Construction et exploration de corpus à partir du Web à l'aide d'une plate-forme logicielle de cartographie documentaire

Thibault Roy & Pierre Beust – {troy, beust}@info.unicaen.fr
Laboratoire GREYC - Université de Caen / Basse-Normandie

Problématique générale

Dans le cadre du Traitement Automatique des Langues (TAL), nous nous intéressons à la thématique des documents électroniques présents sur le Web. Nous présentons ici les principes de fonctionnement et les intérêts de la plate-forme logicielle ProxiDocs[1] permettant de construire des représentations cartographiques d'ensembles de documents électroniques, et plus particulièrement d'ensembles de pages provenant de sites Web, à partir de thématiques choisies et définies par l'utilisateur.
Nos objectifs à travers cette plate-forme sont multiples. Tout d'abord, afin de faciliter l'accès à l'information présente sur le Web, nous cherchons à obtenir une « vue globale » sur les documents présents sur Internet répondant à certains critères donnés. Pour cela, nous exploitons plusieurs moteurs de recherche et nous produisons différentes cartes représentant les ensembles documentaires retournés par ces moteurs. Notre second objectif est l'aide à la construction de corpus à partir du Web. Les usages de tels corpus peuvent alors être multiples. Dans le cadre de nos travaux, nous les exploitons pour mener des analyses thématiques sur des documents récents et d'origines diverses, mais aussi pour obtenir du lexique attesté afin de construire des ressources lexicales exploitées dans différentes analyses linguistiques.
La première partie de cet article présente l'approche adoptée et le fonctionnement de la plate-forme ProxiDocs permettant la cartographie d'ensembles de pages Web. La seconde partie propose un exemple d'utilisation de la plate-forme portant sur l'exploration d'un ensemble documentaire provenant du Web et traitant de la constitution européenne.

Plate-forme ProxiDocs : approche adoptée et mise en œuvre logicielle

Cartographie d'ensembles documentaires

Afin d'avoir une vue plus globale sur les informations contenues dans un ensemble documentaire et de faire intervenir une notion de proximité entre éléments de cet ensemble, il est intéressant de proposer aux utilisateurs une représentation de l'ensemble analysé sous forme de cartes. Une carte d'un ensemble de documents met en évidence des proximités et des liens entre entités textuelles (comme par exemple des mots, des documents, etc.) au sein de cet ensemble (un peu à la manière d'une carte routière mettant en évidence des proximités et des liens entre villes). Depuis 2001, les deux métamoteurs de recherche cartographiques KartOO (Chung et al., 2002) et Mapstan (Spinat, 2002) sont disponibles sur le Web[2]. En réponse à une requête de l'utilisateur, ces deux outils retournent des cartes représentant les sites proposés en réponse à cette requête. Les sites jugés similaires par le système sont alors situés à proximité sur les cartes et il est ainsi possible de distinguer les grandes « familles » d'informations proposées en réponse à la requête de l'utilisateur.

---

[1] La plate-forme ProxiDocs (Roy, 2005) est un outil logiciel *open-source*, développé en Java et disponible avec sa documentation sur le Web : http://www.info.unicaen.fr/~troy/proxidocs. La version de l'application en ligne présentée dans cet article exploite des *servlets Java Server Pages* et implique l'utilisation d'un serveur *Apache Tomcat*.

[2] Moteurs respectivement disponibles : http://www.kartoo.fr et http://www.mapstan.net

Approche orientée thématique

A la manière des outils KartOO et Mapstan, la plate-forme ProxiDocs que nous proposons met visuellement en évidence des similarités et des différences entre documents électroniques d'un même ensemble. La principale différence que nous cultivons avec ces outils est que ProxiDocs va chercher à produire des cartes thématiques (mettant en évidence les principaux thèmes ou sujets abordés) uniquement à partir des centres d'intérêt de l'utilisateur (ou d'un groupe d'utilisateur) dans le contexte de sa tâche. De la même manière que (Pichon et Sébillot, 1999), nous entendons par « thèmes », les sujets abordés dans un texte.

Pour construire les cartes thématiques d'un corpus, ProxiDocs prend en entrée des descriptions du lexique pertinent sélectionné par l'utilisateur. Par exemple, un utilisateur pourra définir la liste des 12 lexies (mots ou mots composés) suivantes correspondant à une description succincte du thème « Aviation » (l'utilisateur ne saisit que les formes lemmatisées des lexies qu'il souhaite faire intervenir dans ses définitions de thèmes) : *avion*, *appareil*, *Airbus*, *A320*, *vol*, *pilote*, *pilotage*, *passager*, *Boeing*, *décollage*, etc. Techniquement, les thèmes définis par le ou les utilisateurs sont enregistrés dans un fichier XML à l'aide du logiciel ThemeEditor (Beust, 2002).

Traitements réalisés

La première tâche à réaliser par l'utilisateur est donc la construction des thèmes qu'il souhaite voir émerger dans ses analyses. Une fois une base de thèmes définie par l'utilisateur, ce dernier va choisir, via la plate-forme, les moteurs de recherche à interroger[3], la langue des pages[4] et le nombre de pages à retourner. Dans le cas d'une interrogation pourtant sur plusieurs moteurs de recherche, la plate-forme distribue la recherche sur les différents moteurs et uniformise les liens obtenus afin d'éliminer la redondance et d'obtenir le nombre de pages souhaité par l'utilisateur. Pour réaliser de telles interrogations sur les moteurs de recherche, soit des API propres aux moteurs sont exploitées (uniquement Google et Yahoo pour l'instant), soit des parcours des pages des liens retournées sont réalisés (pour les autres moteurs).

Une fois l'ensemble documentaire collecté à partir de ces différents paramètres, nous mettons en œuvre dans la plate-forme ProxiDocs différents traitements statistiques (Roy et Beust, 2004) afin d'opérationnaliser la construction des cartes de l'ensemble. Le premier de ces traitements est un comptage des occurrences de lexies de chaque thème et de leurs formes fléchies dans chaque document du corpus. Ce comptage est pondéré selon la taille des textes. Un vecteur de réels de dimension égale au nombre de thèmes est alors associé à chaque texte. Afin de visualiser les vecteurs représentant les textes du corpus sur des espaces en 2 dimensions (qui seront les cartes), il faut réaliser une étape de projection. Les cartes présentées ci-dessous ont été obtenues à l'aide d'une Analyse en Composantes Principales (ACP) (Bouroche et Saporta, 1980, p. 17). Afin de mettre automatiquement en évidence des regroupement entre documents sur les cartes, nous exploitons également une méthode de catégorisation : la Catégorisation Hiérarchique Ascendante (CHA) (Bouroche et Saporta, 1980, p. 54). Une fois l'ensemble de ces traitements réalisés, les cartes de l'ensemble documentaire sont retournées à l'utilisateur. Ces cartes, au format SVG, offrent à l'utilisateur un grand nombre de possibilités d'interaction, telles des zooms, des déplacements et différentes interrogations portant sur les documents et groupes de documents figurant sur les cartes. À ce moment, la possibilité de sauvegarder la totalité ou une partie du corpus représenté par les cartes est accessible à l'utilisateur.

---

[3] Plusieurs moteurs sont accessibles : Google, Tiscali, Yahoo, Altavista, Aol, Msn Search et Lycos.

[4] La restriction sur la langue ne peut se faire actuellement que sur le français et l'anglais.

Exemple d'utilisation de la plate-forme

Afin d'illustrer les possibilités offertes par la plate-forme, nous présentons un exemple d'utilisation. L'objectif visé dans cet exemple est l'exploration d'un corpus de documents électroniques provenant de sites français traitant de la constitution européenne. La possibilité de constitution d'un corpus est également évoquée. La requête fourni à la plate-forme est « constitution européenne », le nombre de documents souhaité est de 100 et les moteurs interrogés sont Google et Yahoo (la figure 1 illustre l'étape de configuration de la plate-forme).

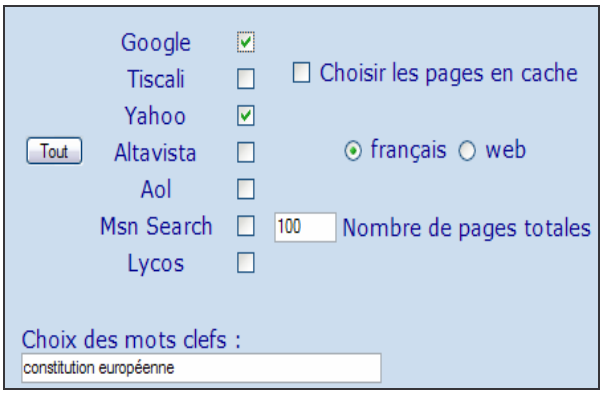

Figure 1 : Interface de configuration de la recherche

Le jeu de thèmes considéré est général et propose des « descriptions » des 12 thèmes généralistes suivants : la *justice,* la *religion,* la *violence,* l'*éducation,* l'*agriculture,* la *sécurité routière,* l'*économie,* l'*aérospatial,* la *guerre,* l'*informatique,* la *pollution* et le *travail*. Un tel ensemble de thèmes a été choisi afin de mettre en évidence les thématiques secondaires gravitant autour de la constitution européenne. La construction des cartes peut alors être lancée. L'utilisateur obtiendra alors au bout d'un certain délai (variant de quelques secondes à plusieurs minutes selon le nombre de pages demandées) un écran lui fournissant des informations sur les traitements réalisés (durée du traitement, requête et moteurs de recherche utilisés, nombre de pages collectées[5], etc.) et donnant accès aux différentes cartes construites. Dans notre exemple, le traitement a duré 42 secondes, une ACP a été réalisée (les composantes 1 et 2 ont été choisies) et une CHA a mis automatiquement en évidence 10 groupes de pages. La figure 2 ci-dessous présente une carte obtenue dans notre exemple :

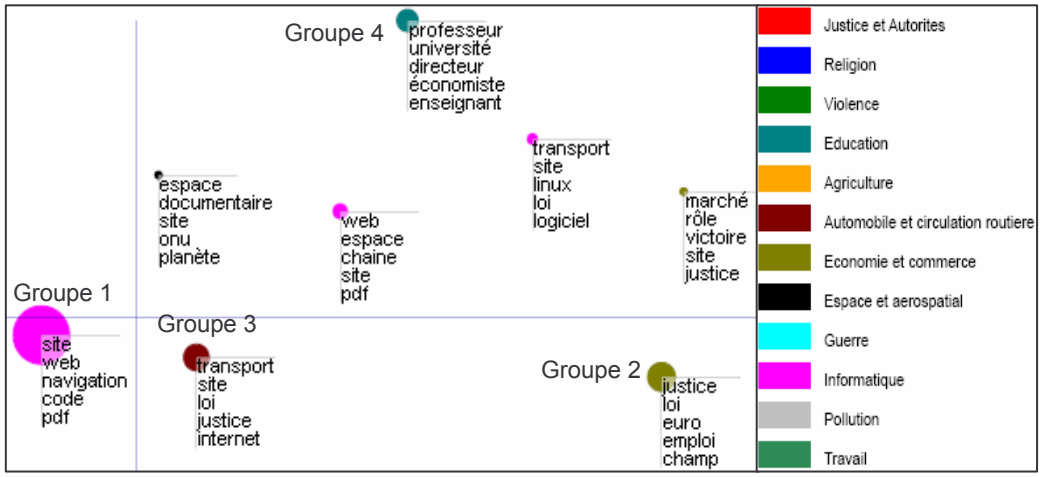

Figure 2 : Carte d'un ensemble documentaire obtenu à l'issue d'une recherche sur le terme « constitution européenne ». Des groupes ont été marqués manuellement sur la carte.

Cette carte représente des groupes de pages constitués à partir de l'ensemble documentaire obtenu. Chaque groupe est représenté par un disque de taille proportionnelle à la cardinalité

[5] Il arrive que l'utilisateur demande un nombre de pages plus important que le nombre de pages proposé par les moteurs interrogés (par exemple, si la requête est trop précise). Dans un tel cas, les cartes sont tout de même construites avec l'ensemble les pages disponibles.

du groupe, un disque est également un lien hypertexte vers un rapport détaillé sur les pages contenues dans ce groupe (ce rapport contient une répartition des thèmes dans les pages du groupes, un classement décroissant des lexies des thèmes apparaissant dans le groupe, la liste des pages contenues dans ce groupe et la possibilité d'y accéder, etc.). Chaque groupe est étiqueté avec les cinq lexies des définitions de thèmes les plus fréquentes dans les pages qu'il rassemble. Les couleurs associées aux groupes de pages sur les cartes correspondent aux thèmes majoritairement représentés. Une analyse rapide de la carte révèle que la thématique de l'informatique semble souvent abordée (groupe 1). En effet, un très grand nombre de pages de l'ensemble obtenu propose des versions électroniques de la constitution européenne ainsi que des outils informatiques en permettant différentes consultations. D'autres thématiques sont également présentes, telles, sans surprise, l'économie dans le groupe 2 (rassemblant des pages traitant des enjeux économiques de la constitution), plus étonnamment, la circulation routière dans le groupe 3 (les pages de ce groupes contiennent des textes et des discussions sur la problématique du transport routier en Europe) et l'éducation dans le groupe 4 (ce groupe contient essentiellement des pages décrivant des enseignements universitaires en économie et en sciences politiques).
Outre les possibilités d'exploration de l'ensemble documentaire obtenu, sa sauvegarde complète ou partielle est également possible. À partir des cartes thématiques mettant en évidence des groupes de documents (telle celle présentée en figure 2), l'utilisateur peut choisir de sauvegarder sur son disque dur l'ensemble documentaire ou seulement certains groupes de son choix. Une archive au format *zip* est alors créée et contient l'ensemble des pages choisies par l'utilisateur (un fichier HTML est alors créé pour chaque page Web et seules les parties textuelles des pages sont sauvegardées). Cette archive est alors exploitable par toutes applications traitant des documents au format texte.

Conclusion

A travers un exemple concret nous avons montré ici quelques unes des possibilités offertes par la plate-forme ProxiDocs pour l'exploration et la constitution d'ensembles documentaires provenant du Web. ProxiDocs fait encore l'objet de développements et d'expérimentations. Nos travaux actuels portent notamment sur l'exploitation des corpus issus du Web, matériaux attestés révélant des usages quotidiens de la langue, afin de construire mais aussi de faire évoluer des ressources lexicales personnalisées (tels les thèmes exploités afin de construire la carte de la figure 2).

Références

BEUST P. (2002). Un outil de coloriage de corpus pour la représentation de thèmes, Actes des *6èmes JADT*, 161-169.

BOUROCHE J.-M. et SAPORTA G. (1980). *L'analyse des données*, Collection Que sais-je ?, PUF.

CHUNG W. et al. (2002). *Business Intelligence Explorer: A Knowledge Map Framework for Discovering Business Intelligence on the Web*, Proceedings of the 36th Hawaii Conference on System Sciences.

PICHON R. et SÉBILLOT P. (1999). Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience, Actes de *TALN 1999*, 279-288.

ROY T. (2005). Une plate-forme logicielle dédiée à la cartographie thématique de corpus, Actes des *RECITAL 2005*, 545-554.

ROY T. et BEUST P. (2004). *ProxiDocs, un outil de cartographie et de catégorisation thématique de corpus*, Actes des *7èmes JADT*, 978-987.

SPINAT E. (2002). Pourquoi intégrer des outils de cartographie au sein des systèmes d'information de l'entreprise ?, Colloque *Cartographie de l'Information*, Paris.