Extraction de segments thématiques pour la construction de résumé multi-document orienté par un profil utilisateur

Sana-Leila Chaar

Université de Marne la vallée
CEA – LIST 18, rue du Panorama
BP 6 92265 Fontenay-aux-Roses Cedex
[sana.chaar]@cea.fr
Date de la thèse : juin 2004

Mots-clefs – Keywords

Extraction d'information, profil utilisateur, résumé multi-document
Information extraction, user profile, multi-document summarization

Résumé – Abstract

Dans cet article, nous présentons une méthode qui vise à donner à un utilisateur la possibilité de parcourir rapidement un ensemble de documents par le biais d'un profil utilisateur. Un profil est un ensemble de termes structuré en sous-ensembles thématiquement homogènes. L'analyse des documents se fonde pour sa part sur l'extraction des passages les plus étroitement en relation avec ce profil. Cette analyse permet en particulier d'étendre le vocabulaire définissant un profil en fonction du document traité en sélectionnant les termes de ce dernier les plus étroitement liés aux termes du profil. Cette capacité ouvre ainsi la voie à une plus grande finesse du filtrage en permettant la sélection d'extraits de documents ayant un lien plus ténu avec les profils mais davantage susceptibles d'apporter des informations nouvelles et donc intéressantes. La production du résumé résulte de l'appariement entre les segments délimités lors de l'analyse des documents et les thèmes du profil.

In this article, we present an information extraction method that selects from a set of documents their most significant excerpts in relation to an user profile. This method relies on both structured profiles and a topical analysis of documents. The topical analysis is notably used for expanding a profile in relation to a particular document by selecting the terms of the document that are closely linked to those of the profile. This expansion is a way for selecting in a more reliable way excerpts that are not strongly linked to profiles but that may bring new and interessant information about their topics.

1 Introduction

La veille stratégique, commerciale, scientifique et technique est une préoccupation majeure pour toute organisation grande ou petite. Les décideurs sont submergés d'information, ne

disposent pas du temps nécessaire pour compiler toutes les sources qui pourraient leur permettre de prendre les meilleures décisions. L'exploitation des documents récupérés pour réaliser une synthèse ou un résumé automatique est devenue un thème qui occupe de plus en plus les chercheurs. On s'est d'abord intéressé au résumé d'un document unique puis, pour les besoins de la veille, l'un des thèmes de recherche principal est devenu le résumé multi-document. Cette nouvelle problématique fait l'objet de nombreux travaux, en particulier dans le cadre de l'évaluation DUC (Document Understanding Conference) (Over, 2001). Le travail que nous présentons dans cet article rejoint donc la problématique du résumé multi-document.

Actuellement les systèmes de résumé, qu'ils soient mono ou multi-document, fonctionnent essentiellement par extraction de passages ou de phrases. Différentes approches ont été développées au sein de ce paradigme. L'une des plus répandues consiste à exploiter des critères essentiellement statistiques ou probabilistes. C'est le cas notamment de la méthode MMR-MD (Maximal Marginal Relevance – Multi-Document) (Goldstein et al., 2000), qui définit une métrique d'intérêt des passages de texte, du système développé par TNO-TPD (Kraaij et al., 2001) qui attribue un score à des segments textuels en combinant un modèle de langage uni-gramme et un modèle bayésien, ou encore le systèmes décrit dans (Boros et al., 2001) qui fait appel à des méthodes de classification. L'exploitation de critères plus linguistiques constitue une deuxième grande approche parmi les systèmes de résumé multi-document. Elle est représentée par des travaux tels que ceux de Radev et McKeown (Raved et al. 1998), de Barzilay (Barzilay et al., 1999) ou encore de McKeown (McKeown et al., 2001). Les systèmes concernés font appel à des traitements linguistiques plus ou moins élaborés (extraction de termes, reconnaissance d'entités nommées, analyse syntaxique, …). Dans cette même perspective, des travaux tels que (Mani, 1999) ou (White et al., 2001) utilisent des méthodes d'extraction d'information pour produire des résumés multi-documents en fonction d'une requête ou d'un profil utilisateur. Enfin, la problématique du résumé multi-document rejoint et peut se combiner à celle de la visualisation et du parcours d'un ensemble de documents ainsi qu'en témoigne le système développé par Rie Kubota Ando (Ando et al , 2000).

Le travail que nous présentons dans cet article vise pour sa part à donner à un utilisateur la possibilité de parcourir rapidement un ensemble de documents selon un point de vue particulier, par exemple à la suite d'une requête réalisée auprès d'un moteur de recherche. Ce point de vue est représenté par le biais d'un profil. Le parcours se fonde quant à lui sur l'extraction des passages les plus étroitement en relation avec ce profil en s'appuyant sur une analyse thématique des documents. L'utilisation d'un profil rapproche ce travail de (Mani, 1999) et (White et al., 2001) tandis que l'importance accordée à l'analyse thématique lui fait entretenir de forts liens avec (Ando et al , 2000).

2 Analyse des documents à résumer

L'analyse des documents a pour rôle de mettre en évidence dans les documents à résumer : le vocabulaire susceptible d'être comparé au profil et les structures thématiques pouvant être mises en correspondance avec les thèmes constituant ce profil.

2.1 Normalisation des documents

La première étape de l'analyse des document vise à normaliser le vocabulaire des documents. Cette normalisation consiste à associer à chaque mot d'un document son lemme.

Pour cela nous utilisons les modules d'analyse morphologique et d'étiquetage morpho-syntaxique du logiciel *SPIRIT* (Fluhr, 1994). A la fin du processus d'analyse et d'étiquetage, seuls les mots non grammaticaux susceptibles d'apparaître dans les profils sont sélectionnés, c'est-à-dire les noms, les verbes et les adjectifs. Ce prétraitement linguistique vise donc à normaliser le vocabulaire et à faciliter ainsi la comparaison avec un profil. Malgré la présence dans les profils de termes complexes, nous ne faisons volontairement appel ni à un extracteur terminologique généraliste, ni à un outil de reconnaissance de variantes terminologiques. En dépit de l'intérêt de ces outils pour la mise en évidence des termes complexes des profils dans les documents, leur fragilité rend leur utilisation délicate dans un contexte, comme celui du filtrage, où chaque occurrence d'un terme peut être importante. Nous détaillerons à la section 4.2 la méthode que nous utilisons pour identifier dans les documents les termes complexes d'un profil.

2.2 Analyse thématique

L'analyse thématique des documents a dans le cas présent pour rôle de segmenter les documents en unités thématiquement homogènes, tâche que l'on nomme segmentation thématique. Les segments ainsi délimités constituent les unités textuelles de base qui sont comparées aux profils. Nous avons choisi d'utiliser le système *TextTiling* de Hearst (Hearst, 1997) en l'adaptant afin de prendre en entrée le résultat du prétraitement des textes décrit à la section précédente et d'aligner les bornes de segment sur des frontières de phrase. Le choix de *TextTiling* s'appuie sur le bon rapport complexité / efficacité de son algorithme : une fenêtre glissante est déplacée sur le texte à segmenter. Pour chaque position du texte ainsi parcourue, on compare les deux parties de la fenêtre situées de part et d'autre de cette position à l'aide d'une mesure de similarité s'appuyant sur les mots présents dans la fenêtre. Les valeurs de cette mesure pour l'ensemble du texte permettent de situer les zones de faible cohésion lexicale. Celles-ci sont assimilées à des ruptures thématiques et donnent lieu ainsi à la définition de segments thématiques. Ces segments ont une taille moyenne aux alentours d'une centaine de mots, cette valeur variant bien entendu en fonction des variations thématiques effectivement rencontrées dans les textes.

Nous n'avons abordé dans cette partie que l'analyse thématique réalisée en dehors du contexte d'un profil particulier. Les opérations d'identification thématique[1] que nous décrirons dans la section 4. permettant d'établir qu'un segment s'apparie avec un profil ou plus globalement que le thème principal d'un texte s'apparie avec un profil relèvent également de l'analyse thématique. Cependant, elles ne sont pas réalisées ici selon une perspective générique qui consisterait en premier lieu à caractériser le thème d'un segment ou d'un texte par ses mots les plus caractéristiques pour ensuite comparer ceux-ci avec les mots du profil.

3 Description des profils utilisateur

Par profil utilisateur, nous entendons une représentation a long terme des centres d'intérêt de l'utilisateur, exprimés sous forme d'un ensemble de termes ou d'expressions jugés les plus représentatifs de son domaine. Ces termes peuvent être simples ou complexes. La constitution d'un profil demande un investissement de la part de l'utilisateur par rapport aux requêtes ponctuelles qu'il a le plus souvent l'habitude de formuler pour effectuer une recherche d'information via un moteur de recherche.

[1] L'identification thématique est la partie de l'analyse thématique visant à déterminer le thème d'une unité textuelle.

Nous avons plus précisément choisi d'adopter une structure thématique pour les profils, c'est-à-dire de regrouper les termes qui les composent en sous-ensembles thématiquement homogènes. Chacun de ces sous-ensembles représente un thème du profil renvoyant au point focal de ce dernier ou bien à son contexte. Par exemple, le profil d'un utilisateur recherchant de l'information sur le problème de l'*utilisation des piles à combustibles pour la propulsion automobile* pourra être décomposé en deux sous-thèmes : l'un relatif au *monde de l'automobile* et l'autre relatif aux *piles à combustibles*. Chacun de ces sous-thèmes regroupe un ensemble de termes significatifs. Par exemple pour le premier sous-thème nous aurons : *pile à combustible, pile à hydrogene, PAC, SOFC,…* et pour le second un autre ensemble de termes : *moyen de transport, véhicule, auto, voiture*, … Cette structuration répond au souci d'améliorer la précision de l'extraction des passages pertinents en ne mettant pas sur le même plan tous les termes composant un profil. Les documents les plus pertinents seront ceux dans lesquels le vocabulaire lié au *monde de l'automobile* et celui lié aux *piles à combustibles* seront simultanément présents. Un document ne comportant que des termes liés au *monde de l'automobile*, même s'ils sont présents en grand nombre, n'a ainsi que peu de chance d'être intéressant. Seule la séparation au niveau du profil des termes liés respectivement à chacun de ces deux sous-thèmes permet d'écarter lors du filtrage un document très marqué par l'un des deux seulement au profit d'un document qui comporte des termes du profil en moins grand nombre mais répartis de façon plus équilibrée entre les deux sous-thèmes qui le caractérisent. Outre l'amélioration de la précision du filtrage qui en résulte, adopter une telle structuration des profils se justifie par son adéquation avec la nature des demandes de recherche d'information émanant des utilisateurs. Celles-ci sont en effet fréquemment définies par un recoupement de plusieurs thèmes et non par la donnée d'un seul grand thème.

4 Filtrage des documents en fonction d'un profil utilisateur

Notre travail met l'accent sur l'utilisation conjointe de profils structurés et d'une analyse discursive des documents pour extraire des passages de textes en rapport avec les attentes de l'utilisateur. Dans le cadre du processus de filtrage, l'analyse des documents est suivie d'une étape d'appariement entre les segments délimités et les sous-thèmes du profil considéré, appariement se fondant sur la similarité de leurs vocabulaires respectifs.

4.1 Sélection des documents

Afin de déterminer les documents pertinents nous appliquons en premier une fonction de similarité entre les termes normalisés des documents (cf. section 2) et les termes du profil utilisateur (cf. section 3). Le résultat de cet appariement permet de définir si un document présente ou non un intérêt du point de vue du profil. Trois cas de figure peuvent apparaître :

- le document est globalement en relation avec le profil, même s'il peut comporter des parties abordant des thèmes en dehors de ceux du profil ;
- une partie seulement du document est en relation avec le profil. Ce dernier correspond à un thème secondaire du document et n'est donc évoqué que ponctuellement au sein de celui-ci ;
- le document n'a pas de relation avec le profil, même à un niveau local.

L'étape de sélection des documents vise à séparer les documents relevant des deux premiers cas de figure de ceux relevant du dernier. Nous considérons par ailleurs que les documents

globalement en relation avec le profil ont nécessairement une partie s'appariant avec ce profil (cf. section 3). Par conséquent, le critère de sélection des documents que nous appliquons est lié au deuxième cas de figure : un document est sélectionné à condition qu'au moins l'un de ses segments (issus de l'analyse thématique) s'apparie avec le profil considéré.

4.2 Selection des extraits de document en fonction du profil utilisateur

Ainsi que nous l'avons indiqué, la structuration thématique des profils vise à éviter qu'un document ou une partie de document ne soit sélectionnée alors qu'elle n'aborde qu'une des dimensions d'un profil. Il apparaît donc logique d'imposer qu'un segment de texte ne puisse s'apparier avec un profil que si chacun des sous-thèmes constituant ce profil est représenté dans le segment. Les segments étant en moyenne de petite taille, nous considérons que la représentation d'un thème dans un segment se manifeste par la présence au sein de celui-ci d'un terme caractérisant ce thème au niveau du profil. Ce critère peut apparaître au premier abord un peu faible à l'échelle d'un seul thème mais il est beaucoup plus significatif à l'échelle d'un profil regroupant plusieurs thèmes. Bien que les termes complexes soient en général plus informatifs que les termes simples, nous n'imposons pas que le terme représentant un thème dans un segment soit un terme complexe. Par ailleurs, les termes complexes *TC* posent un problème de reconnaissance que nous avons choisi de résoudre en adoptant un compromis moyen entre précision et rappel. La reconnaissance stricte d'un terme *TC* du profil dans un segment obéit plus précisément à la série d'heuristiques suivantes :

- les termes simples TS_i composant *TC* doivent apparaître dans le segment dans le même ordre qu'au sein de *TC*. La reconnaissance des TS_i est directe puisqu'ils sont normalisés de la même façon dans les documents et dans le profil ;
- soit *N,* nombre des TS_i. L'espace occupé par une occurrence de *TC* dans le segment ne peut dépasser 1,5**N* mots pleins. Ainsi nous prenons en compte d'éventuelles variations syntaxiques de type insertion ;
- si *TC* comporte des prépositions, celles-ci doivent être également présentes au niveau de ses occurrences dans les documents, ceci en respectant leur position par rapport aux TS_i. Par ailleurs, une occurrence présumée de *TC* ne doit contenir aucun signe de ponctuation.

Le terme complexe *TC* peut également être reconnu lorsque seulement un de ses sous-termes *ST* est présent. On parle alors de reconnaissance approchée. Trois conditions doivent être remplies pour ce faire :

- *ST* doit regrouper au moins 50% des termes simples de *TC* ;
- 1 occurrence de *ST* est reconnue si elle respecte les trois contraintes de reconnaissance stricte d'un terme ;
- *TC* doit avoir été reconnu de façon stricte au moins une fois dans le document auquel appartient le segment considéré.

Dans le cadre de l'identification d'un thème dans un segment, la reconnaissance d'un terme caractéristique de ce thème peut être stricte ou bien approchée. Le fait qu'un document a été sélectionné renvoie à deux cas de figure comme nous l'avons vu en section 4.1 : le document s'apparie globalement avec le profil considéré ou bien seulement une partie de ce document s'apparie avec ce profil. Dans le second cas, la thématique du profil n'étant évoquée que

secondairement dans le document, seul les segments qui s'apparient avec le profil sont sélectionnés. Dans le premier cas au contraire, il peut être intéressant d'élargir le champ de la sélection à des segments ne répondant pas strictement aux critères d'appariement avec le profil mais contenant des termes du document jugés liés à ceux du profil. Plus précisément, dans ce cas de figure, nous effectuons une analyse plus approfondie des documents, afin de sélectionner de nouveaux termes pouvant s'inscrire dans l'un des sous thème de profil.

4.3 Profil enrichi

Un terme nouveau est un terme pouvant s'inscrire dans l'un des sous-thème du profil. La liaison entre un terme nouveau et un terme du profil s'appuie sur une série de cooccurrences au sein du document. Plus précisément, soit $\{tp_{Ki}\}$, l'ensemble des termes du thème K appartenant au profil qui sont présents dans le document. On définit l'ensemble $\{td_{Ki}\}$ des termes du document tels que td_{Ki} cooccurre avec un terme tp_{Ki} (pas nécessairement le même d'un segment à l'autre) dans un segment et ce, dans une proportion suffisamment importante de segments du document. Cette proportion a été fixée dans le cas présent à 1/3. Les termes td_{Ki} constituent ce que nous avons appelé ci-dessus des termes nouveaux. Ces termes nouveaux représentent une forme adaptation des profils par rapport aux documents auxquels ils sont confrontés dans le cadre du processus de filtrage. Les profils ont tendance à donner une description assez générale des thèmes qu'ils recouvrent. Au sein de chaque document, on retrouve les termes de cette description mais on y trouve également une caractérisation beaucoup plus spécifique qu'il est nécessaire de prendre en compte pour analyser finement l'organisation thématique du document. La mise en évidence de ces termes nouveaux présente à cet égard une certaine parenté avec le processus de blind relevance feedback utilisé en recherche documentaire. Reprenons l'exemple de *« l'utilisation des piles à combustibles pour la propulsion automobile»*, après avoir confronté les documents au profil défini par l'utilisateur, nous constatons que les segments sélectionnés contiennent un ensemble de termes qui peuvent s'inscrire dans les sous-thèmes du *monde de l'automobile* et des *piles à combustibles*. Ces nouveaux termes sont donc proposés à l'utilisateur sous forme d'un profil appelé profil enrichi, et grâce au processus de blind relevance feedback, nous recherchons les nouveaux segments pertinents qui peuvent correspondre au besoin de l'utilisateur.

Thème *Piles à combustibles*	**Thème** *Monde de l'automobile*
Pile à combustible	Moyen de transport
Pile à hydrogène	Véhicule
PAC	Auto
PEMFC	**Véhicules hybrides**
AFC	**Constructeur automobile**
Propulsion électrique	**Toyota**

Figure 1 : Exemple de profil enrichi sur l'utilisation des piles à combustible dans le domaine des transports[2] (termes nouveaux en gras)

4.4 Appariement d'un profil et d'un document

Dire qu'un document s'apparie globalement avec un profil revient à dire que le thème principal du document correspond au thème représenté par ce profil[3]. Bien que le problème de la formalisation de l'organisation thématique des textes ait été peu exploré, les travaux sur le

[2] La figure 1 ne donne que quelques uns des termes composant un tel profil.

[3] Le thème représenté par un profil correspond à la réunion des thèmes qui le composent, ce qui présuppose une certaine compositionnalité des thèmes.

résumé automatique laissent entrevoir une définition opérationnelle de la notion de thème principal que l'on peut énoncer comme suit : le thème principal d'un texte est le thème abordé en début ou/et en fin de texte et qui est l'objet d'une partie importante de celui-ci. En transposant cette définition dans notre contexte, déterminer si le thème principal d'un document s'apparie avec un profil revient à vérifier les deux conditions suivantes :

- le profil doit s'apparier avec le premier et/ou le dernier segment du document ;
- plus globalement, l'ensemble des segments du document s'appariant avec le profil doivent représenter une part significative de l'ensemble des segments du document.

La première condition fait appel à la notion d'appariement entre profil et segment développée à la section 4.2 La seconde s'appuie quant à elle sur la version élargie de cet appariement développée à la section 4.3 et prenant en compte les termes nouveaux, c'est-à-dire des termes du document liés aux termes du profil.

5 Fusion des segments sélectionnés

Les segments ainsi sélectionnés, qu'ils appartiennent à un même document ou à des documents différents, peuvent être redondants, *i.e.* une même information est susceptible d'apparaître dans plusieurs segments sous des formes légèrement différentes. Une évaluation de la similarité des segments sélectionnés est donc réalisée, d'abord au sein d'un même document puis entre documents, en utilisant une mesure de similarité adaptée à la comparaison d'unités textuelles. Pour un ensemble de segments similaires, seul est retenu le segment supposé contenir l'information la plus complète, *i.e.* celui dont le vocabulaire couvre le plus complètement l'ensemble des segments concernés. Le rôle de l'étape de fusion vise donc à minimiser les redondances d'information.

5.1 Fusion intra-document

Du point de vue de la détection des redondances entre segments, nous distinguons deux classes de segments :

- les segments dont l'appariement avec le profil se fonde sur des termes du profil reconnus de façon stricte dans le segment ;
- les segments dont l'appariement avec le profil se fonde au moins partiellement sur des termes nouveaux.

Les premiers, appelés segments cœur, représentent plutôt l'instanciation dans un document particulier de l'information déjà contenue dans le profil tandis que les seconds, appelés segments extension, sont davantage voués à apporter de l'information nouvelle en relation avec le profil. Ces deux classes étant complémentaires, nous ne chercherons pas à fusionner les informations issues d'une classe avec celles provenant de l'autre classe. Au sein de chacune de ces deux classes, la détection des redondances entre segments s'appuie sur une mesure de similarité à laquelle est associé le seuil fixé *a priori* $S_{segment}$. Nous avons classiquement choisi la mesure du cosinus, qui s'avère particulièrement bien adaptée à la comparaison d'unités textuelles. La similarité entre deux segments S_1 et S_2 est donnée par :

$$sim(S_1, S_2) = \frac{\sum_i nbOcc(t_i, S_1) \cdot nbOcc(t_i, S_2)}{\sqrt{\sum_i nbOcc(t_i S_1)^2 \cdot \sum_i nbOcc(t_i, S_2)^2}} \qquad (5.1)$$

avec $nbOcc(t_i, S_{\{1,2\}})$, le nombre d'occurrences du terme t_i dans le segment $S_{\{1,2\}}$. Les termes t_i considérés ici correspondent aux lemmes issus du pré-traitement linguistique des documents. Si cette mesure de similarité dépasse le seuil $S_{segment}$, les deux segments sont jugés similaires et sont donc supposés contenir en première approximation les mêmes informations. Un seul des deux segments peut donc représenter les deux. Dans le cas contraire, on conserve les deux segments. Plus globalement, la mesure de similarité (5.1) est évaluée entre tous les segments au sein de chacune des deux classes distinguées ci-dessus (segments cœur et segments extension). Ces segments sont ensuite regroupés en fonction de la valeur de cette mesure : chaque segment est associé au segment avec lequel sa similarité est la plus forte, à condition toutefois que cette similarité soit supérieure au seuil $S_{segment}$. On obtient ainsi un ensemble de regroupements disjoints de segments similaires. Dans le cas où la similarité entre tous les segments est assez forte, une classe peut ne comporter qu'un seul groupe de segments. Un représentant est ensuite sélectionné pour chaque groupe de segments ainsi défini. Nous nous appuyons pour ce faire sur le vocabulaire caractérisant le groupe, vocabulaire défini comme l'ensemble des termes simples communs à au moins deux de ses segments. Le représentant du groupe est plus précisément le segment abritant la plus large proportion de ce vocabulaire.

5.2 Fusion inter-document

À l'issue de l'étape de fusion intra-document, chaque document est représenté par deux ensembles rassemblant des segments non similaires selon (5.1). La fusion inter-document commence par l'union des ensembles de même type des différents documents considérés. Au sein de chacun de ces deux ensembles, nous appliquons le même algorithme de regroupement et de sélection d'un représentant que celui appliqué dans le cadre de la fusion intra-document. En finale, nous obtenons donc un ensemble de segments cœur et un ensemble de segments extension.

6 Classement des extraits de documents

Afin de faciliter l'exploitation ultérieure des deux ensembles de segments issus du filtrage, en particulier leur visualisation, un ordre de pertinence est défini sur leurs éléments. En revanche, ces deux ensembles sont conservés disjoints dans la mesure où chacun d'eux rend compte d'une dimension particulière du filtrage, dimension qui sera ou non exploitée suivant l'application dans laquelle celui-ci vient s'insérer. Plus précisément, chaque segment se voit attribuer un score calculé sur la base du vocabulaire qui le compose. Ce score prend en compte à la fois la présence des termes du profil, sous une forme stricte ou approchée, ainsi que celle des termes communs au groupe de segments dont celui considéré est le représentant. Ce score est donnée par :

$$score(S) = 1{,}0 \cdot \sum_i nbOcc(tps_i, S) + 0{,}75 \cdot \sum_i nbOcc(tpa_i, S) + 0{,}5 \cdot \sum_i nbOcc(tcg_i, S) \qquad (6.1)$$

où tps_i est un terme du profil reconnu de façon stricte, tpa_i est un terme du profil reconnu de façon approchée et tcg_i est un des termes communs aux segments formant le groupe dont le

segment considéré est le représentant. Ces pondérations favorisent la proximité par rapport au profil tout en accordant une place significative aux termes qu'il semble raisonnable d'associer au profil (termes tcg_i). Les segments ayant des tailles assez similaires, nous n'avons pas adopté de normalisation en fonction de la taille des segments. Finalement, le résultat du processus global se présente sous la forme de deux listes, celle des segments cœur et celle des segments extension, ordonnées suivant l'ordre décroissant du score de leurs segments.

7 Conclusion

Nous avons proposée dans cet article une méthode qui permet d'extraire d'un ensemble de documents ses parties les plus significatives relativement à un profil utilisateur structurés sur le plan thématique. Cette méthode fait l'objet d'une implémentation au sein d'un système qui est actuellement en cours de réalisation. Dans ce travail nous faisons la distinction entre les parties de document en relation directe avec un profil et les parties de document représentant potentiellement des informations nouvelles liées à ce profil, ce qui présente un intérêt particulier pour des applications de veille technologique. Plus classiquement, les extraits similaires au sein de chacune de ces deux catégories sont regroupés et désignés par leur représentant le plus caractéristique. Bien que le filtrage soit ici conçu comme une activité récurrente pouvant justifier un certain investissement de la part d'un utilisateur, la nécessité de définir une structuration de ce type peut apparaître comme rédhibitoire vis-à-vis de l'utilisation d'une telle méthode. Il convient néanmoins de souligner que cette structuration n'est pas une condition indispensable à la mise en œuvre de la méthode. Ainsi que nous l'avons vu à la section 4.2, la structure des profils intervient lors de la sélection des segments d'un document afin d'améliorer la pertinence de l'évaluation de la proximité entre un segment et un profil. Au prix d'une baisse de précision, il est cependant tout à fait possible de réaliser cette évaluation avec un profil non structuré. Par ailleurs, une aide à la structuration thématique des profils peut être apportée à l'utilisateur. Si celui-ci dispose d'un ensemble de documents relatifs aux thèmes du profil qu'il souhaite définir, l'utilisation conjointe, d'une segmentation thématique (cf. 2.2) et d'une méthode de classification non supervisée permet de proposer à l'utilisateur une représentation des thèmes de ces documents qu'il pourra ensuite modifier manuellement s'il le souhaite.

Outre la nature des profils, une méthode de extraction se caractérise également par le type des textes qu'elle se propose de traiter. Dans le cas présent, les moyens d'analyse mobilisés, en particulier l'analyse thématique, ne s'appuient que sur la forme de surface des textes et ne font pas intervenir de connaissances externes à ceux-ci de nature sémantique ou pragmatique. Compte tenu de ces moyens, le filtrage de textes dans lesquels une notion est évoquée sous des formes très diverses, comme c'est souvent le cas dans les textes journalistes ou narratifs, est *a priori* moins bon que le filtrage de textes où la récurrence du vocabulaire, à l'instar des textes techniques, est assez élevée, en particulier du fait de son caractère très spécifique.

Il faut d'autre part remarquer que ce relatif minimalisme au niveau des moyens d'analyse rend assez léger la transposition d'une langue à une autre[4], ce qui répond à une partie de nos préoccupations. Une des façons d'améliorer l'analyse des textes tout en préservant cette capacité d'adaptation rapide à une autre langue est de faire appel à des connaissances pouvant être construites automatiquement à partir de corpus. Nous envisageons d'adapter ces principes à la méthode que nous avons présentée ici après validation de celle-ci sous sa forme actuelle.

[4] Cette transposition dépend plus précisément des moyens d'analyse morpho-syntaxique.

Nous projetons en particulier de l'évaluer dans le cadre proposé par les conférences DUC, soit au travers d'une participation à la tâche dédiée au résumé multi-document[5], soit par une utilisation des données de référence qui en sont issues et des outils définis dans ce même contexte afin de réaliser une évaluation automatique.

Références

Ando, Rie Kubota, , Boguraev K., Byrd, Roy J., Neff, Mary S. (2000) Multisummarization by Visualizing Topical Content, *ANLP/NAACL Workshop Automatic Summarization*

Barzilay R., Mc Keown, Kathleen R., Elhadad, M. (1999) Information fusion in the context of multi-document summarization, *37ème Annual Meeting of the ACL*.

Boros, Endre, Kandor, Paul B., Neu, David J. (2001) A clustering based approach to creating multi-document summaries, *ACM SIGIR'01 Workshop on Text Summarization*.

Fluhr C. (1994) SPIRIT : un système d'exploration de données textuelles, *Le Traitement Informatique des Corpus Textuels*, INALF.

Goldstein J., Mittal V., Kantrowitz M., Carbonell J. (2000) Multi-Document Summarization By Sentence Extraction, *ANLP/NAACL Workshop on Automatic Summarization*.

Hearst M. (1997) TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, *Computational Linguistics*, Vol. 23, n° 1, p. 33-64.

Kraaij, Wessel, Spitters, , et al. (2001) Combining a mixture language model and Naive Bayes for multi-document summarisation, *ACM SIGIR'01 Workshop on Text Summarization*.

Mani I., Bloedorn E. (1999) Summarizing similarities and differences among documents, *Information Retrieval*, Vol. 1, n° 1, p. 1-23.

McKeown, Kathleen R., Barzilay, Regina, Evans David, Hatzivassiloglou, Vasileios, Kan, Min-Yen, Schiffman, Barry, Teufel, Simone (2001) Columbia Multi-document Summarization : Approach and Evalution , *ACM SIGIR'01 Workshop on Text Summarization*.

Over P. (2001) Introduction to DUC-2001: an Intrinsic Evaluaton of Generic News Text Summarization Systems, *ACM SIGIR'01 Workshop on Text Summarization*.

Radev, Dragomir R., McKeown, Kathleen R. (1998) Generating natural language summaries from multiple on-line sources, *Computational Linguistics*, Vol. 24, n° 3, p. 469-500.

White, Michael, et al., (2001) Detecting Disrepancies and Improving Intelligibility : Two Preliminary Evaluations of RIPTIDES, *ACM SIGIR'01 Workshop on Text Summarization*.

[5] Une telle participation demandera évidemment une adaptation au niveau de l'étape de fusion afin de considérer des phrases et non plus seulement des segments.