des individus et les systèmes comme des variables. La mesure étudiée est le taux de rappel. Le dendrogramme correspondant est présenté figure 5. La deuxième classe en partant de la gauche (requêtes 312, 405, 432, 420) correspond à celle des requêtes les plus difficiles: celles pour lesquelles la moyenne de taux de rappel est la

plus faible (de 0,03 à 0.16). La classe la plus à droite correspond au contraire aux requêtes faciles (365, 416, 394, 406, 411, 326, 409): celles pour lesquelles la moyenne du rappel est comprise entre 0,47 et 0,60.

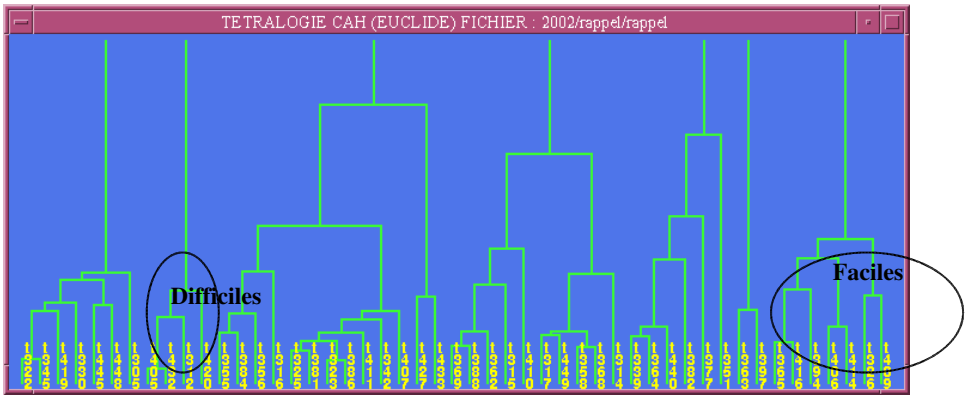
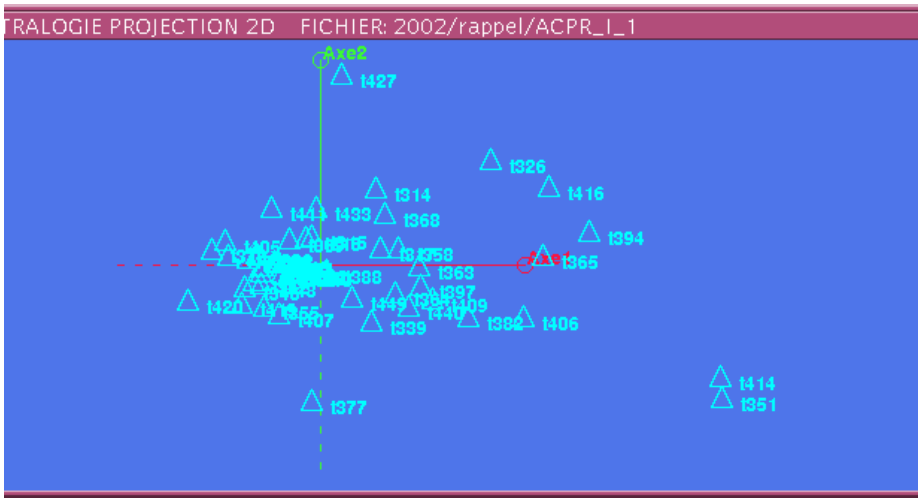
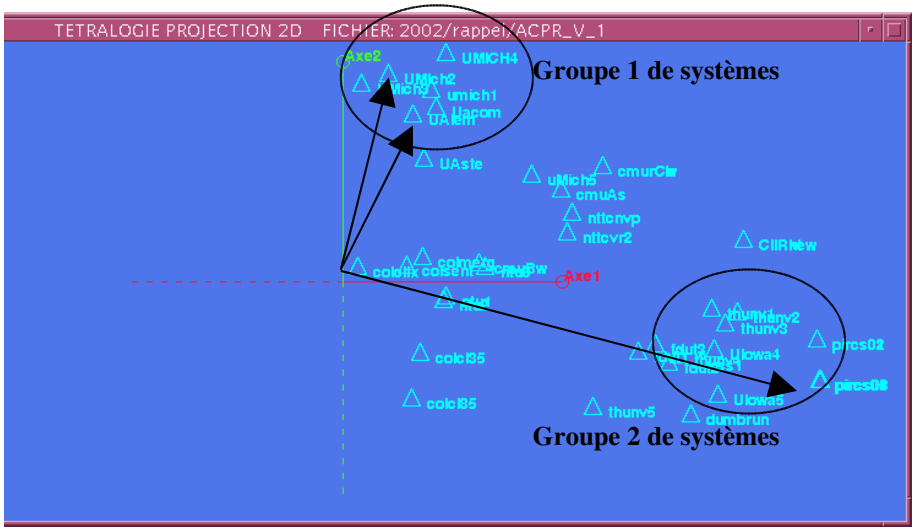


Figure 5 : Classification des requêtes (individus: requêtes, variables: systèmes, mesure: taux de rappel)

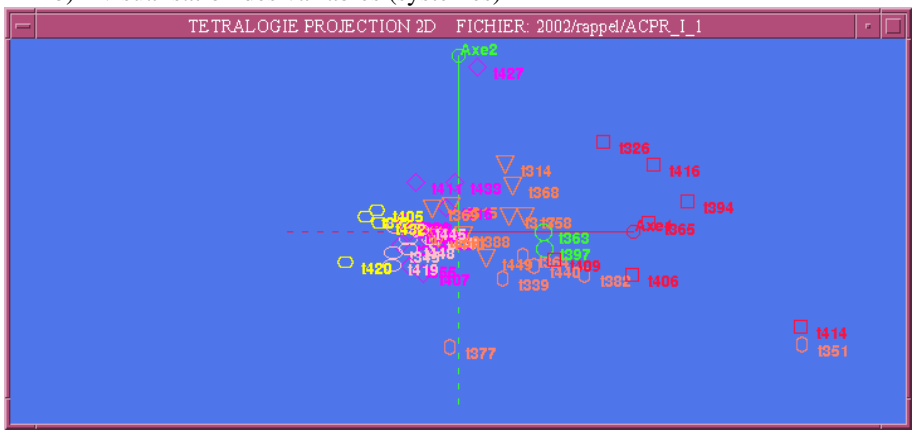
Les figures 6a et 6b présentent les résultats (6a individus, 6b variables) de l'ACP sur le même jeu de données. Seuls les deux premiers axes sont représentés. Ils correspondent à 50 % de l'inertie totale. La figure 6c combine la figure 6a et la figure 5 : à chaque groupe de requêtes est associée une couleur et une forme de point (la version de l'article étant en noir et blanc, les couleurs ont été transformées de façon à ce que les figures restent lisibles).



a) Visualisation des individus (besoins d'information)



b) visualisation des variables (systèmes)



c) Visualisation des individus et de leur classe

Figure 6 : ACP (individus: requêtes, variables: systèmes, mesure: taux de rappel) selon les deux premiers axes.

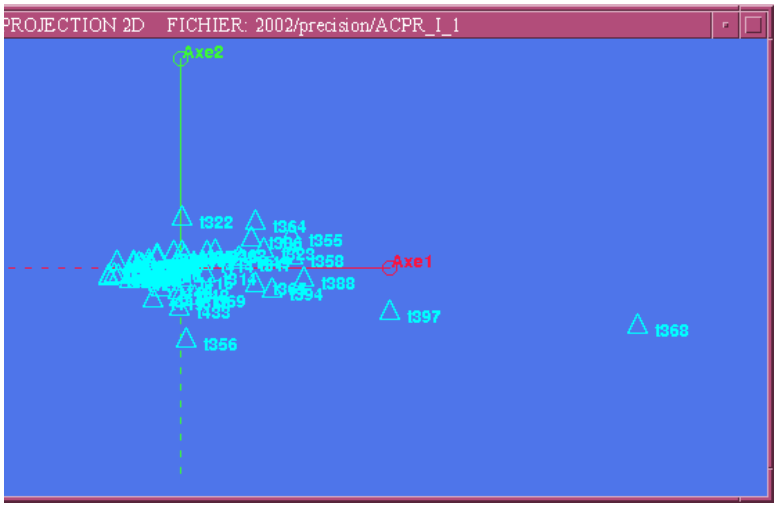
Les variables en périphérie (figure 6b) (celles qui sont sur l'hyper-sphère) nous permettent de distinguer deux groupes de systèmes, que nous nommons arbitrairement N°1 et 2.

Les figures 6 montrent que le groupe de systèmes N°1 fonctionne plutôt bien avec le besoin d'information N°427 (l'individu 427 contribue positivement au vecteur moyen associé au groupe de variables) alors qu'il fonctionne plutôt mal

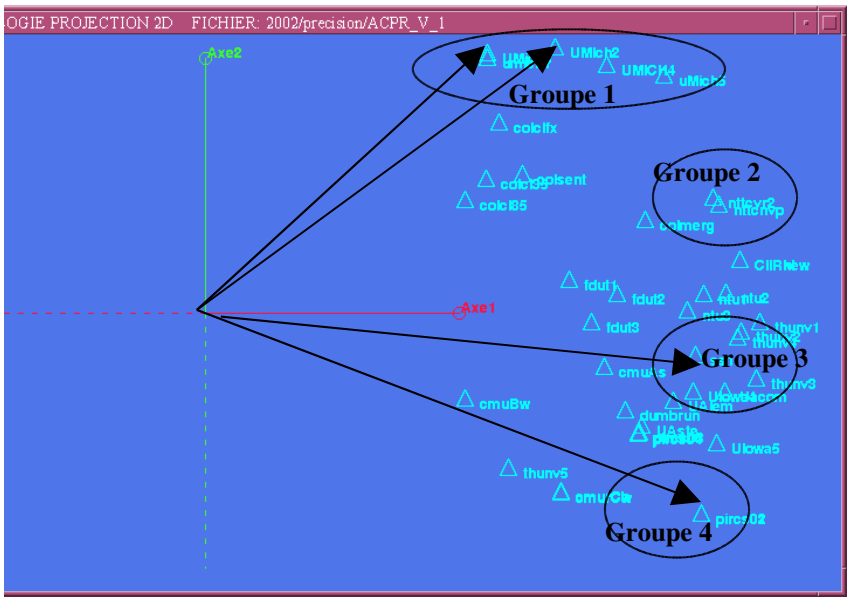
avec le besoin N°377. Ces deux besoins d'information ont une variation importante dans les résultats obtenus : le taux de rappel varie de 0 à 0,91 pour l'un et de 0 à 0, 67 pour l'autre. Ce groupe est constitué des différentes variantes de l'outil UMICH (UMICH1 à 4) et de celles de l'outil UA (UAcom, Ualem, Uastem). Ces figures montrent également que les systèmes du groupe 2 fonctionnent plutôt bien avec les besoins d'information 414 et 351. Ces deux besoins ont également des variations importantes au niveau du taux de rappel : de 0,12 à 1 pour l'un et de 0 à 1 pour l'autre. Les besoins 406 et 382 contribuent également de façon positive. Ces systèmes correspondent aux variations des outils PIRC, Thunv, Durun et Ulowa.

La figure 6c permet de faire le lien avec les classes de requêtes obtenues précédemment. Les requêtes faciles (en rouge dans la version couleur et associées à un symbole carré) se trouvent bien évidement du côté des coordonnées positives (à droite de l'axe vertical). On note cependant leur dispersion dans l'espace ; aucune toutefois ne contribuant aux systèmes du groupe N°1. Les requêtes difficiles (en jaune et associées à un rond) se trouvent bien évidement à l'opposé des requêtes faciles. On note une moindre dispersion. Cela nous conduit à penser qu'il y a plus d'homogénéité dans la difficulté que dans la facilité des requêtes pour les systèmes.

Le même type d'analyse appliqué à la mesure Précision est illustré dans les figures 7.



a) Visualisation des individus (besoins d'information)



b) visualisation des variables (systèmes)

Figure 7 : ACP (individus : requêtes, variables : systèmes, mesure : taux de précision) selon les deux premiers axes.

Les requêtes 397 et 368 contribuent le plus à l'axe associé au groupe 3 de systèmes, constitué des différentes versions de l'outil Thunv (Thunv1, 2 et 4). Les systèmes du groupe 1 répondent mal à ces requêtes en termes de précision.

Ces premiers résultats (figures 6a à 7b) montrent également que les différentes versions d'un même outil impliquent moins de variabilité dans les résultats que deux systèmes différents. En effet, ces différentes versions d'un outil se trouvent généralement dans le même groupe détecté, ou sur le même axe.

6.3 Classification des systèmes

Le commentaire précédent peut être vérifié par une classification des systèmes. Les résultats de la classification sont présentés dans la figure 8, et celle de l'ACP dans la figure 9. Ici, les individus sont les systèmes, les variables correspondent aux requêtes et la mesure correspond à la mesure F. Dans ces figures, il est possible de voir que les versions d'un même outil se trouvent très proches les unes des autres. Par exemple, UAStem, UALem et UACom sont trois versions d'un même système. Ce système a été développé par l'Université d'Amsterdam et ses trois versions correspondent à l'utilisation de radicaux (stem), de lemmes (Lem) ou une

combinaison des deux techniques (Com) lors de l'indexation des documents. Le premier élément de regroupement est donc l'outil utilisé, même lorsque ses différentes versions semblent éloignées en termes de méthodes utilisées. Nous pouvons toutefois noter quelques exceptions comme le cas de Thunv5 qui n'est pas regroupé avec les autres versions de Thunv. Les auteurs du système indiquent qu'il s'agit d'un nouveau système, sans préciser la différence avec le système précédent.

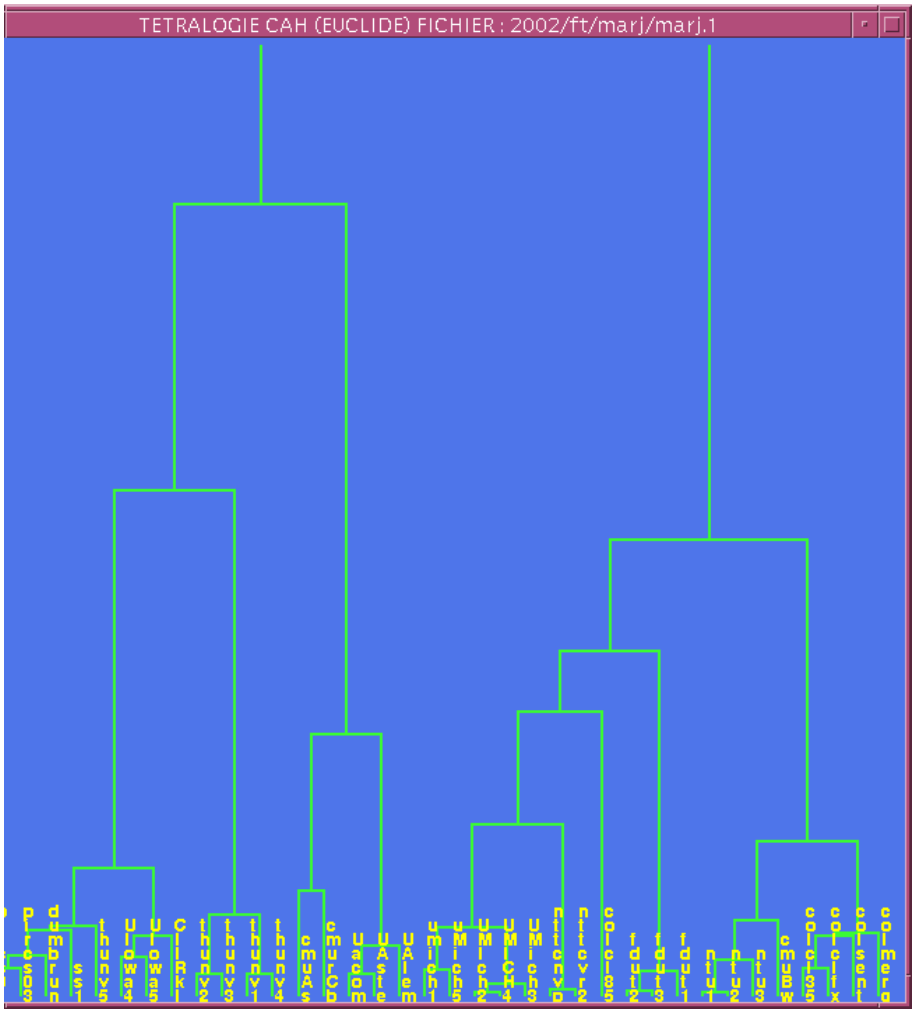


Figure 8 : *Classification des systèmes (individus: systèmes, variables: requêtes, mesure: mesure F)*

La position des versions de FDU est à noter (cf figure 9, bas de l'axe 3). Globalement, ces versions n'obtiennent pas de bons résultats, cependant, les résultats numériques détaillés montrent que cet outil permet d'obtenir de très bon scores sur des requêtes pouvant être qualifiées de difficiles. Cet outil permet par exemple d'obtenir une mesure F de 0,17 par rapport à une requête (315) alors que la valeur moyenne sur l'ensemble des systèmes est de 0,09 et que le meilleur système (celui qui obtient la valeur maximale de mesure F sur l'ensemble des requêtes) obtient pour cette même requête 0,08. Cette même observation peut être faite sur d'autres requêtes (356, 362, 410 par exemple).

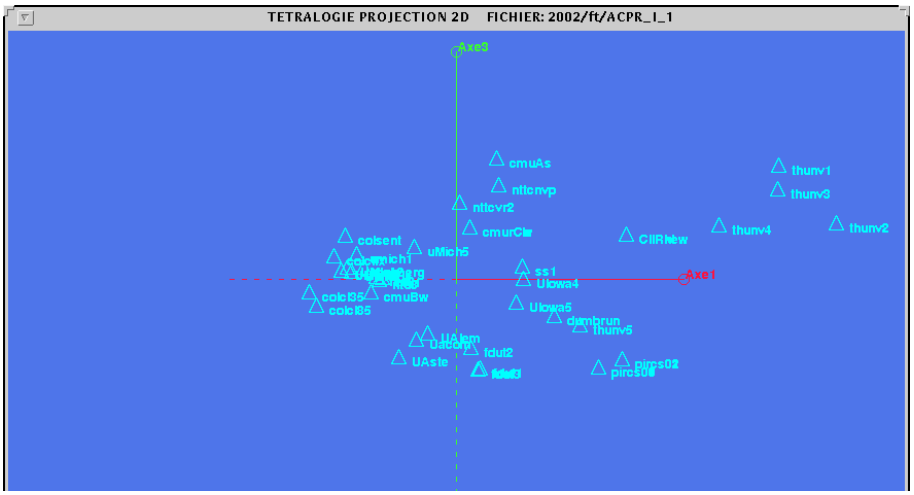


Figure 9 : *Visualisation des variables de l'ACP (individus: systèmes, variables: requêtes, mesure: mesure F), axe 1 et 3.*

7. Discussions et conclusion

Dans cet article, nous avons présenté une analyse des résultats obtenus par 43 systèmes (ou version de systèmes) utilisés pour la détection de la nouveauté sur une collection de test commune en se basant sur des méthodes d'analyse de données. Les premiers résultats ont montré la difficulté de la tâche de détection de la nouveauté.

Sur le jeu de test proposé en 2002, il s'avère même que l'ensemble des modules utilisés pour détecter la nouveauté ont échoué, par une sélection trop stricte des éléments.

Nous avons également obtenu une première classification des requêtes et des systèmes.

L'élément important que nous pouvons conclure est que le paramètre le plus important est l'outil utilisé, alors même que beaucoup d'outils utilisent les mêmes éléments de base de la recherche d'information (représentation de chaque phrase du document par une liste de termes simples extraits selon des techniques "classiques" par exemple). Ainsi, la variabilité dans les résultats (rappel, précision pour différentes requêtes) est plus liée aux outils qu'à leurs paramètres testés dans la tâche. En effet, les différentes versions d'un même outil ont des comportements similaires lors du traitement des requêtes. Ce résultat est important car il indique que les améliorations de performance qui ont été montrées dans la technique de fusion de données peuvent être augmentées en choisissant bien les systèmes à fusionner. Il est en effet important de fusionner plutôt des systèmes qui sont complémentaires. Nos travaux futurs vont donc s'orienter vers la sélection automatique des systèmes à fusionner en fonction des tâches à accomplir.

D'autre part, la grande variabilité des résultats montre qu'il est nécessaire de prolonger cette analyse des résultats, par exemple en étudiant chaque groupe de requêtes et de systèmes. L'objectif de cette étude sera d'analyser chaque groupe de requêtes extrait par notre méthode afin de savoir s'il existe d'autres caractéristiques communes à ces requêtes (type de requêtes, généricité des termes utilisés, type et quantité de résultats attendus, etc.). Nous envisageons d'étudier pour cela différentes caractéristiques linguistiques qui pourraient être extraites directement des formulations des besoins d'information et permettre une typologie pertinente.

L'objectif à terme est de savoir s'il serait envisageable de créer un système adaptatif qui décide lui-même du ou des mécanisme(s) à appliquer en fonction du type de requête rencontré.

8. Références

J. Allan, C. Wade, A. Bolivar, Retrieval and Novelty Detection at the Sentence Level, Research and Development in Information Retrieval, SIGIR'03, p 314-321, 2003.

C. Buckley, G. Salton, J. Allan, Automatic Retrieval with Locality Information Using Smart, Text REtrieval Conference, p 59-72, 1992. (trec.nist.gov)

J-P Benzécri, L'Analyse des Correspondances, Paris, Dunod, Tome 1 et 2, 1973.

H. Binsztok, P. Gallinari, Un algorithme en ligne pour la détection de nouveauté dans un flux de documents, Journées Internationales d'Analyse Statistique des Données Textuelles, JADT, 2002. (www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/tocJADT2002.htm).

C. Buckley, D. Harman, Reliable Information Access Final Workshop Report, Janvier 2004, nrrc.mitre.org/NRRC/Docs_Data/RIA_2003/ria_final.pdf.

C. Buckley et H. Voorhees, Evaluating evaluation measure stability, pp 33 – 40, ACM Conference on Research and Development in Information Retrieval, 2000.

J.-M. Bruneau, IDELIANCE, logiciel de rupture pour l'intelligence économique ? Un cas d'application sur les signaux faibles, Veille Stratégique, Scientifique et Technologique, VSST, 2001.

J. Carbonell, J. Goldstein, The use of MMR, diversity-based reranking for reordering documents and producing summaries Full text, ACM Conference on Research and Development in Information Retrieval, p 335-336, 1998.

K. Collins-Thompson, P. Ogilvie, Y. Zhang, J. Callan, Information filtering, Novelty detection, and named-page finding, Text Retrieval Conference, p 107-118, 2002.

M.-L. Corral, J. Mothe, How to retrieve and display long structured documents ?, congrès Basque International Workshop on Information Technology, BIWIT'95, p 10-19, 1995.

T. Dkaki, J. Mothe, Novelty track at IRIT-SIG, actes de Text REtrieval Conference, p 332-336, 2002.

D. Harman, C. Buckley, "RIA and "Where can IR go from here?"" Workshop à SIGIR 2004, 2004.

D. Harman, Overview of the TREC 2002 novelty track, actes de Text Retrieval Conference, p 46-55, 2002. (trec.nist.gov)

D. Hull et G. Grefenstette, A detailed analysis of English stemming algorithms, Xerox Technical Report, 1996, http://www.xrce.xerox.com/publis/mltt/mltt-023.ps.

H. Kazawa, T. Hirao, H. Isozaki, E. Maeda, A machine learning approach for QA and Novelty tracks: NTT system description, Text Retrieval Conference, p 472-475, 2002.

K.L. Kwok, P. Deng, N. Dinstl, M. Chan, TREC 2002 Web, Novelty and Filtering Track Experiments using PIRCS, Queens College, CUNY , actes de Text Retrieval Conference, 2002. (trec.nist.gov).

L. Lebart, A. Morineau, M. Piron, Statistique exploratoire multidimensionnelle, Dunod, 1995.

M. Nelson, The effects of query characteristics on retrieval results in the TRES retrieval tests, Actes de la conférence CAIS/ACSI, 1995.

J.M. Ponte, W.B. Croft, A language modelling approach to information retrieval, Research and Development in Information Retrieval, SIGIR'98, p 275-281, 1998.

C. Roux, B.Dousset, Une méthode de détection des signaux faibles: application à l'émergence des dendrimères Veille Stratégique, Scientifique et Technologique, VSST, 2001.

G. Salton, J. Allan, C. Buckley, Automatic structuring and retrieval of large text files, communication de l'ACM, 37(2), p 97-108, 1994.

B. Schiffman, Experiments in Novelty Detection at Columbia University, actes de Text Retrieval Conference, p 188-196, 2002. (trec.nist.gov).

E. Sormunen, A retrospective evaluation method for exact-match n best-match queries applying an interactive query performance analyser, Actes de la conférence ECIR, 2002.

C. Stanfill, D.L. Waltz, Statistical methods, artificial intelligence, and information retrieval, Text-based intelligent systems: current research and practice in information extraction and retrieval, Ed. P.S. Jacobs, p 215-226, 1992.

Trec 2002, annexes, http://trec.nist.gov/pubs/trec11/t11_proceedings.html, APPENDICES, 2002.

R. Wilkinson, Effective retrieval of structured documents, actes de la 17ième Conference ACM-SIGIR, Research and Development in Information Retrieval, p 311-317, 1994.

M. Zhang, R. Song, C. Lin, S. Ma, Z. Jiang, Y. Jin, Y. Liu, L. Zhao, et S. Ma, Expansion-based technologies in finding relevant and new information: THU TREC2002: Novelty Track Experiments, actes de Text Retrieval Conference, p 586-590, 2002. (trec.nist.gov).