EXTRACTION DE CONNAISSANCES À PARTIR DE DONNÉES TEXTUELLES – VUE D'ENSEMBLE

Bruno Agard
Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal,
C.P. 6079, succ. Centre-ville, Montréal (Québec), H3C 3A7, Canada
bruno.agard@polymtl.ca

Résumé:

Le document propose de montrer une vue générale de la recherche en ce qui concerne la fouille de données (data mining) lorsqu'il s'agit de traiter des données textuelles (text mining). La fouille de données textuelles se déroule en deux étapes : (1) – la première étape permet de construire une représentation structurée du texte (qui par nature est moins structurée), (2) – la seconde étape exploitera l'information contenue dans les modèles structurés, pour rechercher de l'information pertinente pour l'utilisateur. L'accent sera porté sur le type d'information utilisé et la manière de traiter cette information afin de faire ressortir les lacunes et montrer les directions à prendre afin d'enrichir les techniques existantes. Les principaux outils énoncés seront illustrés par une mise en œuvre à l'extraction de connaissance à partir de l'ensemble des résumés du 9e Colloque National AIP-PRIMECA, *La Plagne, France, 5-8 Avril 2005.*

Mots clés: fouille de texte, data mining, extraction de connaissances.

1 Introduction

De nos jours nous disposons d'une quantité énorme d'écrits sur de très nombreux domaines scientifiques, techniques, littéraires, journalistiques, historiques, … De plus en plus ces écrits sont disponibles ou existent quelque part sous format électronique. Que ce soit sur le web, sur le serveur intranet d'une institution, que l'information soit destinée à la planète entière ou à un groupe précis de personnes, nous sommes de plus en plus submergés d'information. Il nous est cependant de moins en moins possible de lire l'ensemble des textes qui nous intéressent. Malgré cela nous cherchons à tirer l'information essentielle afin de prendre des décisions éclairées et ne pas refaire ce qui a déjà été fait, d'éviter des erreurs. Est-il possible, sans lire un ensemble de textes, d'en extraire l'information essentielle, les tendances, les associations entre les idées ? Peut-on découvrir des connaissances/informations non triviales, implicites, non connues, potentiellement utiles et compréhensibles à partir d'une grande masse de données textuelles ? C'est ce que nous allons tenter de montrer dans le présent article.

Deux grandes directions sont à l'étude pour la recherche de l'information cachée dans les données textuelles : (1) – développer des méthodes de fouilles pour des objets non structurés (extraction de concepts, constructions d'ontologies), axe suivi par les linguistes, les analystes du langage naturel et de la sémantique ; (2) – utiliser les outils de fouille (data mining) qui fonctionnent avec des données structurées, après avoir traité l'information qui est contenue dans les textes pour la mettre sous une forme adéquate (on parle alors de méta données, qui elles seront structurées), axe privilégié par les coutumiers du data mining traditionnel. Nous nous placerons dans le cadre du second axe.

Même si le terme de « text mining » est récent, les outils employés sont plus anciens et ont tous connus leur heure de gloire : classification de documents (fin 1950), mesure de distance entre

documents et regroupements (début années 70), représentation vectorielle d'un document avec mesure de fréquences (milieu années 70) et intelligence artificielle (années 80) [5]. Cependant le faible coût d'accès à des documents numériques ainsi que la puissance de calcul disponible sur le moindre ordinateur rendent possible l'utilisation des résultats précédents à grande échelle.

La fouille de données textuelles (text mining) a pour but d'extraire des patrons intéressants à partir d'un ensemble de données textuelles. Malheureusement les données textuelles sont moins structurées que les habituelles bases de données, ou fichiers Excel. Les informations ne signalent pas leur présence et peuvent être cachées à différents endroits. Aussi le même concept peut être énoncé de différentes manières avec du vocabulaire différent.

La fouille de données textuelles s'effectue en deux étapes. La première vise à extraire de chaque texte une information pertinente. Il s'agit principalement de filtrer le texte (suppression des termes vides de sens : le, la, un, une …) et de compter les termes identiques (avec conjugaison, synonymes, …) afin de construire un vecteur pondéré qui contient seulement les mots « pertinents » du texte.

Ce vecteur est ensuite exploité pour [1 ; 5] :

- Classer les documents en fonction de leur contenu. Les catégories de classifications doivent cependant être définies *a priori* par l'utilisateur. Après une phase d'apprentissage sur un ensemble de documents classifiés, il est possible de catégoriser de nouveaux documents.

- Regrouper les documents de contenu proche et organiser les documents en hiérarchies. Les documents étant représentés par des vecteurs, il suffit de définir des notions de distance de *document à document* et *document à ensemble de documents* pour être capable de retrouver tous les documents de contenu similaire et/ou classer ces documents selon des arborescences de proche en proche.

- Retrouver un document à partir de son contenu. Le vecteur de description de chaque document peut être facilement parcouru pour retrouver l'ensemble des documents pertinents vis-à-vis d'un ensemble de mots clés, les techniques d'indexation permettant une efficacité accrue. Cependant il est aussi possible de retrouver des documents similaires, même s'ils ne partagent pas forcément un vocabulaire similaire, si le vecteur s'appuie sur des dictionnaires (généralement thématiques).

- Dresser des relations entre les personnes/lieux/organisations/concepts… Nous avons vu précédemment que le texte non structuré était représenté sous forme de vecteur (structuré), afin de pouvoir utiliser les méthodes traditionnelles du data mining. Il est cependant relativement aisé d'extraire de l'information enrichie des textes. Les auteurs, les adresses, les courriels, les dates, les citations… sont autant d'informations structurées dans les textes, qu'il suffit de *reconnaitre*. Ces informations, riches de signification, portent de l'information supplémentaire, généralement non ambigüe, qu'il faudra traiter de manière adaptée (par exemple : Citeseer traite les citations).

- Extraire des caractéristiques dans les textes : associations d'idées, structures communes à différents documents… Le *vecteur* de description peut révéler des associations entre les termes utilisés, ces associations donnent une mesure de la force du lien entre deux (ou plus) termes. Par exemple « objet intermédiaire » et « conception » peuvent avoir un lien plus ou moins fort et notamment la force de ce lien peut évoluer dans le temps. À un moment donné, il est possible d'observer une carte des concepts (choisis ou non) dans une base de texte particulière. Cela peut faire ressortir des « trous » qui sont autant d'opportunités de recherche à évaluer.

Suite à cette introduction permettant de donner une vue sur les possibilités offertes par la fouille de données textuelles, le chapitre suivant va insister sur les outils et méthodes utilisés. Le chapitre 3

montrera des exemples d'application, notamment une illustration utilisera l'ensemble des résumés du 9e Colloque National AIP-PRIMECA, La Plagne, France, 5-8 Avril 2005. Le chapitre 4 présentera quelques conclusions et perspectives de travail dans le cadre de la recherche d'information à partir de données textuelles.

2 Outils et méthodes

Cette section présente le processus d'extraction d'information à partir de textes (2.1) ainsi que quelques éléments nécessaires à la compréhension du fonctionnement des techniques de fouille utilisées (2.2 à 2.5).

2.1 Processus d'extraction d'information à partir de textes

Le processus d'extraction d'information à partir de textes est en deux étapes. La première étape vise à extraire des méta données sur chaque texte. Ces méta données, sous forme de vecteur, auront pour but de représenter le texte.

Ensuite ces meta données seront exploitées avec des outils de fouilles structurés. Losiewicz et al. [3] considèrent trois étapes dans la recherche d'information (Cf Figure 1).

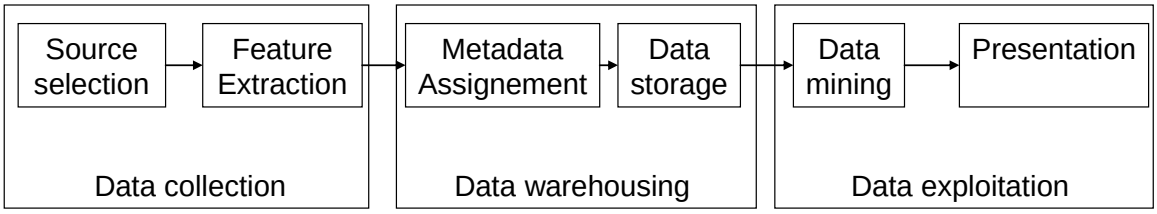


Figure 1. Processus d'extraction de l'information à partir de textes [3]

La première étape (Data collection) à pour but de rassembler les sources documentaires pertinentes (Source selection) et de prétraiter les textes pour retirer les informations non utilisables pour ne garder que l'essentiel du document, là où est sensée être l'information recherchée (Feature extraction). La sélection d'un corpus de textes est cruciale, autant le corpus de textes est adapté, autant les réponses seront pertinentes. Il ne sert à rien de rechercher des informations politiques dans les catalogues IKEA. Aussi les documents doivent être datés, car une information, une relation évolue en fonction du temps (par exemple les relations entre les hommes politiques avant et après les élections).

La seconde étape, actuellement la plus cruciale et la plus difficile est aussi la plus sujette à la critique. Il s'agit d'extraire de chaque texte l'ensemble de son contenu (Metadata assignement) et de le représenter sous une forme adaptée à la future fouille (Data storage). En pratique l'ensemble des mots du texte est supposé apporter de l'information. On extrait donc chaque mot du texte que l'on gardera dans une base de données. Cependant quelques transformations peuvent être effectuées pour enrichir cette étape. Toutes les transformations ne sont systématiquement pas employées, cela dépend des textes utilisés et de ce que l'on cherche à prédire. Notons les transformations suivantes : retirer les mots peu significatifs (le, la, les, du, des, …), transformer les majuscules en minuscules, conjuguer les verbes à l'infinitif, regrouper les mots de même racine, regrouper les synonymes … Chacune de ces transformations doit se faire avec un traitement spécifique (Cf [3] chapitre 2). Des codes informatiques sont disponibles sur internet pour chacune de ces tâches, mais dépendent de la langue considérée. Toutes les méthodes ont pour but de réduire le nombre de mots utilisés pour représenter un texte, simultanément la fréquence des mots qui restent augmente.

Finalement l'exploitation des données (Data exploitation) se divise en deux étapes : la fouille (Data mining) va permettre de construire des modèles, de faires ressortir des liens entre les « données » ; les résultats de la fouille ne seront souvent utilisables qu'avec une présentation adéquate à l'utilisateur (Presentation).

2.2 Format des documents

Les documents à explorer peuvent avoir plusieurs formats, il peut s'agir par exemple d'un format propriétaire plus ou moins facile à exploiter, d'un fichier ASCII, ou d'une image si le texte est simplement scanné sans autre traitements. La plupart des travaux supposent que les documents sont formatés selon la convention XML. Ce format est couramment employé, notamment sur le web. Un document XML contient des données de structure telles que <TITLE>, <AUTHOR>, <ASTRACT>, <TEXT> … qui serviront à délimiter les différents éléments du texte. Par la suite ces différents éléments pourront être traités différemment en fonction de ce que l'on cherche à produire. Notamment les mots apparaissant dans les différentes sections ne seront pas pondérés de la même manière.

2.3 Comment représenter un texte par un vecteur !

Chaque texte sera représenté par un vecteur. Ce vecteur devrait contenir la totalité (ou un maximum) de l'information pertinente contenue dans le texte. Un certain nombre d'informations peuvent être extraits avec peu ou pas de perte d'information : l'auteur, l'institution à laquelle appartient l'auteur, le format du document (article, livre, résumé, page web, …), la date, les mots clefs, la liste des références utilisées par l'auteur, la revue, les numéros de page…

Il reste cependant encore à extraire l'information contenue dans le reste du texte (la plus grande partie du document!). En pratique, on procède par l'extraction de tous les mots contenus dans le texte. Les termes vides de sens sont retirés (par exemple : le, la, les, du, des, …). Toutes les majuscules deviennent des minuscules. Toutes les conjugaisons sont converties à l'infinitif. Les mots de même racine et les synonymes peuvent être traités. Ensuite chaque texte est représenté par un vecteur qui décrit l'ensemble des mots contenus dans le texte (après traitement).

Le vecteur est construit dans le but de représenter le contenu du texte. Chaque coordonné du vecteur représente un mot du texte :

$$d=(w_1, w_2, w_3, \dots w_n) \qquad (1)$$

w_i est le poids du i^{ème} terme, et se calcule comme suit :

$$w_i = tf_i.\log\left(\frac{N}{n_i}\right) \qquad (2)$$

Avec :

- tf_i la fréquence du terme i dans le document d
- N le nombre total de documents dans l'ensemble de textes à étudier
- n_i le nombre de documents dans l'ensemble de textes qui contiennent le terme i

w_i montre l'importance du terme i dans le texte d. tf_i permet de prendre en compte l'utilisation du terme i dans le document d. Le log signifie qu'un terme qui apparait systématiquement dans tous les textes n'apporte rien au texte i. Différentes techniques peuvent enrichir ce processus de base [7]. La pondération d'un mot peut être influencée par sa position dans le texte : titre, sous-titre, début ou fin de chapitre, ailleurs dans le texte, proche d'un autre mot important …

2.4 Dictionnaires

Faire appel à un dictionnaire des synonymes permet de regrouper les termes qui ont la même signification. Cependant il faut être prudent, dépendamment du contexte le même mot peut dramatiquement changer de sens. Par exemple le mot « bâtiment » peut se référer à de nombreux types de construction d'habitation, de bureaux, d'usine, d'hôpitaux … pour un architecte, ou bien il peut s'agir de différents types de bateaux de guerre pour un militaire.

On utilisera donc plutôt des dictionnaires thématiques que des dictionnaires des synonymes à vocation générale. Par exemple la marine et l'armée de l'air Américaine sont convenues d'un dictionnaire des synonymes pour la fouille de données [3]. D'un point de vue global, cela permet à l'ensemble de leurs utilisateurs de partager (d'apprendre, d'imposer ?) le même vocabulaire. Tout le monde est donc sensé trouver l'information pertinente vis-à-vis du domaine du groupe.

En pratique ces dictionnaires des synonymes thématiques sont réalisés par apprentissage. L'utilisateur (ou le groupe d'utilisateur) définit lui-même en fonction de son contexte (de vie, de travail, de culture …) son dictionnaire des synonymes. Il s'en suit que pour un utilisateur donné (ou un groupe d'utilisateur donné), plus il cherche, mieux il trouve, puisque les informations sont toujours plus pertinentes en fonction de l'apprentissage fourni au dictionnaire des synonymes. En revanche si un utilisateur d'un groupe cherche, avec les mêmes mots dans la même base de données, sur un dictionnaire différent, il ressortira des informations différentes.

2.5 Notions de distance

Afin de traiter les informations de similarité entre documents il a fallu définir des notions de métrique. Ces notions pourront être utilisées pour comparer des documents, évaluer la pertinence d'un document vis-à-vis d'un ensemble de mots clés, regrouper les documents similaires, classifier un document nouveau …

Différentes notions de distances sont définies.

– La distance entre deux textes est donnée par la mesure suivante :

$$\cos(d_1, d_2) = \frac{d_1.d_2}{\|d_1\|.\|d_2\|} \qquad (3)$$

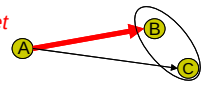
Ceci représente le produit scalaire de d_1 avec d_2, divisé par le produit des normes des vecteurs.

– La distance d'un texte à un ensemble de texte peut se calculer de différentes manières :

Plus proche voisin

$$D = \min\left(d(P_1, P_j), P_1 \in G_1, \forall P_j \in G_2\right)$$

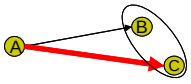
Similarité lien-complet



Voisin le plus éloigné

$$D = \max\left(d(P_1, P_j), P_1 \in G_1, \forall P_j \in G_2\right)$$

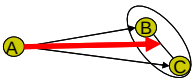
Similarité lien-simple



Distance moyenne

$$D = \frac{\sum_{j}^{n_2} d(P_1, P_j)}{n_2}$$

Similarité lien-moyen



Distance des centres de gravité

$$D = d(\theta_1, \theta_2)$$

(4)

Toutes ces mesures sont appliquées et permettent des regroupements différents.

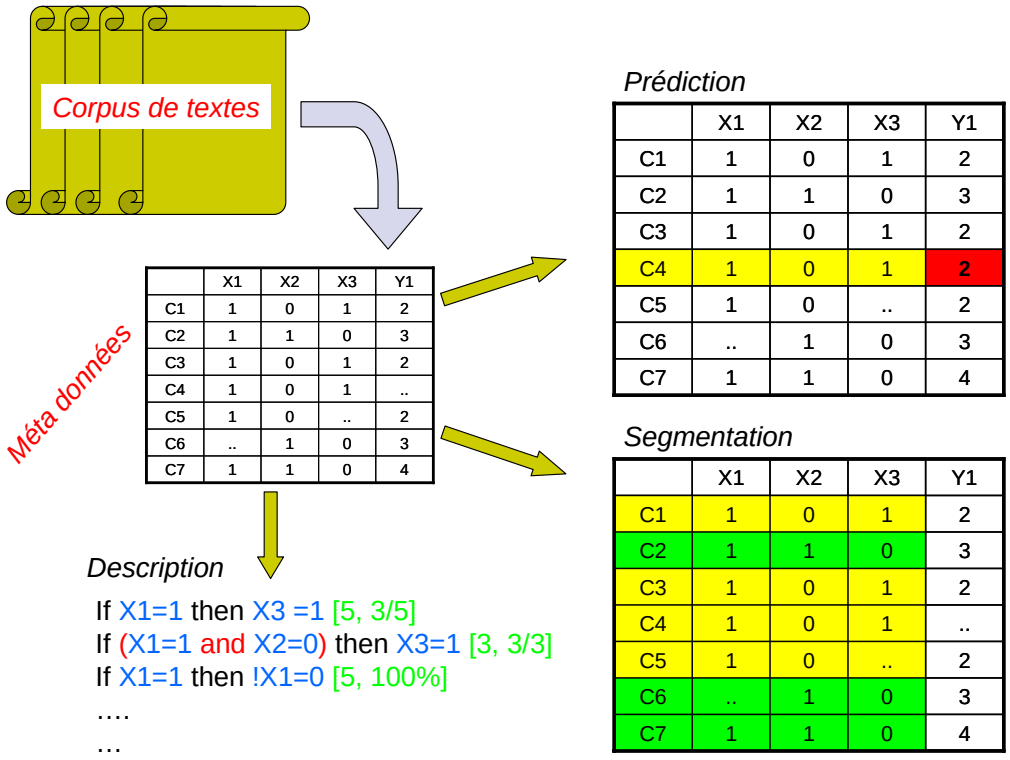
3 Exemples d'applications

Dans la section précédente, il a été présenté le processus général d'extraction d'information à partir de textes, ainsi que quelques éléments de base pour la production des métadonnées relatives à un texte. De manière simplifiée, les informations à traiter sont désormais sous la forme présentée en Figure 2.

	Montréal	Trudeau	La_fontaine	fable	morale	indépendance	label
#123	1	1					Histoire du Québec
#112			1	1	1		Livre de fable
…						…	…
#1212		1				1	Histoire du Québec

Figure 2. Exemple simplifié des vecteurs représentants des textes classifiés

Cette forme beaucoup plus structurée, va pouvoir être exploitée avec l'ensemble des outils de data mining. Bien que la dimension des vecteurs représentants les textes peut être immense ils sont principalement composés de « 0 ». Les algorithmes restent donc très rapides et utilisables en pratique. La Figure 3 représente les possibilités de fouilles offertes suite à la transformation de données.



	X1	X2	X3	Y1
C1	1	0	1	2
C2	1	1	0	3
C3	1	0	1	2
C4	1	0	1	..
C5	1	0	..	2
C6	..	1	0	3
C7	1	1	0	4

	X1	X2	X3	Y1
C1	1	0	1	2
C2	1	1	0	3
C3	1	0	1	2
C4	1	0	1	**2**
C5	1	0	..	2
C6	..	1	0	3
C7	1	1	0	4

	X1	X2	X3	Y1
C1	1	0	1	2
C2	1	1	0	3
C3	1	0	1	2
C4	1	0	1	..
C5	1	0	..	2
C6	..	1	0	3
C7	1	1	0	4

Figure 3. Possibilités de fouille suite à la transformation des données

3.1 Classification de textes, prédiction

La classification de textes (ou la prédiction de classe) s'effectue en deux étapes (Cf Figure 4). La première étape, d'apprentissage, s'effectue à partir d'un ensemble de textes déjà classifiés, elle a pour but de construire un modèle de prédiction. Le modèle de prédiction sera ensuite exploité pour le classement d'un texte nouveau en fonction de son contenu.

Il est possible pour cela de s'appuyer sur différentes techniques : arbres de classification, réseaux de neurones, classificateurs bayésiens…. Les premiers sont interprétables par l'utilisateur, ils vont permettre de définir un label pour chaque texte, en général seuls quelques mots seront nécessaires à la classification. Les seconds peuvent facilement classer un document dans différentes catégories simultanément (c'est un texte de science et de médecine), ils sont en revanche non interprétables par l'utilisateur, aussi un prétraitement sera nécessaire pour définir l'ensemble des termes pertinents sur la couche d'entrée, afin de ne pas construire un réseau de neurone ingérable.

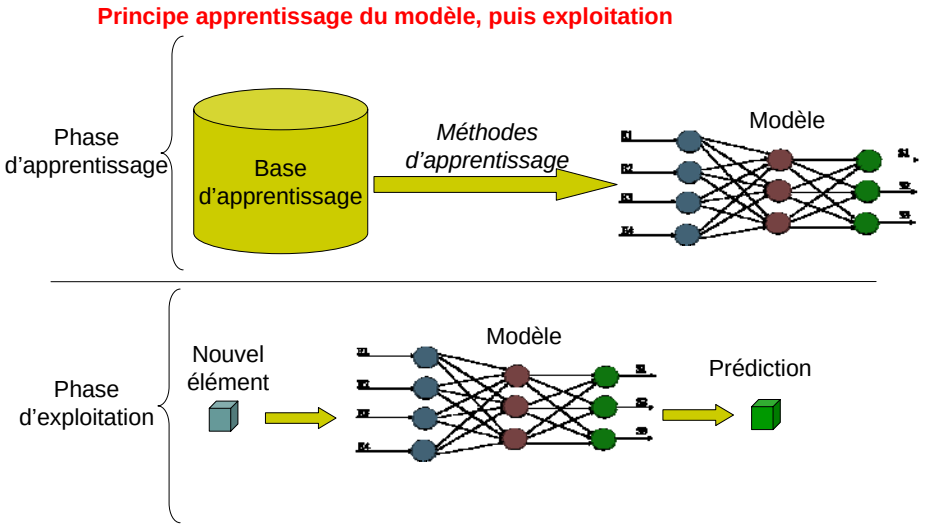


Figure 4. Principe de la classification

Puisqu'il s'agit d'apprentissage supervisé, une base d'apprentissage est nécessaire, où les textes sont déjà classifiés et il faut définir préalablement les catégories. Les bibliothèques disposent d'une base d'apprentissage importante et très structurée depuis longtemps.

3.2 Segmentation de textes

Parfois, pour un ensemble de textes, il se peut que nous ne disposions pas de base d'apprentissage, aucune catégorie de texte n'ayant été définie a priori. Dans ces cas, les méthodes d'apprentissage non supervisées sont avantageuses, elles n'ont pas besoin de prédéfinir de catégories. Les catégories sont un résultat. En revanche ces méthodes ont besoin que soit définies des notions de distance (Cf 2.5). Il est ensuite relativement facile de définir les groupes de textes similaires avec des algorithmes de partition (ex: k-mean), avec des algorithmes hiérarchiques (ex: par agglomération, par division), avec des méthodes par densité, par grilles, …

Chaque ensemble de textes formé pourra ensuite recevoir un label. Le nombre de regroupement dépendra de la finesse et de la précision recherchées par l'utilisateur.

3.3 Descriptions, cartes de concepts

Les associations entre mots peuvent révéler une information *a priori* intéressante pour l'utilisateur. Les liens entre personnes, organisations, concepts peuvent se révéler une information primordiale plutôt que le contenu des textes en eux même.

Illustrons cela par un exemple : nous recherchons des informations sur le *data mining*. Premier réflexe d'un étudiant quelconque à notre époque : son ami *Google*. La réponse de *Google* à la requête de l'étudiant sur « data mining » lui aura révélé le 12 décembre 2006 **48 300 000** réponses, toutes de la forme représentée (Figure 5).

Notre étudiant est à cette étape peu avancé et va devoir parcourir un à un les liens qui lui paraissent pertinents, redéfinir ses critères de recherche, puis … abandonner.

Malgré tout la réponse renvoyée par *Google* contient de l'information et elle est sous format texte. Il est alors possible de fouiller les réponses de *Google* et autres moteurs de recherche pour évaluer les liens entre les mots, les sites et retourner une réponse plus parlante pour notre étudiant. En une étape sur *Kartoo.com* il apprendra que (Figure 6) le *data mining* est tr`s fortement lié à *knowledge discovery*, qu'il trouvera des *information* dans des *free encyclopedia (en.wikipedia.org)*, des *resources* chez *www.kdnuggets.com* et le nom de quelques sociétés (*www.sas.com*, *www.spss.com*) qui œuvre dans le domaine. Mais aussi, malheureusement, des liens publicitaires …

Figure 5. Recherche de « data mining » dans Google.com

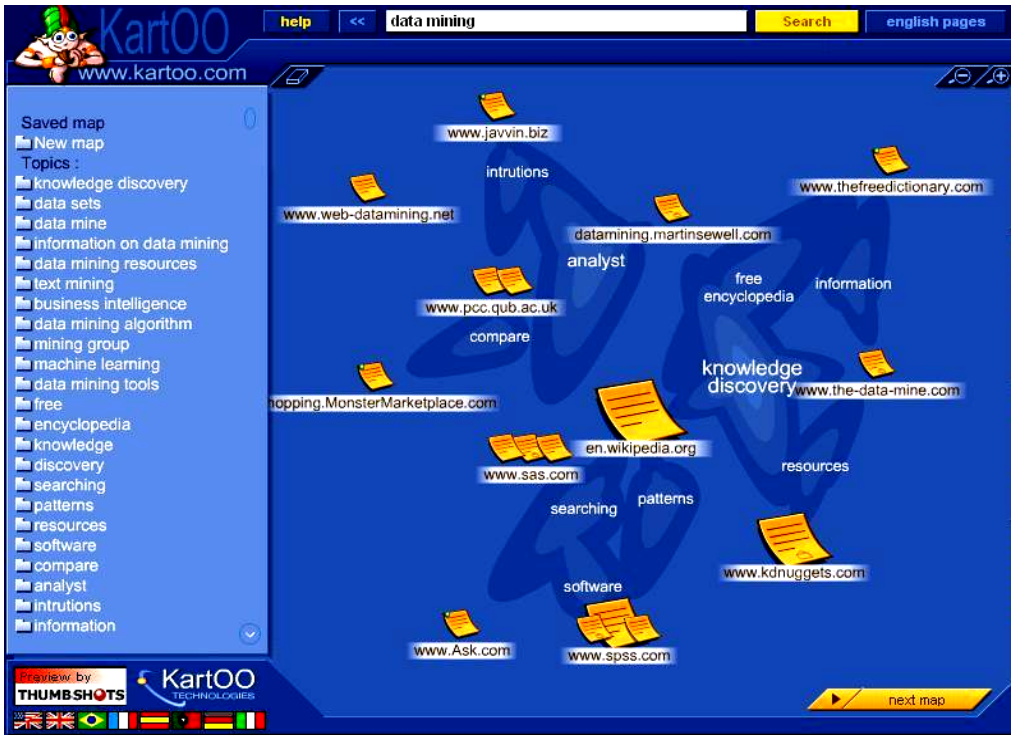


Figure 6. Recherche de « data mining » dans Kartoo.com

Kartoo fournit une liste de sujets: *knowledge discovery, data sets, data mine, information on data mining, data mining resources, text mining, business intelligence, data mining algorithm, mining group, machine learning, data mining software*, … qui se sont révélés pertinents vis à vis de la recherche. L'étudiant est ainsi guidé pour préciser ces critères de recherche.

Feldman et Hirsh [2] ont ainsi étudié la cooccurrence entre les mots utilisés pour établir des liens entre les idées. Wong et al. [6] se sont intéressés à la représentation graphique des informations contenues dans les règles d'association.

3.4 Mise en œuvre

Dans cette section les principaux outils énoncés sont illustrés par une mise en œuvre à l'extraction de connaissance à partir de l'ensemble des résumés du *9e Colloque National AIP-PRIMECA*, La Plagne, France, 5-8 Avril 2005.

Étape 1 : l'ensemble des articles présentés lors du colloque sont assemblés (60 documents), les documents des sessions plénières sont retirés (il reste alors 54 documents). Le titre, les auteurs, le résumé et les mots clé de chaque article sont extraits, le tout sauvegardé au format XML. Ce sera alors la base de recherche.

Étape 2 : Il s'agit tout d'abord de constituer le dictionnaire qui contiendra l'ensemble des mots pertinents pour le corpus de texte. Tous les mots des textes sont extraits, rapportés au singulier ou à l'infinitif avec inclusion des mots composés (dans notre cas : les ensembles de deux mots qui se retrouvent souvent dans un intervalle de 6 mots). On ne conserve que les éléments dont la fréquence d'utilisation dépasse certain seuil fixé à 6 dans notre cas. Un dictionnaire de mots vides de sens (par rapport à un but déterminé) est utilisé pour retirer les mots jugés peu porteurs d'information (le, la, du, c'est…). On peut à loisir ajouter, retirer des mots dans le dictionnaire construit, en fonction de ce que l'on souhaite observer comme information (n'étant pas intéressé aux relations avec les auteurs, ceux-ci ont été retirés). Un dictionnaire de 221 mots est ainsi obtenu. Chaque texte est finalement représenté sous forme de vecteur, avec les fréquences respectives d'utilisations des mots présents dans le dictionnaire (Cf. Fig. 7).

	conception	produit	modèle	outil	méthode	…
Résumé 1	1	1	1	1	0	…
Résumé 2	0	0	2	0	0	
Résumé 3	0	0	0	0	1	
Résumé 4	4	0	0	3	1	
Résumé 5	4	5	7	5	0	
…	…					
Utilisation globale de chaque terme	106	73	61	49	48	…

Figure 7. Représentation vectoriel du corpus de textes (avec total des fréquences)

Étape 3 : Analyse des résultats. Nous pouvons tout simplement observer les mots les plus utilisés, les associations entre les termes. Voici dans l'ordre de fréquence décroissante les mots qu'utilisent les auteurs du colloque : *conception, produit, modèle, outil, méthode, projet, plus, analyse, connaissance, usinage, modélisation, simulation, processus, pièce, entre, étude, permet, coût, système, assemblage, article, objectif, mécanique, contrainte, motsclés:conception, problème, structure, caractéristique, phase, résultat, numérique, industriel, travail, donnée…*Quelqu'un d'extérieur à la communauté peut se faire une idée général du colloque. De la même manière, un rapide calcul des règles d'associations (avec l'algorithme *a priori*) fait ressortir que 20% des articles ont « conception » dans les mots clés et que cela s'est traduit par l'apparition des mots « conception » et « produit » dans 61% des résumés concernés. De la même manière, 22% des résumés utilisent simultanément les termes « conception » et « outil », parmi ceux-ci 67% utilisent aussi le mot « produit ». De manière similaire, « outil » est lié à « méthode », « outil » est lié à « conception », « méthode » ET « analyse » sont liés à « conception »… Connaissant les fréquences d'utilisation de chaque terme et les autres termes qui y sont liés, il est alors possible d'utiliser une représentation graphique judicieuse comme représenté Figure 6.

4 Conclusions et perspectives

Cette vue d'ensemble du domaine de la fouille de données textuelles nous permet d'affirmer que ce domaine de recherche est mûr, il est maintenant possible, en utilisant différents outils et avec un peu d'imagination, de faire ressortir des informations de manière automatique à partir d'un ensemble de textes, sans même les lire. Le processus d'extraction, une fois le corpus de textes défini, est en deux étapes : (1) – structurer l'information contenue dans les textes sous forme de vecteur, (2) – fouiller l'ensemble de vecteurs à la recherche de patrons fréquents.

Les outils de data mining, en évolution rapide, sont déjà capables de traiter des masses considérables de données, une fois les données convenablement formatées. La fouille par elle-même

n'est plus réellement un problème, bien que des améliorations permettent à chaque jour d'obtenir des résultats plus pertinents, plus rapidement. Les méthodes disponibles permettent d'extraire des patrons fréquents, d'isoler les éléments rares, de classer, de prédire,…

S'il n'est pas forcément besoin d'être linguiste pour traiter le modèle structuré, en revanche l'interaction avec des linguistes permettra d'enrichir grandement le contenu du modèle structuré, et par la même occasion la pertinence des résultats. Actuellement les plus grosses lacunes sont dans la perte d'information lors du passage du document de son format texte à son format vecteur. Les recherches effectuées indépendamment par les linguistes sont assez éloignées des problématiques présentées. Les patrons et structures recherchées ne sont pas les même et le but est différent.

Les premières applications sont à portée de la main, Trappey et al. [4] par exemple sont capable de rechercher l'information essentielle contenue dans les brevets et de produire automatiquement des résumés forts utiles pour la classification de ces brevets et la recherche d'antécédence. Les moteurs de recherche tels que *Kartoo* ne se contentent plus de la recherche d'information (qu'ils n'effectuent même pas !) leur point fort est de rendre l'information des moteurs de recherche intelligible, proposer de la connaissance plutôt que de l'information. Aussi les bibliothèques de part leur classification très structurée et leurs mots clés prédéfinis permettent de retrouver des documents non pas seulement sur les mots clés utilisés, mais sur leur signification (recherche par thème, recherche par voisin).

Malheureusement certaines informations sont actuellement inexploitables, et notamment l'information contenue dans leurs structures particulières les rendent difficiles à traiter, ce sont les formules mathématiques, les tableaux et les figures qui malgré leur richesse perdent actuellement une grande partie de leur information.

À l'étape actuelle, nous ne devons plus avoir peur des gros volumes de données à traiter, mais plutôt nous devons trouver les manières pertinentes pour transformer les données non structurées en données structurée. Ensuite la recherche d'information puis la représentation de connaissance sont sur le point de fonctionner.

De nombreuses perspectives s'offrent à nous et il n'est que de s'avoir les attraper. Il n'est plus qu'à considérer tels quels les corpus d'une application particulière et voir ce qu'ils peuvent nous apprendre.

Références

[1] M. Ben-Dov, R. Feldman (2005), Chapter 38: Text Mining and Information Extraction, *in Oded Maimon and Lior Rokach (Ed.), The Data Mining and Knowledge Discovery Handbook*, Springer, ISBN 0-387-24435-2, pp. 801-831.

[2] R. Feldman, H. Hirsh, (1998), Mining Text Using Keyword Distributions, *Journal of Intelligent Information Systems*, vol 10, pp. 281-300

[3] P. Losiewicz, D. Oard, R. Kostoff, (2000), Textual Data Mining to Support Science and Technology Management, *Journal of Intelligent Information Systems*, vol 15, pp. 99-119

[4] A.J.C. Trappey, C.V.Trappey, B.H.S. Kao (2006), Automated Patent Document Summarization for R&D Intellectual Property Management, *10th International Conference on Computer Supported Cooperative Work in Design*, pp.1-6.

[5] S.M. Weiss, N. Indurkhya, T. Zang, F.J. Damerau (2005), *Text Mining – Predictive Methods for Analyzing Unstructured Information*, Springer.

[6] P. C. Wong; P. Whitney, J. Thomas, (1999), Visualizing association rules for text mining, *Proceedings IEEE Symposium on Information Visualization (Info Vis '99)*, 24-29 Oct., pp.120-123.

[7] H. Xu, J. Li, P. Xu (2005), Extract list data from semi-structured document using clustering, *Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering. IEEE NLP-KE '05*, 30 Oct.-1 Nov., pp.559-564