Recherche de la nouveauté dans les textes: une tâche difficile

Taoufiq Dkaki, Gilles Hubert**, Josiane Mothe *,**, Eric Orain ****
{dkaki, hubert, mothe}@irit.fr tél: 05 61 55 63 22

(*) *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*

Mots-clés :

Recherche d'information, détection de la nouveauté, analyse de résultats, typologie de requêtes

Key words :

Information retrieval, novelty detection, analysis of IR results, typology of queries

Palabras claves :

Recuperación de datos, detección de la novedad, análisis de resultados, tipología de preguntas

Résumé

Les systèmes de recherche d'information visent à restituer l'information répondant à un besoin d'information que l'utilisateur exprime au travers d'une requête. Dans cet article, nous nous intéressons à la tâche de détection de la nouveauté dans les textes. Nous présentons les éléments d'une première étude sur les résultats de l'évaluation d'un ensemble de systèmes répondant à cette tâche dans le cadre du programme d'évaluation *Text Retrieval Conference*.

1 Introduction

Les systèmes de recherche d'information visent à restituer l'information répondant à un besoin d'information que l'utilisateur exprime au travers d'une requête. La majorité des systèmes restitue une liste de références à des documents auxquels l'utilisateur choisit ou non d'accéder pour en lire le contenu. 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. Les moteurs s'attachent plutôt à limiter le bruit dans les réponses fournies ainsi que le silence. Pourtant, la satisfaction de l'utilisateur par rapport aux réponses d'un système dépendent de l'objectif de l'utilisateur : veut-il vérifier une hypothèse ? Connaître la réponse à une question précise ? ou 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.

C'est dans ce contexte de systèmes adaptatifs que nous proposons dans cet article les résultats d'une première étude sur le problème de la détection de la redondance ou le problème associé : la détection de la nouveauté. Ce problème de la détection de la nouveauté n'est pas nouveau. En veille, il peut se traduire par la détection des signaux faibles ou des phénomènes atypiques [Bruneau, 2001], [Roux et Dousset, 2001]. Cependant, l'évaluation de ces méthodes reste difficile. Deux programmes d'évaluation internationaux s'intéressent à ces problématiques. Le programme TDT (Topic Detection and Tracking[1]) 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. Le programme TREC (Text Retrieval Conference[2]) 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 des documents qui sont pertinents par rapport à une requête, d'autre part à déterminer parmi ces passages lesquels apportent de la nouveauté.

Dans cet article, nous rapportons les résultats d'une première analyse des résultats obtenus par les différents participants à cette dernière tâche pour essayer de mesurer sa difficulté 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 proposons une revue des mécanismes de détection de la nouveauté dans la tâche TREC. La section 3 est dédiée à la présentation de la tâche nouveauté de TREC et aux mesures utiliser pour évaluer la performance des systèmes. Nous discutons ces résultats dans les sections 4 et 5 afin d'essayer de dégager des éléments utiles pour la construction de nouvelles méthodes de détection de la nouveauté.

2 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.

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

[2] trec.nist.gov

Dans le cadre de la tâche définie par TREC (détection des phrases nouvelles), différentes approches ont été proposées. 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, 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[3]. 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, Mothe, 2003] 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. On peut toutefois noter qu'un workshop associé au congrès ACM-SIGIR (Sheffield, Juillet 2004) est proposé cette année sur ce thème : *RIA and "Where can IR go from here?"*. Ce workshop se propose d'étudier certains phénomènes locaux [Harman, Buckley, 2004] en se basant sur les différences entre systèmes (modèles et techniques de recherche d'information utilisés).

3 La tâche nouveauté de TREC

3.1 Recherche des phrases nouvelles

L'étude présentée dans cet article s'intéresse à la détection de passages de documents potentiellement nouveaux pour l'utilisateur. 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.

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

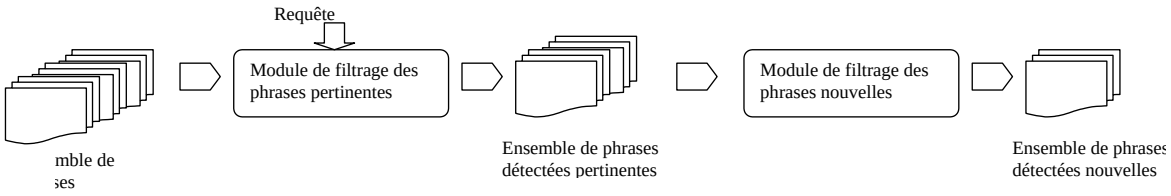

Figure 1: Phases de détection de la nouveauté

3.2 Collections d'évaluation

Caractéristiques de la collection de test

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.

	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

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

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

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

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

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.

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'* un outil et les paramètres associés.

4 Analyse des résultats soumis à TREC

4.1 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 terme de précision et de rappel.

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).

4.2 Requêtes difficiles?

Cette section s'intéresse à étudier globalement 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 2 indique pour chacune des requêtes le nombre de systèmes pour lesquels la mesure F est nulle (43 systèmes ont participés). La valeur apparaît en noir. Par exemple, pour la requête 1, 3 systèmes indiquent une valeur nulle de la mesure F. Ce même graphique indique la valeur moyenne de la mesure F. Pour cette mesure, l'échelle est à diviser par 100. Par exemple, pour la requête 1, la mesure F moyenne sur l'ensemble des systèmes est de 0,01.

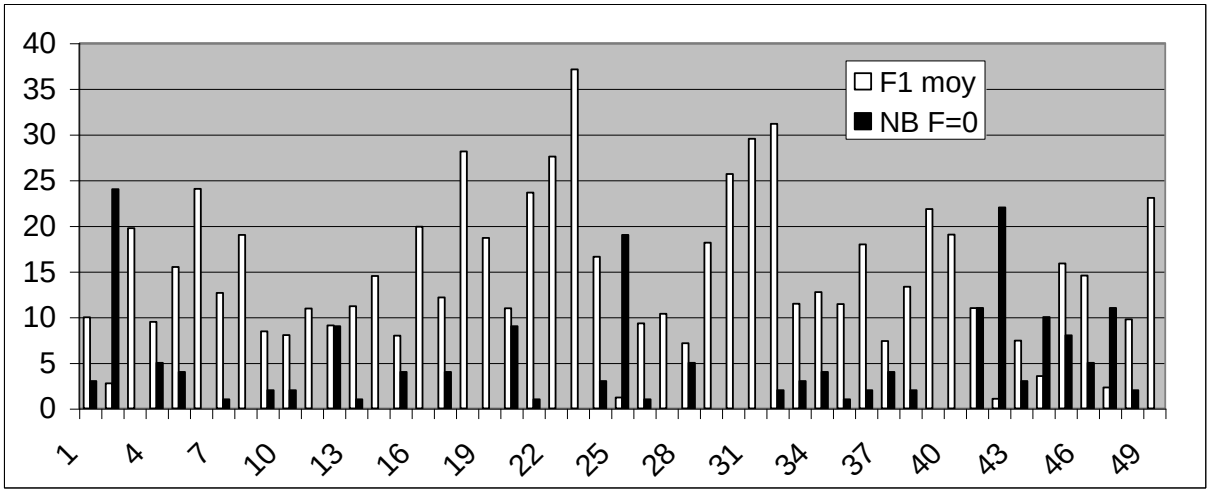

Figure 2: Nombre de systèmes pour lesquels la mesure F est nulle et mesure F moyenne pour chacune des requêtes.

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.

Parmi les requêtes pour lesquelles la mesure F moyenne est la plus faible figurent 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.

Requête	Pert	%Total	Nouv	%Pert
305	15	2.01	15	100
314	25	2.35	25	100
316	22	0.99	18	81.82
322	34	4.57	34	100
325	21	1.25	21	100
330	29	3.49	27	93.1
342	17	2.17	17	100
351	6	0.75	5	83.33
356	10	1.2	9	90
362	47	5.15	46	97.87
364	42	3.5	42	100
368	71	4.63	66	92.96
377	3	0.19	3	100

Requête	Pert	%Total	Nouv	%Pert
312	5	0.87	5	100
315	18	3.08	11	61.11
317	23	4.19	23	100
323	65	4.57	60	92.31
326	10	1	8	80
339	12	1.6	11	91.67
345	47	5.19	47	100
355	103	3.94	78	75.73
358	40	4.8	37	92.5
363	11	2.08	10	90.91
365	34	2.8	34	100
369	13	1.79	12	92.31
381	19	1.38	19	100

382	41	1.83	24	58.53
386	43	4.3	41	95.35
394	21	2.45	21	100
405	40	3.75	37	92.5
407	32	2.46	29	90.62
410	17	1.92	15	88.24
414	29	3.62	25	86.21
419	50	3.32	36	72
427	14	0.31	11	78.57
433	11	1.72	7	63.64
445	10	1.07	5	50
449	57	3.65	57	100

384	23	1.41	23	100
388	56	4.57	56	100
397	29	6.18	28	96.5
406	10	1.79	10	100
409	17	3.26	12	70.59
411	21	1.86	19	90.48
416	36	1.96	30	83.33
420	18	1.4	18	100
432	9	0.96	8	88.89
440	19	1.35	19	100
448	20	2.25	20	100

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

4.3 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 obtenu 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).

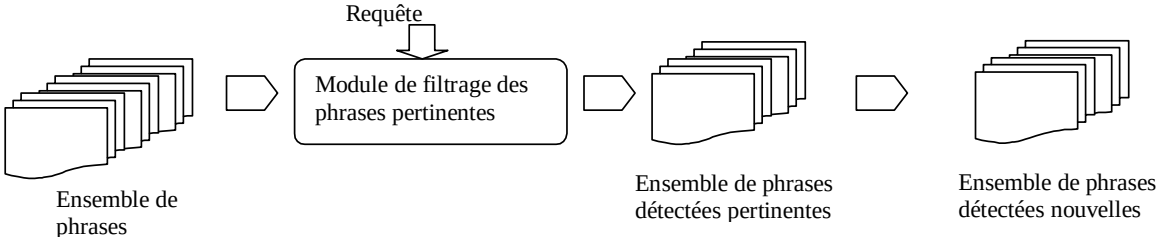

Figure 3: Phases de sélection des phrases nouvelles modifiée

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 (puisqu'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.

Les plus fortes améliorations (de l'ordre de 500%) sont obtenus 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 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 obtient une amélioration de plus de 13% de la mesure F. Aucun système ne détériore globalement les résultats obtenus.

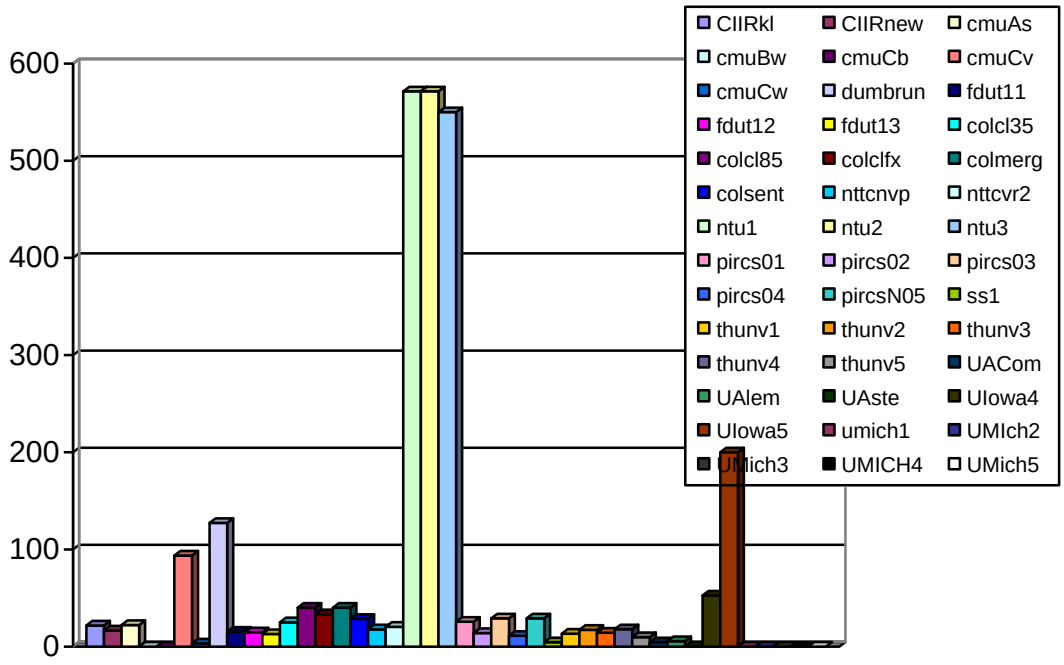

Figure 4 : Amélioration de la mesure F obtenue en considérant toutes les phrases supposées pertinentes comme nouvelles

5 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 sur la mesure F. En effet, ce critère correspond à un indice de performance global. Nous avons donc essayer d'étudier les corrélations qui peuvent exister entre systèmes ou entre requêtes.

5.1 Méthodologie

L'analyse porte d'abord sur la classification des requêtes. Nous essayons de regrouper les requêtes en fonction des performances obtenus 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?

L'analyse porte ensuite sur la classification des systèmes par rapport aux résultats obtenus. L'idée est d'essayer de détecter les groupes de systèmes qui aboutissent aux mêmes types de résultats. A terme, l'étude conjointe des mécanismes utilisés dans ces systèmes et des types de requêtes doit être evisagée.

Nous utilisons pour cela les méthodes de classification hiérarchique et l'analyse en correspondance principale.

5.2 Classification des requêtes

La classification des requêtes est obtenue en considérant les requêtes comme des individus et les systèmes comme des variables. La mesure étudiée est la mesure F. Le dendrogramme correspondant est présenté figure 5. La classe la plus à droite dans la figure 5 (requêtes 420, 377, 445, 312) correspond aux requêtes difficiles: celles pour lesquelles la majorité des systèmes échouent. La deuxième classe en partant de la droite correspond également à des requêtes pour lesquelles la mesure F est faible.

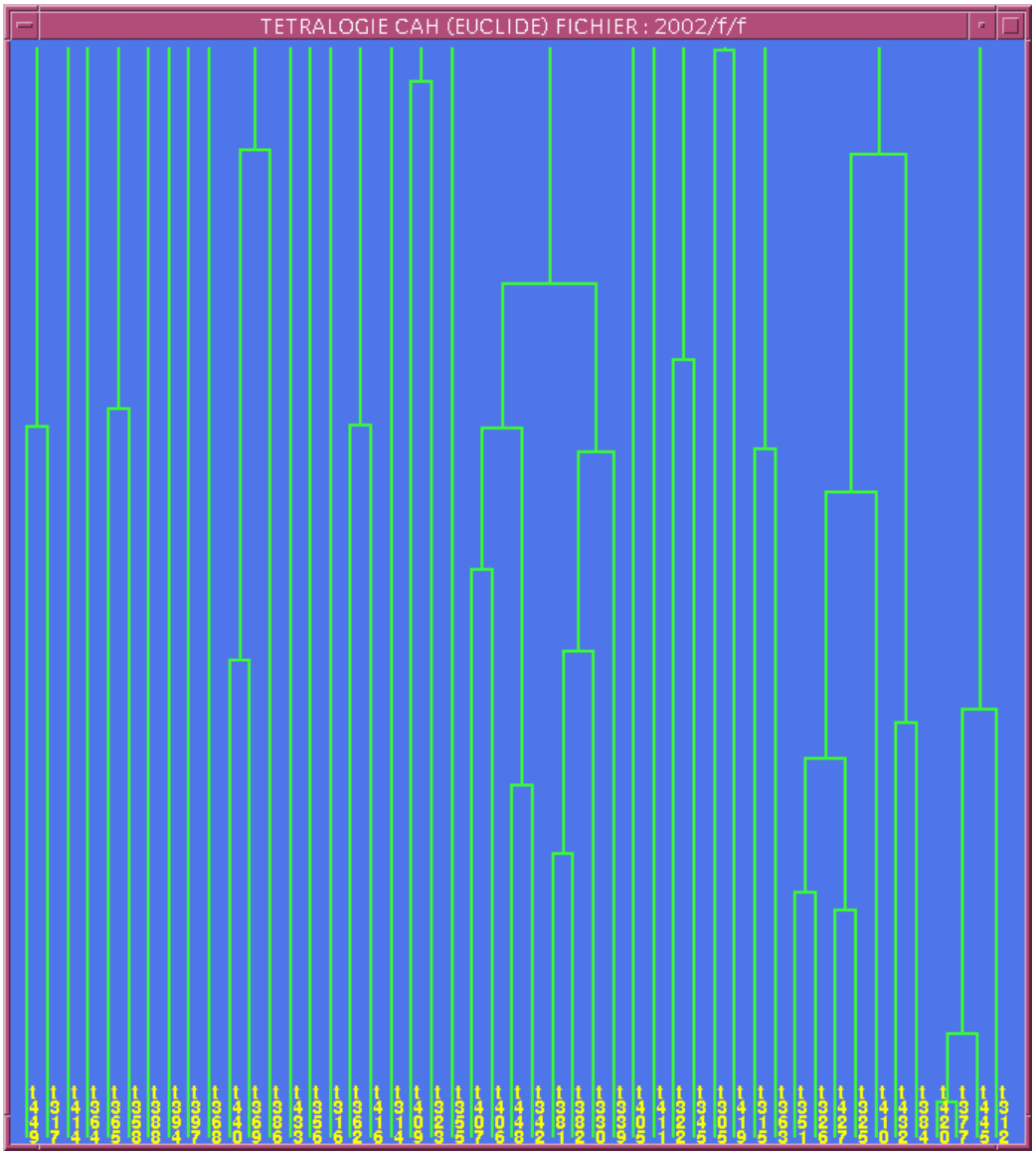

Figure 5 : Classification des requêtes
(individus: requêtes, variables: systèmes, mesure: mesure F)

La figure 6 présente les résultats (individus) de l'ACP sur le même jeu de données.

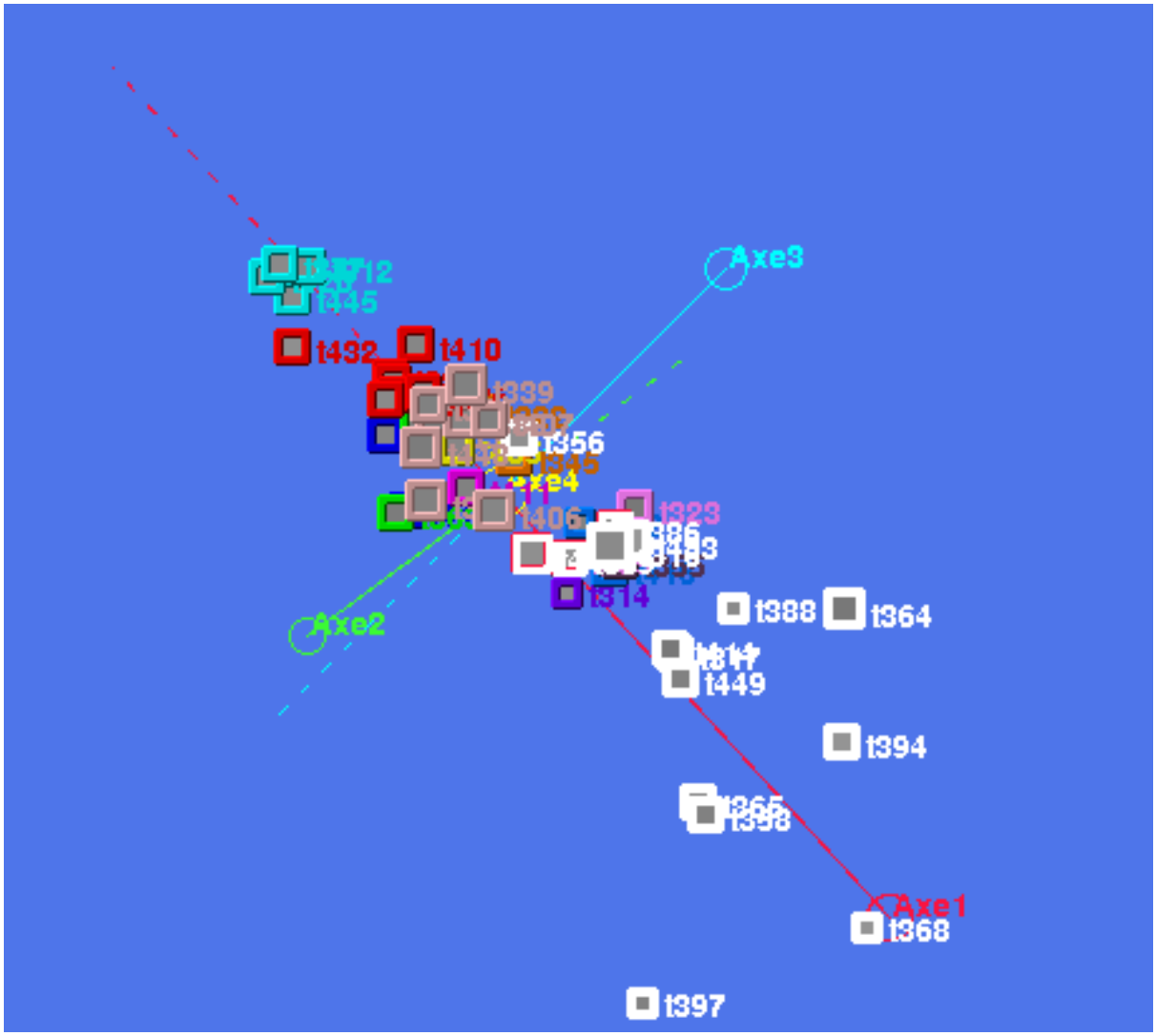

Figure 6 : Classes de requêtes visualisée sur l'ACP associée

5.3 Classification des systèmes

Cette section s'intéresse à la classification des systèmes par rapport aux résultats qu'ils ont obtenus pour les différentes requêtes. La figure 7 présente les résultats de l'ACP précédente (visualisation des variables). La classification présentée figure 8 est basée sur une classification ascendante hiérarchique dans laquelle les individus sont les systèmes, les variables correspondent aux requêtes et la mesure correspond à la mesure F.

Dans les figures 7 et 8, les noms de systèmes donnent une indication sur l'outil utilisé à travers les premières lettres du nom. 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 de termes complets (Com) lors de l'indexation des documents.

La classification des systèmes montre clairement que les différentes versions d'un même outil ont le même comportement et se trouvent dans une même classe.

La position des versions de *fdu* est à noter (cf figure 7, coin bas gauche). Les résultats numériques 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 système

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).

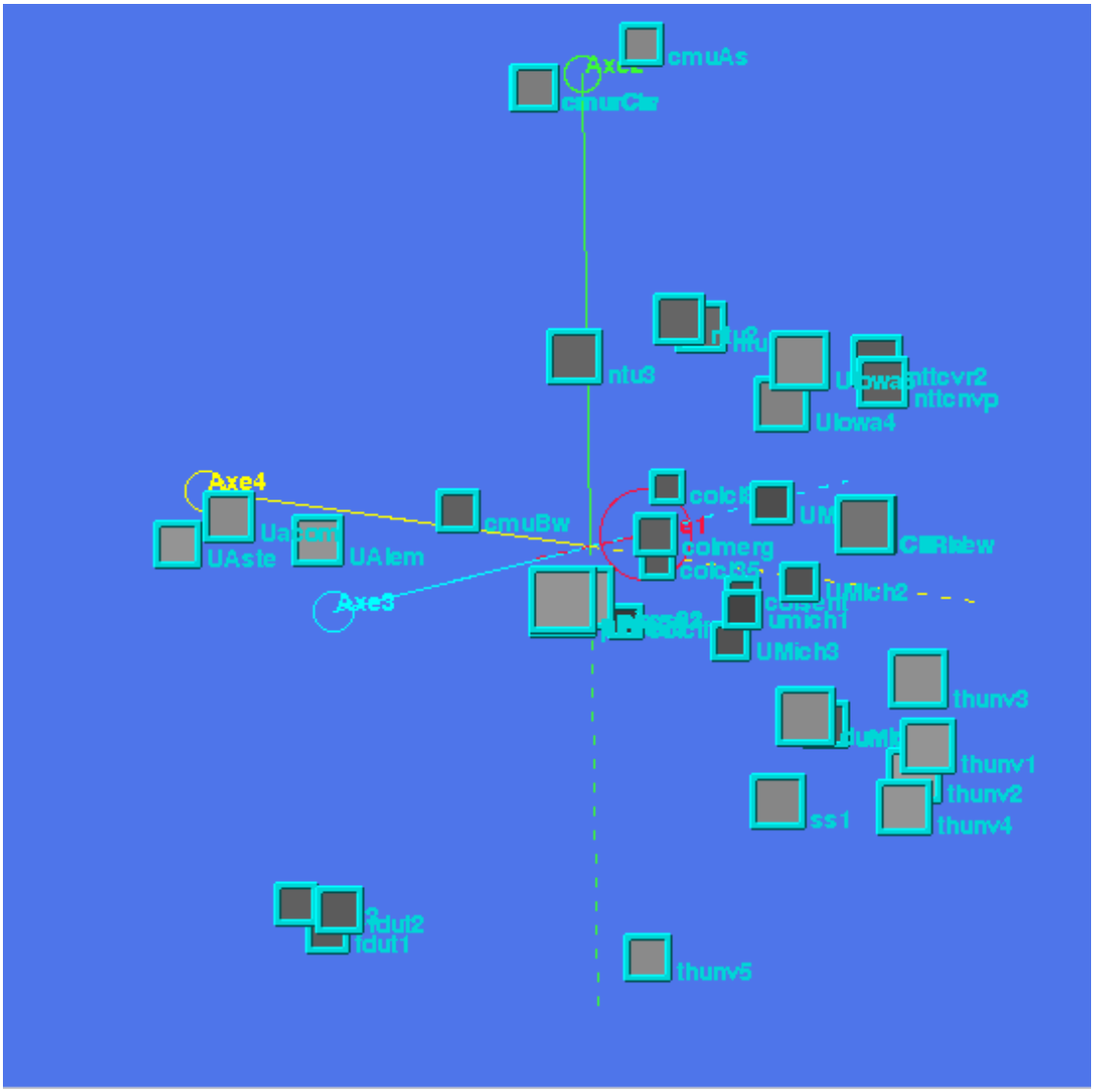

Figure 7 : Visualisation des variables de l'ACP (individus: requêtes, variables: systèmes, mesure: mesure F)

Les classes obtenues figure 8 montrent que le premier élément de regroupement est l'outil utilisé, même lorsque ses différentes versions semblent éloignées en terme de méthodes utilisées. Par exemple dans le cas de *UACom*, *UAStem* et *UALem* , la représentation de l'information est différente dans la mesure où des traitement linguistiques sont ou non appliqués. Cependant, ces trois versions sont très proches.

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.

Les figures 9a et 9b détaillent les résultats obtenus par quelques systèmes choisis en fonction de leurs résultats ou de leur typologie. Pour chaque requête, elles indiquent la valeur de la mesure F obtenue, en la positionnant par rapport à la moyenne des systèmes. Les requêtes sont disposées sur l'axe des ordonnées par ordre de moyenne de mesure F obtenue

sur l'ensemble des systèmes. *Thumv1* correspond au système ayant obtenu la meilleure mesure F, en moyenne sur l'ensemble des requêtes. *Colcl35* est le système qui a obtenu la moyenne la plus faible. IRIT correspond aux résultats que nous avons obtenus avec notre propre système. Nous avons ensuite sélectionné trois autres outils qui semblent être les plus caractérisés compte tenu de l'ACP présentée figure 7. Pour chacun, nous avons choisi la version qui permettait d'obtenir une mesure F intermédiaire (ni la plus forte pour cet outil, ni la plus faible).

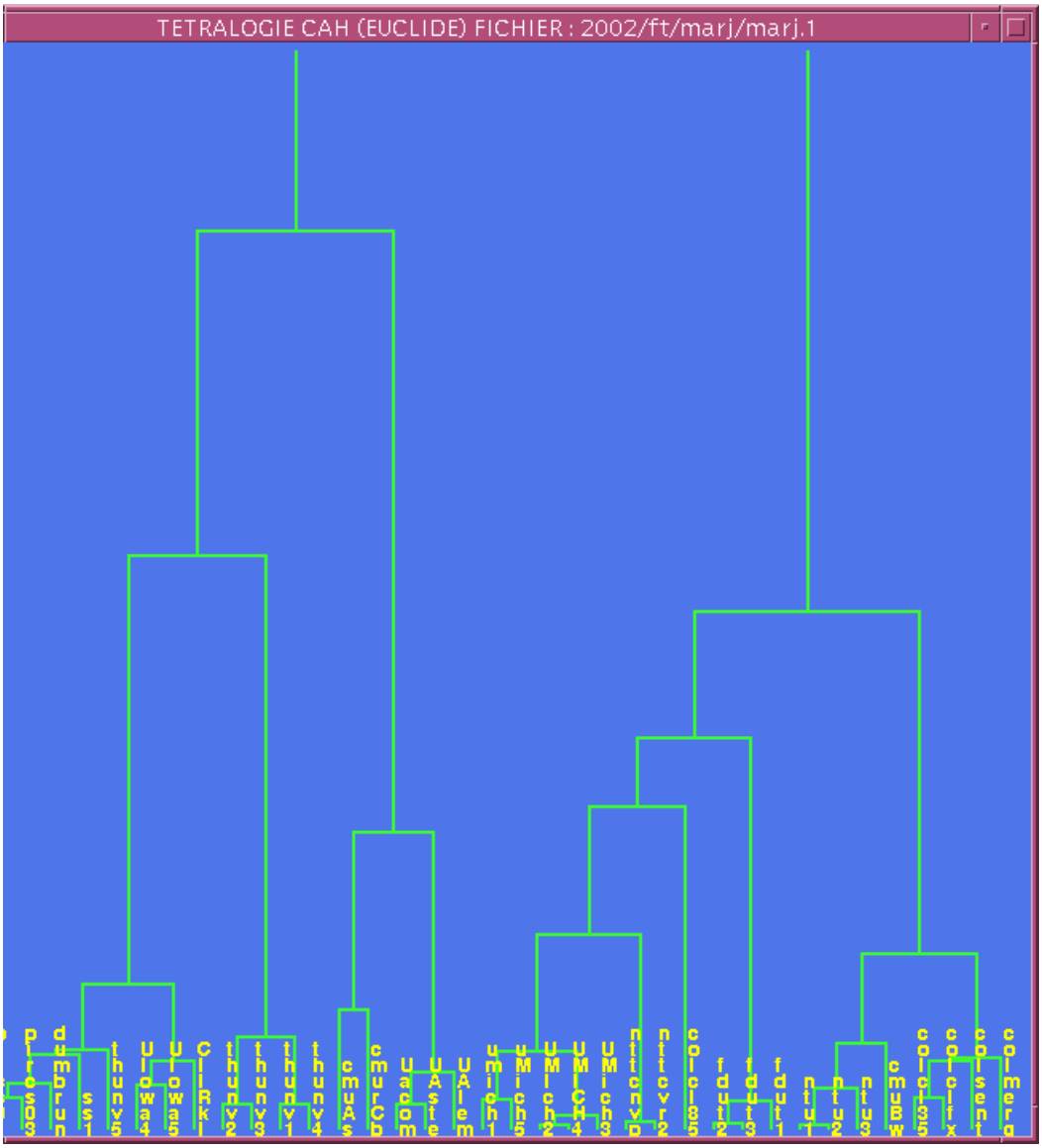

Figure 8 : Classification des systèmes (individus: systèmes, variables: requêtes, mesure: mesure F)

Il est difficile d'extraire une tendance générale des graphiques présentés dans les figures 9a et 9b. Les systèmes ne sont pas systématiquement meilleurs ou moins bons pour l'ensemble des requêtes. Certains systèmes (voir par exemple *cmuAs*) obtiennent de très bons résultats là où d'autres systèmes échouent; mais ces mêmes systèmes échouent pour des requêtes qui globalement semblent assez faciles (voir par exemple cmuAs pour la 6^{ième} et 25^{ième} requête). Même le meilleur système a des résultats très en dessous de la moyenne pour certaines requêtes (par exemple la 4^{ième} requête)

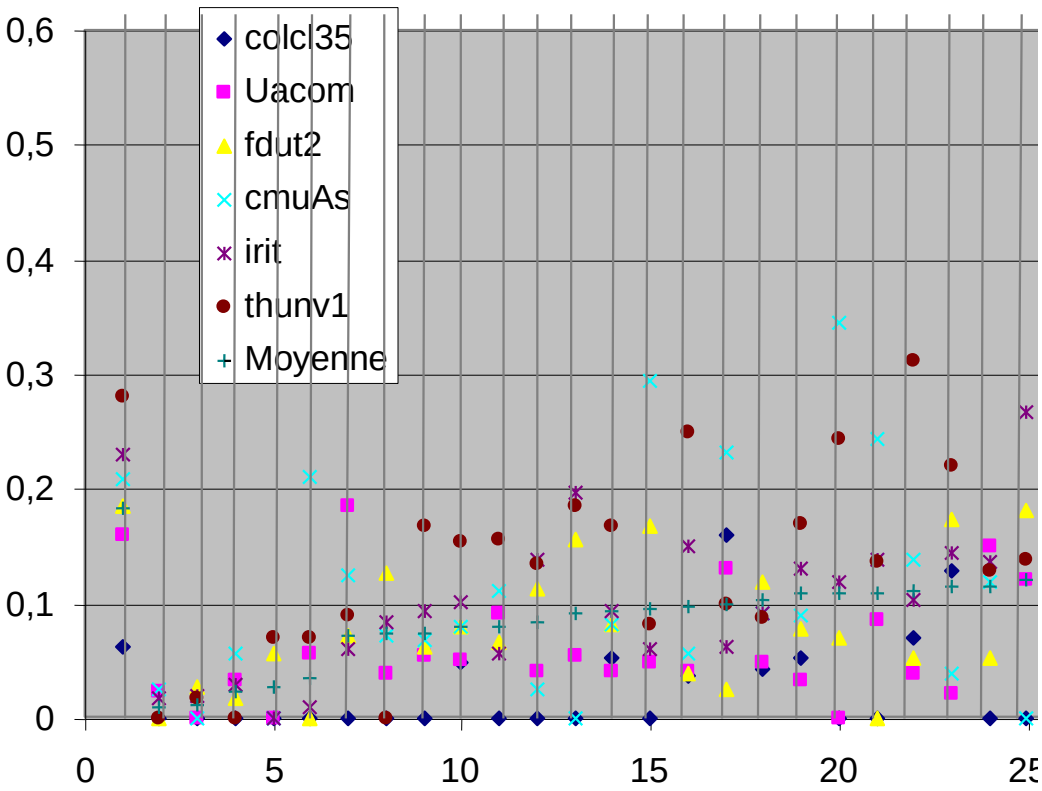

Figure 9a : Mesure F obtenue par quelques systèmes pour chaque requête

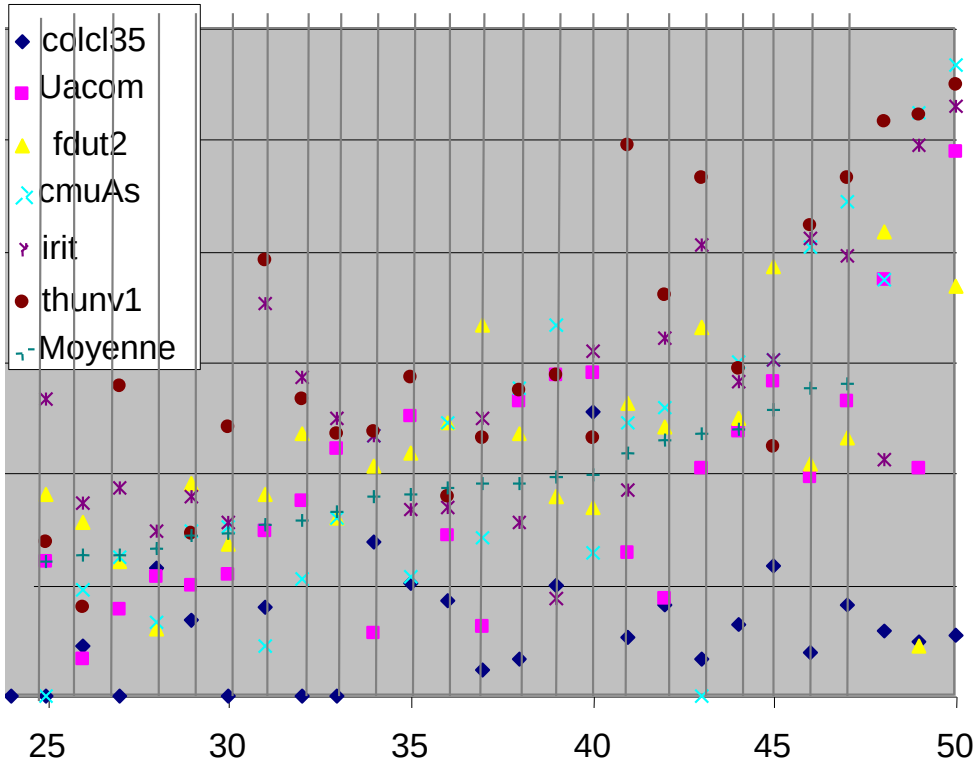

Figure 9b : Résultats de quelques systèmes (suite des requêtes)

6 Conclusion

Dans cet article, nous avons présenté une première analyse des résultats obtenus par 43 systèmes (ou version de systèmes) utilisés pour la détection de la nouveauté sur une collection de test commune. Les premiers résultats ont montré la difficulté de la tâche de détection de la nouveauté. 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 le système utilisé, alors même que beaucoup de systèmes 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).

Ce travail va être prolongé par une analyse plus fine des groupes de requêtes et des groupes de systèmes issus de différents outils. L'objectif de cette étude sera d'analyser chaque groupe de requêtes afin de savoir s'il existe d'autres caractéristiques communes (type de requêtes, généricité des termes utilisés, type et quantité de résultats attendus, etc.). L'objectif à terme est de savoir s'il serait envisageable de créer un système adaptatif qui décide lui-même du mécanisme à appliquer en fonction du type de requête rencontré.

Les premiers résultats obtenus dans cette étude et ceux menés via le programme TREC [Harman, 2004] ne sont pas très optimistes sur ce point, mais nous pensons qu'il est néanmoins intéressant d'analyser plus en détail cette possibilité.

7 Références

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

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. (http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/tocJADT2002.htm)

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, pp 335-336, 1998.

K. Collins-Thompson, P. Ogilvie, Y. Zhang, J. Callan, Information filtering, Novelty detection, and named-page finding, actes de Text Retrieval Conference, p 107-118, 2002. (trec.nist.gov)

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

D. Harman, C. Buckley, “RIA and "Where can IR go from here?"” proposition de workshop à SIGIR 2004, 2004.

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

H. Kazawa, T. Hirao, H. Isozaki, E. Maeda, A machine learning approach for QA and Novelty tracks: NTT system description, actes de Text Retrieval Conference, pp 472-475, 2002. (trec.nist.gov)

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).

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

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

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, pp 586-590, 2002. (trec.nist.gov).