Introduction au text mining pour la veille et l'intelligence économique

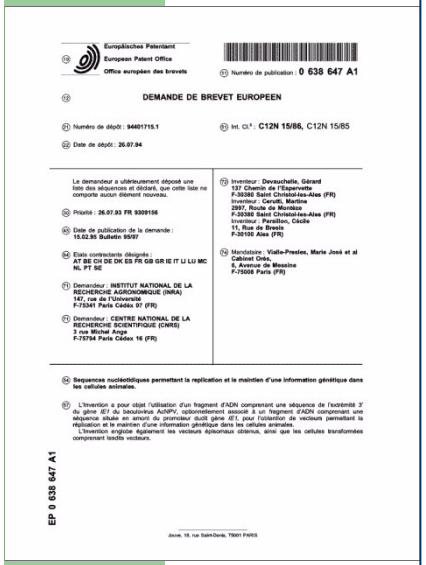
Luc Grivel
Maître de Conférences
Université Paris 1

"Veille et text mining", Rabat, 24-25 mai 2007
Séminaire organisé par ***le Centre National de la Documentation du Maroc***

Plan

- Introduction : processus de veille et text mining
- L'extraction d'information
- La catégorisation (classification supervisée)
- Le clustering (classification non supervisée)
- Conclusion : applications du text mining à l'intelligence économique

Les besoins



Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Numéro de publication : **0 638 647 A1**

DEMANDE DE BREVET EUROPEEN

(21) Numéro de dépôt : 94401715.1
(51) Int. Cl.[6] : C12N 15/86, C12N 15/85
(22) Date de dépôt : 26.07.94

Le demandeur a ultérieurement déposé une liste des séquences et déclaré, que cette liste ne comporte aucun élément nouveau.

(30) Priorité : 26.07.93 FR 9309156

(43) Date de publication de la demande :
15.02.95 Bulletin 95/07

(84) Etats contractants désignés :
AT BE CH DE DK ES FR GB GR IE IT LI LU MC NL PT SE

(71) Demandeur : INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE (INRA)
147, rue de l'Université
F-75341 Paris Cédex 07 (FR)

(71) Demandeur : CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE (CNRS)
3 rue Michel Ange
F-75794 Paris Cédex 16 (FR)

(72) Inventeur : Devauchelle, Gérard
137 Chemin de l'Espervette
F-30380 Saint Christol-lès-Alès (FR)
Inventeur : Cerutti, Martine
Inventeur : Blanc, Stéphane
F-30380 Saint Christol-lès-Alès (FR)
Inventeur : Peralbo, Cécile
11, Rue de Bresis
F-30100 Alès (FR)

(74) Mandataire : Vialle-Presles, Marie José et al
Cabinet Orès,
6, Avenue de Messine
F-75008 Paris (FR)

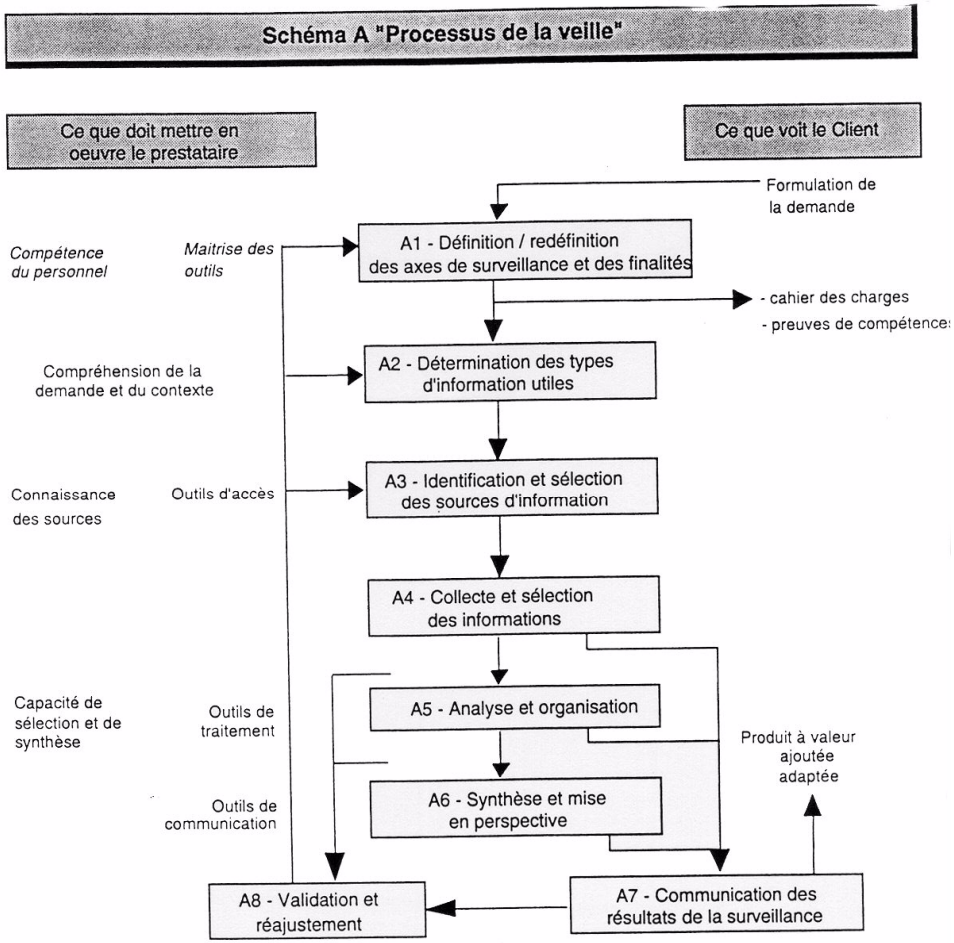
(54) Séquences nucléotidiques permettant la replication et le maintien d'une information génétique dans les cellules animales.

(57) L'invention a pour objet l'utilisation d'un fragment d'ADN comprenant une séquence de l'extrémité 3' du gène IE1 du baculovirus AcNPV, optionnellement associé à un fragment d'ADN comprenant une séquence située en amont du promoteur dudit gène IE1, pour l'obtention de vecteurs permettant la réplication et le maintien d'une information génétique dans les cellules animales.
L'invention englobe également les vecteurs épisomaux obtenus, ainsi que les cellules transformées comprenant lesdits vecteurs.

EP 0 638 647 A1

Afin de limiter les risques d'accident industriel et de pollution, les usines chimiques et les raffineries situées sur le passage du cyclone Rita ont été fermées dès le **22 septembre**.

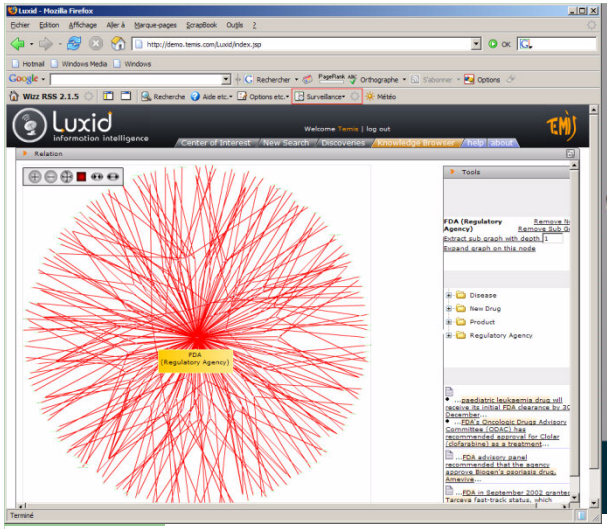
Les unes après les autres, les grandes firmes - **Exxon Mobil, Shell, Dow Chemical, BASF**, etc. - ont **annoncé** qu'**elles arrêtaient** leur **production.**

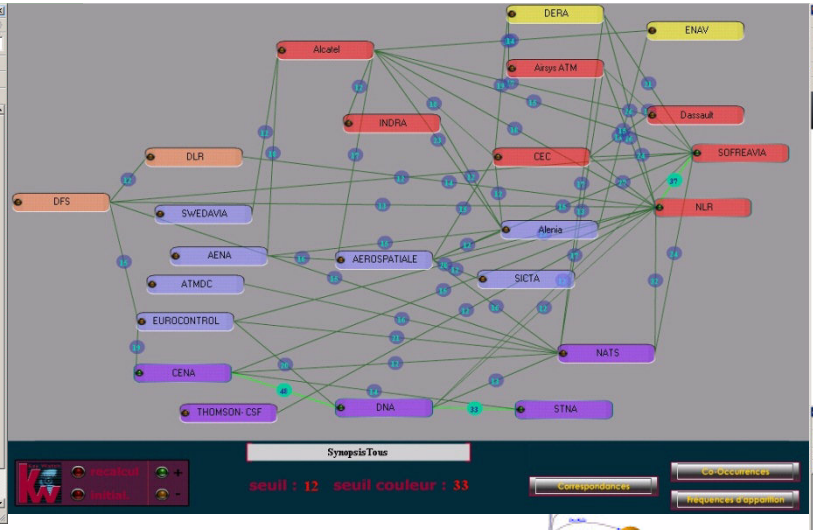


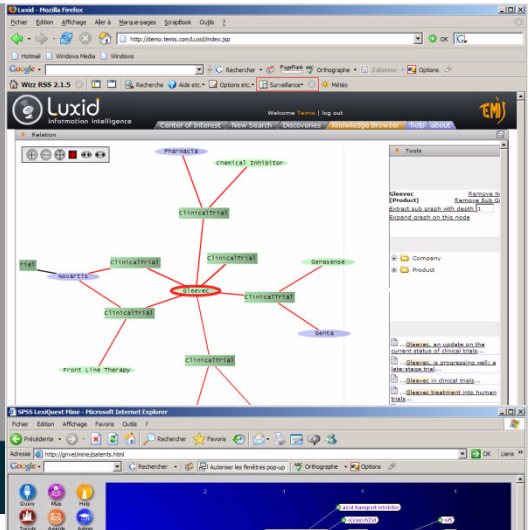
Nature hétérogène de l'information textuelle

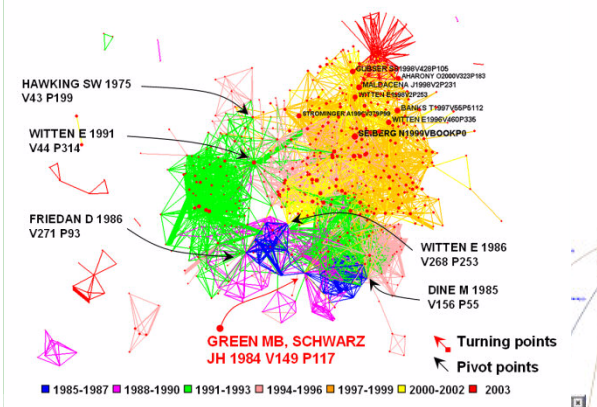
(sémantique /multilinguisme /structure / formats / granularité de l'information

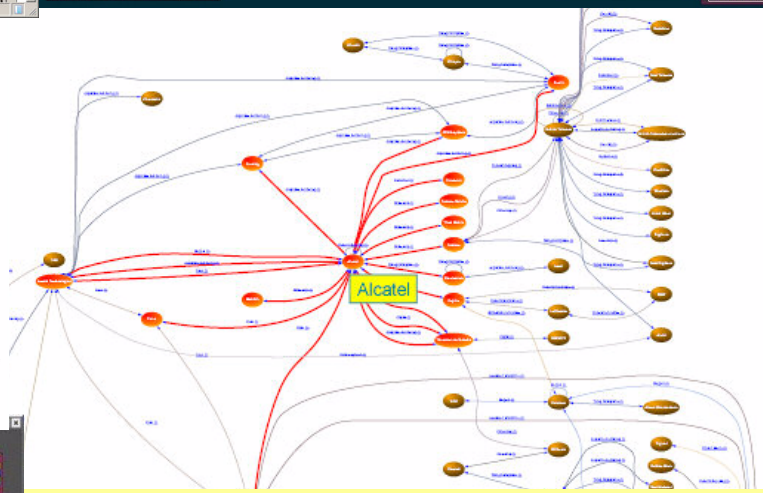
Trop de temps consacré à la collecte de l'information ; pas assez à l'analyse, l'interprétation, la mise en perspective, au partage de l'information

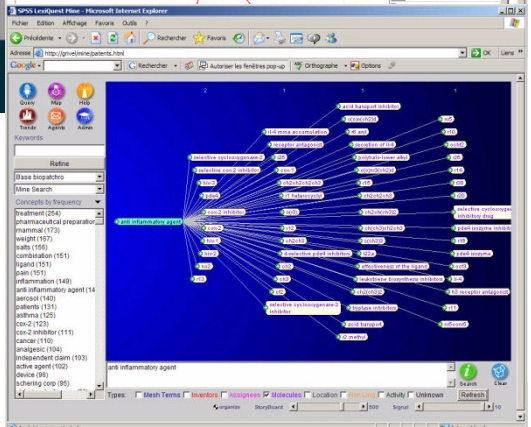


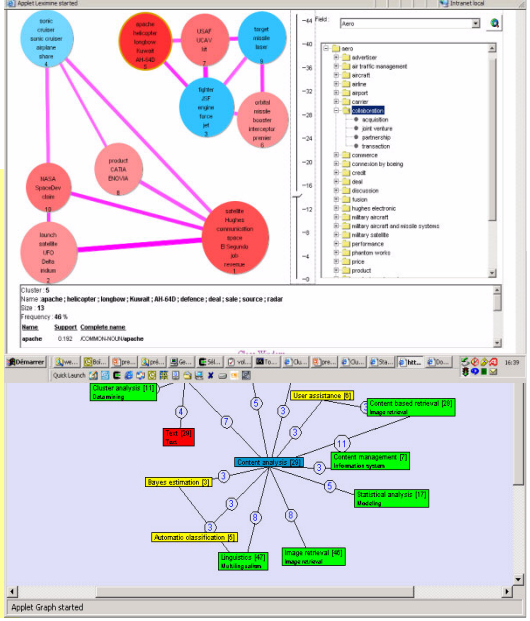


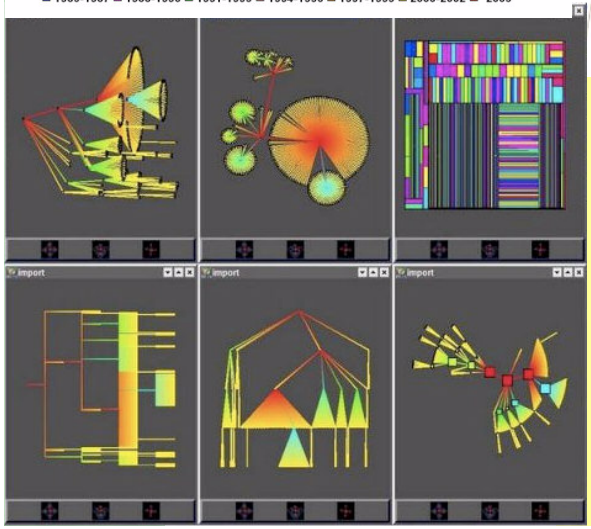












Text Mining : Techniques qui, sous des formes variées, mettent en évidence les éléments jugés importants dans un texte ou dans un corpus de documents

L'objectif est d'organiser l'information pour faciliter la lecture et la compréhension de l'information entrant et sortant de l'entreprise.

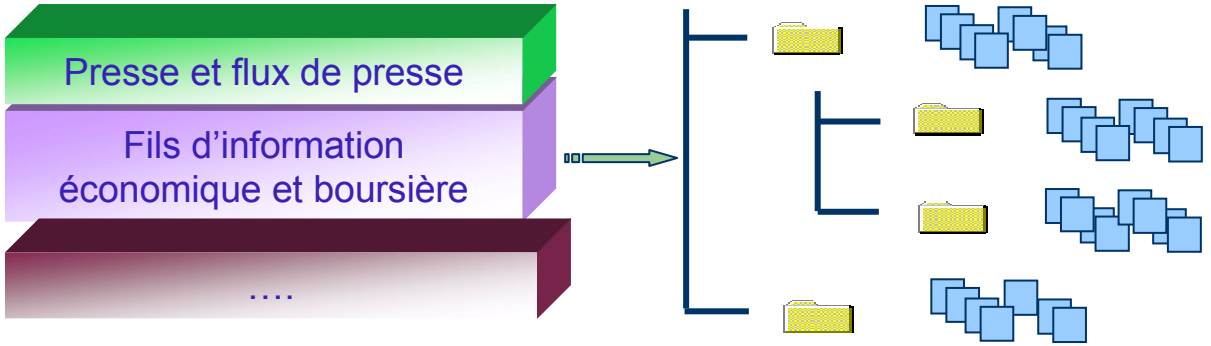
Comment (mieux) exploiter l'information collectée

- Pouvoir organiser les documents comportant l'information qui nous intéresse de manière à pouvoir appréhender globalement l'ensemble documentaire, s'orienter et circuler dans cet espace d'information en suivant des liens pertinents (mettre en relation des concepts, des auteurs, …)
- Accéder plus rapidement au contenu et comprendre le sens de l'information
- Pouvoir mettre en forme et partager le résultat de l'analyse du corpus

Méthodes

- Classer, catégoriser : la classification supervisée
  - Extraction d'information et analyse de texte
  - Apprentissage numérique
- Regrouper, découvrir : le clustering (classification non supervisée)

Catégorisation automatique



Applications :

- classer l'information dans différentes catégories correspondant à des actions ou faits relatifs aux stratégies d'entreprises concurrentes (prise de participation, fusion, acquisition, ouverture de filiales, création de joint venture), commerciales (parts de marché, nombre de clients, nouveau client), etc.
- Routage de brevets vers les examinateurs
- Router des mails

Approches

- Filtrage basé sur un profil de mots-clés (requête)
  - Profil trop précis : beaucoup de documents pas classés
  - Profil trop large : beaucoup de documents mal classés
- Extraction d'information selon une hiérarchie de concepts
- Catégorisation de documents par apprentissage nuén rique

deux types de situations bien distinctes

1. Les catégories ne sont pas trop nombreuses et correspondent à des classes sémantiques ou à une hiérarchie de classe sémantiques (exemples : noms de sociétés, fonctions dans une entreprise, stratégies d'entreprise, …); → la façon d'organiser l'information découlera directement des arbres de concepts qui seront définis
2. Un plan de classement existe, relativement stable et pérenne dans le temps. Il est difficile de définir les concepts qui seront pertinents pour le classement mais il est possible de fournir des exemples de documents représentatifs de chaque classe.

Extraction d'information pour l'IE

- Qui sont les acteurs du secteur d'activité étudié ?
- Quelles sont les actions de ces acteurs, sur quels objets, comment , où et quand ?

Les informations que l'on cherche à repérer peuvent se trouver sous des formes ou des langues diverses.

« Avec une participation de 29.9%dans un important champ pétrolier (Buzzard Oilfield) situé en Mer du Nord, Petrocanada…. »

« Petrocanada has acquired a 29.9% of interest in the Buzzard North Sea oilfield in 2003 »

Prise de participation : acquisition/acquire
Acquéreur : Compagnie/Petrocanada
Cible : Champ pétrolier/Buzzard Oilfield **Montant de part : 29.9%**
Date : **Année/2003**
…

Analyse linguistique d'une phrase

XXL Corp. avait racheté XXS Inc. en 2000

XXL Corp. :	?	+**NOMPROP**
avait :	avoir	**+AUX_I3sg**
racheté :	racheter	***+VPAP***
XXS Inc.:	XXS	***+NOMPROP***
en:	en	***+PREP***
2000:	2000	***+CARD***

1. La segmentation du texte et la détection de langue
2. L'analyse morphosyntaxique
   1. Lemmatisation ou racinisation
   2. Affectation d'une catégorie syntaxique
   3. désambiguisation

Analyse sémantique

- Du mot : sélection du sens des mots individuels (avocat fruit ou avocat juriste) (via un thesaurus ou des règles d'extraction)
- De la phrase : identification des arguments de chaque prédicat et leur rôle sémantique (achat de X par Y)

XXL Corp.	avait racheté	XXX Inc.	en 2000

Cible	**Action**	**Acheteur**	**Date**
Société	*Acquisition*	*Société*	*Année*

Construction de patrons d'extraction

Une tâche guidée par l'objectif à atteindre :

- **"of 1,494 million euros", "for 1,494 million €"**
  - Utiliser les étiquettes (lexicales, syntaxiques et la structure de la phrase (y compris forme passive/active, négations) pour identifier des rôles
  - Utiliser les expressions régulières pour couvrir le + de cas possibles

```
<concept name= "revenue_expression">

   (#ADJ|#ADV)* / ~revenue / #PREP / (#NUMBER)+ / ~Money
 | (#NUMBER)+ / ~Money / #PREP / (#ADJ|#ADV)* / ~revenue
 | (#ADJ|#ADV)* / ~revenue / (#AUX) / #PREP / (#NUMBER)+ /
   ~Money
 | ...

</concept>
```

Analyse de la structure thématique d'un corpus

- **Définition** : classification dans des classes (catégories) pré-existantes.
- **Objectif :** associer une liste de catégories à chaque document
- **Données** :
  - On dispose d'exemples de documents pertinents pour chaque classe ou catégorie. Ils constituent un **ensemble ou une base d'apprentissage.**
  - **Exemples : classification de dépêches de presse, de rapports, de brevets, de mails…**

- **Techniques :** : arbres de décision, régression linéaire, machines à support vectoriel, réseaux de neurones, algorithmes génétiques, modèle de Markov caché, K plus proches voisins, Rochio, …

Résultat de votre recherche...

Recherche texte libre suivant : "fond de pension" dans l'article
Entre le 11/03/02 et 11/03/03

Affinez votre recherche

Concepts liés
- fonds de pension
- Calpers
- durée de cotisation
- systèmes de retraite
- prestations définies
- fonds de retraite
- régimes de retraite
- taux de remplacement
- épargne retraite
- futurs retraités

Hommes
- 9% FRANÇOIS FILLON
- 8% JEAN-PIERRE RAFFARIN
- 4% ALAIN JUPPÉ
- 3% JEAN-LUC CAZETTES
- 3% FRANCIS MER
- 3% XAVIER BERTRAND

Sources
- 21% La Tribune
- 20% Les Echos
- 15% L'AGEFI
- 13% Le Figaro
- 11% Le Monde
- 11% Le Figaro-Eco

Résultats de votre recherche

Cliquez-ici pour effectuer une nouvelle recherche

Votre recherche a abouti à plus de 300 réponses, classées par date de parution

71 % - Le fonds de pension de GE s'évapore
paru le 11/03/03 - (2059 signes)

63 % - L'immobilier professionnel attire de nouveau les investisseurs
paru le 11/03/03 - (7307 signes)

61 % - Repères
paru le 11/03/03 - (3138 signes)

55 % - La Bourse de Tokyo au plus bas depuis 1983
paru le 11/03/03 - (3653 signes)

54 % - Un plus-bas de vingt ans à Tokyo
paru le 11/03/03 - (1830 signes)

38 % - Les fédérations de fonctionnaires appellent à une mobilisation nationale sur les retraites
paru le 11/03/03 - (3165 signes)

38 % - Les vétérans du Golfe s'inquiètent du sort réservé aux soldats américains
paru le 11/03/03 - (6718 signes)

59 % - Les rescapés de la Net-économie
paru le 10/03/03 - (11935 signes)

35 % - Gravelines a le masque
paru le 10/03/03 - (6209 signes)

66 % - L'épargne européenne en patchwork
paru le 08/03/03 - (3991 signes)

56 % - Le Nikkei au plus bas depuis 20 ans
paru le 08/03/03 - (4955 signes)

40 % - Retraites : l'UMP commence sa campagne de terrain
paru le 08/03/03 - (4527 signes)

39 % - Le songe romain de l'art français
paru le 08/03/03 - (4870 signes)

Affinez votre recherche

Sociétés
- 4% Calpers
- 2% Electricité de France
- 2% Standard & Poors
- 2% Securities and Exchange Commission
- 2% Medef
- 2% Banco Bilbao Vizcaya Argentaria

Themes
- 15% Social
- 10% Vie politique
- 9% Retraite
- 7% Conjoncture
- 6% Marchés action
- 6% Analyse

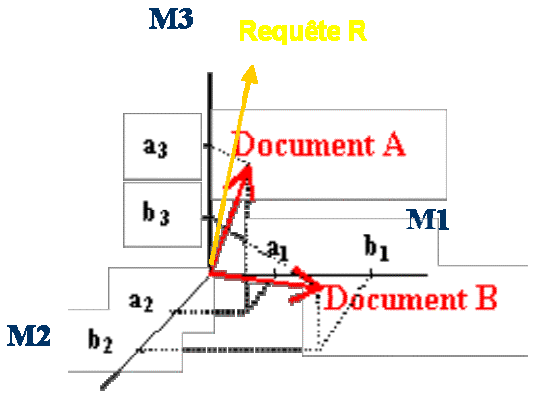
Secteurs
- 11% Banque
- 9% Retraite
- 7% Institutions
- 6% Gestion d'actifs
- 4% Capital investissement
- 4% Énergie - environnement - utilities

K plus proches voisins

Un document devrait être classé dans la même classe que ces K-plus proches voisins dans le corpus d'apprentissage.

- calcul de la similarité entre le document et les exemples du corpus
- les K éléments les plus proches sont sélectionnés et le document est assigné à la classe majoritaire

Calcul de similarité



	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	…
Mot1	2	0	0	1	0	
Mot2	0	3	1	1	0	
Mot3	0	1	0	0	3	
Mot4	0	2	2	1	0	
Mot5	2	0	1	0	0	
…						

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	…
Mot1	0.33	0	0	0.17	0	
Mot2	0	0.36	0.12	0.12	0	
Mot3	0	0.17	0	0	0.51	
Mot4	0	0.02	0.04	0.02	0	
Mot5	0.5	0	0.3	0	0	
…						

- Normalisation
  - Des lignes : pondération par le Tf*Idf
  - Des colonnes par la norme du vecteur pour pouvoir comparer les documents longs et les courts
- Similarité= $Cos(\vec{A},\vec{B}) = \frac{\vec{A}\bullet\vec{B}}{\|\vec{A}\|\|\vec{B}\|} = \frac{\sum_{k=1}^{m} a_k b_k}{\sqrt{\sum_{k=1}^{m}(a_k)^2}\sqrt{\sum_{k=1}^{m}(b_k)^2}}$

	Doc 1	Doc 2	Doc 3	…
Doc 1	1	0	0.15	
Doc 2	0	1	0.33	
Doc3	0.15	0.33	1	

Evaluation du modèle d apprentissage

Catégorie : passenger compartment

L'évaluation de la qualité du modèle construit s pas participé à l'apprentissage

Corpus

- **Fichier:** C:\dev\PSA\ADEC\run\work\tmp33916\extraction\ART_THA.tmx

Statistiques

- **Precision:** 76.41%
- **Recall:** 76.41%
- **Quality:** 76.41%

a c b

Categorie	Precision	Recall	Corrects	Missed	False	Tests
Noise	89.06	91.93	57	5	7	62
Condensation	25.0	66.66	2	1	6	3
Visibility	91.66	84.61	44	8	4	52
Smells	70.21	94.28	33	2	14	35
Air movement	60.0	32.14	9	19	6	28
Slow to heat up	79.31	85.18	46	8	12	54
Slow to cool	35.0	70.0	7	3	13	10
Breakdown	90.0	60.0	18	12	2	30
Regulation	66.66	51.85	14	13	7	27

N.A. means *Non Applicable*

Terminé

Poste de travail

précision = proportion de documents pertinents parmi les documents affectés à la catégorie par le sysê me

Rappel = proportion de documents pertinents trouvés dans l'ensemble des documents attendus pour la catégorie

a nombre de documents corrects dans la catégorie

b nombre de documents mal classés dans la catégorie

c nombre de documents manquants dans la catégorie

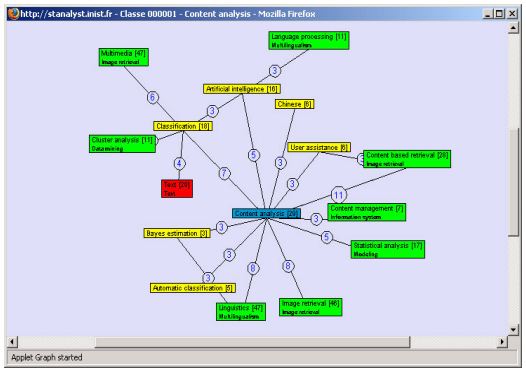
$P = a / (a + b)$

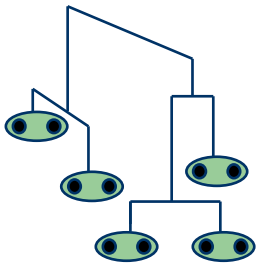
$R = a / (a + c)$

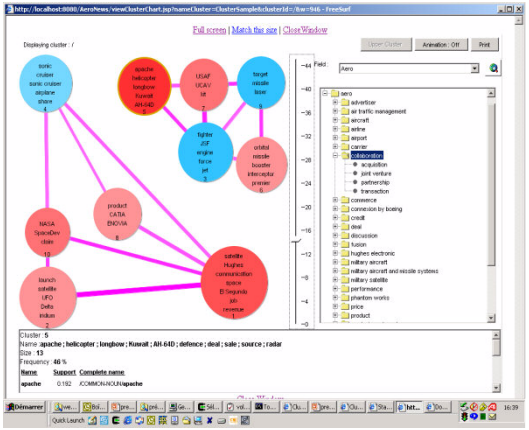
La Classification non supervisée ou clustering

Principe : comparer les données pour regrouper les plus similaires, chaque groupe devant être le plus homogène possible, et les groupes devant être les plus différents possibles entre eux

- Appliqué aux documents : Ex Kmeans axiales
- Appliqué aux mots-clés : Ex. Mots associés

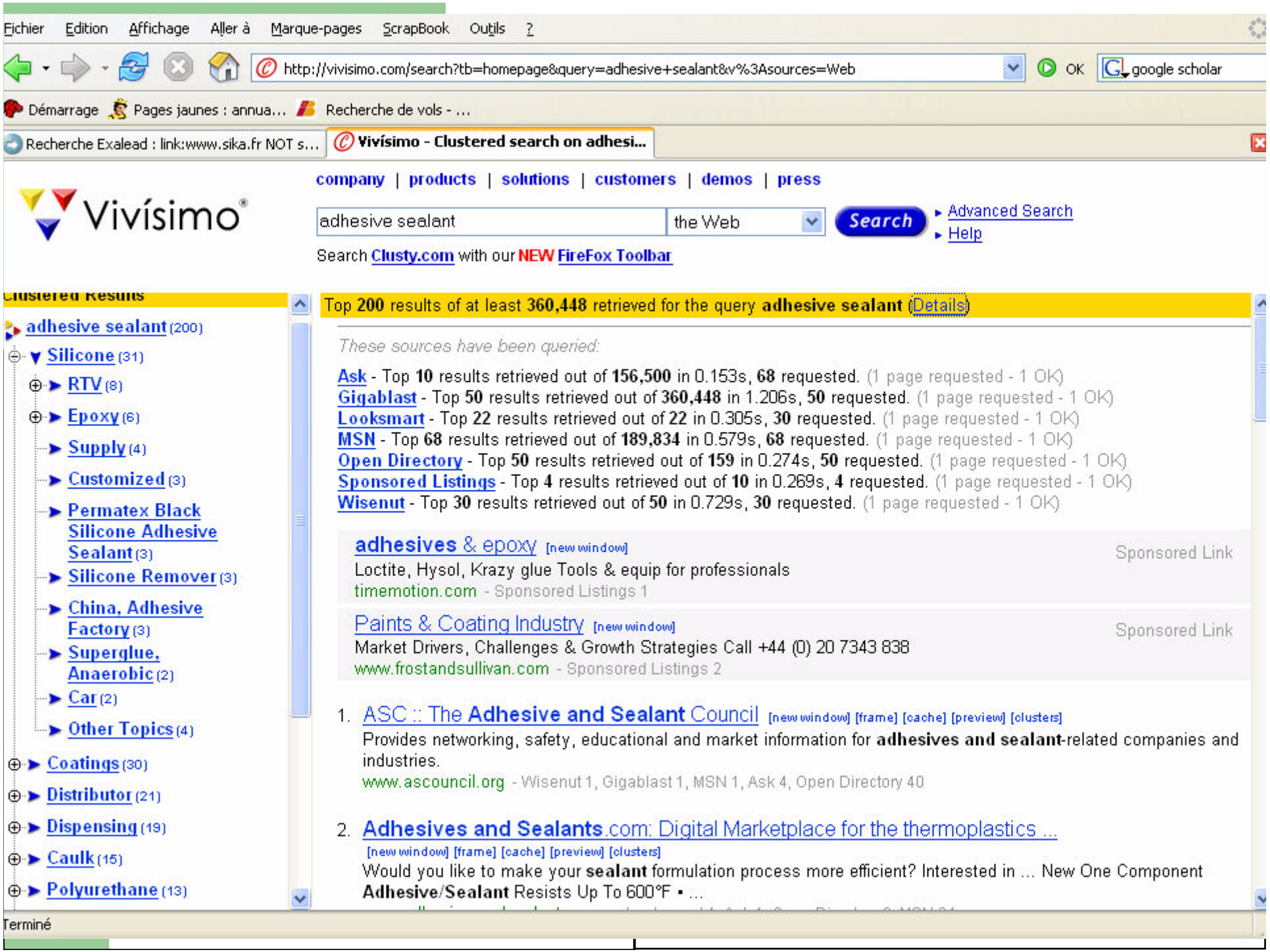






Intérêt du clustering

1) Désambiguiser le résultat d'une recherche (Ex. Trouver des brevets sur '*Antilock Braking Systems' ?* → En spécifiant 'ABS' dans la recherche, on obtient des milliers de réponses
   - Antilocking Braking System
   - AcrylonitrileB utadiene Styrene
   - Air Bearing Surface
2) Avoir une vue d'ensemble sur un sujet et positionner les acteurs (Ex. les thèmes abordées par une revue scientifique, **les équipes de recherche relatives à un thème**, …)
3) Proposer une taxonomie (en amont de la catégorisation)
4) Identifier des doublons (Ex. mesurer l'impact d'une campagne de presse)
5) Se faire une idée de la terminologie d'un domaine


Fichier Edition Affichage Aller à Marque-pages ScrapBook Outils ?
http://vivisimo.com/search?tb=homepage&query=adhesive+sealant&v%3Asources=Web
OK
google scholar
Démarrage
Pages jaunes : annua...
Recherche de vols - ...
Recherche Exalead : link:www.sika.fr NOT s...
Vivísimo - Clustered search on adhesi...
Vivísimo
company | products | solutions | customers | demos | press
adhesive sealant
the Web
Search
Advanced Search
Help
Search Clusty.com with our NEW FireFox Toolbar
Clustered Results
adhesive sealant (200)
Silicone (31)
RTV (8)
Epoxy (6)
Supply (4)
Customized (3)
Permatex Black Silicone Adhesive Sealant (3)
Silicone Remover (3)
China, Adhesive Factory (3)
Superglue, Anaerobic (2)
Car (2)
Other Topics (4)
Coatings (30)
Distributor (21)
Dispensing (19)
Caulk (15)
Polyurethane (13)
Top 200 results of at least 360,448 retrieved for the query adhesive sealant (Details)
These sources have been queried:
Ask - Top 10 results retrieved out of 156,500 in 0.153s, 68 requested. (1 page requested - 1 OK)
Gigablast - Top 50 results retrieved out of 360,448 in 1.206s, 50 requested. (1 page requested - 1 OK)
Looksmart - Top 22 results retrieved out of 22 in 0.305s, 30 requested. (1 page requested - 1 OK)
MSN - Top 68 results retrieved out of 189,834 in 0.579s, 68 requested. (1 page requested - 1 OK)
Open Directory - Top 50 results retrieved out of 159 in 0.274s, 50 requested. (1 page requested - 1 OK)
Sponsored Listings - Top 4 results retrieved out of 10 in 0.269s, 4 requested. (1 page requested - 1 OK)
Wisenut - Top 30 results retrieved out of 50 in 0.729s, 30 requested. (1 page requested - 1 OK)
adhesives & epoxy [new window]
Sponsored Link
Loctite, Hysol, Krazy glue Tools & equip for professionals
timemotion.com - Sponsored Listings 1
Paints & Coating Industry [new window]
Sponsored Link
Market Drivers, Challenges & Growth Strategies Call +44 (0) 20 7343 838
www.frostandsullivan.com - Sponsored Listings 2
1. ASC :: The Adhesive and Sealant Council [new window] [frame] [cache] [preview] [clusters]
Provides networking, safety, educational and market information for adhesives and sealant-related companies and industries.
www.ascouncil.org - Wisenut 1, Gigablast 1, MSN 1, Ask 4, Open Directory 40
2. Adhesives and Sealants.com: Digital Marketplace for the thermoplastics ...
[new window] [frame] [cache] [preview] [clusters]
Would you like to make your sealant formulation process more efficient? Interested in ... New One Component Adhesive/Sealant Resists Up To 600°F • ...
Terminé

Applications du text mining à l'Intelligence économique

- ♣ Identifier les actions ou faits relatifs aux stratégies des entreprises
- ♣+♦ Classer des news, rapports, brevets, etc. selon des rubriques métiers ou selon des profils de veilleurs
- ♣+♥+♠ Identifier les thématiques de recherche sur un domaine, identifier les relations existants entre les acteurs (co-auteurs, citations, co-citations)
- ♣+♥+♠ protéger son image et sa réputation en analysant les propos tenus, les opinions émises sur les médias et notamment Internet et les forums
- ♣ Analyser la portée juridique d'un brevet en extrayant automatiquement les revendications indépendantes, les liens avec d'autres brevets,

…

♣	**Extraction d'information**
♦	**Catégorisation**
♥	**Clustering**
♠	**Bibliométrie**