Cartographie de Textes: Une aide à l'utilisateur dans le cadre de la découverte de nouveaux domaines

Isabelle Debourges, Sylvie Guilloré-Billot, Christel Vrain
Laboratoire d'Informatique Fondamentale d'Orléans
Bâtiment IIIA
Rue Léonard de Vinci – B.P. 6759
F-45067 Orléans Cedex 2
{debourge, guillore, cv}@lifo.univ-orleans.fr

Résumé - Abstract

Nous présentons les avancées d'un projet dans un thème que nous qualifions de Cartographie de Textes qui permet à l'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 extraites du corpus les lient aux mots 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.

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

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.

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

1 Introduction

Les utilisateurs ont aujourd'hui besoin d'outils qui leur permettent de retrouver ce qu'ils cherchent au sein de grandes sources d'informations. La Recherche Documentaire et l'Extraction d'Information sont les deux grands types d'aide mis à leur disposition.

La **Recherche Documentaire** (Salton, 1995; Fondin, 1998) nécessite la saisie de mots clés par l'utilisateur pour produire une sélection de textes jugés pertinents vis à vis de cette requête. Mais 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.

L'**Extraction d'Information** (Wilks, 1997) 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. L'utilisation

d'un outil d'Extraction d'Information nécessite que l'utilisateur 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** (Riloff, 1996; Califf, 1998) 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 connaît pas le domaine .

Ainsi, un utilisateur qui souhaite découvrir un nouveau domaine ne peut pas accéder aisément aux informations qu'il recherche sur celui-ci. Les aides proposées par la Recherche Documentaire et l'Extraction d'Information sont 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.

2 Architecture 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 **??** 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.

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
- 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 aux requêtes antérieures.

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

2. l'émergence des relations sémantiques et lexicales présentes au sein du corpus, entre les mots clés de la requête et les concepts extraits.

Travaux apparentés

Le formalisme graphique que nous utilisons pour les *cartes conceptuelles* peut être apparenté à celui des graphes conceptuels (Sowa, 1984), eux même dérivés des réseaux sémantiques (Quillian, 1968). La différence majeure réside dans le fait que nous ne représentons pas les informations présentes au niveau d'une phrase, mais plutôt une généralisation des informations disponibles dans le corpus, autour d'un thème représenté par le mot clé. De plus, dans les cartes conceptuelles que nous construisons, les concepts constituent les noeuds tandis que les relations étiquettent les liens.

3 Prétraitements et Extraction de Concepts

3.1 Prétraitements

Une phase de prétraitement du corpus permet de réaliser des traitements simples utilisés au cours des traitements plus lourds que sont l'extraction des concepts et l'émergence des relations. Le texte est découpé en phrases, les mots vides et caractères spéciaux sont omis. Le texte est lemmatisé et une catégorie grammaticale est associée à chaque mot (Debourges et al., 2001).

3.2 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, les mots clés sont tout d'abord étendus en ajoutant les synonymes/hyponymes/hyperonymes 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 les mots 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 (Debourges et al., 2001). Ces résultats mettent en évidence la dépendance des cartes par rapport à l'évolution de la requête.

3.3 Exemple

Nous nous intéressons ici à des résultats obtenus sur un corpus décrivant 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[2]. 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 1. La figure 2 montre la carte obtenue avec le mot clé initial *RESEARCH*.

4 Emergence de Relations

4.1 Motivations

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 donne une mesure numérique fondée sur le dénombrement des apparitions conjointes au point fixe.

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

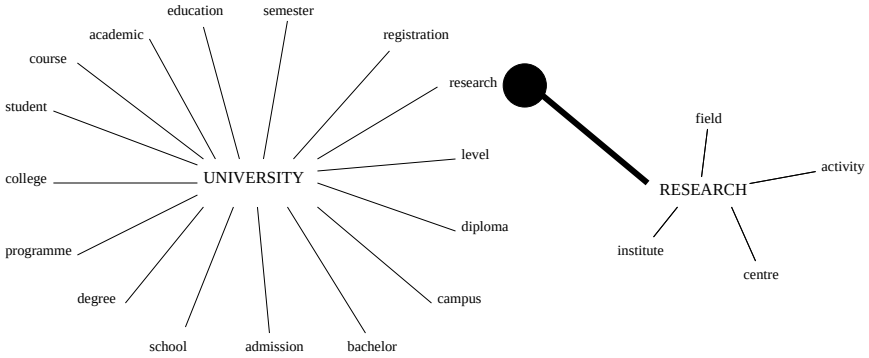

Figure 1: Mot clé initial *UNIVERSITY* et raffinement sur *RESEARCH*

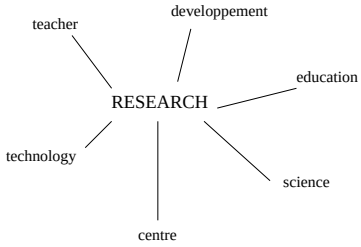

Figure 2: Mot clé initial *RESEARCH*.

Des étiquettes sémantiques/lexicales sur un lien entre un mot clé et un concept émergent permettent à l'utilisateur de connaître les idées qui relient ces deux notions, au sein du corpus. L'utilisateur découvre de façon plus complète le contenu du corpus qu'il a fourni, sans avoir à regarder les textes originaux.

Dans un premier temps nous étiquetons le lien entre un mot clé et un concept par la liste des unités lexicales (composées de un ou plusieurs mots) qui apparaissent fréquemment avec ces deux mots. L'algorithme est donné en 4.2.

Par exemple, les expérimentations menées sur le corpus des universités nous montrent, entre autres, que le concept *ADMISSION* émerge avec le mot clé *UNIVERSITY* (figure 1). L'utilisateur a, à ce niveau, conscience de l'existence d'un lien fort entre ces deux notions. Ces mêmes expérimentations nous permettent alors d'expliciter les liens entre *UNIVERSITY* et *ADMISSION* (figure 3-a) avec l'émergence des unités lexicales *APPLICANT, STUDENT, ENTRANCE, DEGREE, REQUIREMENT*. On retrouve alors les idées fortes concernant l'entrée à l'Université:

Pour devenir étudiant^{+} (entrer à l'Université$^{@}$), un certain niveau$^{!}$ est requis$^{\&}$ pour les postulants*.

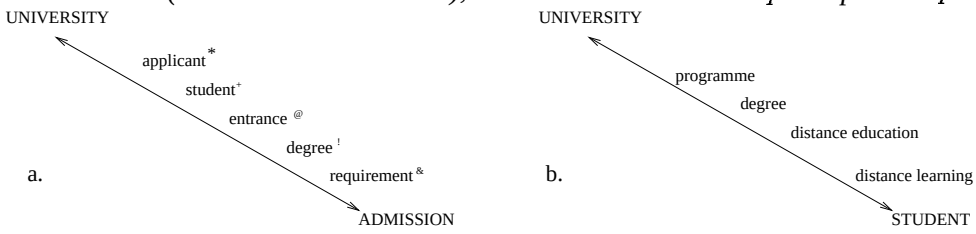

Figure 3: Exemples de relations issues d'expérimentations menées sur le prototype

4.2 Algorithme

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 (Srikant et al., 1996; **?**; Agrawal et al., 1994). Ces algorithmes sont fondés sur une notion de support définie comme suit:

> *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[3] 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 simultanément le mot clé $C1$ et le concept $C2$ 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 ensemble fréquent d'items 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 intéressés par les relations sémantiques et/ou lexicales les plus pertinentes. Elles sont sélectionnées par une grammaire 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 $(adverbe + adverbe)$ moins pertinente, qui sera abandonnée.

4.3 Exemples

La figure 3-b présente les relations composées de un à deux mots qui émergent entre le mot clé *UNIVERSITY* et le concept *STUDENT*, au sein du corpus. En effet, dans le cas présent, il existe des ensembles fréquents S_i de taille 3 et de taille 4 qui permettent l'extraction des relations *programme, degree* et *distance education, distance learning*. L'ordre de présentation des mots qui composent les unités polylexicales a ici été choisi manuellement. Un tel choix sera prochainement inspiré par une consultation de leur ordre classique d'apparition au sein du corpus. En particulier, on voit ici que le corpus s'intéresse aux programmes d'éducation à distance (*distance education* et *distance learning*). Cette particularité du corpus s'explique par son objectif initial: proposer des formations dans des universités des cinq continents à l'ensemble des habitants de la planète.

5 Perspectives

Evolution de l'algorithme 4.2. L'algorithme d'Agrawal ne se préoccupe pas de l'ordre des items dans les transactions. Ainsi, l'ordre d'apparition des mots dans une phrase n'est pas pris en considération dans l'algorithme dérivé que nous proposons. Pourtant, lorsque l'étiquette obtenue pour la relation est une unité polylexicale, cet ordre a un rôle important à jouer. Cette prise en compte de l'ordre d'apparition des mots dans la phrase constitue une prochaine étape dans l'extraction des relations pour la construction des cartes conceptuelles.

Proposition de patrons d'extraction d'information. Le fait d'avoir découvert un domaine (concepts et relations) pourra permettre à l'utilisateur de définir des patrons d'extraction d'information. L'une des applications dérivées des cartes conceptuelles est donc d'aider l'utilisateur

[3] Un ensemble S_i est *fréquent* si son support est plus grand qu'un seuil donné: $support(S_i) \geq min\_support$

dans cette tâche. En effet, s'il dispose d'un corpus contenant des informations de même type, une aide à la généralisation des cartes peut contribuer à obtenir des patrons d'extraction d'information. Cette généralisation sera basée sur deux types d'unification:

- les concepts émergents (qui apparaissent dans les cartes) peuvent être regroupés au sein d'un concept plus général en utilisant une ontologie
- deux relations identiques sur des liens différents issus d'un même mot clé constituent un indice de regroupement des concepts.

6 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 des heuristiques inspirées du Traitement Automatique des Langues d'une part, et l'utilisation de l'Apprentissage Automatique d'autre part. Nous avons délibérément privilégié des méthodes simples pour des questions d'efficacité et de portabilité.

Les algorithmes présentés pour la construction des *cartes conceptuelles* sont indépendants du domaine étudié, et portables d'une langue à une autre. De plus ils ne requièrent pas d'analyse fine du texte, d'où une efficacité accrue.

Références

Agrawal R., Srikant R. (1994), Fast Algorithm for Mining Association Rules, Actes de *The VLDB Conference*.

Califf M-E. (1998), *Relational Learning Techniques for Natural Language Information Extraction*.

Debourges I., Guilloré S., Vrain C. (2001), Cartographie de Textes. Une nouvelle approche pour l'exploration sémantique des corpus homogènes de grande dimension, RR-2001-01 du Laboratoire d'Informatique Fondamentale d'Orléans.

Fondin H. (1998), Les modèles de recherche documentaire, *Le traitement numérique des documents* Editions Hermes.

Novak J. D. (1993), How do we learn our lesson? : Taking students through the process, *The Science Teacher 60(3)*.

Quillian M.R. (1968), Semantic Memory, dans *Semantic Information Processing*, M.I.T. Press.

Riloff E. (1996), Automatically Generating Extraction Patterns from Untagged Text, Actes de *Thirteenth National Conference on Artificial Intelligence, AAAI-96*, 1044-1049.

Salton G., McGill J. (1995), *Introduction to Modern Information Retrieval* Mc Graw Hill.

Sowa J.F. (1984), *Conceptual Structures. Information Processing in Mind and Machine*, Addison Welsey.

Srikant R., Agrawal R. (1996), Fast Discovery of Association Rules, Actes de *Advances in knowledge discovery and data mining*, AAAI Press, 307-328.

Wilks Y. (1997), Information Extraction as a Core Language Technology, *Information Extraction 1997, LNCS 1299*, Springer, 1-9.