“ Les mots sortent

Iris Eshkol, Jean-Paul Taravella

« Les mots sortent du lot » : l'extraction d'information,

Iris Eshkol, Université d'Orléans et Jean-Paul Taravella

L'extraction d'information est une offre qui se structure depuis quelques années en France. Cette technologie répond à un besoin essentiel : permettre de repérer et d'extraire, à moindre coût, certains éléments structurés d'information à partir d'un texte non structuré On connaissait l'extraction principalement par le terme générique de « text mining » (fouille de texte) qui diffère du « data mining » (analyse de base de données structurées). Aujourd'hui les technologies sont matures et les applications sont nombreuses et ciblées. Citons :

- L'indexation automatique : Il s'agit d'extraire des mots représentatifs du contenu du texte et d'indexer relativement finement de très gros volumes documentaires (comme par exemple en rétro conversion, ce qui permet aux personnels de se consacrer à d'autres tâches) ; Le classement automatique (sans plan de classement prédéfini) : il s'agit de classer les documents sous différentes thématiques qui sont générées et arborées automatiquement,
- L'élaboration d'ontologies, de thésaurus, de terminologies, éventuellement multilingues : il s'agit grâce à l'extraction de mots et leur mise en relation avec des liens typés, d'aider le documentaliste à la découverte d'un domaine et à une première proposition de structuration en « thésaurus ».
- Le résumé automatique : l'objectif est d'obtenir une représentation synthétique du texte, avec l'extraction de phrases significatives,
- L'aide à la lecture : il s'agit de repérer rapidement dans des gros documents et/ou une grande volumétrie de documents, les mots et les faits recherchés, en contexte (mise en évidence du mot ou du fait dans un paragraphe ou une phrase)
- La recherche et la navigation en moteur de recherche : il s'agit d'identifier des catégories de mots présents dans la liste des résultats produits lors d'une recherche (toutes les « organisations », toutes les « personnes », etc. présentes dans cette liste) afin d'affiner ou d'étendre la recherche [1]

Les grands domaines utilisateurs sont donc ceux qui réclament une analyse de gros volumes textuels et une restitution sous forme structurée et synthétique : la « Relation Client », opérée par la direction Marketing ou Qualité qui souhaite une analyse automatique des e-mails ou des forums par exemple, afin de distinguer automatiquement des tendances d'expression positives et négatives sur un sujet ou un produit. Une autre application encore peut être l'analyse prédictive sur le client, en utilisant toutes les informations disponibles structurées et désormais non structurées sur le client,. Les enquêtes, les CVs, peuvent également être traités de cette même manière en exploitant les rubriques non structurées. Le deuxième grand domaine utilisateur est la veille, opérée par la direction Stratégie ou Recherche qui souhaite découvrir des mots et de relations entre ces mots, qu'il n'attendait pas (« joystik » lié à «mode de conduite» sur le site d'un constructeur automobile) ou qu'il poursuivait expressément (SociétéX *rachète* SociétéY ; Mr John *quitte* la SociétéX).

Dans tous les cas, si la Direction informatique reste encore l'acheteur principal, les utilisateurs finaux sont très vite concernés pour la mise au point des règles d'extraction selon le domaine concerné (cf. Principes technologiques) et pour l'interface métier selon leur attente de restitution des données extraites.
Aujourd'hui les principaux clients sont encore le monde de la santé, les éditeurs de contenu et celui du renseignement qui surconsomme de l'information textuelle. Mais nul doute que la

[1] article Que choisir ? sur les moteurs de recherche

performance des technologies et la mise en concurrence de plus en plus forte d'éditeurs européens ou mondiaux permettront d'intéresser un panel plus large d'entreprises et d'administrations. Avantage de taille, l'extraction d'information peut démontrer un vrai retour sur investissement. D'autres technologies comme le moteur de recherche ont un impact qui reste plus « qualitatif ». L'extraction d'information et les applications permises (indexation, classement, résumé … automatiques) peuvent être mises en rapport avec le fonctionnement actuel, d'un point de vue qualitatif et *quantitatif*.

L'offre se distingue selon deux axes. D'une part les éditeurs, spécialisés sur l'extraction qui proposent un produit et des services autour de cette technologie. Il s'agit principalement de Clearforest (http://www.clearforest.com/), Inxight (http://www.inxight.com/), Lingway (http://www.lingway.com/), SPSS (http://www.spss.com/), Temis (http://www.temis-group.com/). Ces éditeurs mettent généralement en oeuvre des grammaires pour des traitements fins. Leur présence est européenne, voire mondiale (Clearforest, Inxight, SPSS) . Ils peuvent également offrir du service pour contextualiser leur produit. Certains comme Inxight ou Temis ont une stratégie de vente de leur technologie à d'autres éditeurs non spécialisés. Le second axe est constitué par les éditeurs de moteur de recherche qui proposent une fonction d'extraction d'information principalement à des fins de navigation dans la liste des résultats (cf Articles moteur de recherche).

Les tendances de cette offre sont :

- le développement d'applications verticales car les règles et les vocabulaires sont en général spécifiques d'un domaine. Par ailleurs la technologie d'extraction ne suffit pas à rendre le service attendu par l'utilisateur qui attend généralement une solution applicative (veille, analyse prédictive, etc.),
- la mise à disposition d'outils permettant la modification des règles d'extraction afin que les utilisateurs finaux gagnent en autonomie. La mise à jour de ces règles réclamme les compétences d'un connaisseur du domaine linguistique et la connaissance du formalisme infomatique utilisé, souvent les expressions régulières (cf. Principes technologiques),
- la poursuite du rapprochement des mondes du data mining et du text mining, pour offrir une analyse complète et cohérente des données structurées et des données non structurées (rachat de Lexiquest par SPSS; partenariat d'Inxight et de SAS)
- la stabilisation de l'offre en terme de technologie source. Inxight et Temis par exemple proposent une technologie à base des outils développés initialement de Xerox. Le ticket d'entrée pour développer des analyseurs est très élevé et il y a fort à parier qu'il y aura peu de (purs) nouveaux entrants.

Les outils fonctionnent aujourd'hui sous Windows, Unix, traitent les document de plus de 200 formats différents, produisent du XML en sortie et coûtent à partir de 20-50 Keuros. Les critères discriminants sont plutôt les suivants :

- Le paramétrage et contextualisation par le client : le système est ouvert ou fermé. La notion de technologie employée n'est ici pas neutre car les outils sont soit des outils basés sur de la statistique (cooccurence avec analyseur morphologique) qui sont plutot fermés, soit des outils mettant en œuvre des ressources et fonctions linguistiques avancées (grammaires, vocabulaires) avec la prise en main possible ou non par l'utilisateur de ces données linguistiques. Dans le cas où l'éditeur ne proposerait pas ce type de fonction, l'utilisateur prendra soin d'observer les marchés déjà traités par l'éditeur afin de garantir que ce dernier connaît déjà les règles et les vocabulaires du domaine.

- La valeur de la technologie de base : c'est en effet la qualité de l'analyseur linguistique qui fera la qualité des mots extraits, avant leur comptage. Mais cette valeur reste difficile à évaluer car intégrée dans une chaine de traitement plus complexe.
- La scalabilité : la résistance à la charge, au volume,
- L'ergonomie de restitution des résultats : une cartographie par exemple peut être esthétique mais inexploitable car ne permettant pas de repérer les éléments souhaités
- Le nombre de langue traitées : au delà du français, de l'anglais, l'allemand , l'espagnol.

Les principes de l'extraction

On peut distinguer deux types d'extraction. En premier lieu l'extraction « libre ». L'information a extraire n'est pas définie auparavant. L'objectif est de décrire et représenter le contenu global des documents (l'extraction de mots ou de phrases « représentatifs» du texte). L'étape de l'extraction doit être suivie par celle de filtrage d'information qui permet de nettoyer les résultats des mots non représentatifs. En second lieu l'extraction « normalisée » : L'objectif est de repérer les mots et les extraire selon des catégories prédéfinis (personne, organisation, lieu, produit, etc).

Les algorithmes que l'on utilise pour les deux types d'extraction sont fondés sur des connaissances statistiques (on exploite la fréquence des unités textuelles) et des connaissances linguistiques (on exploite des dictionnaires et des règles d'extraction)

Quelquesoit les outils (analyseur statistique ou analyseur linguistique) l'extraction est toujours dépendante de la langue. En effet les règles d'extraction peuvent être basées sur :

- les listes « d'amorces », construites préalablement (les amorces sont des mots qui peuvent exprimer la catégorie recherchée : « société », « compagnie », « filiale ».) et les règles de leur comportement syntaxique (« société de + [Groupe Nominal]). Ces informations permettront de repérer et de typer une chaine comme « société de service informatique »; Dans un autre exemple « Président John *Withmann* », le mot inconnu « Withmann » aura de bonnes chances d'être un nom de personne, si la règle dit qu'une fonction (Président) et un prénom (John) préalablement identifié dans le dictionnaire sont généralement suivi d'un nom. Ces amorces et leur position (avant ou après le mot à catégoriser) sont souvent spécifique de la langue : ainsi on ne déterminera pas de la même façon la catégorie « lieu » en anglais (Orléans city) ou en français (la ville d'Orléans).
- la reconnaissance des verbes qui permettent la mise en relation de concepts et la recherche et l'extraction de faits : '*Spigadoro Inc.* <u>*acquires*</u> *largest European private label pasta company »*
- *la* prise en compte de « marqueurs » linguistiques (« *ainsi », « donc », « en premier lieu », etc.)* pour l'extraction de phrases importantes

Ces règles d'extraction sont souvent écrites en « expressions régulières ». Il s'agit de la combinaison de caractères littéraux et spéciaux permettant d'identifier une chaine de caractères répondant à ce modèle. Ainsi l'expression régulière « NP / (Coord / NP) ? / airport » permettra de reconnaître la chaîne « Heathrow and Gatwick airports » (Heathrow (*NP*) / (and (*Coord*) / Gatwick (*NP*)) / airports)

On comprend alors que ces règles peuvent être spécifiées après l'analyse du corpus et de ses caractéristiques[2]. C'est pourquoi la plupart des traitements reposent sur la langue mais plus encore sur des ressources propres au domaine et même propres à l'application. Le corpus de référence doit ainsi être le plus représentatif que possible du domaine, ce qui implique des difficultés supplémentaires pour la constitution de ce type de corpus qui est généralement fourni par le client.
Les limites des approches utilisées en extraction d'information proviennent dans la majorité de cas de la linguistique : l'auteur n'utilise jamais le même terme lexical pour désigner un même référent (synonymes, métaphore, etc.) ; l'emploi des termes anaphoriques. De fait l'extraction d'information normalisée est utilisée surtout dans des domaines de spécialités limités vu la difficulté plus générale du traitement automatique du langage : la richesse de son vocabulaire, l'ambiguïté des mots (Calvin Klein est un nom de personne mais aussi d'entreprise), l'utilisation des métaphores, etc.

[2] Il est à remarquer que les règles et le vocabulaire permettant l'extraction peuvent être eux-memes produits par l'apprentissage automatique du corpus de référence.