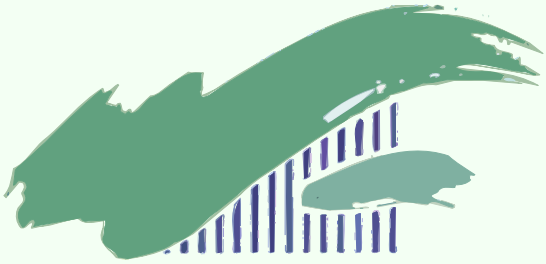

UNIVERSITE D'ORLEANS
Faculté des Sciences
LIFO

Laboratoire d'Informatique Fondamentale d'Orléans
4, rue Léonard de Vinci, BP 6759
F-45067 Orléans Cedex 2
FRANCE

Rapport de Recherche

www : http://www.univ-orleans.fr/SCIENCES/LIFO/

CARTOGRAPHIE DE TEXTES
Une nouvelle approche pour l'exploration sémantique des corpus homogènes de grande dimension

Isabelle Debourges, Sylvie Guilloré-Billot, Christel Vrain
Université d'Orléans, LIFO

Rapport N° **2001-01**

Résumé

Nous présentons les avancées d'un projet dans un thème que nous qualifions de *Cartographie de Textes* (ou *Text Mapping*) qui permet à un utilisateur novice d'explorer un nouveau domaine par navigation au sein d'un corpus homogène grâce à des cartes conceptuelles interactives. Une carte est composée de concepts pertinents relativement à la requête initiale et à son évolution, au sein du corpus; des relations sémantiques ou lexicales, extraites du corpus, les lient aux mots clés de la requête. Des techniques d'Apprentissage Automatique sont combinées avec des heuristiques statistiques de Traitement Automatique des Langues pour la mise en évidence de collocations afin de construire les cartes. Un prototype implantant ces idées a été écrit et les expérimentations menées sur différents corpus de langues anglaise et française montrent des résultats encourageants.

Mots Clés: Cartographie de Textes, Recherche Documentaire, Extraction d'Information, Apprentissage Automatique.

Abstract

We present an ongoing research project on the new field of Text Mapping that allows a novice user to explore a new domain by navigation through an homogeneous corpus thanks to interactive conceptual maps. A map is composed of concepts (the nodes) depending on the user's request and its evolution, and semantic/lexical relations (the links). Machine Learning techniques are combined with Natural Language Processing methodologies to build the maps. The algorithms are implemented in a prototype. Experiments on different english and french corpora show encouraging results.

Keywords: Text Mapping, Information Retrieval, Information Extraction, Machine Learning.

1 Introduction

La quantité des textes électroniques disponibles sur les réseaux s'est accrue à une vitesse vertigineuse ces dernières années. Les utilisateurs ont aujourd'hui besoin d'outils qui leur permettent de retrouver ce qu'ils cherchent au sein de cette gigantesque source d'informations. La Recherche Documentaire et l'Extraction d'Infor-mation sont les deux grands types d'aide mis à la disposition de l'utilisateur pour l'exploitation des larges ressources textuelles électroniques.

La **Recherche Documentaire** [29, 13, 30] nécessite la saisie de mots clés par l'utilisateur pour produire une sélection de textes jugés pertinents vis à vis de cette requête. Certains outils proposent une évaluation de la pertinence de chaque texte sélectionné par rapport à la demande initiale. Cependant, bien que la quantité des textes sélectionnés soit déjà un premier filtrage relativement à la quantité globale des textes disponibles, de cette étape ressortent encore de nombreuses références. Ainsi, l'utilisateur n'a que très rarement la possibilité de parcourir lui-même l'ensemble des textes sélectionnés par un outil de recherche documentaire aussi performant soit-il. Le manque de temps et/ou le volume des textes à exploiter en sont les principales causes. Si de plus la requête est un peu vague, le phénomène sera encore aggravé, et la tâche deviendra insurmontable à la main. Un utilisateur qui souhaite découvrir un domaine ne trouvera donc pas ici un outil suffisamment adéquat à son problème.

L'**Extraction d'Information** [35, 14, 19] peut, à partir d'un ensemble de textes traitant d'un thème commun et de patrons d'extraction (ou *templates*), fournir des instanciations de ces patrons à partir des informations contenues dans les textes du corpus fourni. Cette approche des grandes ressources textuelles permet d'extraire, à partir d'un modèle, des informations de même type, en repérant les instanciations de celui-ci présentes dans les textes. Cette méthode de traitement permet une recherche rapide d'informations de nature semblable. Un tel besoin n'est pourtant pas quotidien chez un utilisateur lambda. De plus, avant l'utilisation d'un outil d'Extraction d'Information, il est nécessaire que l'utilisateur définisse un patron d'extraction. Il faut donc qu'il soit suffisamment familier du thème traité par le corpus pour être capable de définir ce patron. On pourrait envisager que l'**acquisition automatique de patrons** [27, 6, 7, 18] puisse aider notre utilisateur. Mais ces outils demandent une validation des patrons qu'ils proposent par l'utilisateur... ce qui est difficile si l'utilisateur ne connait pas le domaine .

Ainsi, un utilisateur qui souhaite découvrir un nouveau domaine ne peut-il pas accéder aisément aux informations qu'il recherche sur celui-ci. En effet, il ne pourra pas exploiter l'ensemble des documents qu'il obtiendra par Recherche Documentaire. Il ne sera pas non plus capable de définir des champs pertinents pour des patrons d'Extraction d'Information puisqu'il ne connaît pas bien le domaine. Les aides proposées à l'utilisateur de grandes ressources textuelles électroniques, représentées par la Recherche Documentaire et l'Extraction d'Information sont donc une première avancée. Mais elles restent largement insuffisantes dans le cas où la recherche porte sur un domaine où le nombre des textes disponibles n'est pas restreint. Le problème subsiste pour un corpus ne

contenant pas d'informations correspondant à un même modèle et donc pour lequel on ne peut pas définir de patron d'extraction.

1.1 Proposition d'une solution

Afin de s'attaquer à ce problème de recherche d'information pertinente dans des collections de textes, nous proposons une méthode innovante pour la construction de *cartes conceptuelles* de textes. L'architecture générale présentée à la figure 1 est la suivante. L'utilisateur constitue, à l'aide d'un outil de recherche documentaire, un corpus homogène[1] spécialisé dans le domaine qui l'intéresse. Afin de lui permettre d'extraire les connaissances présentes au sein de son corpus, la *Cartographie de Textes* lui propose des cartes conceptuelles interactives présentant les concepts les plus pertinents ainsi que les relations qui les lient, au sein des textes. Les *cartes conceptuelles* proposées dépendent directement de la requête et de son évolution. Ainsi l'utilisateur peut-il demander une carte plus spécifique autour d'un concept qu'il sélectionne. La carte obtenue tiendra alors compte des requêtes précédentes.

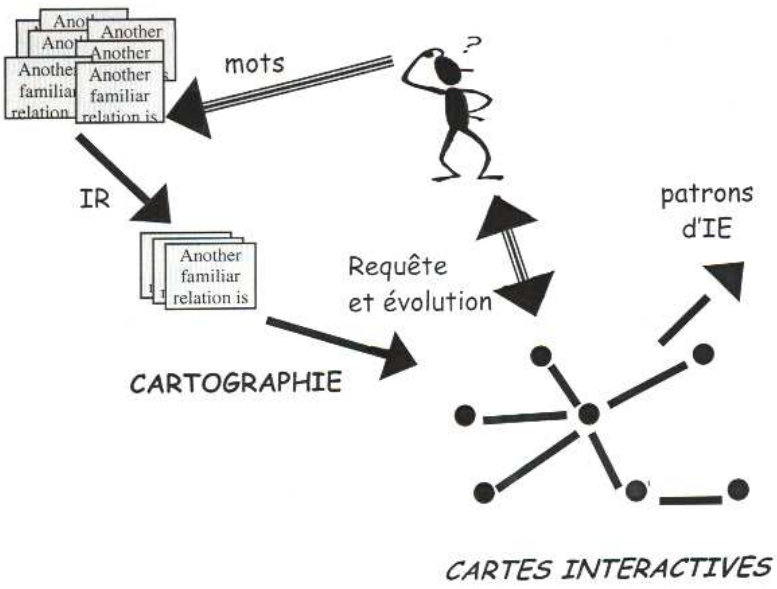

FIG. 1 – *Architecture générale proposée.*

Une *carte conceptuelle* est un réseau dynamique construit autour des mots de la requête. Elle est constituée:

- de concepts pertinents par rapport à la requête courante, en tenant compte des requêtes précédentes

1. il est préférable que le corpus soit *homogène* du point de vue du *genre* des textes et de la *langue* utilisée

– des relations sémantiques ou lexicales qui émergent au sein du corpus entre les concepts extraits et les mots de la requête.

Comme les cartes sont interactives et dépendent de l'évolution de la requête de l'utilisateur, des informations plus fines peuvent être obtenues à partir de la carte courante.

Ainsi, un utilisateur novice qui souhaite découvrir un nouveau domaine peut-il constituer un corpus homogène spécialisé en utilisant un outil de recherche documentaire. Il peut ensuite, au fil des cartes et des nouvelles requêtes explorer les notions fortes contenues au sein des textes. Il prend ainsi connaissance du contenu sémantique du corpus sans avoir à procéder à une lecture exhaustive du texte. De plus, les idées qui ressortent dans les cartes peuvent lui permettre de retrouver rapidement les séquences de texte qui y sont liées, par des pointeurs vers le texte original. Les cartes obtenues et les connaissances acquises lui permettront de construire, s'il le souhaite, des patrons d'extraction pertinents par rapport au corpus. La cartographie de textes constitue en ce sens un lien nouveau entre recherche documentaire et extraction d'information.

Nous proposons une méthode de construction automatique des cartes conceptuelles. Elle est composée de trois grandes étapes. Le corpus est tout d'abord prétraité pour permettre des calculs plus rapides au cours des étapes suivantes. La construction d'une carte comprend:

1. l'extraction des concepts pertinents au sein du corpus, par rapport à la requête courante et les requêtes antérieures.
2. l'émergence des relations sémantiques et lexicales présentes au sein du corpus, entre les mots clés de la requête et les concepts extraits.

Dans la seconde partie de ce papier, nous proposons un algorithme pour chacune des phases de construction des cartes conceptuelles. Ces algorithmes sont inspirés de différents courants tels que les statistiques, l'apprentissage automatique ou le traitement automatique des langues. En effet, l'extraction des concepts utilise les phénomènes de cooccurrence et de collocation au sein des phrases contenant des mots sémantiquement proches de ceux de la requête. L'émergence des relations est basée sur un algorithme d'apprentissage automatique dérivé de celui de Agrawal & al [32, 1, 2].

La troisième partie est consacrée à la présentation de résultats obtenus sur des corpus homogènes (du point de vue du contenu sémantique, mais aussi du genre du discours) par le prototype implantant ces algorithmes.

1.2 Travaux apparentés

La cartographie de textes est originale dans sa façon d'aider l'utilisateur de grandes ressources documentaires, et dans le lien qu'elle tisse entre Recherche Documentaire et Extraction d'Information [15].

Pourtant, cette carence de méthodes et outils dans le domaine de la recherche d'information au sein des immenses ressources textuelles disponibles a déjà été constatée et a donné lieu à des méthodes hypertextuelles avec la création automatique de liens. Ainsi que le souligne Jean-Marie Viprey [34],

la complexité inhérente et la polysémie des objets sont telles que d'autres méthodes doivent être développées. Il propose alors d'appliquer la statistique multidimensionnelle endogne du vocabulaire, et les graphes qu'elle produit, en vue d'une cartographie hypertexte ouverte, interactive, dynamique, pertinente son objet, modle et mmoire des parcours critiques effectus et faire.

Par rapport à notre approche, mme s'il garde un historique des parcours effectus, il ne propose pas d'en tenir compte dans les parcours venir. [16, 3, 25] se sont également intéressés à la génération automatique de liens hypertextes.

Dans[22], A. Lelu M. Halleb, B. Delprat se sont interessés la "recherche d'infor-mation et cartographie dans des corpus textuels à partir des fréquences de n_grammes". Cette approche leur permet d'identifier des motifs à l'aide d'une métrique qu'ils ont définie, afin de pouvoir travailler sur des textes écrits dans d'autres alphabets que l'alphabet latin. Ils donnent les résultats de premières expérimentations sur des textes russes, chinois ou écrits en glyphes mayas. La méthode utilisée s'apparente à l'analyse des données textuelles.

D'autre part, la cartographie de concepts (Concept Mapping[2]) est une technique de représentation de la connaissance sous forme de graphes, développée dès les années 1960 par Prof. Joseph D. Novak à Cornell University [26]. Les graphes de connaissances sont des réseaux de concepts. Les réseaux sont constitués de noeuds (les concepts) et de liens (les relations entre concepts). Les concepts, et quelquefois les liens, sont étiquetés. Les liens sont soit non-, uni-, ou bi-directionnels. Ce travail est basé sur les théories de David Ausubel, qui a souligné l'importance de connaissances minimales pour être capable d'apprendre de nouveaux concepts. La cartographie de concepts peut être utilisée à plusieurs fins:

- pour générer des idées (brain storming ...)
- pour représenter une structure complexe (textes longs, grands sites web ...)
- pour communiquer des idées complexes
- pour aider l'apprentissage en intégrant explicitement de nouvelles et d'anciennes connaissances
- pour évaluer le compréhension ou diagnostiquer l'incompréhension

La différence entre les cartes de concepts de D. Novak et les "cartes de l'esprit" ("mind maps") de Tony Byzan [5] est que les mind maps ne traitent que d'un concept principal, alors que les cartes de concepts peuvent traiter simultanément de plusieurs concepts principaux. Ceci se traduit par le fait qu'une mind map peut être représentée comme un arbre, alors qu'une carte de concepts peut nécessiter une représentation sous forme de réseau.

2. The Concept Mapping homepage: http://www.to.utwente.nl/user/ism/lanzing/cm_home.htm

2 Algorithmes pour la construction des cartes conceptuelles

Pour chaque requête (initiale ou raffinement), les concepts pertinents sont tout d'abord extraits: ils deviennent les noeuds de la carte conceptuelle, autour des mots clés de la requête. Ensuite les relations sémantiques/lexicales entre les mots clés et les concepts sélectionnés sont extraites. Elles étiquettent les liens de la carte conceptuelle.

La construction des cartes conceptuelles que nous proposons utilise deux algorithmes successifs: un algorithme pour l'extraction des concepts, et un algorithme pour l'émergence des relations sémantiques et lexicales. Un prétraitement initial du corpus facilite ces deux étapes.

2.1 Le prétraitement du corpus

Le prétraitement du corpus rassemble des petits traitements simples bien que fort utiles pour la construction des cartes conceptuelles.

Le texte est découpé en phrases ou segments compris entre deux ponctuations fortes. Les caractères spéciaux sont supprimés et le texte est entièrement écrit en lettres minuscules. Le texte est ensuite lemmatisé et une catégorie grammaticale est associée à chaque mot.

Pour les textes de langue anglaise, nous utilisons les ressources de WordNet[3] [23] ainsi que la grammaire simple, donnée ci-dessous, du même genre que celle de Jacques Vergne [33].

si *word* existe en temps que ADVERBE alors tiquette ← Adv
sinon si *word* existe en temps que NOM alors tiquette ← Nom
sinon si *word* existe en temps que ADJECTIF alors tiquette ← Adj
sinon si *word* existe en temps que VERBE alors tiquette ← Verbe
sinon tiquette ← Inconnu[4]

La catégorie grammaticale sélectionnée est alors "collée" au mot concerné. La lemmatisation est utilisée pour une amélioration de la précision dans l'extraction des concepts. En effet, les formes fléchies d'une même racine doivent être dénombrées ensemble, de même que les diverses conjugaisons d'un verbe au sein des textes (cf 2.3). Les étiquettes grammaticales seront ensuite utilisées pour la sélection des meilleures relations sémantiques/lexicales (cf 2.6).

Les mots vides sont omis, en utilisant une stop-liste. Ce traitement, souvent considéré comme trop "radical" par les linguistes prendra une dimension plus modérée dès que les unités polylexicales seront mises en évidence [11] avant

3. WordNet est un système lexical de référence disponible par ftp anonyme depuis *clarity.princeton.edu cd pub/wordnet* [23] ou *http://www.cogsci.princeton.edu/~wn/*. Sa structure est inspirée des théories psycholinguistiques sur la mémoire lexicale humaine. WordNet a été développé par un groupe de psycholinguistes et linguistes depuis 1985 à l'Université de Princeton.

4. L'étiquette Inconnu signifie que ce mot n'apparaît pas du tout dans la base WordNet

une telle suppression. Dans l'état actuel des choses, les unités polylexicales apparaissent, mais elles sont disloquées.

2.2 De cooccurrences à collocation

Les termes de cooccurrence et collocation, que nous utilisons dans la suite du propos, demandent peut-être une mise au point quant à leur définition [8, 9]. En effet, le terme collocation a été introduit dans les années 1930 pour caractériser certains phénomènes linguistiques de cooccurrence, mais il reste tout de même un peu flou. Nous utilisons les mots *cooccurrence* et *collocation* dans un cadre où ils peuvent être définis de la façon suivante:

- La cooccurrence de deux unités linguistiques $U1$ et $U2$ caractérise le fait qu'il existe des séquences de texte (de une à plusieurs) dans lesquelles $U1$ et $U2$ apparaissent simultanément, sans notion d'ordre, au sein du corpus.
- La collocation de $U1$ et $U2$ est le reflet d'un phénomène remarqué de cooccurrences régulières et en grand nombre de $U1$ et $U2$.

Par extension, *la collocation* est ici vue comme l'apparition régulière et en grand nombre de deux unités lexicales dans les mêmes énoncés. Tandis que *la cooccurrence* se restreint au fait que deux unités lexicales existent ensemble dans une même unité de contexte. Ainsi, la cooccurrence est-elle un élément nécessaire mais non suffisant à la propriété de collocation. Dans la littérature, l'attention étant bien souvent attirée vers le dénombrement des cooccurrences, celles-ci deviennent le centre du propos. La collocation devient alors l'Arlésienne de la linguistique [37].
Notre approche est donc apparentée à la tendance contextualiste de Sinclair [SIN 91] [31] pour qui *la collocation est un phénomène textuel défini en fonction de l'apparition de cooccurrences significatives à l'intérieur d'une fenêtre.*

2.3 L'Extraction des Concepts

L'algorithme de sélection des concepts les plus pertinents par rapport à la requête (ensemble de mots clés) est basé sur la volonté de sélectionner les unités lexicales (ou polylexicales) qui apparaissent dans le contexte de la requête. Pour cela, l'ensemble des mots clés est tout d'abord étendu en ajoutant les synonymes/hyponymes/hyper-onymes[5] ... afin de sélectionner les phrases traitant de la même idée, même si l'auteur a fait un effort pour ne pas répéter exactement le même mot. Ensuite, on sélectionne les phrases qui contiennent au moins un élément de l'ensemble des mots clés étendus; l'ensemble des termes les plus fréquemment rencontrés `dans le contexte de la requête` est généré: les concepts qui apparaissent le plus avec les mots clés sont sélectionnés. Les itérations suivantes du processus (sélection de phrases, et calcul de fréquents) `propagent` la sélection des mots apparaissant dans le contexte de la requête. Le processus s'arrête quand le point fixe est atteint.

Cet algorithme d'extraction des concepts est original sous deux aspects: l'extraction des concepts dépend directement des mots clés fournis par l'utilisateur et l'heuristique est construite sur une recherche de point fixe.

5. pour les textes de langue anglaise, nous utilisons le produit libre WordNet

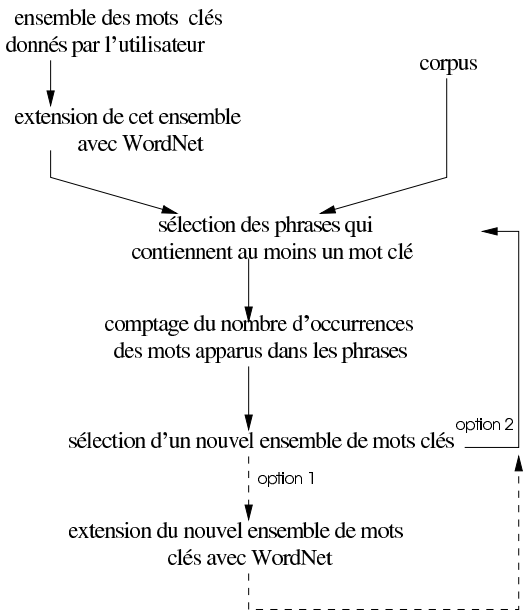

FIG. 2 – *Algorithme d'Extraction des Concepts.*

1. L'ensemble des mots clés fournis par l'utilisateur est étendu: les synonymes, hyponymes, hyperonymes, ... sont ajoutés, relativement aux choix faits par l'utilisateur dans la définition des paramètres. Un nouvel ensemble de mots clés est alors constitué, appelé $Keys$
2. chaque phrase du corpus qui contient au moins un mot de $Keys$ est sélectionnée. Un poids est associé à chaque phrase sélectionnée: $weight(s) = n$ où n est le nombre d'éléments de $Keys$ qu'elle contient
3. une mesure est associée à chaque mot w:

$$m(w) = \sum_{s_i \ contenant \ w} weight(s_i) \tag{1}$$

4. un nouvel ensemble $Keys$ de mots clés est sélectionné, par rapport aux poids (par exemple, ceux dont la mesure est la plus élevée)
5. par rapport à l'option d'extension (choisie par l'utilisateur au début), l'ensemble des mots clés $Keys$ est ou n'est pas étendu
6. le processus est réitéré, à partir du point 2, en utilisant le corpus entier.

Le processus n'est plus itéré à partir du moment où l'ensemble des mots obtenus et leurs poids respectifs a convergé vers un point fixe. (Les expérimentations actuelles atteignent un point fixe après quatre à sept itérations). On observe que pour atteindre le point fixe: 1) l'ensemble des concepts émergents et leur ordre d'apparition dans la liste classée par poids décroissants se stabilisent, 2)les poids associés aux concepts se stabilisent.

Insistons bien sur le fait que les poids des phrases changent à chaque nouvelle itération puisque l'ensemble des mots clés a évolué.

Raffinement A la fin du processus, l'utilisateur peut demander une vue plus fine sur un concept spécifique lié à un mot clé dans la carte courante. Dans ce cas précis, le corpus initial est réduit à l'ensemble des phrases contenant les mots clés étendus de la carte.

2.4 Exemple

Le tableau ci-dessous[6] présente les résultats obtenus sur le corpus constitué du livre "Introduction au langage C" de Bernard Cassagne. Les mots clés initiaux fournis portent sur le thème de l'affectation.

6. Les mots sont indiqués en **gras** lors de leur première apparition

Itération	Mots clés courants et poids	Seuil de sélection
0	affecter, affectation, affecte, affectations	0
1	39: affecter 36: affectation 35: **valeur** 25: **type** 19: **expression** 17: **peut** 15: **variable** 14: **tableau** 13: **oprateur**	13
2	673: type 577: valeur 348: peut 331: tableau 276: expression 212: **si** 209: **int** 206: variable 198: oprateur 184: **l** 168: **variables**	168
3	1322: type 1060: valeur 720: int 692: si 661: peut 658: tableau 630: l 555: expression 494: **d** 489: **fonction** 462: **vers** 454: **pointeur** 453: **comme** 434: **deux** 422: variables 379: variable 346: oprateur	330
4	1948: type 1441: valeur 1021: int 999: si 970: tableau 960: l 952: fonction 889: peut 875: vers 867: comme 838: pointeur 822: d 776: deux 688: expression 564: variables 522: variable 501: **paramtre** 488: **exemple**	487
5	1995: type 1470: valeur 1047: int 1032: fonction 1020: si 997: tableau 974: l 905: vers 903: peut 901: comme 856: pointeur 848: d 788: deux 690: expression 647: exemple 620: paramtre 570: variables 552: variable	498
6	1995: type 1470: valeur 1047: int 1032: fonction 1020: si 997: tableau 974: l 905: vers 903: peut 901: comme 856: pointeur 848: d 788: deux 690: expression 647: exemple 620: paramtre 570: variables 552: variable	498

Le lecteur aura certainement remarqué que ce texte de langue française n'a pas été lemmatisé au cours du prétraitement. Les résultats obtenus peuvent en être légèrement modifiés [4]. Pourtant, ils permettent de constater que l'ensemble des mots clés obtenus évolue à chaque itération. La première itération donne des concepts très proches de ceux de la requête. Puis, la propagation joue son rôle, et d'autres concepts entrent dans la sélection. Le point fixe est atteint après six itérations. Parmi les concepts émergents, *peut* et *comme* ne sont pas

très pertinents. Ils apparaissent du fait de leur omniprésence au sein du corpus. Pourtant, nous avons choisi de ne pas les faire entrer dans la stop-liste d'élimination des mots vides car ils peuvent cacher un degré de précision dans le discours que nous voulons pouvoir exploiter dans l'émergence des relations sémantiques et/ou lexicales.

La carte partielle résultante est la suivante (figure 3):

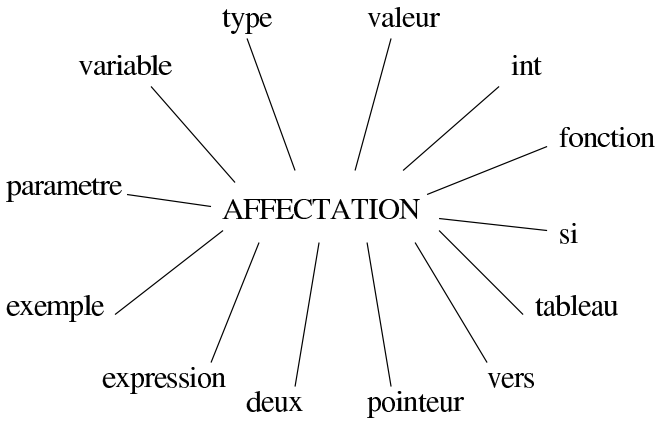

FIG. 3 – *Mot clé initial =AFFECTATION*

dans laquelle "*comme*", "*peut*", "*d*" ont été omis.

2.5 Remarques

□ L'omission des mots vides est nécessaire pour le bon déroulement de notre algorithme. En effet, leur présence engendre une propagation non justifiée pour la sélection des phrases. L'utilisation de poids relatifs (fréquence d'apparition dans les phrases sélectionnées par rapport à la fréquence d'apparition dans le corpus global) permet d'éviter une tel inconvénient. Mais les traitements impliqués sont encore plus lourds.

□ L'algorithme que nous proposons agit de la façon suivante:

```
Tant qu'il existe une propagation possible
     l'ensemble des phrases slectionnes volue,
     et ventuellement l'ensemble des mots cls.
```

où la propagation est due à la sélection de nouvelles phrases et la sélection de nouveaux mots clés produits.
La propagation ainsi définie reste finie puisque le corpus est lui même fini. L'algorithme proposé permet donc d'atteindre un point fixe après un nombre fini d'itérations pour un seuil de sélection calculé de façon régulière au cours du processus. C'est à dire que la règle de calcul du seuil doit rester la même tout au long du processus.

2.6 L'Émergence des Relations

L'extraction des concepts donne une approximation de l'intensité des liens qui existent entre un mot clé ($C1$) et un concept extrait ($C2$): elle est basée sur la valeur de la mesure du mot au point fixe (cf 2.3).

Afin de proposer des étiquettes sémantiques et/ou lexicales aux liens entre mots clés et concepts, nous avons étudié l'apport d'algorithmes initialement dédiés à l'extraction de connaissances dans des bases de données transactionnelles, permettant de faire émerger des associations entre items d'une même transaction [32, 1, 2]. Ces algorithmes sont fondés sur une notion de support définie comme: *l'ensemble S_i d'items a pour support s dans l'ensemble des transactions D si s% des transactions de D contiennent l'ensemble S_i.*

Seuls les ensembles d'items fréquents[7] sont recherchés. Nous avons implanté un tel algorithme où une phrase est vue comme un ensemble d'items représentés par les mots. La recherche s'effectue de la façon suivante:

1. chaque phrase contenant les mot clé et concept $C1$ et $C2$ est sélectionnée et devient une transaction (sélection des cooccurrences de $C1$ et $C2$)
2. l'algorithme d'Agrawal est exécuté sur les transactions et ne conserve que les ensembles fréquents (émergence des collocations)
3. pour tout S_i, on définit $R_i = S_i \backslash \{C1, C2\}$ comme une relation candidate entre $C1$ et $C2$, composée de un ou plusieurs mots.

On obtient ainsi les ensembles de mots qui apparaissent fréquemment avec les deux concepts $C1$ et $C2$ entre lesquels on cherche les relations existantes au sein du corpus. Pourtant, ces ensembles ont besoin d'être filtrés, puisque nous sommes plus précisément intéressés par les relations sémantiques et/ou lexicales. C'est pourquoi les relations sémantiques et lexicales les plus pertinentes sont sélectionnées par une grammaire de choix basée sur les combinaisons d'étiquettes grammaticales apposées aux mots au cours du prétraitement. Par exemple, parmi les relations composées de deux éléments nous préférerons la combinaison ($verbe + adverbe$) à la combinaison ($adverbe + adverbe$) qui sera abandonnée.

2.7 Remarque générale

L'extension de la requête permet la sélection des phrases dans lesquelles apparaissent des thèmes proches de ceux évoqués dans la requête fournie par l'utilisateur. La recherche de point fixe avec l'itération du processus a pour objectif de continuer ce processus de recherche par "propagation". Le risque encouru est que le point fixe atteint soit toujours le même quelle que soit la requête. En effet la définition de celui-ci est très étroitement liée au choix du seuil de sélection des "mots de plus grand poids". Or, la définition du seuil est envisageable sous deux angles: soit en se basant sur les scores des mots et en ne conservant que les plus performants (score le plus élevé diminué d'une marge

7. Un ensemble S_i est *fréquent* si son support est plus grand qu'un seuil donné: $support(S_i) \geq min_support$

à définir), soit en définissant une quantité (qui peut être dynamique) de mots à sélectionner à chaque itération. Dans notre prototype, ce choix est laissé à l'utilisateur. Il peut choisir de sélectionner les n meilleurs, les $n\%$ meilleurs, ou le meilleur et ceux dont la mesure est supérieure ou égale à la mesure du meilleur moins $n\%$.

3 Résultats

Dans cette section, nous présentons des résultats que nous avons obtenus avec le prototype issu de l'implantation en langage C des algorithmes présentés ci-dessus. Il est composé de trois modules, chacun dédié à une des grandes étapes du traitement. Il est ainsi possible de régler tous les paramètres dont nous avons parlé tels que liste des mots vides, extension de la requête, seuil de sélection... Une interface graphique est en cours de réalisation.

Nous avons fait des expérimentations sur trois corpus anglais. Le premier est un ensemble de textes descriptifs des Universités du monde entier et leurs conditions d'admission (560 335 mots). "Introduction to WordNet: an on-line Lexical Database" (37 792 mots) et "The Little Prince" d'Antoine de Saint Exupéry (15 901 mots) sont les deux autres corpus.

3.1 Corpus des Universités

Nous nous intéressons ici à des résultats obtenus sur le corpus des universités. Ce corpus est composé de près de 1000 textes descriptifs d'autant d'universités réparties sur les 5 continents. Les textes ont été recueillis sur le web[8]. Chaque texte est long d'une à deux pages.

Les concepts obtenus avec le mot clé initial *UNIVERSITY* et un raffinement sur *RESEARCH* sont présentés dans la figure 4. La figure 5 montre la carte obtenue avec le mot clé initial *RESEARCH*.

8. http://www-icdl.open.ac.uk/icdl/, site de l'*International Centre for Distance Learning*

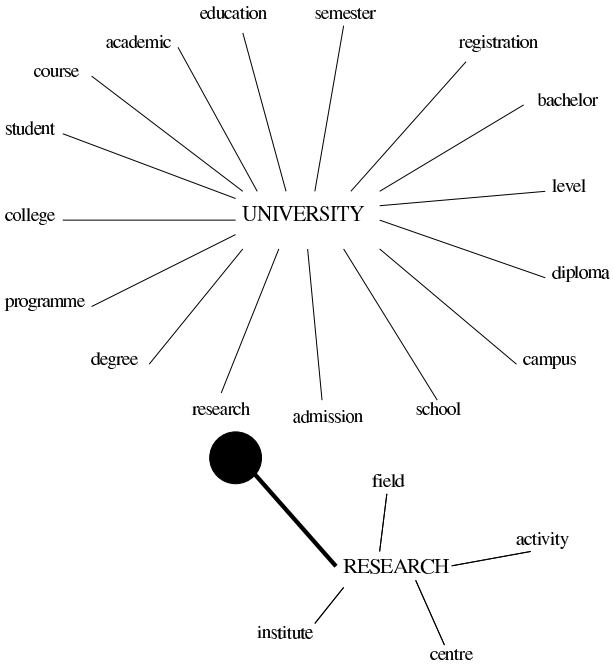

FIG. 4 – *Mot clé initial UNIVERSITY et raffinement sur RESEARCH*

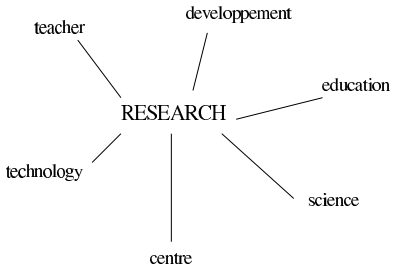

FIG. 5 – *Mot clé initial RESEARCH.*

Ces résultats mettent en évidence la dépendance des cartes par rapport à l'évolution de la requête.

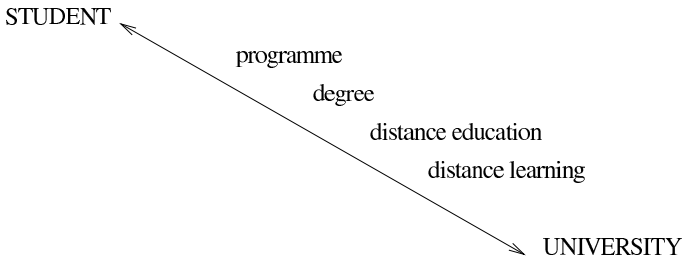

FIG. 6 – *Relations entre UNIVERSITY et STUDENT.*

La figure 6 présente les relations composées de un à deux mots qui émergent entre le mot clé *UNIVERSITY* et le concept *STUDENT*, au sein du corpus.

3.2 Corpus "Introduction to WordNet"

Ce corpus est constitué des articles de présentation de WordNet, rédigés par George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, et Katherine Miller.

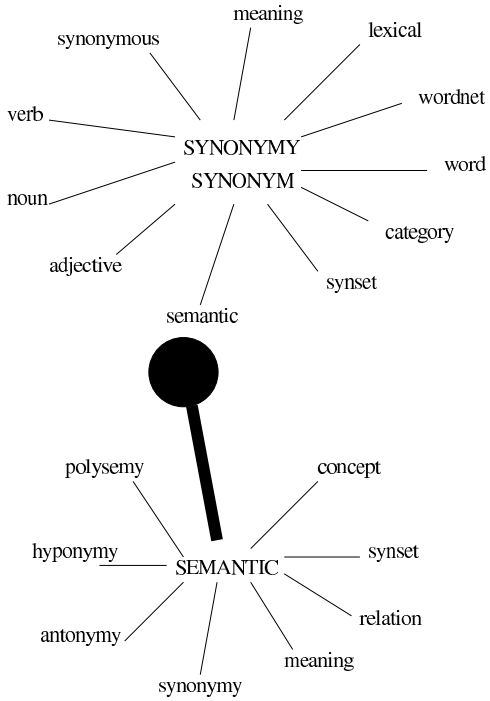

FIG. 7 – *Mots cls initiaux SYNONYMY, SYNONYM et raffinement sur SEMANTIC.*

La carte présentée à la figure 7 a été obtenue sur ce corpus en utilisant les mots clés initiaux *SYNONYMY* et *SYNONYM*; un raffinement a ensuite été demandé sur *SEMANTIC*.

La première carte met bien en évidence les liens synonymiques proposés au sein de WordNet entre verbes (*verbs*), noms (*nouns*) et adjectifs (*adjectives*). Notons que *adverbs* n'apparaît pas. Ceci s'explique par le fait que cette relation

n'est que très peu évoquée au sein du corpus. La seconde carte, qui présente la notion de *SEMANTIC* dans le contexte de la *SYNONYMY* rappelle que la base WordNet contient les liens de *polysemy, hyponymy, antonymy, synonymy.*

3.3 Corpus "The Little Prince"

Les résultats obtenus sur le corpus "The Little Prince" [9] d'Antoine de Saint Exupéry ne répondent pas aussi bien à nos attentes que ceux évoques précédemment. En effet, on obtient une carte comme

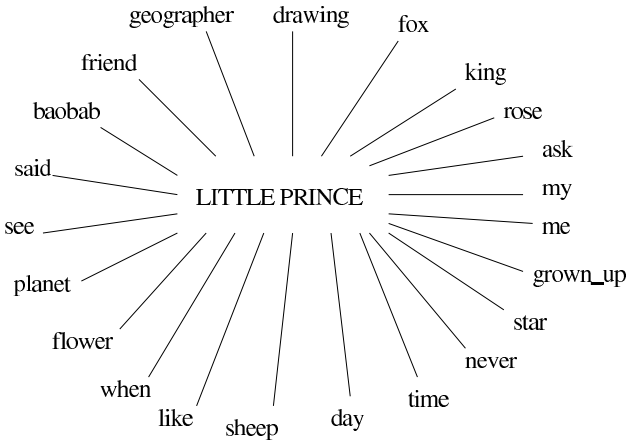

FIG. 8 – *Mot clé initial =LITTLE PRINCE.*

où des éléments importants de l'histoire comme le `sheep`, la `rose`, le `baobab`, le `geographer`... parviennent à émerger. Pourtant, ces résultats ne peuvent être obtenus que grâce à des seuils de sélection faibles. Ceci s'explique en partie par la faible longueur des phrases telles que *Draw me a sheep* qui implique que cette phrase ne parviendra pas à obtenir un poids suffisant pour faire émerger les mots qu'elle contient si les seuils sont trop élevés. En effet, les poids les plus forts sont détenus par des mots de liaison tels que `you, one, but, from, have, they, with, him, what...` que nous n'avons volontairement pas éliminés au cours du prétraitement. On peut alors se demander si, en général, un texte littéraire contient suffisamment d'éléments de répétition d'un même thème pour pouvoir le faire émerger. L'utilisation de fréquences relatives sur la globalité du texte ne permet pas non plus de résoudre cet inconvénient. Par contre, l'utilisation de fréquences relatives par rapport à une "partie du texte" (paragraphe ou chapitre...) pourrait peut-être permettre d'obtenir des résultats plus justes.

4 Prochaines évolutions

□ Ce travail propose une aide à l'utilisateur pour découvrir un nouveau domaine par le biais des *cartes conceptuelles*. L'une des prochaines étapes concernera la génération automatique de patrons d'extraction d'information.

9. http://www.geocities.com/Athens/Rhodes/1916/lepetitprince/index.html

□ Pour une extraction des concepts encore plus performante, nous envisageons l'utilisation de la relation d'hyperonymie au sein du calcul des poids des mots. En effet une heuristique comme: pour trois mots A, B, C avec les mesures $m(A)$, $m(B)$ et $m(C)$, si

$$\begin{aligned} & m(A) \leq \mathit{seuil} \leq m(C) \\ \mathit{et} \quad & m(B) \leq \mathit{seuil} \leq m(C) \\ \mathit{et} \quad & m(A) + m(B) \geq m(C) \\ \mathit{et} \quad & \exists \mathit{\ un\ concept\ hyperonyme}(A,B) \end{aligned} \tag{2}$$

alors $\mathit{hyperonym}(A,B)$ sera sélectionné et C pourra, en fonction de l'evolution de la valeur du seuil, être sélectionné ou non. Les concepts A et B seront représentés par leur ancêtre commun.

5 Conclusion

La Cartographie de Textes, telle que nous la proposons, établit un lien entre Recherche Documentaire et Extraction d'Information en proposant une méthode pour accéder à l'information recherchée au sein de larges collections de textes. Un utilisateur peut ainsi découvrir un domaine sur un corpus obtenu par recherche documentaire et définir des patrons d'Extraction d'Information.

Les algorithmes que nous proposons pour la cartographie de textes sont basés sur deux utilisations différentes de la collocation, dues à des objectifs distincts représentés par l'utilisation du Traitement Automatique des Langues d'une part, et l'utilisation de l'Apprentissage Automatique d'autre part.

Les deux algorithmes présentés pour la construction des *cartes conceptuelles* sont portables d'une langue à une autre. Un prétraitement différent doit éventuellement être envisagé pour chaque langue.

L'extraction des concepts clés du document peut aussi être utilisée dans le cadre d'une indexation automatique de documents.

Remerciements

Nous remercions Bernard Cassagne d'avoir bien voulu nous confier les sources latex de son ouvrage "Introduction au Langage C", disponible a l'adresse
http:// clips.imag.fr/ commun/ bernard.cassagne/ Introduction_ANSI_C.html.

Références

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD Conference*, pages 207–216, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithm for mining association rules. In *Proceedings of the VLDB Conference*, 1994. Also published as expanded version available as IBM Research Report RJ9839, June 1994.

[3] J. Balpe, A. Lelu, and F. Papy. *Techniques avancées pour l'hypertexte.* Editions Hermes, 1996.

[4] E. Brunet. Qui lemmatise dilemme attise, 2000. Revue Lexicometrica.

[5] T. Buzan. *Mind Map Book.* Dutton, 1994.

[6] M. E. Califf. Relational learning techniques for natural language information extraction, 1998.

[7] M. E. Califf and R. J. Mooney. Relational learning of pattern-match rules for information extraction, 1998.

[8] B. Daille and G. Williams. Journée d'etude de l'atala - la collocation. Technical Report 00-13, Institut de Recherche en Informatique de Nantes, 2001.

[9] I. Debourges. Cartographie de textes: Une nouvelle approche pour l'exploitation de grands corpora. Technical Report 00.13, Journée d'Etudes de l'ATALA - La collocation. Institut de Recherche en Informatique de Nantes, 2001.

[10] I. Debourges, S. Guilloré, and C. Vrain. Text mapping: A new approach to large corpus. In *Proceedings of the TELRI-II seminar, Ljubljana sept 2000*, 2000.

[11] G. Dias, S. Guillore, J.-C. Bassano, and J.-G. P. Lopes. *Extraction Automatique d'Unités Lexicales Complexes: un enjeu fondamental pour la recherche documentaire*, volume 41:2. Editions Hermes, 2000. Revue T.A.L.

[12] C. Enguehard and L. Panthera. Automatic natural acquisition of a terminology. In *Journal of Quantitative Linguistics*, volume 2, pages 27–32, 1995.

[13] H. Fondin. *Les modèles de recherche documentaire*, chapter 23. Editions Hermes, 1998. Le traitement numrique des documents.

[14] R. Gaizaukas, T. Wakao, K. Humphreys, H. Cunningham, and Y. Wilks. Description of the lasie system as used for muc-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 207–220. Morgan Kaufmann, 1995.

[15] R. Gaizauskas and S. Robertson. Coupling information retrieval and information extraction: A new text technology for gathering information from the web. In *Proceedings of RIAO 97: Computer-Assisted Information Searching on the Internet*, pages 356–370, 1997.

[16] S. J. Green. *Building Hypertext Links By Computing Semantic Similarity*, volume 11-5. 9, 10 1999.

[17] B. Habert, A. Nazarenko, and A. Salem. *Les linguistiques de Corpus.* Armand Colin, 1997.

[18] S. Huffman. Learning information extraction patterns from examples. 1996.

[19] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. University of sheffield: Description of the lasie-ii system as used for muc-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7).* Morgan Kaufmann, 1998.

[20] D. Landau, R. Feldman, Y. Aumann, M. Fresko, Y. Lindell, O. Lipshtat, and O. Zamir. Textvis: An integrated visual environment for text mining. In *Principles of Data Mining and Knowledge Discover*, 1998.

[21] L. Lebart and A. Salem. *Statistique Textuelle.* DUNOD, 1994.

[22] A. Lelu, M. Halleb, and B. Delprat. Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de n_grammes. In *Proceedings of the JADT 1998*, pages 391–400, 1998.

[23] G. A. Miller and al. Introduction to wordnet: An on-line lexical database, five papers. 1993.

[24] E. Morin. *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques.* PhD thesis, Institut de Recherche en Informatique de Nantes, 1999.

[25] G. Nault, V. Rialle, and J. Meunier. *PROGEN: a Genetic-Based Semi-Automatic Hypertext Construction Tool - first steps and experiments.* 1999.

[26] J. D. Novak. *How do we learn our lesson? : Taking students through the process*, pages 50–55. 1993.

[27] E. Riloff. Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence, AAAI-96*, pages 1044–1049, 1996.

[28] L. Romary. *Outils d'accès à des ressources linguistiques*, chapter 8. Editions Hermes, 2000.

[29] G. Salton and J. McGill. *Introduction to Modern Information Retrieval.* Mc Graw Hill, 1995.

[30] H. Samier and V. Sandoval. *La recherche intelligente sur l'internet. Outils et méthodes.* Editions Hermes, 1998.

[31] J. Sinclair. *English Lexical Collocations: A study in computationnal linguistics*, chapter 2. Uni Press, 1996.

[32] R. Srikant and R. Agrawal. Fast discovery of association rules. In *Advances in knowledge discovery and data mining*, pages 307–328. AAAI Press, 1996.

[33] J. Vergne. http://www.limsi.fr/tlp/grace/www/ sys.id.html#aleat, 1999.

[34] J.-M. Viprey. Hypertexte de corpus littéraire: cartographie et statistique multidimentionnelle. In *JADT 2000, Actes des 5èmes Journées Internationales d'analyse statitstique des données textuelles, 9-11 mars 2000*, pages 535–538. Publications de l'Ecole Polytechnique fédérale de Lausanne, 2000.

[35] Y. Wilks. Information extraction as a core language technology. *Information Extraction 1997, LNCS 1299*, pages 1–9, 1997.

[36] Y. Wilks. *Information Extraction as a Core Language Technology*, pages 1–9. Number LNCS 1299. Springer, 1997.

[37] G. Williams. Collocational networks for prototypical thematic grouping in a french monosource corpus. In *Proceedings of the TELRI-II seminar, Ljubljana sept 2000*, 2000.