Extraction d'informations pour la veille technologique avec le système VIGITEXT

GOUJON Bénédicte

Équipe LaLIC du CAMS (UMR 17) - Bureau Van Dijk Ingénieurs Conseils

96 bd Raspail, 75006 Paris - 57 bd de Montmorency, 75016 Paris

goujon@msh-paris.fr

Cet article présente la démarche suivie pour mettre en place un logiciel d'aide à la veille technologique. Pour l'analyse de documents très techniques, les veilleurs utilisent des outils d'infométrie, qui sont pertinents sur les données structurées, mais qui ne sont pas adaptés pour l'exploitation des informations textuelles. Nous avons donc réalisé un logiciel d'extraction d'informations, nommé VIGITEXT. Notre approche, basée sur la définition de notions indépendantes du domaine comme l'/amélioration/, l'/augmentation/ ou l'/utilisation/, permet d'extraire des informations textuelles à partir d'abrégés descriptifs de brevets rédigés en anglais sans utiliser de lexique technique ou de calculs statistiques. De plus, cette méthode est opérationnelle pour tous les sujets de veille, et les résultats, qui sont les extraits organisés selon les notions, sont simples à utiliser par des veilleurs.

Dans cet article, nous décrivons les particularités de la veille technologique, et les limites des logiciels généralement utilisés. Ensuite, nous détaillons l'exploitation de notions générales basée sur la définition de connaissances linguistiques et qui met en œuvre la méthode d'exploration contextuelle. Nous présentons enfin le prototype VIGITEXT, avec ses spécificités et ses utilisations possibles dans une démarche de veille.

1. Introduction

Cet article présente la démarche suivie pour réaliser un logiciel d'aide à la veille technologique. L'observation des particularités de la veille, de ses outils, et du comportement des veilleurs nous a amené à exploiter des notions indépendantes du sujet, comme l'/amélioration/, l'/augmentation/, l'/utilisation/. Le prototype opérationnel VIGITEXT repère dans les données, en utilisant la méthode d'exploration contextuelle, des extraits textuels exprimant ces notions.

2. La veille technologique et les outils

2.1. L'information et la veille

Dans une démarche de veille technologique, il est nécessaire d'exploiter un maximum de sources d'informations, afin de ne pas passer à côté d'informations capitales. La typologie de l'information présentée par François Jakobiak [Jakobiak 95], qui propose une partition en classes, formes, types et supports, montre bien la diversité des informations qui peuvent être exploitées pour la veille. Cependant, selon les spécialistes de la veille technologique, comme Daniel Rouach [Rouach 96], les documents les plus importants dans une démarche de veille technologique sont les brevets, rédigés en anglais, obtenus principalement sous forme d'abrégé descriptif (ou référence) par l'interrogation de banques de données spécialisées. Le logiciel que nous réalisons doit être adapté à ces sources d'information.

Par ailleurs, il est nécessaire d'utiliser plusieurs outils d'analyse de données, pour exploiter au mieux les informations disponibles. Nous avons donc analysé les principaux outils adaptés à la veille, soit les outils infométriques, et nous avons cherché à définir une nouvelle méthode d'analyse des données textuelles offrant un résultat complémentaire pertinent pour le veilleur.

2.2. Outils infométriques pour l'analyse d'abrégés descriptifs de brevets

Les abrégés descriptifs de brevets contiennent des informations structurées, dans des champs " auteur ", " date ", " mots-clés ", et des informations textuelles non structurées dans les champs " titre " et " résumé ". Les outils infométriques utilisés pour la veille s'appliquent surtout sur les données structurées.

Comme le décrit Françoise Rousseau-Hans [Rousseau-Hans 98], l'utilisation d'outils infométriques dans une démarche de veille technologique permet une approche globale de l'information contenue dans un corpus. Ces outils découpent d'abord les données en unités (mots, dates ou chaînes de caractères), puis appliquent des calculs mathématiques afin d'obtenir sous forme de graphiques ou de cartes une représentation des unités en fonction de relations ou proximités calculées. Par exemple, les outils Dataview, développé au CRRM [Dataview], ou Tétralogie, développé à l'IRIT [Tétralogie], sont adaptés pour l'analyse des parties structurées des abrégés descriptifs de brevets.

Cependant ces outils ne sont pas vraiment adaptés pour l'analyse du texte libre. En effet, la plupart des outils infométriques attribuent de l'importance aux unités fréquentes. Or, le veilleur doit pouvoir identifier des indices de veille, qui sont des informations rares mais justement très importantes. Par ailleurs, certains outils se basent sur des terminologies prédéfinies. Or les informations étonnantes ou nouvelles s'expriment rarement avec des termes ou à partir de concepts connus à l'avance. Donc, pour offrir un accès pertinent au contenu textuel des abrégés descriptifs de brevet, nous avons choisi de mettre en place un outil d'extraction d'informations.

2.3. L'extraction d'informations

Les outils d'extraction terminologique ou d'acquisition de connaissances sont les principaux outils qui permettent l'extraction d'informations. Par exemple, Coatis [Garcia 98] est un outil d'aide à l'élaboration d'une terminologie de l'action liée à un domaine, qui utilise la stratégie de l'exploration contextuelle. Papins [Pugeault 95] est un outil d'extraction de connaissances à partir de textes, basé sur des descriptions de la sémantique lexicale. Reader [Delisle 96] est un outil de création semi-automatique de lexique pour l'acquisition de connaissances lexicales et conceptuelles. Ces trois outils, qui permettent l'extraction de connaissances à partir de textes plutôt techniques, visent cependant à être utilisés par des terminologues ou ingénieurs de la connaissance. Or notre objectif est de réaliser un outil qui soit utilisable directement par un veilleur.

3. Nouvelle approche : exploitation de notions générales intéressantes

Nous avons choisi les spécificités suivantes pour caractériser notre logiciel d'extraction d'informations : pas d'utilisation de lexiques techniques prédéfinis, pas de calcul de fréquence des termes pour évaluer l'intérêt des résultats obtenus, et résultats exploitables par des veilleurs.

Par ailleurs, nous avons observé une démarche de veille, qui a utilisé entre autre un corpus d'environ 2000 abrégés descriptifs de brevets en anglais sur le domaine des plantes transgéniques. Ces sources d'informations textuelles ont été difficiles à exploiter. Le problème que l'on cherche à résoudre est de faciliter l'accès aux informations pertinentes pour un veilleur dans une telle base documentaire.

Pour obtenir les résultats suivants, nous avons étudié un premier corpus de 30 documents de type abrégé descriptif de brevets en anglais sur le thème des plantes transgéniques [Goujon 99].

3.1. Identification des notions

En observant quelques résumés de brevets, nous avons identifié des notions qui ne sont pas liées au domaine, comme le /changement/, l'/utilisation/, l'/amélioration/, et qui reviennent fréquemment dans les textes. Cela est en fait logique, car la description d'une innovation doit mettre en valeur ce qui est nouveau ou amélioré ou différent.

Ces notions présentent donc plusieurs intérêts dans une démarche de veille : elles sont fréquentes dans ce type de source d'information (parmi de nombreux termes techniques), elles apportent une information intéressante sur ce qui est décrit dans chaque document, et elles sont utilisables quel que soit le sujet traité.

3.2. L'extraction d'informations à partir de notions générales

Nous manipulons actuellement onze notions, organisées en deux ensembles : d'une part un ensemble cohérent de notions exprimant un /changement/ (avec la notion générale /changement/, et quatre sous-notions /amélioration/, /détérioration/, /augmentation/ et /diminution/), d'autre part un ensemble de notions diverses introduisant un résultat (/production/, /résistance/, /utilisation/, /contrôle/, /identification/, /effet causé/). Ces notions sont associées à des ensembles d'indicateurs linguistiques en anglais (*transform, increase, application, resistance*, ...) qui, associés à des indices (prépositions, ...) vont permettre d'identifier les occurrences de ces notions dans les textes.

L'extrait final que l'on veut obtenir doit contenir au moins l'indicateur de la notion, et un complément d'information. Voici la description du complément d'information principal selon quelques notions :

• /changement/ (et notions dérivées) : expression de ce qui subit le changement. Par exemple, l'extrait suivant : " *Alfalfa protoplasts transformed* " contient un indicateur de /changement/ " *transformed* " et un complément d'information principal " *Alfalfa protoplasts* ".

• /utilisation/ : expression de l'application, du résultat de l'utilisation. Par exemple, " *used to modify the sensitivity of a plant to light* " contient le complément d'information principal " *modify the sensitivity of a plant to light* ".

• /résistance/ : expression de l'élément qui a été contré par une résistance. Dans l'extrait " *increase their resistance to pathogens* ", le complément d'information principal est " *pathogens* ".

3.3. Intérêt des extraits pour le veilleur

Cette approche, qui n'est pas basée sur une terminologie liée au domaine, permet l'extraction d'informations diverses, qui ne sont peut-être pas toutes pertinentes, mais qui vont au moins informer le veilleur sur le contenu réel des textes (les extraits sont présentés sans formalisme ni organisation complexe), et peut être mettre en valeur des utilisations étonnantes ou des modifications très prometteuses.

Ce type d'information ne peut pas être obtenu avec un moteur de recherche, car un tel logiciel est surtout utilisé pour rechercher des informations précises, et pas des notions. De plus, le résultat est généralement le nombre de documents correspondant à la requête, ce qui oblige l'utilisateur à consulter plusieurs documents.

3.4. L'exploration contextuelle

Nous avons utilisé dans notre approche la méthode d'exploration contextuelle, mise au point par Jean-Pierre Desclés et Jean-Luc Minel [Desclés *et al.* 94], qui a déjà été utilisée avec succès pour d'autres tâches (résumé automatique avec Seraphin [Berri 96], repérage d'actions dans les textes avec Coatis [Garcia 98], ...) sur le français. Voici un exemple de règle formalisée associée à la notion de /résistance/ :

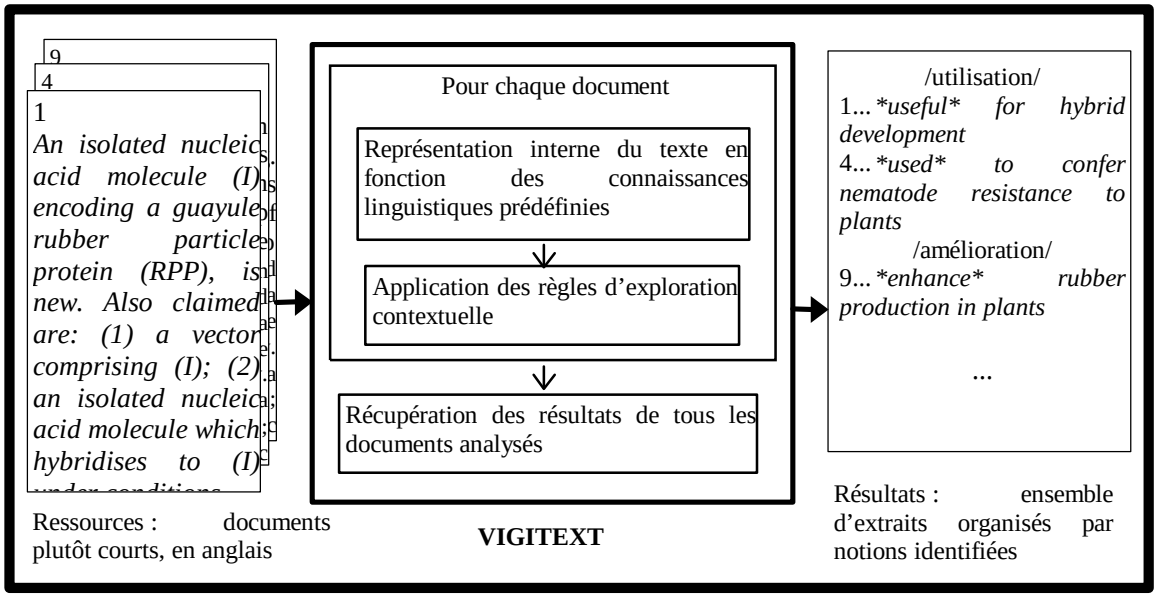
Conditions := ∃ declencheur ∈ {*resistant*, *tolerant*}, ∃ indice ∈ {*to*, *against*} / distanceEnMots(declencheur,indice) = 1 et Position(declencheur) < Position(indice)
Actions := CreerCIPApres(indice), CreerExtraitAvantApres(declencheur).

Cette règle, appliquée au texte suivant : " *... The DNA can be used to produce transgenic plants more resistant to plant pests. ...* ", permet d'identifier l'extrait : " *transgenic plants more *resistant* to plant pests* " avec le complément d'information principal " *plant pests* ". L'extrait obtenu est délimité à gauche par le verbe " *produce* " (verbe associé à la notion de /production/) et à droite par la ponctuation.

L'identification du complément d'information principal permet de ne pas tenir compte de cas tels que : " *... due to modification.* ", où il est question de /changement/, mais sans aucun détail supplémentaire.

4. Le prototype opérationnel VIGITEXT

4.1. Architecture générale de VIGITEXT



Comme le montre ce schéma, l'analyse d'un document consiste à construire dans un premier temps une représentation interne du texte en fonction des connaissances linguistiques, puis à appliquer dans un deuxième temps les règles d'exploration contextuelle. Chaque contexte identifié entraîne la création d'un résultat (extrait avec notion exprimée et référence du document). Tous les résultats sont regroupés par notion sur l'interface utilisateur.

Vigitext est écrit avec le langage de programmation orienté objet Java. Il analyse des documents définis dans une base de données Access, et permet à un utilisateur de naviguer dans le corpus par l'intermédiaire d'une interface de visualisation des extraits identifiés. Un module de regroupement des extraits permet à l'utilisateur d'organiser les résultats selon ses propres connaissances.

Actuellement ce système exploite environ 68 indicateurs (soit près de 200 formes) liés aux notions (principalement des verbes, mais aussi quelques noms et adjectifs), et une centaine d'indices linguistiques (prépositions, articles, formes de l'auxiliaire être, etc.)

Nous avons défini environ 80 contextes qui sont reconnus à l'aide des règles d'exploration contextuelle.

4.2. Exemples d'extraits obtenus

Voici, pour illustrer, quelques extraits obtenus, associés à la notion de /diminution/, précédés des numéros des documents sources :

4	*The amount of CAB protein (and hence chlorophyll content) is *reduced**
8	**decreased* degree of branching of amylopectin starch*
10	**reduce* the time for germination of seeds*
23	**reduces* the quantum efficiency of photosystem II (PSII)*
23	**decrease* of zeaxanthin and antheraxanthin levels*

À l'affichage, les extraits correspondant à une notion sélectionnée ne sont pas ordonnés par pertinence, car nous n'avons pas encore identifié de moyens permettant d'attribuer à un extrait un niveau de pertinence. Nous pensons que seul le veilleur peut évaluer la pertinence d'un extrait.

4.3. Combinaison possible avec un outil d'infométrie

Un outil infométrique peut dans un premier temps être utilisé pour repérer les principaux déposants et les thèmes généraux contenus dans le corpus. En parallèle, VIGITEXT est utilisé pour permettre au veilleur d'accéder aux descriptions contenues dans les brevets, qui sont organisées selon les notions indépendantes du domaine.

Par ailleurs, le veilleur peut regrouper des extraits qui lui semblent concerner un même sujet. L'application d'un outil d'infométrie sur l'ensemble des documents correspondant lui permet alors d'obtenir les principaux déposants travaillant sur ce sujet, ou de voir l'évolution dans le temps des travaux concernés.

5. Conclusion et perspectives

Une évaluation sur 30 documents nouveaux de type abrégé descriptifs de brevets a permis d'obtenir 131 extraits. Le taux d'extraction pertinente, qui correspond à la moyenne entre le taux de rappel et le taux de précision, est de 0,85. Les mauvais résultats correspondent à des extraits mal délimités, comme : *" e.g. caterpillars and is *used* to control "*. Un essai sur un corpus de 10 résumés d'articles concernant la santé a donné 27 extraits, dont : *" *reduced* risk of fractures "*, *" *production* of odontoblasts "*, *" *treatment* of osteoporosis "*.

Vigitext est donc adapté pour l'analyse de résumés de brevets ou d'articles en anglais sur n'importe quel domaine.

Du point de vue linguistique, il nous reste à affiner les lexiques et les règles associés aux notions prédéfinies.

Du point de vue informatique, il reste à prendre en compte les remarques de quelques veilleurs pour l'optimisation des interfaces.

Références

BERRI J. (1996), *Contribution à la méthode d'exploration contextuelle. Application au résumé automatique et aux représentations temporelles. Réalisation du système SERAPHIN*, Thèse de Doctorat, Université de Paris-Sorbonne, Paris.

Dataview : http ://crrm.univ-mrs.fr/commercial/software/dataview/dataview.html

DELISLE S. (1996), *Le traitement automatique du langage naturel au service de l'ingénieur de la connaissance : le système Reader*, in TAL+AI 96, pp.60-66.

DESCLES J.-P., MINEL J.-L. (1994), " L'exploration contextuelle " in *Le résumé par exploration contextuelle*, rapport interne du CAMS n°95/1, recueil des communications effectuées aux rencontres Cognisciences-Est, 25 novembre 1994, Nancy, pp.3-17.

DKAKI T., DOUSSET B., MOTHE J. (1997), *Mining information in order to extract hidden and strategic information*, RIAO 97, PP.32-51.

GARCIA D. (1998), *Analyse automatique des textes pour l'organisation causale des actions. Réalisation du système informatique Coatis*, Thèse de Doctorat, Université Paris-Sorbonne, Paris.

GOUJON B. (1999), *Extraction d'informations techniques pour la veille par l'exploitation de notions indépendantes d'un domaine*, TIA'99, in Terminologies Nouvelles.

JAKOBIAK F. (1995), *L'information scientifique et technique*, PUF, n°3015, p.8.

PUGEAULT F. (1995), *Extraction dans les textes de connaissances structurées : une méthode fondée sur la sémantique lexicale linguistique*, Thèse de Doctorat, Université Paul Sabatier, Toulouse.

ROUACH D. (1996), *La veille technologique et l'intelligence économique*, PUF, p.37.

ROUSSEAU-HANS F. (1998), *L'analyse de corpus d'information comme support de la veille stratégique*, in Document numérique Vol.2, n°2/1998, p.189.

Tétralogie : http ://atlas.irit.fr