Chapitre VII. Segmentation thématique: État de L'art

1. Introduction

Avec l'augmentation de la taille et de la complexité des documents traités, la recherche et l'extraction de passages pertinents ont connu un essor particulier depuis une dizaine d'années. Il devient de plus en plus important d'aider les utilisateurs à accéder plus rapidement à l'information recherchée et à développer de nouveaux outils de recherche d'information. La segmentation thématique pourrait être considérée comme un outil d'aide pour différentes tâches d'accès à l'information. Elle peut être employée conjointement à l'utilisation de moteurs de recherche conventionnels pour aider les utilisateurs à évaluer rapidement la pertinence des documents retournés en réponse à une requête, ou encore pour faciliter la navigation à travers un corpus conséquent. Le résumé de texte est une autre tâche dont le résultat peut être amélioré par la segmentation thématique. Un résumé peut ainsi être généré à partir de différentes thématiques pertinentes identifiées par un système de segmentation. La plupart des approches pour la segmentation proposées reposent sur des méthodes statistiques ou linguistiques. Les méthodes statistiques, reposent le plus souvent sur l'approche de 'sac de mots' (bag of words) pour représenter une unité textuelle (paragraphe ou phrase). Récemment des méthodes basées sur l'apprentissage ont été proposées pour la segmentation. L'avantage de ces méthodes est d'être capable de s'adapter à des conditions opérationnelles bien plus diverses et en particulier de s'adapter à différents types de corpus, exemple [Amini et al., 1999] qui s'appuient sur des modèles de Markov cachés, ou bien [Caillet et al., 2004] qui proposent une classification des termes de même que [Mekhaldi et al., 2004].

Dans la première partie de ce chapitre, nous introduisons la problématique de la segmentation thématique des textes, sa définition, ses objectifs et ses approches. Dans la deuxième partie, nous décrivons ses trois grandes familles de méthodes, à savoir : les techniques qui fonctionnent à partir du calcul de similarités, les techniques basées sur la répétition des mots et celles basées sur les chaînes lexicales.

2. Définition de la segmentation thématique

La définition de la segmentation thématique n'est pas unique. Elle est différente pour chaque type de corpus. On prend comme exemple les types suivants:

- Pour les discours politiques, la segmentation thématique est basée sur la structure thématique des discours mis en ligne sur le site de référence. Chaque discours a été divisé en paragraphes thématiques lors de leur écriture ou lors de la constitution des corpus mis en ligne par l'organisme en charge de cette tâche.
- Pour l'ouvrage scientifique, les segments thématiques à retrouver sont les différentes sections, à savoir les chapitres, sections, sous-sections et sous-sous-sections. Le but est donc de déterminer la première phrase de chaque section.

On peut la définir de manière plus précise comme le découpage des textes en segments thématiquement homogènes. Elle consiste à rechercher les ruptures de thèmes dans le texte, afin de déduire et de trouver les zones de texte ayant une sémantique de thèse commune.

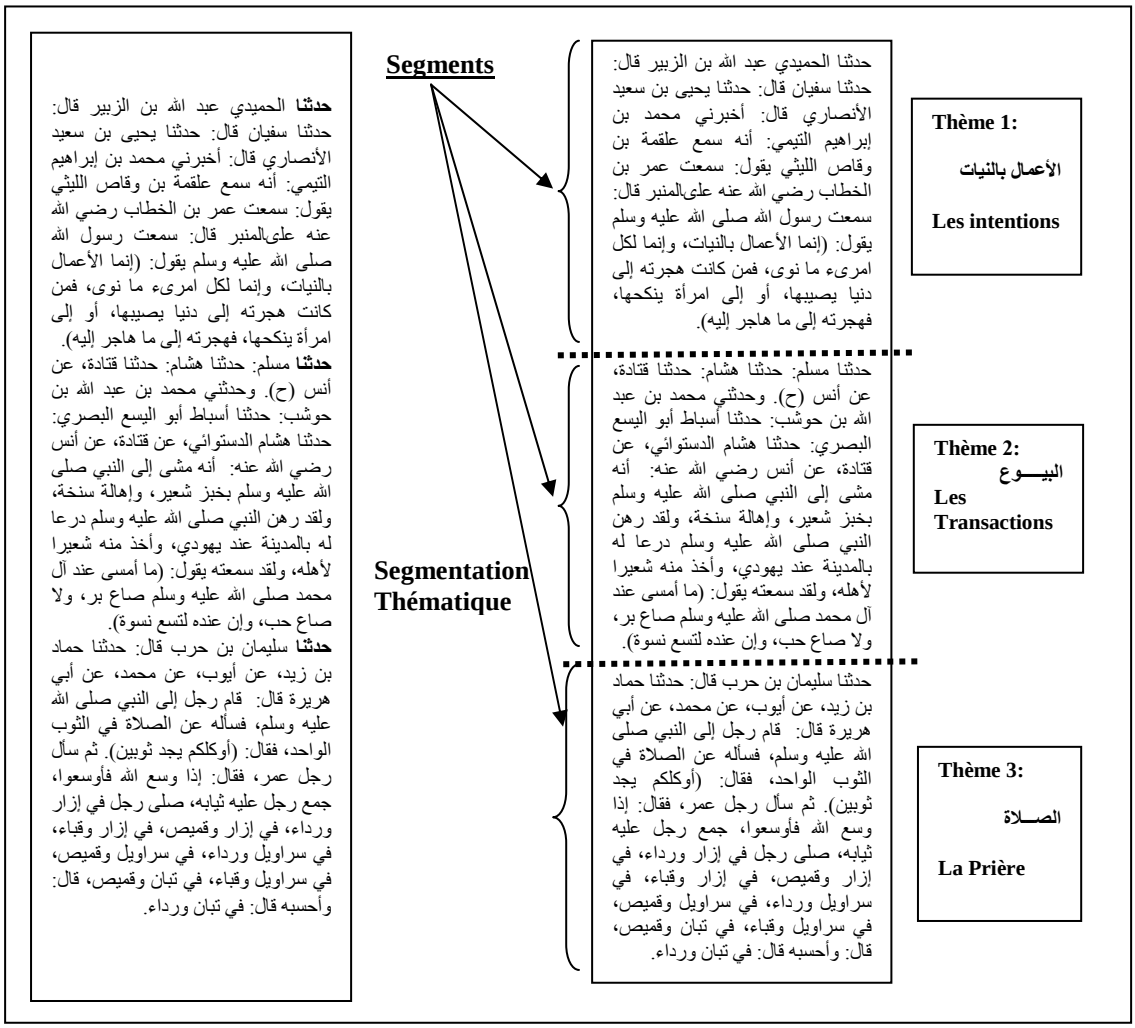

Fig. 7.1. Principe de la segmentation thématique.

3. Objectif de la segmentation thématique

La segmentation thématique peut être utilisée pour différents objectifs:

- La segmentation permet, par exemple, d'isoler des zones répondant précisément à une requête. Ceci est particulièrement utile dans un système de recherche d'informations.
- La segmentation peut également être utilisée pour l'indexation de textes.
- Les méthodes de classification de documents peuvent également s'appuyer sur la segmentation de textes.
- Les approches de résumés de textes peuvent utiliser les informations liées à la segmentation thématique.

4. Les différentes segmentations possibles

Beaucoup de moyens ont été mis au point pour segmenter un texte en thèmes cohérents. D'un coté, il existe une segmentation qui basée sur des méthodes supervisées ou non supervisées. D'un autre cote une segmentation ou le segment sont linéaires ou hiérarchiques [Khalis, 2006].

4.1. Supervisées ou non supervisées

La principale différence entre ces méthodes tient au fait qu'elles sont ou non supervisées. Les méthodes supervisées reposent sur une information *a priori* sur les classes importantes (leur nombre, leur signification, leurs caractéristiques statistiques), soit issue de bases de données, soit acquise lors d'une étape d'apprentissage. Parmi les méthodes supervisées, il y a par exemple PLSA [Brants et al., 2002] qui apprend des probabilités d'appartenance des termes à des classes sémantiques. Par opposition, les méthodes non supervisées qui ne nécessitent pas (ou très peu) d'information *a priori*, le caractère non supervisé porte sur l'estimation des caractéristiques statistiques des mots par des processus mathématiques de regroupement de données l'interprétation des résultats obtenus n'est alors effectuée qu'*à posteriori*. Parmi les méthodes non supervisées, on peut citer la méthode TextTiling [Hearst, 1997], la méthode C99 [Choi, 2000], Dotplotting [Reynar, 1998] et Segmenter [Kan et al., 1998].

4.2. Linéaire ou hiérarchique

Il existe également deux manières de segmentation soit linéairement (les segments de texte trouvés sont adjacents), soit hiérarchiquement (on cherche à repérer les phrases correspondant à un même thème).

5. Les méthodes de segmentation thématique

Cette partie présente un état de l'art des méthodes de segmentation thématique, on distingue trois grandes familles: les méthodes basée sur la répétition des mots, les méthodes basées sur le calcul de la similarité et les méthodes basées sur les chaînes lexicales.

5.1. Les méthodes basées sur une mesure de similarité

Les méthodes de segmentation à base de similarité considèrent les différents segments du texte à traiter comme autant de vecteurs. Les composantes des vecteurs étant, dans la plupart des cas, les fréquences d'apparition des mots au sein d'un segment du texte, après que celui-ci soit nettoyé des mots vides (mots jugés comme peu porteurs de sens). Parfois, cette fréquence des mots est pondérée par un IDF (*Inverse Document Frequency*), pour renforcer l'importance des mots supposés thématiquement saillants. L'objectif de ces méthodes est donc de mesurer la proximité ou l'éloignement des portions de textes étudiés grâce à l'angle que forment leurs vecteurs représentatifs. Elles s'appuient donc en général sur le cosinus de cet angle, qu'elles considèrent comme la similarité. La similarité est ensuite exploitée de diverses manières. Ces méthodes bien qu'efficaces deviennent rapidement inutilisables à mesure que le volume de données augmente.

5.1.1. La méthode de Salton étendue

[Salton et al., 1996] proposent une stratégie de segmentation mixte (*global-local Text comparaison*) calculant des similarités entre des zones de textes étendues (les paragraphes par exemple). Cette méthode utilise l'hypothèse que si les représentations vectorielles de deux extraits ont une faible similarité alors ces extraits ont un faible lien thématique. Ainsi deux extraits peu similaires donneront lieu à une segmentation du document qui les contient. L'algorithme de Salton procède à la décomposition des textes en segments et en thèmes où un segment est un bloc de texte contigu traitant d'un seul sujet et un thème est un ensemble de tels segments. Dans cette approche, le processus de segmentation commence au niveau des paragraphes. Ce choix d'unité minimale peut se justifier par le fait que les auteurs d'un texte exposent en général un point de vue par paragraphe.

Plus précisément la méthode de Salton pour la décomposition d'un document en thèmes s'articule autour des points suivants:

- Calculer les similarités entre différents paragraphes du document et retenir celles qui sont supérieures à un certain seuil (seuil1). Construire le graphe de similarité et en extraire les triangles. Un triangle est un ensemble de trois paragraphes fortement liés les uns aux autres, et donc susceptible de représenter une thématique cohérente.
- Pour chaque triangle construire un vecteur centroïdes (représentation vectorielle) qui est la moyenne des trois vecteurs représentant les paragraphes du triangle.
- Fusionner les triangles dont la similarité des vecteurs centroïdes est supérieure à un deuxième seuil (seuil2). Réitérer la fusion jusqu'à satisfaction d'un critère de convergence.

La méthode de Salton permet de décomposer chaque document d'un corpus en thèmes cohérents.

5.1.2. La méthode de Kozima

La méthode proposée dans [Kozima, 1993] se base sur la proximité sémantique entre les mots, calculée par une mesure de distance au sein d'un réseau sémantique. La préexistence d'un réseau sémantique adapté demeurant une contrainte très lourde. La mesure de cohésion lexicale définie par Kozima est appelée le profil lexicologique de cohésion LCP (*lexical cohesion profile*), qui localise des frontières de segment dans un texte. Les mots dans un segment sont liés ensemble par l'intermédiaire des relations lexicologiques de cohésion. Le LCP enregistre la similitude mutuelle des mots dans un ordre de texte. Kozima définie la cohésion lexicale en tant que similitude sémantique entre les mots, et a proposé une méthode pour la mesurer. La similitude entre les mots est calculée en écartant L'activation sur un réseau sémantique qui est systématiquement construit d'un dictionnaire anglais.

Les scores de LCP pour un mot particulier sont la somme des scores de similarité sémantiques qui résultent de la comparaison de ce mot à chaque mot dans une fenêtre des mots précédents. Kozima a essayé de prouver que les minimums locaux de ces scores de similarité correspondraient aux positions des frontières de thème en texte. Il a comparé les frontières identifiées en utilisant le LCP à la segmentation identifiée par 16 sujets qui ont marqué un texte dont les frontières de paragraphe avaient été enlevées. Kozima n'a pas évalué l'exécution de son algorithme, mais a déclaré que la segmentation produite par le LCP a ressemblé à celui qui produits par les annotateurs humains.

5.2. Les méthodes basées sur la répétition de termes

En passant par une représentation graphique des termes, il est plus facile de visualiser leur répartition le long du document étudié. Ainsi la méthode du nuage de points, présentée par Helfman [Helfman, 1996] emploie cette représentation pour la recherche d'information. Le principe est de positionner sur un graphique chaque occurrence des termes du document. Cette approche visuelle de la représentation d'un texte a été reprise et adaptée à la segmentation thématique par Reynar [Reynar, 1998] dans son algorithme Dotplotting.

5.2.1. La méthode Dotplotting

L'algorithme est proposé par [Reynar, 2000]. Il se base sur une représentation graphique du texte par les positions des occurrences des termes du texte à segmenter sont représentées sur un graphe. Lorsqu'un terme apparaît à deux positions du texte x et y, les quatre points (x, x), (x,

y), (y, x) et (y, y) sont représentés sur un graphe, ce qui permet de déterminer visuellement les zones du texte où les répétitions sont nombreuses. Cette méthode est adaptée par [Reynar. 2000] à la segmentation thématique de textes. Les positions de début et de fin des zones les plus denses du graphe sont les limites des segments thématiquement cohérents. La densité est calculée pour chaque unité d'aire en divisant le nombre de points d'une région par l'aire de cette région. A partir de là, deux algorithmes peuvent déterminer les frontières thématiques: identifier les limites en maximisant la densité au sein des segments, ou de repérer la configuration qui minimise la densité des zones entre les segments.

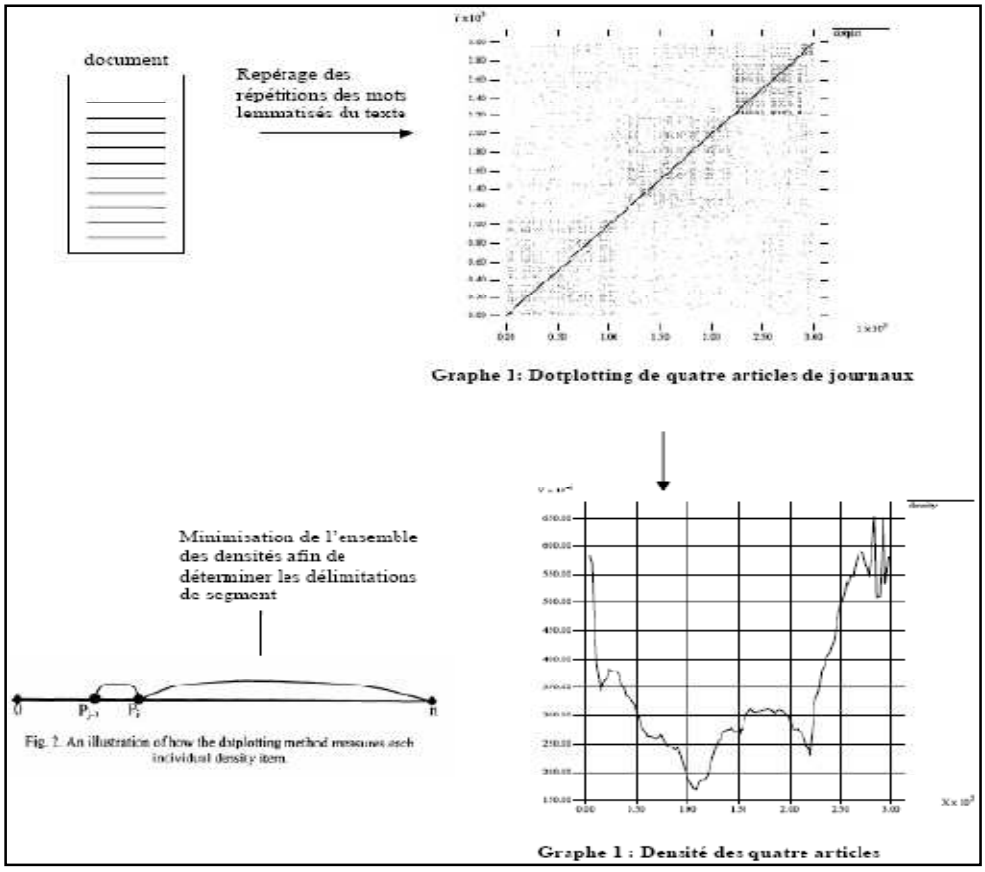

Fig. 7.2. Principe de Dotplotting [Reynar 2000]

5.2.2. La méthode de Masson

C'est la méthode de segmentation initiale [Masson, 1995] fondée simplement sur l'analyse de la distribution des occurrences des mots d'un texte. Une unité textuelle minimale, par exemple le paragraphe est représenté par un vecteur de la forme $G_i = (g_{i1}, g_{i2}, \ldots, g_{it})$ Où g_i représente le nombre d'occurrences d'un descripteur donné dans le paragraphe G_i. Les vecteurs sont constitués des descripteurs retenus à la suite du prétraitement du texte. Ils sont pondérés par le facteur *tf.idf* qui estime l'importance d'un descripteur par rapport à sa distribution dans le texte. …Il est défini par:

$$tfidf = tf_{ij} \cdot log \frac{N}{df_i} \qquad (7.1)$$

Où tf_{ij} est le nombre d'occurrences du descripteur T_j dans un paragraphe *i*; df_j est le nombre de paragraphes dans lesquels T_j apparaît et *N* le nombre total de paragraphes dans le texte. Ce facteur de pondération permet de renforcer les descripteurs, même rares, dont la distribution est centrée sur un ou quelques paragraphes. Inversement, les descripteurs, même très fréquents, dont la distribution est quasiment homogène sur l'ensemble du texte seront inhibés. Les ruptures thématiques sont ensuite déterminées par une mesure de distance vectorielle entre chaque paire adjacente de paragraphes (le paragraphe 1 est comparé au paragraphe 2, le 2 au 3 et ainsi de suite). La mesure de distance retenue est le *coefficient de Dice* défini pour deux vecteurs *X =(x1, x2,..., xt)* et *Y =* (*y*1, *y*2,..., *yt)* par :

$$c(x,y) = \frac{2\sum_{i=1}^{t} w(x_i)w(y_i)}{\sum_{i=1}^{t} w(x_i)^2 + \sum_{i=1}^{t} w(y_i)^2} \qquad (7.2)$$

Où *w*(*xi)* est le nombre d'occurrences d'un descripteur *xi* pondéré par le facteur *tf.idf.* Les faibles valeurs du coefficient indiquent des ruptures thématiques dans le texte alors que les fortes valeurs indiquent au contraire une cohérence thématique locale. Cette méthode de segmentation n'est applicable qu'à des textes dans lesquels les termes significatifs des thèmes développés sont souvent réemployés tout au long du texte.

5.2.3. Méthode de Richmond, Smith et Amitay

Richmond, Smith et Amitay décrivent également une technique pour localiser des frontières thématiquement [Richmond et al, 1997], leur méthode considère l'importance des mots basés sur leur fréquence sur un document et la distance entre les répétitions. Ils déterminent la similarité entre les régions voisines du texte en additionnant les poids des mots qui se produisent dans les deux régions et puis en soustrayant les poids additionnés de mots qui se produisent seulement dans un segment. Ils normalisent cette figure par la division par le nombre de mots dans chaque section. Leur algorithme a cinq étapes. D'abord, un certain prétraitement de base est fait. Après, ils calculent le poids de chaque mot, qu'ils appellent sa signification [Reynar, 1998]. Ils calculent ces valeurs en utilisant la formule ci-dessous. Les scores de signification pour différentes instances du même type de mot peuvent différer dépendant un contexte.

$$signifiance(x) = \frac{1}{n}\sum_{i=1}^{n} arctan\left(\frac{D_{xi}}{\frac{W}{w}}\right) \qquad (7.3)$$

x représente une marque particulière de mot. *W* est le nombre de marque de mot dans le document et *w* est le nombre d'occurrences des mots du même type que le mot *x*. D_{xi} est la distance entre le mot *x* et la $i^{ième}$ plus proche répétition de ce mot. *N* est le nombre de voisins les plus proches considérés utiles pour le calcul de signification et est déterminé par la formule ci-dessous.

$$n = \left[\frac{8}{1-e^{-200\cdot\left(\frac{w}{W}-0.02\right)}}\right] + 2 \qquad (7.4)$$

Les valeurs de *n* s'étendent de deux à dix. La *signifiance(x)* s'étend au commencement de 0 à $\pi/2$ et est normalisée entre 0 et 1, avec 0 significations minimum de marque. Richmond *et al.* utilisent la valeur de signifiance pour chaque mot, calculant la similarité entre deux régions d'un document. Ils ont déterminé la taille optimale des régions et ont comparé sur la base de quinze phrases. La formule pour la similarité entre les régions, ce qu'ils appellent Correspondance, est présentée ci-dessous:

$$Correspondance = \frac{\frac{|A'|-|A''|}{|A|}+\frac{|B'|-|B''|}{|B|}}{2} \quad (7.5)$$

Dans la formule ci-dessus A est le sac des mots contenus dans la première région et B est le sac trouvé dans la deuxième région. Un type de mot apparaît dans un de ces sacs plus d'une fois s'il se produit plus d'une fois dans la région associée du texte. A' et B' sont les parties de A et B, respectivement, contenant des mots des types qui se produisent dans A et B. A'' et B'' sont les parties de A et B, respectivement, qui contiennent des mots des types qui se produisent seulement dans A ou B. la notation $|A|$ indique l'addition des scores de signification des mots dans A. Richmond *et al.* ont converti les scores de correspondance en utilisant une formule pour calculer la moyenne de convolution (*Smoothing*). Finalement, ils ont placé les frontières thématiques aux endroits où les scores de correspondance sont petits. L'application de leur algorithme a été appliqué au texte des articles à partir d'un journal et d'un article de psychologie. Leurs résultats ont suggéré que l'algorithme est bien exécuté, mais ils n'ont pas réalisé une évaluation systématique en utilisant un corpus.

5.3. Les méthodes basées sur les chaînes lexicales

La segmentation à base de chaînes lexicales relie les occurrences multiples des mots dans un document et estime qu'une chaîne est rompue si la distance entre deux occurrences du même mot est trop importante. Cette distance est généralement exprimée en nombre de phrases. Le concept de chaînes lexicales a été élargi aux chaînes conceptuelles à l'aide de Word Net [Fellbaum, 1998] ou d'autres ressources sémantiques. [Kan et al., 1998] montrent que l'amélioration est très peu significative. Parmi les méthodes qui fondent à ce principe, on peut citer les suivantes.

5.3.1. La méthode Segmenter

Segmenter [Kan et al., 1998] effectue une segmentation linéaire (les segments trouvées sont adjacents) basée sur les chaînes lexicales présentes dans le texte cible. La détermination du changement de thème dépend principalement de l'extraction d'information thématique utile.

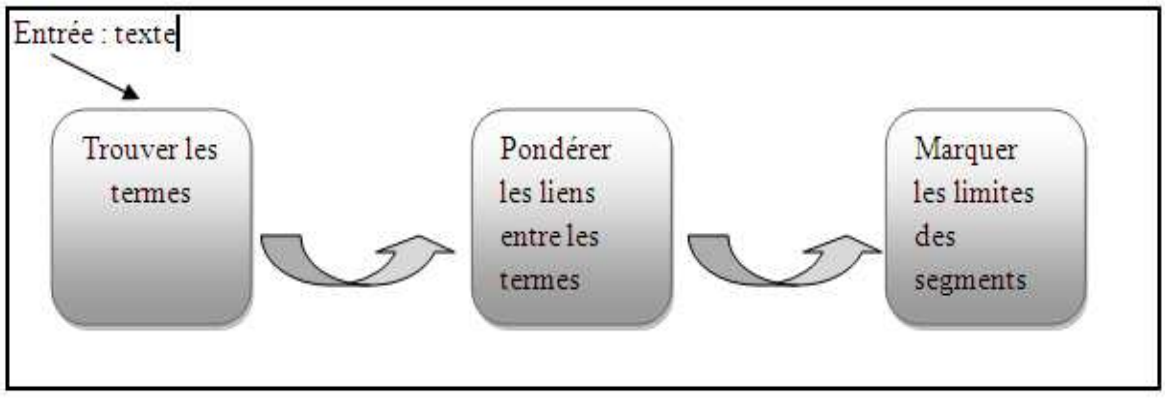

Fig. 7.3. Architecture de segmenter

Un segment est déterminé par un schéma de poids de somme égale à 0 en utilisant les occurrences des groupes nominaux et des formes pronominales. Après l'effectuation d'un prétraitement habituel, l'extraction des termes se fait selon trois catégories d'information:

- Les groupes de noms propres
- Les groupes de noms communs

- Les pronoms personnels et possessifs

Une fois les termes trouvés, les unités similaires sont combinés ensemble. Les groupes nominaux sont mis sous forme canonique selon leurs entêtes. Pour les pronoms possessifs, ils sont associés à leur pronom personnel respectif. Un seuil de filtrage des mots hors de propos est mis en place. Le seuil de fréquence est de deux occurrences pour déterminer la thématique, les groupes nominaux et les pronoms apparaissant une seule fois sont écartes. Ensuite les termes extraits et filtrés, sont évalués afin d'arriver à une segmentation. Etant donné un terme unique, soit un groupe nominal soit une forme pronominale, et la distribution de ses occurrences, les occurrences apparentées sont liées ensemble. La proximité est utilisée comme mesure de l'appariement des occurrences.

Une fois les liens établis, un poids est donné. Les délimitations de paragraphes n'étaient pas prises en compte dans les étapes précédentes, maintenant ils vont être pris en compte par la mise en place d'étiquette. Les étiquettes correspondent aux marquages des liens des occurrences de termes dans les paragraphes concernés. Par rapport à un paragraphe particulier, les résultats de pondération de la 0-somme sont soit négatifs soit positifs. Un score positif indique le début d'un segment et un score négatif la continuité d'un thème. L'utilisation de la 0-somme crée un problème pour trouver un seuil trivial malgré que les données soient normalisées autour de la valeur 0. Dans ce cas, le paragraphe de valeur maximum pour chaque groupe de paragraphe de valeur positive est pris comme délimitation de segment.

5.3.2. La méthode des chaines lexicales pondérées

Proposée par [Sitbon et Bellot, 2005], une chaîne lexicale relie des termes de manière linéaire dans un texte. Les méthodes actuelles de segmentation les utilisent pour relier les occurrences d'un même terme (ou lemme) qui sont "proches". Une chaîne est rompue lorsque le nombre de termes qui séparent deux occurrences dépasse une valeur fixée appelée hiatus. On peut alors recenser pour chaque phrase les chaînes actives.

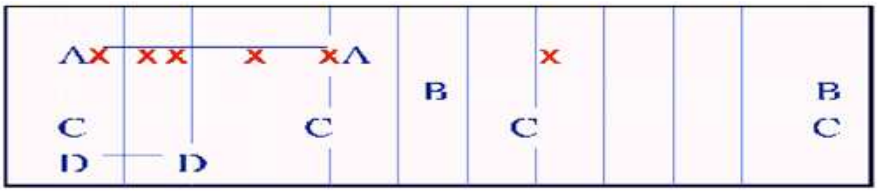

Fig. 7.4. Construction des chaînes lexicales

Les applications des chaînes lexicales utilisent actuellement des hiatus définis de manière empirique, et la notion d'activité d'une chaîne est binaire (elle est active ou non active). Le premier objectif est d'éliminer le caractère empirique du hiatus, afin que notre outil puisse s'adapter à n'importe quel type de texte sans intervention de l'utilisateur. Pour cela on peut imaginer tout simplement la suppression du hiatus. Ceci revient à relier toutes les répétitions de termes par l'utilisation de hiatus locaux : le hiatus moyen est calculé pour chaque lemme. Ainsi si un mot est fortement répété à deux endroits distincts du texte, il y aura automatiquement la création de deux chaines. De plus, s'il est répété trois fois en début de texte, puis une fois à la fin, il n'y aura qu'une seule chaîne comprenant les occurrences de début de texte. Ces techniques créent un déséquilibre dans le sens de l'activité des chaînes en jeu dans le calcul des frontières. Il faut alors pondérer les chaînes, en fonction de leur compacité (ratio entre leur taille et le nombre d'occurrences). [Galley *et al*, 2003] proposent une pondération des chaînes en fonction de la compacité et de la fréquence du terme considéré, et obtient de bons résultats, malgré un

hiatus déterminé de manière empirique. La catégorie des lemmes a été intégrée à cette pondération. Le poids d'une chaîne associée à un terme m est défini par:

$$score(chaine, m) = freq(chaine, m) \times log\left(\frac{l_{texte}}{l_{chaine}}\right) \quad (7.6)$$

Où *freq(Chaine, m)* est le nombre d'occurrences du terme m dans la chaîne, *Ltexte* la longueur du texte, *Lchaine* la longueur de la chaîne (on est alors indépendant de la taille des textes à segmenter), et *max (Chaine, cat(m))* le poids de la forme grammaticale la plus importante parmi les occurrences du terme dans la chaîne. Puis on calcule les similarités à chaque fin de phrase, qui est une rupture thématique potentielle. La similarité est calculée avec:

$$\mathrm{sim}(A, B) = \frac{\sum_m \mathrm{score}(A,m) \times \mathrm{score}(B,m)}{\sqrt{\sum_m \mathrm{score}(A,m) \times \sum_m (B,m)}} \quad (7.7)$$

Où *A* et *B* sont les ensembles de vecteurs représentant les poids des chaînes lexicales actives dans les *n* phrases avant et après (nous avons choisi *n=2*), *score(X, m)* étant le poids maximal du terme *m* dans l'ensemble des vecteurs *X*. Les frontières retenues sont alors celles pour lesquelles la similarité est en dessous d'un seuil déterminé par

$$\mathrm{sim}_{\mathrm{lim\ it}} = \mu + \frac{\delta}{2} \quad (7.8)$$

Où μ et σ Sont la moyenne et la variance de toutes les similarités calculées [Galley et al.2003].

5.3.3. La méthode de Morris et Hirst

Morris et Hirst ont décrit un algorithme de segmentation de discours [Morris et Hirst, 1991] basé sur la cohésion lexicale [Halliaye et Hasan, 1976]. Puisque l'un de leurs buts était de fournir l'appui pour la théorie de Grosz et de Sidner de structure de discours [Grosz et Sidner, 1986], leur algorithme divise le texte en segment qui forme la structure hiérarchique. La première étape dans l'algorithme de Morris et de Hirst est de lier les ordres des mots relatifs d'un document à des chaînes lexicales. Deux mots forment au commencement une chaîne lexicale quand ils sont reliés par cohésion lexicale. Chaque mot additionnel à une chaîne lexicale existante doit participer à une relation de cohésion lexicale avec au moins un mot déjà dans la chaîne. Morris et Hirst ont utilisés le thesaurus de *Roget* [Roget, 1977] pour déterminer si une paire de mot satisfait une de ces relations. Ils sont obligés d'identifier les chaînes lexicales à la main parce que le thesaurus de Roget n'était pas disponible en forme compréhensible par une machine. Morris et Hirst ont utilisé le thesaurus différemment pour identifier les chaînes lexicales. Ils ont décidés que si les paires de mots sont satisfaites pour la cohésion lexicale de Halliday et de Hasan, on examine alors les entrées d'index pour assurer les deux mots. En utilisant leur technique, deux mots sont considérés pour être reliés dans la même chaîne lexicale si l'un quelconque des cas suivants se présente :

- Les mots partagent une catégorie commune.
- Un mot est trouvé dans une catégorie qui contient un indicateur à une catégorie contenant le deuxième mot.
- Un mot est une étiquette d'une catégorie contenant l'autre mot.
- Chaque mot est dans une catégorie contenant une référence à une catégorie commune.
- Les mots sont dans le même groupe de la catégorie.

Morris et Hirst n'ont pas identifié des relations entre les paires de mots qui ont été largement séparés dans le texte. Ils sont manipulés des relations entre les mots éloignés qui, si plus étroits

ensembles, auraient été dans la même chaîne lexical en permettant aux chaînes lexicales d'être liées à une autre. Après qu'ils aient identifié toutes les chaînes lexicales dans un document, ils ont comparé les éléments des chaînes pour déterminer si les chaînes postérieures étaient effectivement des suites. Morris et Hirst ont analysé un nombre restreint de textes en utilisant leur algorithme, mais n'ont pas quantitativement évalué son exécution. Au lieu de cela, ils ont comparé leurs résultats avec la structure qu'ils ont identifiée pour chaque discours selon la théorie de Grosz et de Sidner. Ils ont constaté que les structures identifiées par leur algorithme d'enchaînement lexical étaient semblables aux structures qu'ils ont identifiées à la main.

5.4. Autre méthodes de segmentation thématique

5.4.1. La méthode basée sur l'apprentissage non-supervisé des concepts

C'est une approche basée sur des techniques d'apprentissage pour la segmentation de texte présenté par [Pessiot et al, 2004]. Cette méthode considère comme unité de base le paragraphe. Elle comporte trois étapes successives. On apprend tout d'abord les concepts. Dans une deuxième étape, on caractérise les paragraphes dans l'espace de ces concepts. On trouve finalement les différentes thématiques présentes dans la collection en regroupant les paragraphes "sémantiquement" proches au sens de ces concepts. On notera par $D = \{x_i\}_{[1..n]}$ L'ensemble des *n* paragraphes dans la collection et u_k les coordonnées du $k^{ième}$ centroïdes c_k, de la partition de mots.

- Etape 1

Le concept est définit comme un ensemble de mots qui sont déterminés à partir de l'analyse des cooccurrences de mots dans les paragraphes. Chaque mot w de V (l'ensemble des P mots du vocabulaire) est d'abord représenté par un vecteur $\overrightarrow{W} =< n(w,i) > i \in \{1..n\}$ de dimension n caractérisant le nombre d'occurrences du mot w dans chaque paragraphe x_i. En se basant sur cette représentation mot-paragraphe, on utilise un algorithme de partitionnement sur les mots (l'algorithme x-*moyenne*) pour trouver les différentes classes de mots dans l'espace des paragraphes. Les mots appartenant aux mêmes ensembles auront la même représentation dans le nouvel espace. Cette représentation sémantique, même si elle est relativement fruste, est considérablement plus riche que la représentation 'sac de mot 'des entités que l'on veut comparer. Elle permet également d'identifier très clairement quels sont les différents concepts présents dans un paragraphe. Formellement, on fait l'hypothèse que les mots w sont générés indépendamment par un mélange de densités:

$$p(\overrightarrow{w}) = \sum_{k=1}^{k} \pi_k p(\overrightarrow{w}|c = k) \qquad (7.9)$$

k dénote le nombre de partitions trouvées et π_k représentent les probabilités de classes. Les $\overrightarrow{w}$ sont ici les mots représentés dans l'espace des paragraphes (la modélisation des densités $p(\overrightarrow{w}|c = k)$ par des densités gaussiennes hyper sphériques de matrice de covariance commune $\Sigma = \sigma^2$.I était un bon compromis entre l'efficacité et la complexité). Les estimateurs du maximum de vraisemblance des centroïdes u_k sont :

$$\mu_k = \frac{1}{|c_k|} \sum w_j \ \in c_k \overrightarrow{w_j} \qquad (7.10)$$

$$\sigma^2 = \frac{1}{p-k} \sum (\overrightarrow{w_j} - \mu_{(j)})^2 \qquad (7.11)$$

En utilisant ce modèle de mélange, la log-vraisemblance du centroïdes c_k est :

$$\hat{\imath}_{\,m}(c_k) = \sum_{w_j \in c_k} \left(\frac{1}{\sqrt{2\pi\sigma}} - \frac{1}{2\sigma^2} \left\| \overrightarrow{w_j} - \mu_k \right\|^2 + \log \pi_k \right) \qquad (7.12)$$

Où ||. || désigne la norme euclidienne. Le meilleur modèle ici est celui qui maximise

Le critère CIB défini par

$$CIB = \sum_k \hat{\imath}_m(c_k) - \frac{p_m}{2} log\, p \qquad (7.13)$$

Où, *pm* est le nombre de paramètres du modèle probabiliste i.e. les moyennes et variances des composantes du mélange. L'algorithme *K-moyennes* débute avec *K = 2* sur tout le vocabulaire et avec un nombre maximum d'itérations *T*. Avec cette partition initiale, il utilise alors l'algorithme *2-moyennes* sur chaque classe et vérifie si ce nouveau partitionnement fait croître *CIB*. Si c'est le cas, la classe initiale est remplacée par deux de ses fils. L'algorithme recherche sur le vocabulaire entier la meilleure classe à partitionner

- Etape 2

Les segments de textes, ici les paragraphes, seront maintenant représentés dans l'espace des concepts. Pour cela, un paragraphe x_i sera caractérisé par un vecteur dans l'espace des concepts $\vec{x}_i =< \bar{n}(c,i) >_{c\in\{1,\ldots,|c|\}}$ où la caractéristique $\bar{n}(c,i)$ représente le nombre d'occurrences des mots du concept c dans le paragraphe x_i et $|c|$ est le nombre de concepts c découverts. Les caractéristiques d'un paragraphe dans cette nouvelle représentation traduisent le degré de représentation de chaque concept dans ce paragraphe. Dans la dernière étape les paragraphes seront comparés dans ce nouvel espace de concepts.

- Etape 3

Les paragraphes sont partitionné a l'aide d'une approche dite vraisemblance classifiante (VC) qu'est un formalisme général qui permet formuler et justifier de façon unifiée un grand nombre d'algorithmes de partitionnement qui ont été proposés. La vraisemblance classifiante définie sur les paragraphes est :

$$L_{vc} = \sum_{i=1}^{n} \sum_{k=1}^{\Omega} t_{ki}\, log\ \ p(\vec{x}_i \cdot y = k) \qquad (7.14)$$

Où n est le nombre de paragraphes dans la collection, Ω le nombre de classes (concepts) trouvées et t_{ki} un indicateur de classe qui vaut 1 si x_i appartient à la partition k et 0 si non. Les paragraphes sont supposés être générés indépendamment suivant un mélange de densités:

$$p(\vec{x}) = \sum_{k=1}^{\Omega} p(y = k) p(\vec{x}|y = k) \qquad (7.15)$$

Les paramètres du mélange sont estimés en maximisant le critère *VC*. L'algorithme utilisé pour la vraisemblance classifiante est l'algorithme *CEM* [Celeux, 1992]. C'est une technique itérative similaire à l'algorithme *EM* [Dempster, 1977].

En suivant une approche classique pour l'estimation de densité dans les textes, supposé que pour chaque composante du mélange $p(\overrightarrow{x_i}|y^{(j)} = k)$, les caractéristiques de V sont indépendantes et la densité sera estimée par un modèle baysien naïf.

Les paramètres de ce modèle sont l'ensemble des probabilités de mélanges $\pi_k = (p = k)$ et des coefficients des lois binomiales de chacune des caractéristiques p_{ck} qui dénotent la probabilité d'apparition d'une occurrence du concept c dans un paragraphe. Sous ces hypothèses, $(\vec{x}|y = k) \equiv \prod_{c=1}^{|c|} p_{ck}^{n(c,x)}$ avec $\bar{n}(c,x)$ qui est le nombre d'occurrences du concept c dans le paragraphe. En dérivant (7.15) par rapport aux π_k et p_{ck} et en introduisant des multiplicateurs de Lagrange pour prendre en compte les contraintes suivantes sur les parameters $\sum_k \pi_k = 1\ et\ \forall k, \sum_c p_{ck} = 1$, les estimateurs du maximum de vraisemblance pour π_k et p_{ck} sont:

$$\pi_k = \frac{\sum_{i=1}^{n} t_{kt}+1}{n+\Omega} \qquad (7.16)$$

$$p_{ck} = \frac{\sum_{i=1}^{n} t_{kt}\bar{n}(c,x_i)+1}{|x_i|+|c|} \qquad (7.17)$$

Où$|x_i| = \sum_{c=1}^{|c|} \bar{n}(c, x_i) \forall i$. Cette étape réalise donc une classification automatique du paragraphe dans l'espace des concepts préalablement appris.

5.4.2. Méthode basée sur les réseaux de cooccurrences lexicales

Cette approche a été proposée par [Ferret et al., 1998]. L'idée directrice dans cette approche est l'utilisation d'une méthode existante [Masson, 1995] qui est fondée sur la distribution des occurrences des mots d'un texte, et la renforcer par un réseau de cooccurrences lexicales qui permet de renforcer des descripteurs différents mais fortement liés lorsqu'ils sont employés dans une même unité, ou de créer de nouveaux descripteurs pour des unités. On tient compte ainsi du style du texte et de la manière dont la cohésion y est assurée.

Afin d'améliorer le processus de segmentation sur des textes où les mêmes mots sont peu répétés mais où la notion qu'ils portent apparaît sous la forme de mots pragmatiquement ou sémantiquement proches, ces mots sont alors ajoutés aux vecteurs représentant les paragraphes.

La modification des valeurs des descripteurs vise à rapprocher des paragraphes développant un même sujet, mais où les mots utilisés sont différents, ou peu répétés. L'idée est que si deux mots A et B du texte sont liés dans le réseau, alors "lorsque l'on parle de A, on évoque aussi un peu B, et réciproquement". Donc lorsque deux descripteurs A et B sont liés par une liaison de poids w dans le réseau, alors la présence des deux termes dans les paragraphes est renforcée la où ils sont simultanément présents, et on ajoute le descripteur absent du paragraphe lorsque l'un des deux seulement y figure. Dans le cas du renforcement, avec A présent k fois et B présent n fois, cela se traduit par l'ajout de wn au nombre d'occurrences de A et de wk à celui de B. Lors de l'ajout d'un descripteur, son poids est w fois le nombre d'occurrences du mot lié. Ce processus est effectué pour tous les couples des mots du texte, en utilisant le nombre d'occurrences réel des descripteurs pour calculer les ajouts, et non les valeurs modifiées.

Cette manière favorise l'émergence de descripteurs significatifs. Lorsqu'un ensemble de mots présents dans des paragraphes voisins possèdent des liaisons les uns avec les autres, ils vont ainsi se renforc mutuellement et tendre a rapprocher ces paragraphes. S'il n'y a pas de renforcement mutuel, les modifications apportées ne seront pas significatives et les paragraphes resteront alors séparés. Cet ajustement est effectué avant de pondérer la valeur des descripteurs par un facteur *tf.idf* modifié pour prendre en compte le fait que de nombreux descripteurs ont une tendance à se retrouver présents dans tous les paragraphes liés par nombreuses liaisons, même faibles, entre mots. On modère ainsi l'effet trop brutal de *tf.idf* par l'écart-type de la distribution du descripteur considéré. Le facteur résultant s'écrit:

$$log\left(\frac{N}{df_j} \cdot \frac{\sqrt{\sum_{k=1}^{N}\left(tf_{jk}-\overline{tf_t}\right)^2}}{df_j} + 1\right) \qquad (7.18)$$

Le réseau de cooccurrences lexicales est construit à partir d'un corpus de textes important. Les textes de ce corpus doivent être tout d'abord prétraités est analysés afin d'extraire les cooccurrences lexicales, les détailles de cette étape sont montrés ci-dessous:

- Etape1: Le prétraitement des textes de corpus

Cette étape a pour but de caractériser les textes par leurs mots significatifs (noms, verbes et adjectifs, en éliminant les noms propres, les abréviations et certains verbes). Chaque texte doit être retenu se sa forme canonique. Cette opération est réalisée par une chaîne de traitement ayant pour point de départ des textes sous forme *ASCII* accompagnés d'un balisage SGML élaboré automatiquement suivant le processus décrit dans [Adda et al. 1997]. Les textes sont premièrement segmentés à l'aide du segmenteur *MtSeg* du projet *Multext*. Après segmentation, les textes sont étiquetés à l'aide de l'étiqueteur *TreeTagger* [Stein et Schmid, 1995] afin de lever l'ambiguïté sur la catégorie lexicale des formes fléchies et d'obtenir la forme canonique des mots retenir. Le prétraitement des textes se termine par la sélection des mots qui représenteront les textes.

- Etape 2: Le réseau de cooccurrences lexicales

Le calcul des cooccurrences est réalisé à partir de la méthode décrite dans [Church et Hanks, 1990]. L'évaluation de ces cooccurrences est réalisée en faisant glisser de un mot en un mot sur une fenêtre d'une taille de 20 mots sur les textes du corpus. À chaque position de la fenêtre, on enregistre les cooccurrences entre le mot de tête et les autres mots de la fenêtre.

À la suite du calcul des cooccurrences, il doit opérer une sélection afin de ne retenir que les plus significatives et filtrer ainsi le bruit. Seules les cooccurrences de fréquence supérieure à 5 sont conservées, ce qui représente à peu près 1/3 de celles présentes initialement. Ensuite il faut adopter une estimation de l'information mutuelle comme mesure de la cohésion entre deux mots. L'utilisation de cette valeur de cohésion conduit à la normaliser. Cette normalisation consiste à diviser chaque valeur d'information mutuelle par l'information mutuelle maximale relative au corpus, donnée par la formule suivante:

$$Imax = log_2(N^2(T_f - 1)) \qquad (7.19)$$

Où N est taille du corpus et T_f est la taille de la fenêtre.

5.4.3. Méthode basée sur les réseaux de collocation et les récurrences des mots

Cette méthode de segmentation consiste à traiter les textes linéairement. Elle détecte les changements de thème (les ruptures des segments) à l'aide d'un réseau de mots que l'on appelle réseau de collocations. Cette méthode se déroule à travers le processus suivant: une fenêtre délimitant l'espace de focalisation de l'analyse est déplacée sur le texte considéré. Cette fenêtre contient sous forme lemmatisée les mots pleins du texte issus de son prétraitement. Un contexte thématique est associé à cette fenêtre de focalisation. Il est constituée à la fois des mots de la fenêtre et des mots d'un réseau de collocations. A chaque position de la fenêtre, un contexte thématique caractérise la dimension thématique de l'entité à laquelle est associée la fenêtre par l'intermédiaire de deux vecteurs : le vecteur texte et le vecteur collocation. On sélectionne les mots du réseau de collocations qui sont liés à au moins trois mots de la fenêtre pour évaluer la cohésion lexicale d'un texte.

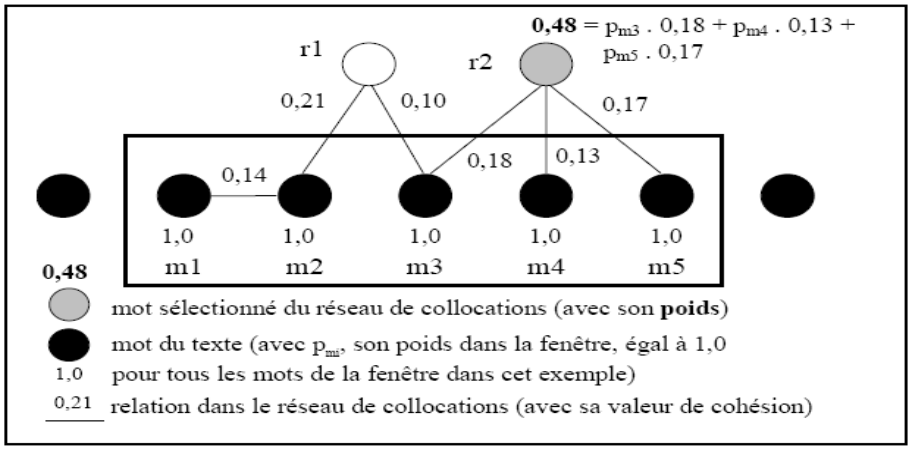

Fig. 7.5. Sélection et pondération des mots du réseau de collocations [Ferret et Grau, 1998].

Un segment de texte est alors défini par l'ensemble des mots successifs correspondant à des positions de la fenêtre ayant une forte valeur de cohésion, Le contexte thématique d'un segment est le produit de la fusion des contextes associés à la fenêtre de focalisation lorsque celle-ci se trouve dans le segment. Cette fusion est réalisée à chaque nouvelle extension du segment. Cette combinaison, réalisée séparément pour les vecteurs textes et collocation, consiste à fusionner deux listes de mots pondérés. Pour déterminer si le contenu de la fenêtre de focalisation est thématiquement cohérent avec le segment courant, on compare les contextes associés à ces deux entités. Cette comparaison est réalisée en deux étapes: une mesure de similarité est d'abord calculée entre les vecteurs des deux contextes; les valeurs obtenues sont ensuite exploitées par une procédure de décision statuant sur la similarité des deux contextes. La mesure du cosinus est utilisée pour évaluer le degré de similarité entre un vecteur du contexte de la fenêtre (*Vf*) et le vecteur de même type dans le contexte du segment(*Vs*)

$$sim(vf_x, vf_y, t) = \frac{\sum_i poids_x(m_i,cs,t)\cdot poids_y(m_i,cf,t)}{\sqrt{\sum_i poids_x(m_i,cs,t)^2 \cdot \sum_i poids_y(m_i,cf,t)^2}} \qquad (7.20)$$

Afin de minimiser le bruit dans les vecteurs, cette mesure ne prend en compte que les mots les plus récurrents des contextes des segments, l'importance d'un mot dans un contexte étant supposée corrélée avec sa récurrence au sein de celui-ci. Cette récurrence est définie comme la proportion, parmi les contextes de fenêtre de focalisation ayant permis de constituer le contexte de segment, de ceux contenant le mot considéré. Si la similarité entre le contexte de la fenêtre de focalisation et celui du segment actif est rejetée on déduit la présence d'un changement de thème à la position correspondante et le segment actif est clos. Sinon, le segment actif est étendu afin d'englober la position courante.

6. Conclusion

La segmentation thématique a connue un grand essor dans les dernières années avec le développement du Web et de la disponibilité de grandes quantités d'informations sous forme de texte. Dans ce chapitre on a essayé de décrire les différentes approches de la segmentation thématique à savoir les approches linéaires ou hiérarchique et les approches supervisées ou non supervisées. L'état de l'art du domaine de segmentation thématique nous a permis de distinguer trois grandes familles de méthodes: les méthodes basées sur la répétition des mots, les méthodes basées sur le calcul de la similarité et les méthodes basées sur les chaînes lexicales. Dans le prochain chapitre, nous d'étudierons l'efficacité des algorithmes à base de cohésion lexicale comme un moyen de segmentation thématique des textes prophétiques.