Restituer l'information utile à l'utilisateur : visualisation de la pertinence et de la nouveauté dans les textes

Taoufiq Dkaki (1,3),Chaoukhi Mhamedi (1), Josiane Mothe (1,2),

(1) Institut de Recherche en Informatique de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 04

(2) Institut Universitaire de Formation des Maîtres, 56 Avenue de l'URSS, 31400 Toulouse

(3) ISYCOM-GRIMM, Toulouse le Mirail, 5 allées Antonio Machado, 31058 Toulouse Cedex 9

{mothe/dkaki}@irit.fr tél: 05 61 55 63 22

RÉSUMÉ.

Les systèmes de recherche d'information restituent généralement des documents entiers en réponse à une requête de l'utilisateur. Ce niveau de granularité n'est pourtant pas satisfaisant en particulier lorsque les documents sont longs. Dans ce cas, l'utilisateur peut être plus satisfait si le système lui permet un accès direct aux passages les plus pertinents. Une autre étape dans l'adaptation aux besoins de l'utilisateur correspond à lui indiquer quels sont les passages redondants ou quels sont les passages qui apportent des éléments nouveaux. Dans cet article nous proposons un mécanisme de sélection et de visualisation des passages pertinents et des passages nouveaux qui permet à l'utilisateur d'être plus efficace lors de la consultation des réponses d'un système de recherche d'information.

MOTS-CLES. :

Rechercher d'information, redondance, nouveauté, pertinence, recherche de passages, mise en évidence de passages

1 Introduction

Les systèmes de recherche d'information ont pour objectif de restituer l'information pertinente par rapport à un besoin d'information exprimé sous forme de requêtes. Traditionnellement, la recherche d'information est synonyme de recherche de documents (Salton, 71), (Rijsbergen, 79). Ainsi, la granularité de traitement de l'information se situe au niveau du document entier à la fois pour l'indexation (la représentation de l'information pour son accès ultérieur), la recherche et la restitution. Les systèmes de recherche d'information restituent donc des listes de documents susceptibles de répondre au besoin exprimé par l'utilisateur. Pourtant ce niveau de granularité n'est pas toujours adapté. Lorsque les documents sont longs par exemple, l'utilisateur peut avoir du mal à se focaliser sur la partie réellement importante par rapport à son besoin. La restitution d'un document complet peut également être inadaptée dans le cas où le document traite de différents aspects ou de différents thèmes (un rapport de projet ou une description des cours dispensés par un organisme par exemples). L'utilisateur peut également avoir des difficultés à comprendre pourquoi un document a été restitué par le système. Les listes de documents traditionnellement restituées par les SRI n'apporte aucune aide. Différentes propositions s'attachent à répondre toutefois à ces problèmes.

Lorsque les documents sont longs, la simple mise en évidence des termes issus de la requête dans les textes restitués offre une aide précieuse à l'utilisateur. En effet, cette mise en évidence explique à l'utilisateur pourquoi le document lui a été restitué et d'autre part le guide dans son choix de prise en compte des documents restitués. Des processus plus sophistiqués remettent en cause la granularité au niveau document. En particulier, le développement de

ressources documentaires structurées (SGML et maintenant XML) a permis un ensemble de propositions qui s'appuient sur la structure générique des documents (Wilkinson, 1994), (Salton, 1994), (Corral, 1995), (Fuhr, 2002). Selon ces approches les parties les plus pertinentes des documents sont restituées à l'utilisateur plutôt que les documents complets. Plus récemment (Harman, 2002), TREC[1] (www.nist.gov) s'est intéressé à un niveau de granularité plus fin dans le cadre de documents non structurés : celui de la phrase. Ce niveau est parfois inadapté car trop fin. Cependant lorsqu'il est combiné à l'étude de la redondance, il peut s'avérer pertinent. En effet, plusieurs parties peuvent être pertinentes pour un utilisateur, mais si elles sont redondantes, une seule peut suffire à répondre au besoin de l'utilisateur. Ainsi la tâche 'Novelty' du programme TREC s'intéresse à étudier la nouveauté dans les parties (phrases) restituées à l'utilisateur ; c'est à dire que les systèmes doivent éviter de restituer des informations que le système considère comme redondantes par rapport à des éléments déjà consultés par l'utilisateur ; les aspects interface ne sont pas considérés dans ce programme d'évaluation.

Nous nous intéressons au double problème de la recherche des parties pertinentes et des parties nouvelles dans le cadre des systèmes de recherche d'information textuelle. Le cadre d'étude que nous avons choisi est celui défini par TREC. Cela nous a permis d'évaluer les mécanismes que nous proposons et de les confronter à la communauté internationale du domaine. Cet article s'intéresse en particulier à l'interface qui est nécessairement associée à ce type de processus. Les problèmes auxquels il faut répondre concernent au moins trois aspects :

- permettre à l'utilisateur d'accéder directement aux parties pertinentes et aux parties nouvelles,
- aider l'utilisateur dans le choix de lecture des éléments restitués,
- permettre à l'utilisateur de garder le contexte des parties restituées pour garder le sens global des éléments qu'il lira.

Cet article comprend différentes sections qui permettent de décrire notre approche. La section 2 présente un état de l'art du domaine. La section 3 présente la méthode que nous avons définie pour détecter les passages pertinents et les passages nouveaux ainsi que les résultats d'évaluation que nous avons obtenus grâce à cette méthode. La section 4 décrit l'interface qui permet un affichage personnalisé en fonction des choix de l'utilisateur.

2 Travaux du domaine

La restitution de passages plutôt que des documents entiers vise à mieux satisfaire l'utilisateur en lui permettant de se focaliser sur les éléments correspondant réellement à sa recherche. Dans (Wilkinson, 1994), les documents SGML sont découpés en unités documentaires en fonction de leur structure induite par la DTD ou structure générique. Chaque unité ainsi obtenue est indexée et le système de recherche d'information défini un calcul de ressemblance entre la requête de l'utilisateur et les parties de documents. (Corral, 1995) est basé sur le même type d'approche mais propose également un mécanisme de navigation dans la structure des documents auxquels appartiennent les parties retrouvées. Ce mécanisme permet à l'utilisateur de connaître le contexte associé aux éléments restitués. Ces

[1] TREC : Text Retrieval Conference est un programme international d'évaluation des systèmes de recherche d'information. Il comprend un certain nombre de tâches ('tracks') qui évoluent au fur et à mesure des éditions.

travaux se basent sur des documents possédant une structure marquée. Dans de nombreux cas, une telle structure n'existe pas.

Les travaux réalisés dans le cadre de TREC s'intéressent à évaluer la détection de la pertinence au niveau des phrases dans des documents dont la structure n'est pas marquée. Il faut noter que TREC n'apporte aucun élément quant aux interfaces qui permettront aux utilisateurs de tirer profit de cette information. Du point de vue de la détection des phrases pertinentes, la majorité des systèmes évalués à TREC ont considéré les phrases comme des documents de petite taille, chacune étant indexée (Harman, 2002). Les mécanismes de recherche sont alors appliqués aux phrases comme s'il s'agissait de documents. (Schiffman, 2002) utilise une phase préliminaire afin 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. (Kwok , 2002) a utilisé WordNet pour réaliser l'expansion à la fois des requêtes et des phrases. Dans (Dkaki, 2002) la sélection de phrases est également basée sur la ressemblance des représentations des phrases et celles des requêtes. Une originalité de l'approche concerne le fait que ces représentations se basent sur une catégorisation de termes en fonction de leur représentativité. Tous ces travaux reposent sur une granularité qui se situe au niveau de la phrase.

Le niveau de granularité inférieur correspond aux mots. C'est ce niveau de granularité qui est généralement utilisé dans les interfaces pour permettre à l'utilisateur de localiser au sein des documents restitués les éléments qui sont susceptibles d'être les plus intéressants. Ainsi, dans certaines interfaces en particulier sur le Web (Copernic[2], Xinquery[3]), les termes de la requête apparaissant dans le contenu des documents sont mis en évidence.

La recherche de la nouveauté quant à elle se base sur différentes techniques. Du point de vue de leur détection, (Allan, 2003) utilise une représentation des phrases à base de langage de modèles (Ponte, 1998). (Dkaki, 2002) utilise une fonction de décision qui combine la similarité de la phrase considérée avec chacune des phrases déjà traitées et avec une phrase virtuelle correspondant à l'union des phrases déjà traitées. (Kazawa, 2002) sélectionne les phrases nouvelles parmi les phrases pertinentes en se basant sur la mesure de la pertinence marginale maximum (MMR).

Les interfaces aidant l'utilisateur à analyser les résultats d'une recherche pour qu'il se centre sur les parties pertinentes ou les parties nouvelles sont peu nombreuses et se restreignent à faire apparaître en surbrillance les termes de la requête qui ont permis de retrouver le document en cours d'affichage.

Notre interface apporte un niveau de visualisation supérieur et permet de mettre en évidence les passages ou les phrases susceptibles d'être les plus intéressantes pour l'utilisateur. Celui-ci dispose d'une interface contextuelle grâce à laquelle il peut décider de son centre d'intérêt en fonction de l'usage qu'il souhaite faire de l'information. Il peut ainsi choisir de n'afficher que les phrases effectivement pertinentes ou bien éviter la redondance en n'affichant que les passages qui apportent de la nouveauté. Enfin, s'il le souhaite, le contexte général de ces éléments peut être retrouvé en affichant les documents entiers tout en gardant la possibilité de mettre en évidence le type de phrases choisies.

---

[2] http://www.copernic.com

[3] http://www.dtsearch.com/PLF_engine_2.html

3 Détection de la pertinence et de la nouveauté

Nous avons étudié différents types de résultats auquel peut s'intéresser un utilisateur :

- les documents supposés pertinents : la majorité des systèmes de recherche restitue ce type de résultat,
- les passages pertinents : il s'agit ici de se focaliser sur les parties de documents qui correspondent vraiment à la requête . Ce type de résultat est particulièrement intéressant lorsque les documents sont longs et qu'ils abordent différents aspects ou différents thèmes plus ou moins proches de la requête,
- les passages nouveaux : il s'agit de permettre à l'utilisateur de s'affranchir de la redondance de l'information afin d'éviter au maximum de le confronter plusieurs fois à la même information.

Les techniques de détection de la pertinence et de la nouveauté ne sont pas traitées en détail dans cet article, nous en donnons tout de même les principes et les formules de calcul. Ces mécanismes ont été développés pour la tâche 'Novelty' de TREC (Dkaki, 2002). Les détails ayant permis de choisir les valeurs de paramètres peuvent être trouvés dans (Dkaki, 2002) et (Dkaki, 2003).

Cette méthode originale permet de détecter les phrases pertinentes et les phrases nouvelles en caractérisant les termes d'indexation selon trois classes : très pertinents, pertinents, non pertinents. Chaque requête et chaque phrase sont automatiquement caractérisées par un ensemble de termes. La fonction de ressemblance entre requête et phrases utilise l'ensemble des termes communs afin de décider de la pertinence des phrases. Des analyses syntaxiques permettent d'améliorer le traitement des textes lors de leur caractérisation. Les phrases nouvelles sont obtenues par filtrage des phrases pertinentes.

L'interface que nous présentons dans la section 4 se base sur les résultats de ces modules.

3.1 Représentation des informations

Une première étape consiste à représenter les textes. Cette phase correspond à l'indexation automatique qui vise à attacher des index (termes ou concepts) représentatifs et permettant une discrimination des textes.

3.1.1 Pré-traitement des textes

Quel que soit le texte (requête, passages de document) il est pré-traité pour en extraire les termes représentatifs. Pour ce pré-traitement, nous nous appuyons sur un processus utilisé en recherche d'information (Baeza, 1999) :

1. Suppression des mots vides en référence à une liste pré-établie,
2. Les mots restants sont normalisés pour obtenir la racine de chacun d'eux. Cette normalisation s'appuie sur un dictionnaire qui contient 21291 entrées.

Dans (Dkaki, 2003) nous avons également étudié l'extraction de groupes de mots plutôt que de mots simples. La différence des résultats en terme de pertinence (silence et bruit) des passages retrouvés n'est pas significative.

3.1.2 Traitement des requêtes

Un exemple de requête est fourni figure 1. Chaque requête est composée de trois parties : titre, description, narration. Il est possible de traiter chaque partie indépendamment ou de les fusionner. Nous avons étudié différents types de combinaison, les meilleurs résultats ont été obtenus en considérant l'ensemble des composants des phrases. De façon générale, une requête est considérée comme un texte et les termes sont extraits comme expliqué dans le paragraphe 3.1.1.

Topic: 35
Title: NATO, Poland, Czech Republic, Hungary
Type: event
Descriptive: Accession of new NATO members: Poland, Czech Republic, Hungary, in 1999.
Narrative: Identity of current and newly-invited members, statements of support for and opposition to NATO enlargement and steps in the accession process and related special events are relevant. Impact on the new members, i.e., requirements they must satisfy, and their expectations regarding the implications for them are relevant. Progress in the ratification process is relevant. Future plans for NATO expansion, identification of nations admitted on previous occasions, and comments on future NATO structure or strategy are not relevant.

Figure 1. Exemple de requête

Chaque terme extrait est pondéré puis catégorisé selon différent groupes : fortement pertinent, faiblement pertinent et non pertinents.

Soient T_k une requête composée d'un titre T, d'un descriptif D et d'une narration N, t_i un terme et $tf_{i,k,P}$ la fréquence de t_i dans la partie P de T_k.

Le poids associé à un terme d'indexation pour la requête est alors calculé comme suit:

$$Poids(t_i,T_k)=\sum_{P\in\{T,D,N\}}\mu_P\cdot tf_{i,k} \text{ si } \sum_{P\in\{T,D,N\}}\mu_P\cdot tf_{i,k}\geq 3 \qquad (1)$$

$$Poids(t_i,T_k)=1 \text{ si } \sum_{P\in\{T,D,N\}}\mu_P\cdot tf_{i,k}>0$$

$$Poids(t_i,T_k)=0 \text{ sinon}$$

où μ_P est une constante dépendant de la partie P considérée.

Un terme assez fréquent possède un poids proportionnel à sa fréquence (μ_p permet de pondérer l'importance de chaque partie de la requête ; typiquement le titre aura plus d'importance que le descriptif : $\mu_T \geq \mu_D$). En revanche, un terme trop peu fréquent à un poids ramené à une constante, ici 1.

Les catégories de termes sont définies à partir de ces poids :

Termes fortement pertinents : ce sont les termes qui ont un poids supérieur à 3,

$$HT_k=\{t_i / t_i\in T_k \text{ et } poids(t_i,T_k)>1\} \qquad (2)$$

Termes faiblement pertinents : ce sont les termes qui obtiennent un poids égal à 1

$$LT_k=\{t_i / t_i\in T_k \text{ et } poids(t_i,T_k)=1\} \qquad (3)$$

Termes non pertinents : ce sont les termes qui ont un poids nul.

La prise en compte de la partie de requête dans laquelle apparaît chaque terme (titre, descriptif, narration) est étudiée dans (Dkaki, 2003). Les termes du titre, si le système leur donne plus d'importance relative, permettent une meilleure sélection des phrases pertinentes.

3.1.3 Traitement des documents

Chaque partie de document (dans les expérimentations, chaque phrase) est considérée comme un texte et les termes sont extraits comme indiqué dans la section 3.1.1.

Le système associe un poids à chaque terme ainsi extrait selon la formule (4) :

Soient S_j une phrase d'un document, t_i un terme et $tf_{i,j}$ la fréquence de t_i dans S_j.

$$poids(t_i,S_j)=tf_{i,j} \quad (4)$$

3.2 Passages pertinents

3.2.1 Besoins utilisateur

La restitution de passages –ou de phrases- pertinents plutôt que des documents entiers est particulièrement intéressante lorsque les documents sont longs ou qu'ils abordent différents aspects dont seuls certains sont liés au besoin d'information. Visualiser des passages permet donc à l'utilisateur d'accéder directement à l'information importante plutôt que d'avoir besoin de parcourir le document complet.

3.2.2 Fonction de sélection des phrases pertinentes

Notre méthode consiste d'abord à calculer la ressemblance entre chaque phrase et la requête puis à ne garder que certaines parmi les plus ressemblantes : celles qui possèdent une proportion suffisante de termes de la requête, la proportion nécessaire dépendant de la catégorie des termes considérés.

Soient une requête T_k et une phrase S_j, la ressemblance entre la requête et la phrase est calculée par :

$$Score(S_j,T_k)=\sum_{ti}(Poids(t_i,S_j)\cdot Poids(t_i,T_k))$$

Une phrase donnée S_j est alors considérée comme pertinente si :

$$Score(S_j,T_k) > f\left(\frac{|LS_j|}{|LS_j|+|HS_j|}\right)\cdot|HT_k| + g\left(\frac{|HS_j|}{|LS_j|+|HS_j|}\right)\cdot|LT_k| \quad (5)$$

où $|X|$ est le nombre d'éléments de l'ensemble X et

$$f(0)=2\ ;\ \forall x\in\ ]0,1],\quad f(x)=1.5$$
$$g(0)=0.85\ ;\ \forall x\in\ ]0,1],\quad g(x)=0.3$$

3.2.3 Résultats

Dans (Dkaki, 2003), nous avons étudié l'influence des différents paramètres de la méthode sur l'efficacité du processus. Nous nous sommes pour cela appuyés sur les collections

fournies par TREC. Les résultats que nous avons obtenus sur les collections de test décrites dans la table 1 et disponibles via le programme TREC sont rapportés dans la table 2 :

	TREC 2002	TREC 2003
Nombre de requêtes	49	50
Nombre de documents pertinents par requête	22,3	25
Nombre de phrases par requête	1321	796,4
Nombre de phrases pertinentes par requête	27,9	311,14
% de phrases pertinentes	2,1	39

Table 1. Description des collections de test

	TREC 2002	TREC 2003
Précision moyenne % de phrases effectivement pertinentes parmi les phrases retrouvées par le système	15 %	62 %
Rappel moyen % de phrases pertinentes retrouvées par rapport à l'ensemble des phrases pertinentes	50 %	64 %
Mesure F (2*Rappel * Précision)/(Rappel+Précision)	0.188	0.553

Table 2. Résultats sur les collections de test – Détection de la pertinence

Les résultats obtenus positionnent notre approche au rang 4 sur 13 participants en 2002 et 5 sur 14 en 2003. Il faut notre que si les résultats sont acceptables pour un utilisateur par rapport à la collection 2003, ceux obtenus sur la collection 2002 sont relativement pauvres. Malgré tout, ces résultats positionnent bien notre approche au niveau international.

3.3 Passages nouveaux

3.3.1 Besoins utilisateurs

Compte tenu de la quantité d'information électronique maintenant disponible, il est fréquent que plusieurs documents ou plusieurs passages de documents soient effectivement pertinents pour une requête mais comportent des redondances. Si cette redondance peut être utile dans certains cas, pour donner une certaine validité à l'information trouvée par exemple, dans d'autre cas cette redondance correspond à du bruit pour l'utilisateur. La détection de passages nouveaux vise donc à s'affranchir de ce bruit.

3.3.2 Détection des passages nouveaux

Pour décider si une phrase S_j doit être considérée comme nouvelle, nous calculons la ressemblance entre la phrase S_j et chacune des phrases S_m traitées précédemment ainsi que la ressemblance entre la phrase S_j et une phrase virtuelle Λconstruite à partir de l'union des S_m :

Soient

$\Omega = \{S_1, S_2, \ldots, S_n\}$ un ensemble de phrases et $\Lambda = \bigcup_{i \in \{1,\ldots,n\}} S_i$, Λ est donc une phrase virtuelle composée de l'ensemble des phrases de Ω,

$Sim(x, y)$ une fonction qui calcule la ressemblance entre x et y et

S_j une phrase pour laquelle le système doit décider si elle est pertinente.

Le système calcule différentes ressemblances qu'il combine pour décider de la nouveauté de la phrase en cours de traitement. Le principe général étant que la phrase en cours de traitement doit être suffisamment différente de la phrase virtuelle constituée des phrases déjà choisies comme pertinentes et suffisamment différentes des dernières phrases traitées. La seconde condition est un cas particulier de la première condition et permet de prendre en compte le fait que des phrases proches dans un document ont plus de chance de parler du même sujet.

Pour réaliser cette sélection de phrases, le système calcule les ressemblances suivantes :

$Sim(S_j, \Lambda) = \alpha_i$ et,

pour chaque $i \in \{1, \ldots, n\}$ $Sim(S_j, S_i) = \omega_{j,i}$ (6)

Le système considère alors les q phrases précédentes les plus ressemblantes à la phrase en cours de traitement et considère les séries $\delta_{j,i}$ *avec* $i \in \{1, \ldots, n\}$ qui sont les séries obtenues en ordonnant les $\omega_{j,i}$ par ordre décroissant. Le système calcule alors

$\beta_j = \sum_{i \in \{1,\ldots,q\}} Sim(S_j, \delta_{j,i})$ (q=4 dans les évaluations que nous avons faites) (7)

S_j est considérée comme redondante (non nouvelle) si :

$\alpha_p \geq T_1$ et $\beta_p \geq T_2$ (8)

où $T_1 = 1$ et $T_2 = 0.6$

Les résultats que nous avons obtenus sur les collections de test décrites dans la table 3 et disponibles via le programme TREC sont indiqués dans la table 4 :

3.3.3 Résultats

	TREC 2002	TREC 2003
Nombre de documents pertinents par requête	22,3	25
Nombre de phrases par requête	1321	796,4
Nombre de phrases nouvelles par requête	25,3	204,5
% de phrases nouvelles	90,9%	65,7%

Table 3. Description des collections de test

	TREC 2002	TREC 2003
Précision moyenne % de phrases effectivement nouvelles parmi les phrases sélectionnées par le système	14 %	43 %
Rappel moyen % de phrases nouvelles sélectionnées par le système parmi l'ensemble des phrases nouvelles	49 %	71 %
Mesure F	0.187	0.477

Table 4. Résultats sur les collections de test – Détection de la nouveauté

Ces résultats positionnent notre approche au quatrième rang pour TREC 2002 et au troisième pour TREC 2003. Comme pour la détection de la pertinence, les résultats sont meilleurs sur la collection 2003. Il faut notre que la tâche sur cette collection est plus facile que dans le cas de la collection 2002, en raison du pourcentage de phrases effectivement pertinentes et effectivement nouvelles. En effet, lorsque plus de 90% des phrases sont nouvelles, détecter celles qui sont redondante s'avère difficile car la fonction de décision risque d'éliminer des phrases qui sont effectivement nouvelles.

4 Interface permettant la contextualisation de l'information

Les copies d'écran que nous présentons sont issues du système que nous avons développé. Il permet de visualiser des phrases qui correspondent à la notion de passage. Ce choix est justifié par le fait que les travaux de détection de passages pertinents et de passages nouveaux ont été réalisés dans le cadre de la tâche 'Novelty' de TREC qui défini un passage comme étant une phrase (Harman, 2002). Cependant, l'interface est suffisamment générique pour permettre la visualisation de tout type de passage.

L'interface que nous proposons vise à permettre à l'utilisateur d'accéder directement à l'information qui lui sera la plus utile (phrases pertinentes ou phrases nouvelles, en fonction de ses choix), mais également à lui permettre de garder le contexte des éléments retrouvés.

4.1 Restitution des passages pertinents ou nouveaux

L'utilisateur peut choisir de visualiser les documents entiers ou seulement les phrases pertinentes ou les phrases nouvelles. Une liste de références aux éléments accessibles est alors affichée comme dans la plupart des systèmes de recherche d'information. Cette liste peut être ordonnée soit par identifiant de document, soit par degré de pertinence/nouveauté, soit par source (auteur). La recherche de la pertinence ou de la nouveauté est calculée au niveau de la phrase. Un document est considéré comme pertinent s'il contient au moins une phrase pertinente ; il est considéré comme nouveau s'il contient au moins une phrase nouvelle.

La figure 2 présente l'interface. Des couleurs sont choisies par défaut, mais elles peuvent être modifiées par l'utilisateur.

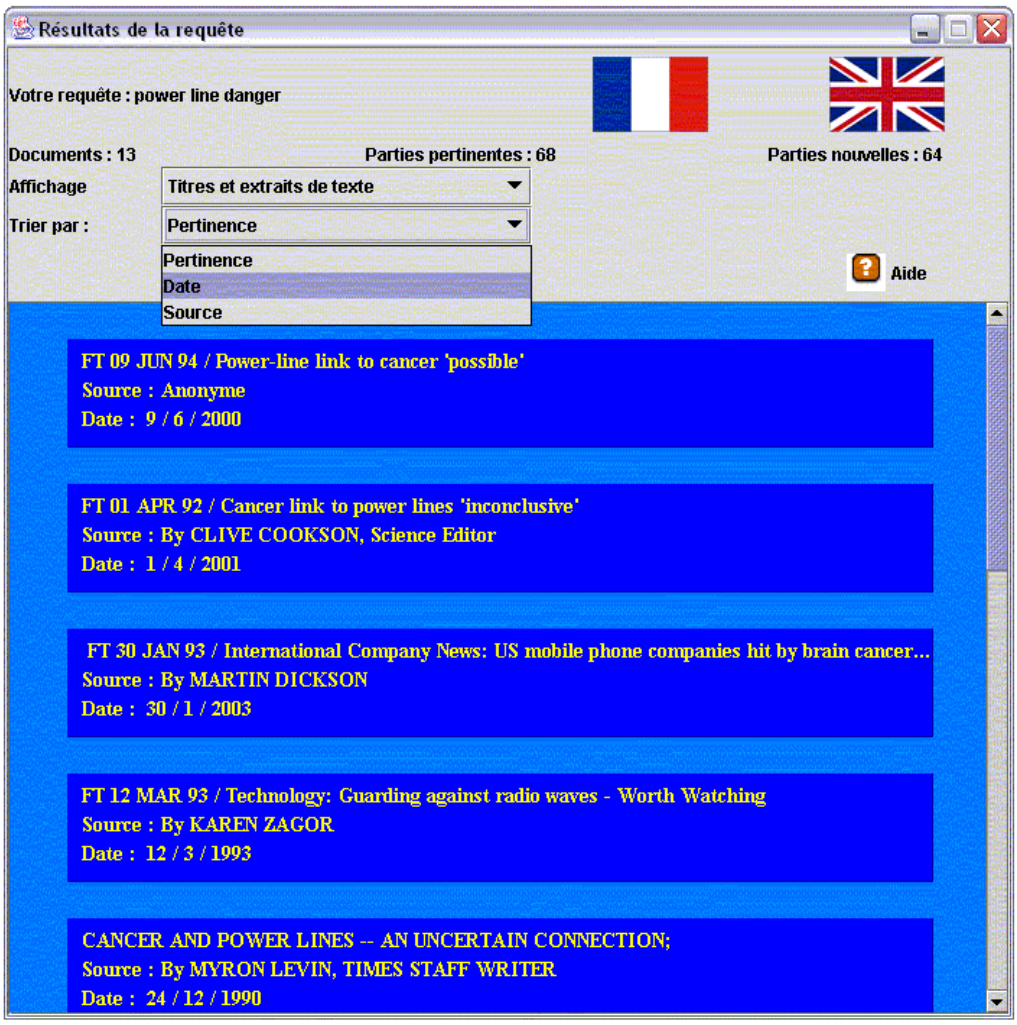

Figure 2 : Restitution des éléments pertinents ou nouveaux

Dans cette fenêtre (figure 2) :

Affichage permet de choisr les éléments à afficher: niveau de granularité (documents ou phrases), type (pertinents, nouveaux).

Trier par: permet à l'utilisateur d'ordonner les éléments restitués par degré de pertinence ou de nouveauté, par date ou par source (auteur).

Il est alors possible pour l'utilisateur de visualiser le contenu d'un élément qu'il aura choisi par un simple clic souris. Cet élément peut être visualisé soit indépendamment du contexte dans lequel il apparaît (figure 3) ; soit en affichant le contexte (figure 4).

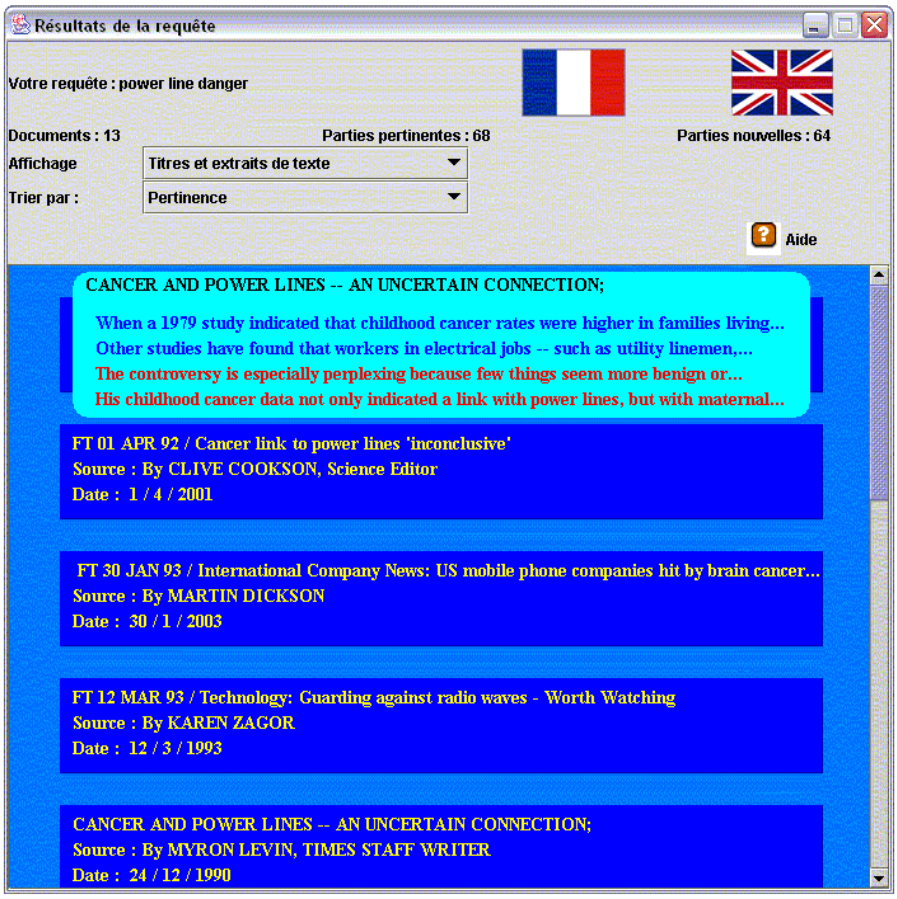

Figure 3 : Visualisation des éléments sélectionnés

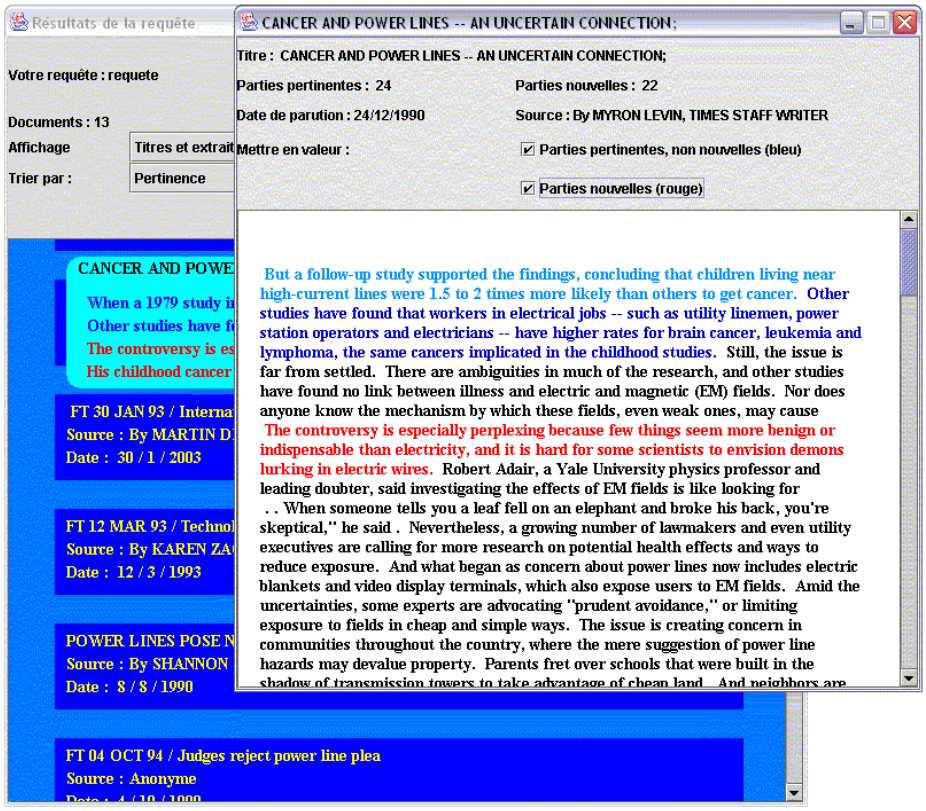

Figure 4 : Visualisation contextuelle des phrases pertinentes et nouvelles

Dans le premier cas (Figure 3), si le niveau de granularité de restitution correspond à la phrase, chaque phrase pertinente ou nouvelle est restituée de façon indépendante du document dont elle est issue.

Dans le deuxième cas, même si le niveau de restitution choisi est la phrase, lors de l'affichage, l'ensemble du document sera visualisé mais la phrase considérée sera l'élément central dans la fenêtre. Elle apparaîtra de plus dans une couleur spécifique.

La figure 4 illustre le cas où l'utilisateur demande d'afficher un document en visualisant les phrases nouvelles (en rouge), celles qui sont seulement pertinentes (en bleu) et les autres (en noir) en affectant à chaque type de phrase une couleur différente. Le degré de pertinence des phrases est rendu par un degré de coloration.

5 Conclusion

La quantité d'information actuellement disponible sous forme électronique rend indispensable l'utilisation de mécanismes d'accès à l'information efficaces et adaptés aux besoins des utilisateurs.

Dans cet article, nous avons proposé un mécanisme permettant de sélectionner des éléments d'information de granularité plus petite que le document. Cette sélection peut être effectuée selon deux critères : le critère de la pertinence (traditionnellement choisi par les systèmes) et celui de la nouveauté. Il s'agit de permettre à l'utilisateur d'accéder directement aux éléments les plus utiles pour lui sans toutefois qu'il perde le contexte des éléments d'information qu'il sélectionne. L'interface que nous proposons permet cette double approche.

Les mécanismes de sélection des éléments (pertinence, nouveauté) ont été évalués dans la cadre du programme international TREC. Nos travaux actuels visent à améliorer cette approche en particulier en utilisant les approches de type analyse factorielle.

6 Références

(Allan, 2003) J. Allan, C. Wade, A. Bolivar, Retrieval and Novelty Detection at the Sentence Level, 26 ème Conférence Annuelle ACM, Research and Development in Information Retrieval, SIGIR'03, pp 314-321, 2003.

(Baeza, 1999) Baeza-Yates, R., Ribeiro-Neto, B., Modern Information Retrieval, Addison-Wesley Ed., ISBN 0-201-39829-X, 1999.

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

(Dkaki, 2002) T. Dkaki, J. Mothe, Novelty track at IRIT-SIG, actes de Text Retrieval Conference TREC'02 , pp 332-336, 2002.

(Dkaki, 2003) T. Dkaki, J. Mothe, Recherche de la pertinence et de la nouveauté dans les textes, soumis à la Conférence en Recherche d'Information et Application CORIA, 2003.

(Fuhr, 2002) N. Fuhr, N. Goevert, G. Kazai and M. Lalmas. INEX: Intitiative for the Evaluation of XML Retrieval, ACM SIGIR Workshop on XML and Information Retrieval, Tampere, Finland, 2002.

(Harman, 2002)D. Harman, Overview of the TREC 2002 novelty track, actes de Text Retrieval Conference TREC'02, pp 17-28, 2002.

(Kazawa, 2002) 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 TREC 2002, pp 472-475, 2002.

(Kwok , 2002) K.L. Kwok, P. Deng, N. Dinstl, M. Chan, Queens College, CUNY TREC 2002 Web, Novelty and Filtering Track Experiments using PIRCS, pp 520-528, 2002.

(Murtagh, 2003) F. Murtagh, T. Taskaya, P. Contreras, J. Mothe, K. Englemeier, Interactive Visual Interfaces: A Survey, Artificial Intelligence Review, 19, 263-283, 2003.

(Ponte, 1998) J.M. Ponte, W.B. Croft, A language modelling approach to information retrieval, actes de la 21 ème Conférence Annuelle ACM, Research and Development in Information Retrieval, SIGIR'98, pp 275-281, 1998.

(Rijsbergen, 1979) C. J. Van Rijsbergen, « Information Retrieval », Butterworths, London, Second Edition, 1979. http://www.dcs.gla.ac.uk/Keith/Preface.html

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

(Salton, 1971) G. Salton, « The SMART Retrieval System », Experiments in automatic document processing, Prentice Hall Inc., Englewood Cliffs, NL, 1971.

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

(Wilkinson, 1994) R. Wilkinson, Effective retrieval of structured documents, actes de la 26 ème Conférence Annuelle ACM, Research and Development in Information Retrieval, SIGIR'94, pp 311-317, 1994.