

CHAPITRE 1

TEXT MINING

1.1. Introduction :

Les textes expriment un grand nombre d'informations de natures diverses mais la manière dont cette information est représentée rend difficile l'analyse automatique ; L'information n'est donc pas structurée (texte libre) ; Cette absence de structure n'autorise pas un accès direct aux informations ; Le volume de données est très important rendant impossible toute analyse par un humain.

1.2. Qu'est ce que le Text Mining ?

Le Text Mining, également appelé fouille de textes ou extraction de connaissances, est un domaine de recherche dont la première définition est donné par (R.Feldman, 1995). Le Text Mining, qui est une des disciplines du traitement automatique du langage naturel (TALN), est en pleine expansion car il permet de traiter un volume important de données textuelles provenant d'internet, de mails, d'enquêtes de satisfaction, de contacts clients... qui ne peuvent être exploitées manuellement. L'objectif du Texte Mining est de faire ressortir, dans une masse très importante de données textuelles, l'information utile afin qu'elle puisse devenir exploitable informatiquement. Il s'agit donc d'extraire de la connaissance de documents sémantiquement proches et de rechercher des relations entre entités textuelles (termes) ou entre documents et de découvrir des tendances, des concepts.... [05]

La fouille de textes est un ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de nouveauté ou de similarité dans des textes produits par des humains pour des humains. Dans la pratique, cela revient à mettre en algorithmes un modèle simplifié des théories linguistiques dans des systèmes informatiques d'apprentissage et de statistiques [05].

Les disciplines impliquées sont donc la linguistique calculatoire, l'ingénierie du langage, l'apprentissage artificiel, les statistiques et bien sûr l'informatique [05].

En effet, "La fouille de textes ou l'extraction de connaissances dans les textes est une spécialisation de la fouille de données et est membre du domaine de l'intelligence artificielle. Cette technique est fréquemment désignée sous l'anglicisme texte mining [05].

1.3. Outils de Text Mining :

Les outils de Text Mining ont pour objectif de faciliter la découverte de connaissances. En première analyse nous pouvons identifier 4 types d'outils de Text Mining [06]:

- les outils de classification ;
- les outils de résumé automatique ;
- les outils d'extractions de connaissances ;
- les suites logicielles de Text Mining.

1.3.1. Outils de classification :

Les outils de classification permettent de réaliser des traitements à haut niveau de valeur ajoutée sur des fonds documentaires. Ils assurent la réalisation des opérations suivantes:

- génération automatique de plans de classement : organisation de façon dynamique et intuitive d'un ensemble non structuré de documents en thèmes et établissement d'une véritable cartographie du fonds documentaire considéré ;
- catégorisation automatique : classement par apprentissage des documents dans un plan de classement préexistant, il est possible à ce niveau de catégoriser des fonds documentaires de natures hétérogènes [06].

1.3.2. Outils de résumé automatique :

L'objectif d'un outil de résumé automatique est de produire, à partir du contenu d'un document, une représentation condensée dans laquelle les informations importantes du texte original sont préservées tout en tenant compte des besoins de l'utilisateur. Il existe deux grandes catégories de techniques pour construire un résumé automatique :

- la reformulation, qui s'attache à comprendre le contenu du document de manière à générer un nouveau texte, contenant de nouvelles phrases, différentes du texte original.
- l'extraction, qui repose sur l'extraction d'information. Le résumé obtenu contient les éléments jugés importants du texte original. [06]

1.3.3. Outils d'extraction de connaissances :

La vocation des outils d'extraction de connaissances est d'identifier l'information pertinente. Ces outils mettent en œuvre une analyse du texte pour interpréter et construire une représentation formelle qui permettra d'apporter automatiquement des réponses précises à l'utilisateur. Il ne s'agit

donc pas simplement de sélectionner un fragment brut du texte, mais de mettre des éléments en relation pour restituer une information complète et structurée à partir d'un patron prédéfini [06].

1.3.4. Suites logicielles de Text Mining :

Les suites logicielles de Text Mining sont de véritables boîtes à outils dont la vocation est de faciliter la découverte de connaissances. Cet ensemble d'outils propose l'ensemble des fonctionnalités qui sont offertes par les différents outils que nous venons de voir [06].

1.4. Tâches principales de la fouille de textes:

Dans cette section, nous allons énumérer les trois principales tâches auxquelles s'attaque la fouille de textes. Chacune de ces tâches sera un cas particulier du schéma général de la figure 1.1, pour lequel nous préciserons [07]:

- La nature des données et des résultats (en particulier, s'il s'agit de textes, quelle représentation est privilégiée).
- La nature des ressources utiles, à titre obligatoire ou facultatif.
- La nature des méthodes utilisées pour la programmer, et si elle peut être abordée par apprentissage automatique.
- Les applications concrètes de cette tâche.

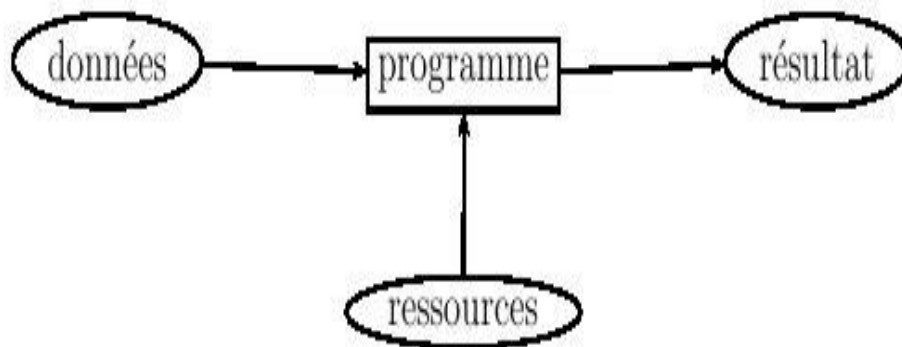


Figure 1.1 : schéma général d'une tâche de fouille de textes [07].

1.4.1. Classification de textes :

La tâche la plus "naturelle" à envisager, étant donnée la section précédente, est la classification de textes. Elle consiste à ranger des textes ou des documents dans des "classes" prédéfinies :

- les données sont donc des textes, la plupart du temps représentés sous la forme de vecteurs.
- Des variantes de ce type de représentation ont été étudiées spécialement pour cette tâche, par

exemple pour donner plus d'importance aux mots présents dans des titres, ou privilégier certaines catégories grammaticales.

- les ressources nécessaires sont celles qui permettent la représentation du texte : anti dictionnaire, lemmatiseur voire analyseur morphologique, compte d'occurrences, étiqueteur "part of speech" si on privilégie certaines catégories...
- cette tâche est presque'exclusivement abordée par apprentissage automatique, à partir d'exemples de textes déjà classés. Parmi les méthodes citées en fin de section précédente, celles basées sur des comptes statistiques ("naïve Bayes") ou sur les techniques SVM donnent les meilleurs résultats.
- il y a de très nombreuses applications concrètes de cette tâche. L'une d'elle fonctionne déjà sur la plupart des gestionnaires de courrier électronique : il s'agit du programme qui suggère que certains des mails reçus sont probablement des "spams" non désirés. Les deux classes sont alors "spam" et "non spam" et l'ensemble des courriers déjà reçus constitue l'échantillon d'apprentissage à partir duquel le programme apprend à poser son diagnostic. De manière générale, la classification automatique de textes par "thème" peut rendre de grands services. On peut aussi utiliser des méthodes similaires pour retrouver l'auteur d'un texte (l'étiquette de la classe est alors un nom d'auteur) à partir d'exemples de textes attribués à coup sûr : des critiques littéraires s'en sont servi pour argumenter que certaines pièces signées par Molière avaient en fait été écrites par Corneille. Enfin, l'autre type d'application en plein développement de la classification est la reconnaissance automatique des opinions véhiculées par un texte : les classes, dans ce cas sont par exemple "favorables" et "défavorable". Certaines sociétés qui reçoivent des courriers électroniques de consommateurs à propos de leurs produits s'en servent pour analyser leur contenu. Dans ce cas, la représentation des textes a intérêt à privilégier les adjectifs et les verbes, qui sont les principaux moyens d'exprimer une opinion [07].

1.4.2. La recherche d'information (ou RI) :

Est l'autre "tâche" générale d'ors et déjà omniprésente dans nos usages quotidiens des ordinateurs. Nous la sollicitons chaque fois que nous recherchons des documents répondant à une "requête".

- la donnée fournie par l'utilisateur est donc une requête. Celle-ci peut prendre des formes diverses, suivant le niveau d'expertise de cet utilisateur et la structure de la base de documents à interroger : simple liste de mots clés, langage de requête structuré (combinaisons de critères booléens, expressions rationnelles, requêtes type SQL...), voire

document "exemple" dont on cherche des exemplaires "proches" parmi un ensemble de textes.

- les ressources sollicitées sont tout d'abord le corpus de textes ou de documents que l'on cherche à interroger. Ce peut être une base d'articles, une encyclopédie, ce peut être Internet... Comme précédemment, il est éventuellement fait appel aux ressources nécessaires à la représentation de la requête par un vecteur. Enfin, quand la requête est réduite à un ensemble de mots-clés, il est courant d'utiliser un thesaurus ou une ontologie pour l'étendre à des mots "proches" (par synonymie, ou par généralisation en "remontant" dans la hiérarchie par exemple).
- on distingue trois familles de méthodes pour aborder la RI :
 - les méthodes *booléennes* fonctionnent à l'aide d'un simple index qui donne, pour chaque unité lexicale figurant dans la requête, la liste des textes où cette unité est présente. Les requêtes acceptées sont alors généralement des combinaisons de critères booléens (avec les opérateurs NON, ET, OU). Des calculs simples permettent d'obtenir la liste des textes où tous ces critères sont satisfaits en même temps.
 - les méthodes *vectorielles*, comme leur nom l'indique, codent toutes les informations (la requête et les documents de la base) sous la forme de vecteurs. La représentation TF-IDF est née dans ce contexte, et y est particulièrement efficace. La RI se ramène alors à trouver les vecteurs les plus "proches" d'un vecteur donné (celui représentant la requête). Pour quantifier ces distances, on utilise souvent des mesures basées sur le cosinus de l'angle qu'ils font entre eux (facile à calculer par des formules mathématiques).
 - les méthodes *statistiques* qui en font reviennent à faire de la classification automatique en supposant que l'on connaît déjà, pour la requête, un ensemble de documents "pertinents" et de documents "non pertinents", et que l'on cherche à trouver tous les documents devant être classés comme pertinents. On le voit, cette méthode n'est pas vraiment comparable aux autres, puisqu'elle fait des hypothèses supplémentaires sur ce qui doit être fourni au système. Mais c'est la seule manière de faire intervenir de l'apprentissage automatique dans la tâche de recherche d'information.
- la recherche sur Internet est, bien sûr, l'application phare de cette tâche. Les moteurs de recherche mettent en œuvre des méthodes booléennes : leur index fait leur force ! Or ces méthodes ne permettent pas de classer en "plus ou moins pertinent" les documents obtenus

(en l'occurrence les sites Web). C'est pourquoi ils doivent employer d'autres techniques (d'où l'importance du fameux "Page Rank" de Google) pour classer par ordre de pertinence ces sites. De nombreux autres logiciels existent pour gérer la documentation d'une société ou d'une organisation : la plupart fonctionnent par des méthodes vectorielles. Des recherches sont en cours pour étendre leur domaine d'application aux documents structurés disponibles sous la forme d'arbres, pour lesquels on veut autoriser des requêtes mélangeant contenu textuel et structure [07].

1.4.3. Extraction d'information :

Est la dernière tâche fondamentale que nous voulons présenter ici. Comme son nom l'indique, elle se fixe comme objectif d'*extraire* de textes des informations factuelles précises. Imaginons par exemple les textes de petites annonces de vente de voitures, rédigées librement. Les informations qu'elles contiennent peuvent se résumer à la valeur de quelques "champs" factuels : qui vend quel type de voiture, de quel kilométrage, à quel prix, etc. On appelle *rappeur* (terme anglais qui signifie "envelopper") un programme capable de remplir automatiquement les valeurs de ces champs à partir du texte initial de la petite annonce. Un wrapper est nécessairement spécialisé dans le traitement d'un certain type de textes : celui qui traite les petites annonces de vente de voiture ne saura pas quoi faire d'annonces de locations d'appartements, et inversement.

- les données d'entrées sont des représentations de textes de même type, où la notion de *séquence* est préservée, elles peuvent aussi être des *documents structurés* (pages HTML ou XML) ; les sorties sont des *données structurées*, en général sous la forme d'une liste d'attributs (prédéfinis) remplie ;
- parmi les informations disponibles au wrapper, on suppose qu'il y a la liste des champs à extraire (cette liste dépend bien sûr du type de textes). Les ressources linguistiques utiles à la réalisation de cette tâche dépendent de la méthode employée : toutes les techniques d'identification d'entités nommées (liste de valeurs possibles, mais aussi expressions régulières ou automates) sont intéressantes car, souvent, la plupart des données à extraire (noms propres ou valeurs numériques) sont des entités nommées. Des étiqueteurs grammaticaux, voire des analyseurs syntaxiques, sont parfois aussi employés.
- pour définir un "wrapper", il est possible de le programmer directement. Les méthodes les plus efficaces font appel, pour chaque champ à remplir, à des automates ou à des expressions régulières qui repèrent les environnements possibles où peut apparaître l'information visée. Mais, depuis quelques années, l'apprentissage automatique de wrappers à partir d'exemples de textes d'où ont été extraites des données factuelles est un thème de

recherche très actif. Certains systèmes ré-exploitent pour cela des techniques de classification automatique, mais en les adaptant à ce nouveau contexte. Des compétitions existent pour comparer les meilleurs programmes.

- Un système d'extraction automatique fournit rapidement un "résumé structuré" d'un texte. Les données "attributs/valeur" qu'il fournit en sortie peuvent facilement alimenter une feuille de calcul ou une base de données relationnelle, ce qui intéresse tous ceux qui doivent manipuler de nombreux exemplaires de documents standardisés. Il existe aussi des "wrappers d'arbres" particulièrement adaptés aux pages HTML, capables d'extraire certaines des informations contenues dans ces pages, en tenant compte de leur environnement à la fois textuel et structurel (balises). Ils sont très utiles aux "veilleurs" chargés de surveiller les sites qui évoluent beaucoup, afin de les aider à se focaliser rapidement sur les données qui les intéressent [07].

1.5. Applications de Text Mining :

On va présenter quelques applications de la fouille de textes [08]. :

1.5.1. Les études :

Dans leur rivalité, la plupart des compagnies offrent constamment plus ou mieux, ce qui fait évoluer les tâches et l'environnement de l'employé. Avant de faire des changements, il serait utile de mesurer les réactions de l'employé. Par exemple, si une compagnie projette de changer les avantages de retraite pour ses employés, une étude peut être conduite sur leurs opinions. Même les questions à choix multiples sont trop fermées pour analyser ces opinions sous chaque angle et en fournir l'état complet. Les questions (ouvertes) qui acceptent des réponses en langue naturelle sont difficiles à traiter.

La méthode de fouille de données travaille bien quand les choix multiples mènent rapidement à des tableaux et peuvent être résumés. Mais face à des milliers de réponses en langue naturelle, c'est très difficile de traiter chaque réponse manuellement et de disposer en tableau les résultats.

1.5.2. Intelligence économique :

Elle est utile aux compagnies d'une même industrie pour suivre les produits et les développements de l'un et de l'autre. Les constructeurs automobiles comme Ford surveillent probablement directement les sites web de Toyota (pour se faire une idée des actions commerciales, de la production, des investissements, développements et recherches) et indirectement Toyota à travers la presse professionnelle, les forums, les nouvelles du METI (Ministry of Economy, Trade

and Industry). Faire ceci sur une base journalière et manuelle est lent. Les sites des grandes compagnies telles que Ford ou Toyota peuvent avoir plusieurs centaines de pages web, sans compter celles des sites plus ou moins connexes. Une approche automatisée qui télécharge périodiquement et analyse les pages relatives à un concurrent est plus effective.

1.5.3. La gestion des clients :

Un client peut appeler un centre d'assistance technique, ou envoyer un mail directement ou via une page web. Ces renseignements peuvent être rassemblés et peuvent être entreposés dans un dépôt pour les analyser. Une vue globale des réactions des clients dans la forme d'une taxonomie rend facile à identifier les classes de produits et les produits individuels objets des plaintes. Le problème le plus commun avec un produit peut ainsi être extrait. La catégorisation des réclamations des clients facilite le routage des messages vers l'expert approprié comme la rédaction d'un rapport sur le problème. Quelques produits sont complexes et la grande compagnie typique a plus d'un niveau de support. Si une question ne peut pas être répondue au premier niveau de support, la requête du client est redirigée vers un spécialiste du niveau suivant.

1.5.4. La recherche médicale :

Localiser des articles qui utilisent certains termes médicaux dans des contextes définis intéresse souvent les chercheurs. Ce type de recherche n'est pas faisable avec un moteur de recherche : il peut y avoir des myriades de combinaisons de termes et ce n'est pas pratique de soumettre manuellement des questions individuelles pour chaque combinaison à un moteur de recherche [09].

1.5.5. La recherche légale :

Il n'est pas toujours facile de corréler des rapports de police, des déclarations écrites ou des actes notariés spécifiques à une affaire, avec le droit (législation, réglementation, jurisprudence). Les outils automatisés qui peuvent extraire et résumer les renseignements ont alors beaucoup d'avantages sur un moteur de recherche, car le chercheur ne peut pas toujours connaître les mots-clés ou la terminologie spécifique touchant les renseignements dont il a besoin [09].

1.5.6. Connaître l'opinion publique :

Savoir « qui pense quoi » au sujet d'une question précise est d'un grand intérêt pour les politiciens et les sociologues. Les outils de fouille de textes peuvent augmenter les résultats des élections.

Les citoyens s'expriment sur le web, et la plupart de ces informations sont disponibles au public.

Malheureusement, ces renseignements sont éparpillés et utiliser un moteur de recherche ne résoudra pas le problème du recouvrement. Après avoir choisi certaines sources de renseignements pertinentes sur le web, les données rassemblées pourraient être analysées pour corréler opinions impliquées, fréquence, et catégories de citoyens. Si une question est fréquemment discutée, comment est-ce que les gens ressentent le sujet ? [09].

1.5.7. Shopping :

Le magasinage sur le web veut trouver le bon produit au bon prix. Les prix varient d'un site à un autre et ce n'est pas pratique de visiter et parcourir chaque site manuellement. Un site comparatif pourra se baser sur des analyses automatiques des sites vendeurs, à partir d'une liste de critères.

1.5.8. La recherche académique :

Un crawler est un logiciel d'indexation conçu pour passer en revue des sites Web et pour télécharger l'information contenue dans ces sites. L'information qu'il télécharge est le code source HTML. Dans le cas des moteurs de recherche, ce code source est stocké dans une grande base de données et plus tard analysé et classé dans l'index des moteurs de recherche.

Il peut visiter pour un département toutes les pages web sur un thème donné, et extraire les titres et autres renseignements pertinents de toutes les publications et rapports rencontrés. Cette liste peut être catégorisée et les auteurs-clés dans un département peuvent être identifiés. La même procédure peut être appliquée à de multiples départements à travers les universités et les résultats peuvent être coulés dans une seule hiérarchie de catégories. Les universités qui ont des intérêts communs peuvent être localisées et les sujets qui paraissent populaires, détectés. Les rapports entre départements peuvent être identifiés à travers les liens hypertextes comme à travers les citations dans les publications [09].

1.5.9. Le triage automatisé :

On se pose la question du classement automatique d'un ensemble de mémoires académiques [10]. Un programme peut-il distinguer automatiquement les bons essais des mauvais, et leur assigner des niveaux basés sur les analyses textuelles ? Les premiers systèmes de triage automatisé basés sur les traits simples traitaient des spams et leur assignaient les plus hauts niveaux à des essais pauvrement écrits. Cependant, il y a une corrélation forte entre une bonne écriture et certains traits du texte. Le problème principal est de trouver automatiquement et d'extraire ces traits.

1.5.10. Catégorisation des textes :

Une des raisons de construire une taxonomie de documents est de trouver plus facilement des documents pertinents.

Le problème de catégorisation peut être décrit comme la classification de documents dans de multiples catégories. Nous avons un ensemble de catégories $n \{c_1, c_2, \dots, c_n\}$ auxquelles nous devons assigner m documents $\{d_1, d_2, \dots, d_m\}$, avec $n \ll m$.

Document \rightarrow descripteur ou profil du document \rightarrow classement, en cherchant la classe dont le profil est le plus corrélé au profil du document.

La catégorisation de textes peut être utilisée sur des flux de renseignements dynamiques qui ont besoin d'être organisés. Nous définissons les renseignements dynamiques comme des mails, des articles et nouvelles, blogs, articles scientifiques, brevets, et données légales. Les applications incluent l'acheminement automatique des questions des clients, les demandes médicales, et la recherche d'entités dans le flux de renseignements. Ce type de renseignements est produit journallement, et l'utiliser est difficile sans quelque catégorisation [09].

1.6. La mise en oeuvre du texte mining :

On peut distinguer deux étapes principales dans les traitements mis en place par la fouille de textes :

La première étape, l'analyse, consiste à reconnaître les mots, les phrases, leurs rôles grammaticaux, leurs relations et leur sens. Cette première étape est commune à tous les traitements. Une analyse sans interprétation n'a que peu d'intérêt et les deux sont dépendantes. C'est donc le rôle de la seconde étape d'interpréter cette analyse.

La seconde étape, l'interprétation de l'analyse, permet de *sélectionner* un texte parmi d'autres. Des exemples d'applications sont la classification de courriers en spam, c'est-à-dire les courriers non sollicités, ou non spam, l'application de requêtes dans un moteur de recherche de documents ou le résumé de texte qui sélectionne les phrases représentatives d'un texte voire les reformule.

Le critère de sélection peut être d'au moins deux types : la nouveauté et la similarité. Celui de la nouveauté d'une connaissance consiste à découvrir des relations, notamment des implications qui n'étaient pas explicites car indirectes ou entre deux éléments éloignés dans le texte. Celui de la similarité ou contradiction par rapport à un autre texte ou encore la réponse à une question spécifique consiste à découvrir des textes qui correspondent le plus à un ensemble de descripteurs dans la requête initiale. Les descripteurs sont par exemple les noms et verbes les plus fréquents d'un texte [05].

1.7. Disciplines connexes :

La fouille de textes se distingue du traitement automatique du langage naturel par son approche générale, massive, pratique et algorithmique de par sa filiation avec la fouille de données. Son approche est moins linguistique. De plus, la fouille de textes ne s'intéresse pas au langage oral comme le fait la reconnaissance vocale.

La fouille de textes recoupe la recherche d'information pour la partie requête sur un moteur de recherche de documents. Par contre, la recherche d'information s'intéresse a priori plus aux types de requêtes possibles et aux indexations associées qu'à l'interprétation des textes.

Et pour information, car on s'éloigne alors du domaine de la fouille de textes, l'interprétation de l'analyse peut aussi *générer* un nouveau texte. Des exemples d'applications sont la correction des fautes d'orthographe, la traduction, le dialogue homme-machine ou l'imitation d'un style d'écriture [05].

1.8. Conclusion :

Dans notre mémoire, on a concentré sur les techniques de la fouille de texte (Text Mining) qui est une partie intégrante très importante du DM. Ces techniques aident les utilisateurs de l'internet, ainsi que toutes les personnes travaillant sur les textes à acquérir la connaissance pertinente à partir de grandes quantités de textes.