

Journée d'étude du Jeudi 5 Décembre 2002 à Rouen,
des professionnels de l'information-documentation
ADBS Normandie, AIVP et GIDE

Indexation automatique et langage naturel

Sylvie Dalbin

Assistance & Techniques Documentaires - DESYBEL GIE
SylvieATD@aol.com

Ce document peut être exploité librement.
Merci de citer auteur et source

Objectifs de l'intervention

Se positionner professionnellement et envisager ce type
de solution dans son environnement de travail

Profiter pleinement des exposés des utilisateurs de ces
systèmes et de leur expérience

***Préciser les concepts-clés de l'indexation texte
intégral et de la recherche en langage naturel***

Contexte

- technologies mal connues (texte intégral, pertinence, langage naturel...), mais dont on pratique l'implémentation dans des systèmes de recherche d'information.... depuis plus de 15 ans
- 3 expériences distinctes relatives au cours de la journée :
contexte, technologies, produits

Indexation texte intégral et recherche en langage naturel

Sommaire

- 1 - Pourquoi ?
 - 2 - Principes généraux de l'indexation & la recherche
 - 3 - Traitements statistiques
 - 4 - Traitements linguistiques et sémantiques
 - 5 - Classification automatique
 - 6 - Marché des logiciels d'indexation et de recherche en texte intégral et LN
 - 7 - Problématiques de l'évaluation
- En conclusion : évolution de nos métiers

1. La recherche en langage naturel : pourquoi ?

Ressources numériques et usages

Développement des ressources numériques

- volumes, flux ; accessibles sur les réseaux

Usages multiples, contraintes d'exploitation

- rapidité de mise à disposition, accès direct, information exploitable facilement par une diversité d'utilisateurs

Élargissement des catégories de documents et supports d'information manipulés

- livres, rapports, articles..., mais aussi articles de forum, sites, bases d'information sur les demandes des clients...enfin documents audiovisuels, multimédia

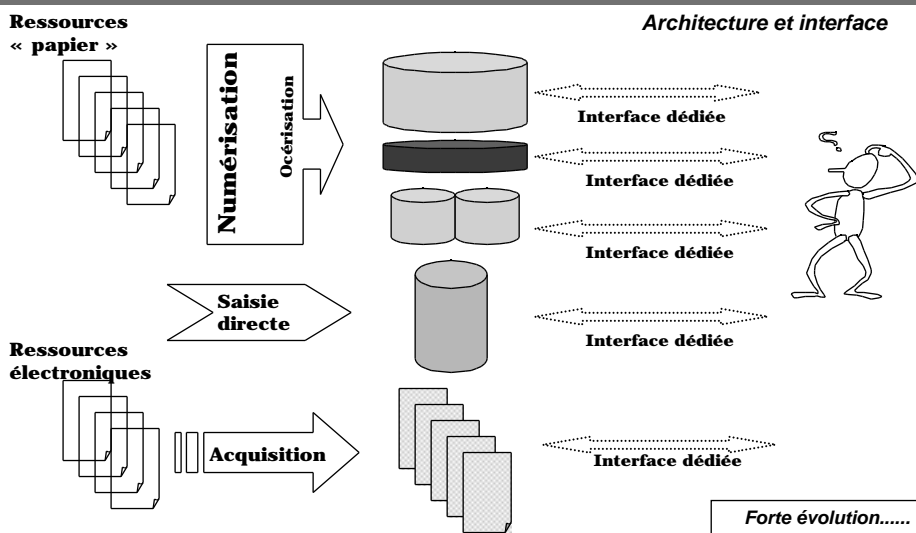
Chaînes de traitement variées en fonction du couple "valeur de l'information"/"coût du traitement"

- par exemple : analyse fine pour une conservation à long terme # diffusion rapide sans traitement

**Usages multiples # ressources diversifiées
traitements diversifiés**

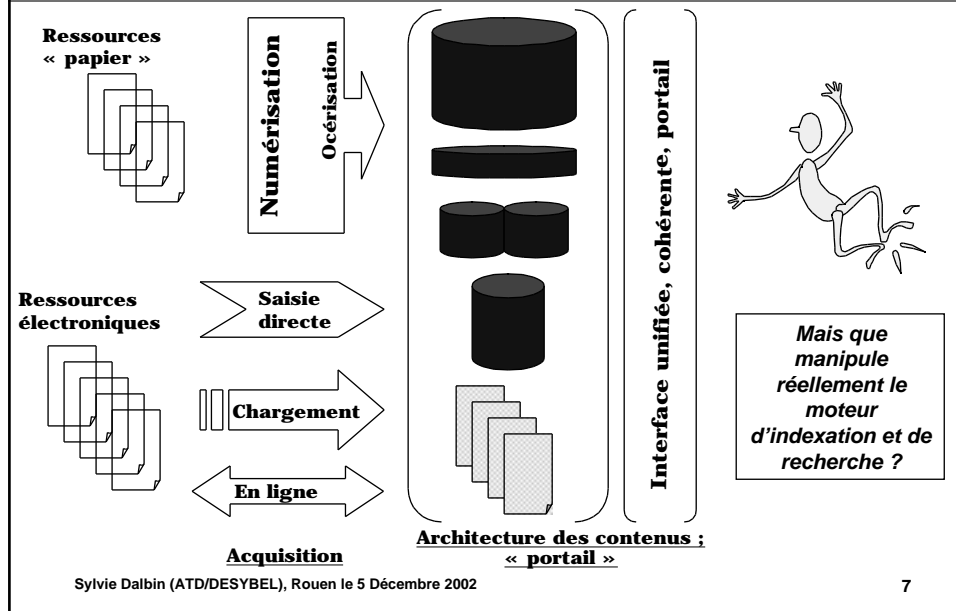
Exploiter le contenu des ressources numériques

Des systèmes de gestion électronique de documents (SGED)...



Acquisition Gestion/stockage / Recherche-Accès

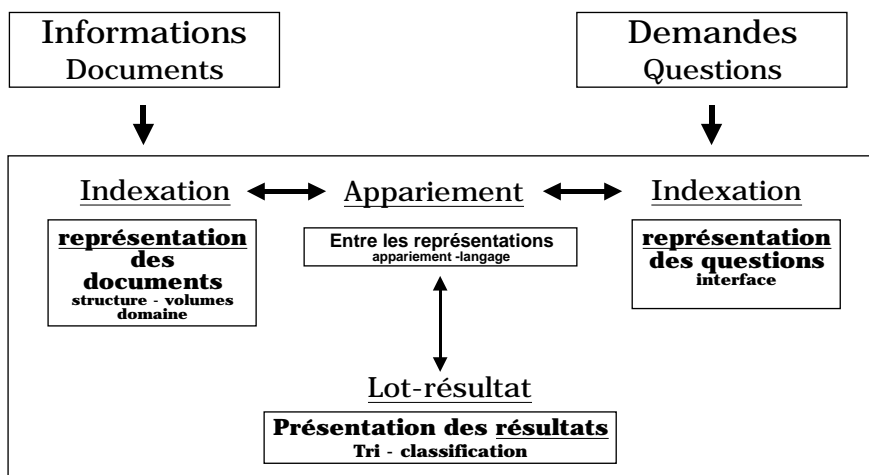
... aux systèmes de gestion et d'accès aux contenus



7

2. La recherche en langage naturel : principes généraux

Schéma fonctionnel d'un système de recherche d'infodoc (SRI)



Sylvie Dalbin (ATD/DESYBEL), Rouen le 5 Décembre 2002

9

Principe de base d'un système de recherche d'information

Indexation

- Etablir une représentation compacte (moins de données, plus de sémantique), significative (relativement au contenu des objets documentaires, aux utilisateurs) et rapides à calculer et à comparer, d'un document ou d'un ensemble de documents ou d'information et des requêtes

Recherche

- Soit appairer les représentations de la requête et celle des ressources. Les représentations des ressources et des questions peuvent être de même nature ou de nature différente (i.e. des outils différents)
- Soit naviguer dans une représentation des ressources informationnelles (arborescence)

Recherche par le contenu

- les clés d'accès aux documents sont obtenues par l'exploitation automatique du contenu de l'objet documentaire
- Valable pour les documents textuels (contenu textuel) ou multimédias (contenu visuel)

Sylvie Dalbin (ATD/DESYBEL), Rouen le 5 Décembre 2002

10

Résultats de l'indexation des index plus ou moins riches en fonction des outils utilisés

Document Texte ou question

l'indexation manuelle ou automatisée s'effectue sur des unités informationnelles diverses : des abstracts, des titres, des textes complets, mais pour l'indexation automatique, toujours sur l'information textuelle

Index

- indexation manuelle
- indexation automatique
- titre
- résumé
- document textuel

"humain"

Index

- abstracts
- automatisée
- complets
- diverses
- indexation
- information
- informationnelles
- manuelle
- s'effectuent
- texte(s)
- text(uelle)
- titre
- unité(s)

*"texte intégral"
brut*

Index

- abstract(s) / résumé(s)
- *document?*
- effectu[er] (*effectuent sur*), *réalis[] par...*
- *index[] manuel[]*
- *index[] automat []*
- *index[] manuel[] et automat []*
- *information*
- *information text []*
- *résumé(s) / abstract(s)*
- *text []*
- *text[] complet(s)*
- *titre(s)*
- *unité(s) / information []*

traitement linguistique brut

La moteur exploite un index dont le contenu est différent

Modes de représentation du contenu d'un document, d'un ensemble de documents, question

Qu'indexe-t-on ? Le contenu textuel :

- des documents et les questions
- des documents seuls
- des documents partiellement

Comment indexe-t-on ?

- humainement, par des mots-clés libres ou pris dans une liste
- automatiquement, par les mots contenu dans le texte
 - ... associés à des traitements statistiques (partie 3)
 - ... associés à des traitements linguistiques et/ou sémantiques (partie 4)
 - ... regroupés par classes/clusters (partie 5)

***La recherche s'effectue sur des index de contenus
et de formes différents***

Possibilité de mixer ces modes

3. Traitements statistiques

Principe général de la notion de « tri par pertinence » (1)

3-Traitements statistiques

L'opération booléenne (tout ou rien)

- est remplacée et/ou complétée par le calcul d'une proximité/distance entre la représentation de la question et celle des textes

La réponse est

- un **ordonnement** des documents
- suivant ce degré de "pertinence" des documents par rapport à la question
 - ce n'est pas strictement un sous-ensemble de la base.

Principe général de la notion de « tri par pertinence » (2)

Cet ordonnancement du lot résultat est possible grâce au :

calcul d'un poids (pondération)

- valeur attribuée aux documents
 - page web, document bureautique sur un intranet,...
- calcul construit à partir de critères essentiellement statistiques
 - occurrence du terme dans le document, proximité et ordre du terme,....; appliqué aux index
- la valeur du poids de chaque document peut être comprise entre une borne inférieure pour un document estimé non pertinent (0 par exemple), et une borne supérieure pour un document estimé tout à fait pertinent.

calcul de proximité entre documents et requête

- degré de similarité= ressemblance= distance

Processus général indexation et traitements statistiques

- Etape 1 :** Indexation des **documents** + calcul complémentaire du **poids** des documents (algorithmes)
>> index enrichi, pondéré
- Etape 2 :** Indexation de la **requête**
>> index
- Etape 3 :** **Appariement** entre les deux représentations (index) requête/documents, modification éventuelle du poids des documents en fonction de la requête
>> calcul de similarité document/requête
>> extraction des « documents pertinents »
- Etape 4 :** Etablissement d'un lot-résultat, en fonction du seuil établi par l'administrateur du système
- Etape 5 :** **Ordonnement** (tri par pertinence = relevance ranking) automatique du lot-résultat, grâce aux pondérations établies dans les étapes précédentes

Pondération relative ou absolue

La valeur du poids attribué aux documents et qui permet l'ordonnement du lot-résultat est :

- soit absolue, c'est-à-dire indépendante de la requête
 - le calcul du poids attribué au document s'effectue au niveau de l'étape 1, en fonction des autres documents du fonds, et n'est pas modifié par la requête.
- soit relative, c'est-à-dire dépendante de la requête
 - le poids attribué au document sera modifié en fonction de la requête (mots et syntaxe de la requête)

Algorithmes de pondération (a)

Valeur absolue (hors requête)

occurrence (fréquence) d'un mot dans le document

occurrence d'un mot dans le document par rapport au nombre de mots du document (densité).

- Un document petit en taille aura une meilleure pondération

occurrence d'un mot dans le document par rapport à son occurrence dans la base (discriminant):

- les mots peu fréquents dans le corpus sont favorisés, les mots "vides " sous-entendus trop fréquents sont soit éliminés soit sous-évalués

localisation d'un mot dans le document (métadonnées, premières lignes du texte, liens...)

typographie du mot dans le document

- un poids plus important peut être donné à un mot en majuscule ou en gras (typographie sur Google) à l'intérieur d'un texte

appartenance d'un mot à une liste contrôlée

Algorithmes de pondération (b)

Valeur absolue (hors requête)

Pondération plus forte pour :

- Des pages de références
 - pages qui sont référencées par d'autres documents, c'est-à-dire qui ont beaucoup d'autres pages/liens qui pointent sur elles
 - technologie Pagerank de Google, appelée " indice de popularité ", calcul qui s'appuie sur le nombre de liens qui pointent sur le document/page
 - Mais pénalise les ressources récentes, non référencées
- Des pages pivots
 - pages contenant de nombreuses références à d'autres documents (nombreux liens sur la page, tels les répertoires de signets)
 - Google : calcul qui s'appuie sur le nombre de liens qui partent du document/page

Autre critères : pondération plus forte pour :

- Pages sélectionnées par des utilisateurs (après lecture) ou simplement cliquées (indice de clic)
- Pages sponsorisées

Mixité des critères

*Difficulté d'identifier les critères mis en oeuvre par les moteurs
Paramétrer en fonction de son contexte les critères, les seuils ?*

Algorithmes de pondération (c) appariement

Valeur relative (par rapport à la requête)

Poids plus important aux documents :

contenant un plus grand nombre de termes de la question

— $A + B + C$; $A + B$; $B + C$; $A \cdot B$; C

dont la proximité (et l'ordre) des termes de la requête se retrouve dans les documents

Pondération possible en fonction d'un poids attribué par l'utilisateur aux termes de sa requête

3. Les traitements linguistiques et la recherche en langage naturel (LN)

A. Recherche d'information et langage naturel : problématiques

B. Ressources linguistiques exploitables par les moteurs d'indexation et de recherche

A - Problématiques

Rappel

- L'indexation manuelle (liste d'autorité, thesaurus...) est effectuée au niveau du concept
- L'indexation « texte intégral », au niveau du mot (en surface)

Problème

- formulations différentes d'une même idée > silence
- ambiguïté : réponses hors sujet > bruit

Solution

- exploiter les techniques automatiques du langage : extraire des mots et des liens sémantiques entre mots
- les outils linguistiques privilégient l'utilisation de dictionnaires, mais pas uniquement

Problèmes liés au langage (a)

Synonymie, totale ou partielle

- Totale : oculiste et ophtalmologiste
- Partielle : logement (terme générique) et maison (terme spécifique); bras, main ou pied (partie de) et corps ; sigle : XML = extended markup language; abréviation :
- Périphrase : Frigidaire et réfrigérateur

Termes complexes et expressions

- Pompe à vélo, pomme de terre ; Garde des Sceaux; bouillon de culture (biologie) ; faire tâche d'huile ; mettre en œuvre

Expressions multiples d'une même idée ou concept

- les "ventes du vin français à l'étranger" = exportations viticoles françaises = ventes françaises à l'étranger dans le secteur du vin
- coût des logiciels de gestion gestion du coût des logiciels logiciel de gestion des coûts (*de l'importante des rôles des « mots vides »?*)
- Fer, alliage ferreux, acier, métaux ferreux, Fe
- Acier anti-corrosion, acier résistant à la corrosion, ...
- cours de maths, cours de mathématiques, enseignement des maths, les maths sont ici enseignées...
- Au = numéro atomique 79 = or

Résolus par des dictionnaires, généralistes ou spécialisés

Problèmes liés au langage (b)

Homographie, homonymie

- avocat (fruit ou droit) ; or (métal ou conjonction) ; bibliothèque (meuble ou bâtiment ou organisme) ; CAP (certificat d'aptitude prof.) cap
- avions (verbe avoir, ou le substantif au pluriel)
Faible statistiquement, mais pouvant avoir un poids informationnel fort
- DSI (2 langues) : Digital speech interpolation diffusion sélective de l'information

Métaphore (effets d'image) et métonymie (glissement de sens)

- la source du Nil - la source d'information -> la source de mon chagrin
- « policier » : appartenant à la police, type de roman
- cours de maths, du dollar

Ellipse

- « entreprises privées et publiques » = entreprise privée, entreprise publique

Paraphrase

- Jean-Paul II = le Saint-Père ; élection du président de la république française; élection présidentielle au suffrage universel; scrutin présidentiel

Anaphore

- « L'Intranet utilisait le robot d'Altavista pour la consultation du Web. Il offrait une sécurité absolue » (qui ? l'intranet ou le robot d'Altavista ?)

Problèmes liés au langage (c)

Dénotation (sens propre) et connotation (sens figuré)

- Sens propre (dénotation) : sens de base d'un mot, stable, analysable hors de son contexte d'usage
- Sens figuré (connotation) pris par un mot dans un contexte particulier
- Synonymies possibles (dénotation=connotation) surtout dans les domaines techniques
- Niveau de langue : chaussure pompe ; cancer carcinome
- Analyses particulières (connotation fréquemment différente de la dénotation) pour les objets audiovisuels, images fixes/animées

Multilinguisme

- Homographies inter-langues : case (emplacement en Fr, cas en En)

Erreurs de frappe, d'orthographe, de grammaire

Résolus par des traitements sémantiques de plus haut niveau

Synthèse : différents niveaux d'ambiguïtés (d)

au niveau du mot pris isolément

- sémantique lexicale

association des mots suivant leur rôle dans la phrase

- sémantique grammaticale

au sein du document ou d'un ensemble de documents

- sémantique contextuelle

au niveau des situations rencontrées dans « la vie »

- "pragmatique"

des requêtes "pauvres"

- reformulation

Nécessité de déployer des traitements linguistiques, sémantiques de haut niveau, mais également statistiques

Quelques exemples de questions posées sur l'intranet des AGF

l'image des assurances en France

la circulation dans les ronds points

Résultats Allianz 2002

Aménagement et réduction du temps de travail, réduction du temps de travail, RTT, ARTT...

VAE, validation des acquis, validation des compétences,...

Des noms propres

- un nom de produit financier, d'assurance
- un nom d'organisme ou de société
- un arrêt Perruche

en quoi la gestion des connaissances peut elle être un facteur de croissance et de développement pour l'entreprise

tempêtes hiver dernier

Chaîne de traitement (a)

1- reconnaître les mots et les normaliser

2- regrouper les termes équivalents sémantiquement

A Extraction des parties textuelles et segmentation du texte en mots

B Ramener les mots à leur forme de base : le lemme

— Traitement des variations morphologiques (flexion, dérivation)

- Exemple : chevaux > chevaux, dérèglent > dérègl(er)
 - Lemmatisation (infinitif pour les verbes,...) pour tenir compte des variations flexionnelles (catégorie grammaticale, genre, nombre)
- Exemple : courageux/courage
 - Analyse des mots dérivés (racinisation=stemming)
- Détection d'erreur, phonétique (type Soundex)

Chaîne de traitement (b)

B Ramener les mots à leur lemme (suite)

—Traitements des variations syntaxiques lexicales

- Exemple : tondeuse à gazon / tondre le gazon / le gazon tondu
 - reconnaissance des locutions et expressions idiomatiques
- Exemple : avocat (fruit) - avocat (acteur, juriste)
 - traitement des homographies par appartenance à leur catégorie grammaticales.
- Reconnaissance des mots composés
- Exemple : Agence de presse, agence soviétique de presse
 - Reconnaissance des expressions contiguës ou disjointes
- Exemple : car (nom ou conjonction)
 - Traitement des mots "vides"

C Traitements syntaxiques

- Déterminer la structure des phrases

D Traitements des variations sémantiques

***Exploitation de ressources terminologiques,
analyseurs et/ou règles***

Désambiguïisation sémantique en exploitant le contexte des mots

Exemples :

- "acheter un chausson à la boulangerie pour le goûter des enfants. La plupart des autres viennoiserie contiennent du chocolat. Au moins les chaussons aux pommes comportent de la compote "
- " acheter une paire de chaussons chez CHAUSSTOUS pour les mettre devant la cheminée ".

Analyse automatique du "contexte"

- "chausson"
 - > chausson__viennoiserie
 - > ou chausson__boulangerie
- dictionnaire : relation existant entre chausson et viennoiserie (équivalent à TG) ou chausson et boulangerie (équivalent à VA)

Exemple : termes non intégrés dans le dictionnaire général Spirit

Exemple tiré de l'index AGF

INTERNET	3290	internet, Internet, INternet, INTERNET
WEB	1215	web, Web, WeB, WEB
TELECOM	541	telecom, télécom, Telecom, Télêcom, Télécom, TELECOM
FIDELISATION	505	fidelisation, fidélisation, Fidelisation, Fidélisation, FIDELISATION
UBS	66	UBS <i>Rajouter : U.B.S./Union des banques suisses</i>
SACHS	63	Sachs
DEMATERIALISATION	60	dematerialisation, dématérialisation, Dématérialisation, DEMATERIALISATION

Remarque : le système a généré automatiquement les autres graphies des termes (accentuation, majuscule/minuscule), mais ne les a pas placés dans leur contexte sémantique (relations).

Les termes dont l'occurrence est élevée peuvent être intégrés dans le dictionnaire général.

B - Outils et ressources linguistiques

Pour opérer ces traitements linguistiques, les moteurs d'indexation et de recherche exploitent des outils spécifiques

des référentiels terminologiques

- Listes de mots "vides" ; anté-dictionnaires (« mots vides »)
- Lexiques ; thésaurus; classification
- Dictionnaires de formes fléchies, ...
- Réseau sémantique, graphe de concepts (reformulation)
 - Ontologie sous forme de graphe de relations lexicales : Worldnet, les travaux de Mémodata (Caen), Topic de Verity, ...
- Base de connaissances

des grammaires ("grammaire linguistique", structure/DTD,...)

des règles (de reformulation, découpage du texte, reformulateur morphologique (racineur), actant/acté (Tropes)...)

- Exemples : les racineurs

Divers autres outils comme les phonétiseurs,...

Dictionnaires

Définition

- "connaissances sur la langue préalablement décrites par un expert humain dans une base de données et utilisées par des automates au moment de l'interprétation du texte à traiter" (Lingway)

Différents types de dictionnaires

- de formes fléchies, de synonymes, d'expressions idiomatiques...
- généraux et/ou spécialisés (privés)

Limites

- pas toujours existantes
- pas toujours complètes
- pas toujours évolutives
- pas portable d'un domaine à l'autre

Exemples de ressources linguistiques

Exemples

- Arisem (KnowledgeBase)
 - Référentiel multilingue (20.000 concepts & 500.000 liens en 5 langues), personnalisable
- Lexiquet
 - 60 000 mots, 500 000 liens, 150 000 concepts sémantiques
 - Lexitrack (outil d'extraction de terminologies) et lexibuild (outil d'administration)
- Spirit
 - plus de 500.000 entrées incorporant les différentes formes fléchies d'un même mot : singulier/pluriel, masculin/féminin, formes conjuguées pour les verbes ;
 - un lexique d'expressions idiomatiques intégrant notamment les sigles (développées), des locutions ("à concurrence de", "à l'issue de"), les mots composés
 - Base de 130.000 règles intégrant notamment les synonymes usuels de la langue française.

Dictionnaire unitermes et locutions (Spirit)

Mot	Mot nominal	Classe	Valeurs Grammaticales	Genre/Nombre	Dérivations
<input checked="" type="checkbox"/> assurance vie	assurance vie	Substantif	S	NS, N'	1

—Intégrer assurance vie

Dictionnaire de reformulation (Spirit)

Mot	Mot nominal	Classe	Valeurs Grammaticales	Genre/Nombre	Dérivations
<input checked="" type="checkbox"/> activité de conseil	activité de conseil	Substantif	S	NS, N'	1

—Intégrer les expressions idiomatiques : activité de conseil, conseil, activité de consultant, consulting

5. Classification automatique

5-Classification automatique **La place de la classification automatique : pourquoi ?**

Présentation des résultats dans un système de recherche

- liste de documents fournis souvent longue
- rarement exploitée dans sa globalité par les utilisateurs : des documents pertinents mais mal positionnés ne sont pas "vus"

La classification automatique améliore la qualité de la recherche en offrant une visibilité :

- sur le fonds interrogé : classification globale de l'ensemble des documents (en amont de la recherche)
- sur le lot résultat : classification dite locale, des documents résultant de la recherche

Classification www.aol.fr (Exalead)

The screenshot shows the AOL search interface. The search bar contains the text "effets et dangers du dopage dans le sport". Below the search bar, there are navigation links for "Internet français" and "moodis". The search results are displayed in a list format, with the first result being "22 documents trouvés". The results are categorized by "Liste des produits dopants", "Taux d'hémoglobine", "Affaire Festina", and "Lutte antidopage". The first result is "Les dangers du sport : le dopage", which discusses the dangers of doping in sports and mentions the 1998 Tour de France. Other results include "dopage", "dopé", and "Le guide de prévention sportive et de lutte contre le...".

Sylvie Dalbin (ATD/DESYBEL), Rouen le 5 Décembre 2002

39

Classification hiérarchisée : Vivísimo

The screenshot shows the Vivísimo search interface. The search bar contains the text "effets et dangers du dopage dans le sport". Below the search bar, there are navigation links for "company", "products", "demos", "partners", and "press". The search results are displayed in a list format, with the first result being "Top 67 documents sélectionnés for the query effets et dangers du dopage dans le sport". The results are categorized by "Lutte contre", "La santé", "Médecine", "Club de l'Éducation", "France", "Antidopage_Science", "Substances", "dopage", "L'Étude", and "Antidopage_Sport". The first result is "22-02-Session", which discusses the effects and dangers of doping in sports. Other results include "Medicine - Dépendances - Drogues : savoir plus, risquer moins" and "International Olympic Academy".

Sylvie Dalbin (ATD/DESYBEL), Rouen le 5 Décembre 2002

40

Classification automatique : principes (a)

Répartition automatique des objets dans des classes

2 catégories de méthode de classification automatique

—classement automatique de documents dans des classes pré-établies

- Classes préexistantes (a priori)
- apprentissage supervisé : les classes constituent un ensemble d'apprentissage
- On assigne aux documents une (plusieurs) catégories existantes.
- Problème : Élaboration et suivi de la liste de classes
- Exemples : Arisem, K2 Enterprise de Verity,

—Regroupement de documents constituant des classes construites dynamiquement, a posteriori

- création automatique de catégories dans lesquelles sont classées les documents. Ces catégories sont établies sur la base de similarités trouvées entre documents (apprentissage, non supervisé)
- Appelé « clusterisation »
- Problèmes : Trouver automatiquement et rapidement des groupes; les nommer
- Exemples : AOL/Exalead ;Fast Topic ; Vivisimo, Autonomie,...

Classification automatique : principes (b)

Usages en recherche

- Aide à la sélection de documents au sein du lot-résultat par le biais de notions non exprimées dans la question
- Aide à l'élimination des corrélations inintéressantes, évidentes mais non repérées ou connues
- Idées nouvelles par la mise en perspective de corrélation non établie par l'utilisateur

Techniques complétées par celles de représentation graphique de l'information

Attention à la terminologie adoptée

6. Les logiciels d'indexation et recherche d'information

Les offres des éditeurs et prestataires

Offre « globale »

Constituée de « briques logicielles »

- adaptées au contexte : volumes/flux et types de ressources
- avec une orientation particulière en terme d'usage : Recherche-Intranet/portail, GED, gestion de contenu, workflow, veille (text-mining), travail de groupe (groupware), portail/diffusion,...

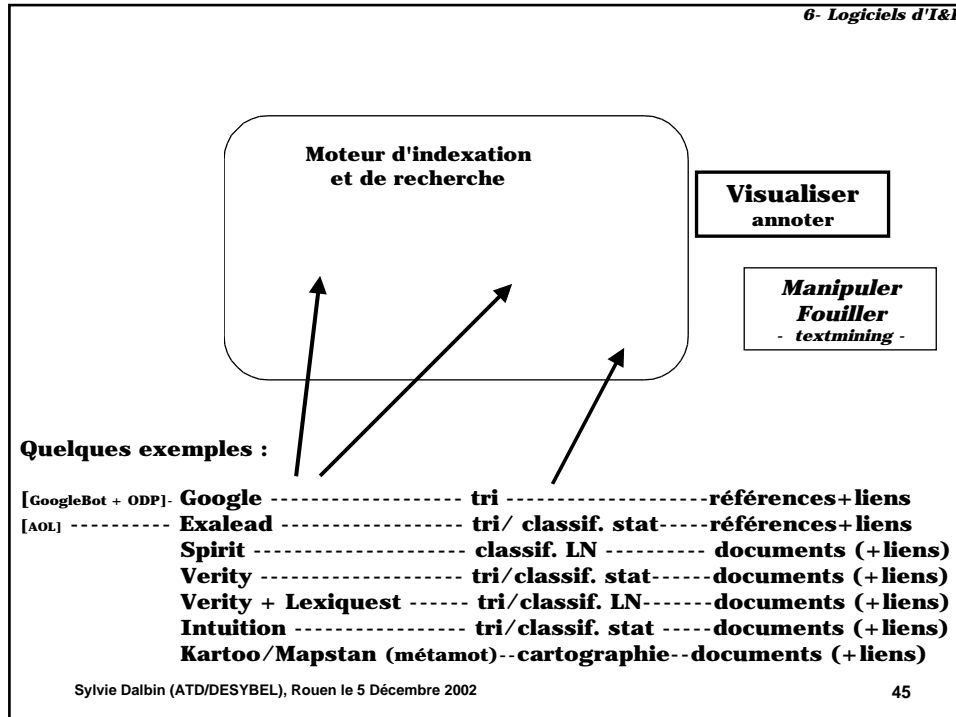
Incluant systématiquement un moteur I&R

- en texte intégral et/ou en langage naturel

Administration des outils linguistiques : faiblesse

Evolution : XML (format d'échange et stockage des données)

Au sein des offres, identifier les moteurs de base...



Une typologie de logiciels d'I&R

A/ Moteur essentiellement statistiques

AIRS d'Euritis, BasisPlus

Text Retrieval (Context) d'Oracle
RetrievalWare (Excalibur) de Convera
 SearchServer (Fulcrum) de Hummingbird /
 Cadic/SearchServer
 Information server de Verity
 ZylIndex de Zylab ...

B/ Moteurs linguistiques/sémantiques

AMI d'Albert
 Fulty de **Lingway** (anciens lexiquest)
 LexiGuide de Lexiquest (>> Ertl)
Intuition (>> Darwin) de **Sinequa**
 Pertimm d'Ogmios
Spirit TGID, ...
 Arisem
 Autonomy
 SmartDiscovery (Inxight)
 RetrievalWare (Excalibur) de Convera ...

Consultation/annotation

Acrobat d'Adobe (PDF) ?

C/ Classification

• *Exploitation d'algorithmes de classification*
 Exalead (Aol.fr)
 Categorizer (Inxight)
 SemProfile d'Arisem
 Verity, ...

• *Exploitation d'une classification*
Tacsy de Lingway (CIB de l'Inpi)
 Topic de Verity, ...

D/ Représentation graphique (carte)

Kartoo, Mapstan sur Internet
 LexiMine de Lexiquest, Text Navigator d'IBM,
 Semantic Map de Datops, Tropes d'Acetic,
 Umap de Trivium, ... VizServer d'Inxight

E/ Résumé automatique

Copernic Summarizer, Pertinence, ...

Quelques critères de distinction des logiciels

Indexation des documents et/ou des requêtes

Traitement des données structurées / non structurées (articulation)

Types de traitements linguistiques mis en oeuvre (voir partie 4)

- Morphologique (*Verity*) ; Morpho-syntaxique; Syntaxique (*Intruition, Spirit...*) ; Sémantique
- Facilité d'administration des dictionnaires
- Possibilité de mettre en oeuvre des traitements spécifiques selon les ressources
Intuition, ExLibris,...

Traitements statistiques (voir partie 3)

Classification des résultats (voir partie 5)

- A priori (*Arise*) ou a posteriori
- À partir des mots de la question (classes de *Spirit*) ou d'autres notions complémentaires portées par les documents (*Exalead*)

Volumes/flux des ressources à traiter, des utilisateurs

Formats en entrée (.doc, html, > XML), pour le stockage des données (texte, XML?)

Prix : à partir de 70 000 euros à 140 000 euros (internet+ 200 en intranet)

7. Problématiques de l'évaluation

Evaluer la recherche d'information

Pourquoi évaluer ?

- Etre en mesure d'adapter le système aux pratiques des utilisateurs

Qu'évalue-t-on ?

- Le logiciel ou le dispositif documentaire dans son entier ?
- Mesure de la pertinence des résultats : bruit et silence.

>> Evaluer le moteur de recherche

- Mais elle ne mesure pas la performance du dispositif face aux besoins des usagers : pertinence du fonds documentaire, interface IHM/portail, exploitation enrichie des résultats de recherche, consultation facilitée des documents, réutilisation de l'information... .

>> Evaluer la réponse du dispositif aux besoins des usagers de l'information

Notions de bruit et silence

Pour évaluer la performance d'un système de recherche d'information, les méthodes "classiques" se basent sur :

- le **bruit** : documents non pertinents trouvés
- Indicateur de mesure du bruit > le **taux de précision**
 - ratio entre le nombre de documents pertinents trouvés et le nombre total de documents trouvés
- le **silence** : documents non trouvés, mais pertinents
- Indicateur de mesure du silence > le **taux de rappel** (recall)
 - ratio entre le nombre de documents pertinents retrouvés et le nombre total de documents pertinents dans le système

L'équilibre entre le rappel et la précision dépend du but visé et du contexte (utilisateur)

Problèmes en recherche : bruit et silence

Les causes des problèmes de bruit et de silence sont multiples.

On peut citer :

	Silence	Bruit
—prise en compte d'un concept inadéquat		*
—non prise en compte d'un concept informatif	*	
—prise en compte d'un concept non informatif		*
—niveau de spécificité mal compris	*	(*)
—mauvaise traduction d'un concept	*	*

Méthodes d'évaluation

Analyser les questions posées par les utilisateurs

Elaborer et mettre en oeuvre des protocoles de test

- "poser" des batteries de questions au système
- analyser les résultats

S'appuyer sur des tests et des évaluations réalisés par des éditeurs ou sociétés spécialisées

- attention aux méthodes employées
- Par exemple :
 - Text Retrieval Conference (TREC). <http://trec.nist.gov/>
 - Classement des automates de recherche/ Marc Duval. [En ligne]. Longueuil, Québec, 2001. <<<http://www.dsi-info.ca/classement-introduction.html>>>

Réaliser des enquêtes périodiques

"Log" des moteurs de recherche : un exemple

6	128.193.224.39	intradoc	article 83 du code des assurances		
10	128.193.224.41	intradoc	kit de ressources	Ouvrage	
1035	130.138.224.31	intradoc	les fauteuils électriques handicapés sont soumis à la rc auto obligatoire . sur quelle base légale ou jurisprudentielle peut-on alléguer cette affirmation ?	assurance automobile	
12	128.193.224.42	intradoc	le multi réseaux dans l'assurance		
14	128.193.224.61	intradoc	assurance vie	Article	
28	128.193.226.235	intradoc	gestion des performances	Ouvrage	INFORMATIQUE
35	128.193.226.67	intradoc	l'image des assurances en France		
50	128.193.227.36	intradoc	serrure deux points		
62	128.193.229.39	intradoc	communication asynchrone	Article	INFORMATIQUE
73	128.193.244.154	intradoc	La circulation dans les ronds points	Ouvrage	DROIT
89	128.193.245.129	intradoc	les contrats multisupports en retraite collective		
91	128.193.245.129	intradoc	Qu'est ce qu'une catégorie ministérielle ? (exemples : 211, 212, 214)		
92	128.193.245.129	intradoc	Régime fiscal et social des prestations forfaitaires prévues dans les contrats Collectifs SANTE	Prévoyance et retraite collective	
233	128.217.224.114	intradoc	okassurance		
241	128.217.224.117	intradoc	pays carte verte		
495	128.65.226.131	breves	filia-MAIF	groupement des sociétés du GEMA	['01/01/2000' ; '01/10/2001']
504	128.65.226.138	intradoc	Article	faillite	Tribune&Assurance
530	128.65.226.189	intradoc	gonfler un cv	Tribune/assurance/entreprise	
638	128.65.49.48	intradoc	En quoi la gestion des connaissances peut elle etre un facteur de croissance et de developpement pour l'entreprise		
Sylvie Dalbin (ATD/DESYBEL), Rouen le 5 Decembre 2002					53

En conclusion

La place de ces techniques et leur évolution

Pluri-modalités de recherche

- Articulation de la recherche sur le texte non structuré, avec une recherche sur zones structurées d'une notice
- Recherche "intelligente" : lexicale, linguistique, sémantique
- Choix d'une classe de documents : classification des résultats
- Navigation dans une arborescence (classification,...) vs expression d'une requête

Architecture fonctionnelles des systèmes

- Gestion de contenu hétérogène :
 - gestion bibliographique >> gestion des ressources numériques >> gestion de documents structurés XML
- Indexer les ressources et les questions automatiquement, avec le même outil ou avec des outils distincts adaptés
- La finalité de la recherche : trouver, d'où l'importance de l'interface homme-machine (ergonomie, classification, cartographie)

La position des professionnels de l'information

Pratiques et finalités de la recherche :
utilisateurs (contenu) # documentalistes (notice)

D'une logique monolithique (une base bibliographique)
à une logique différenciée en fonction des fonds, des
utilisateurs

Problématiques :

- Que devient le métier :
 - Maîtriser parallèlement les techniques de recherche bibliographiques et celles de recherche sur le contenu
 - Des compétences plus poussées en ingénierie linguistique
 - Développer des activités de formation, de conseil, de contrôle
 - Remplacer l'activité d'indexation comme moyen de « connaissance » du domaine de l'activité
- Articuler fonds électronique / fonds papier
- Reprise de l'existant

Annexe 1 : Éléments bibliographiques

Présentation d'expériences

- Le système CIB-LN d'accès aux brevets en langage naturel/ Darrigade S., Lyon-Bougeat M., Marx B., Documentaliste - Sciences de l'Information, 2001, vol. 38, n°2, p.100-110
- Une expérience d'utilisation d'un système d'information documentaire en langage naturel/ Sylvie Dalbin, Bruno Salléras. Documentaliste - Sciences de l'Information, 2000, vol. 37, n° 5-6, p. 312-324
- Indexation manuelle et indexation automatique : dépasser les oppositions / Ghislaine Chartron, Sylvie Dalbin, MG Montell, Monique Vérillon. - Documentaliste-Science de l'information, vol. 26, 1989, n°4-5, p. 187-187.

Méthodes et techniques d'indexation et de recherche automatiques

- Cours Inria 1992 (Interfaces intelligentes dans l'IST), 1994 (Le traitement électronique des documents), 1996 (La recherche d'information sur les réseaux), 2002 (La recherche d'information sur les réseaux. 2) ; édités par l'ADBS depuis 1994 [*Présentation systématique d'articles sur l'indexation automatique et/ou les traitements en langage naturel*]
- Documentations techniques sur les logiciels, produits par les éditeurs (sites) : extensions .com aux noms des sociétés éditrices.
- Recherche d'information dans les documents textuels / Sébillot . - IRISA, février 2002
- Recherche d'information sur les réseaux . cours INRIA, Le Bono, 20 septembre-4 novembre 2002 / coord. Le Moal JC, Hidoine B, Calderan L.Paris, ADBS Éditions, 2002
- Actualités des langages documentaires : fondements théoriques de la recherche d'information/ Jacques Maniez. Paris, ADBS Éditions, 2002. [En particulier les chapitres III, IV et V : notions d'objets informationnels et de document ; typologie des systèmes de recherche d'information]
- Ingénierie des langues / Jean-Marie Pierrel (dir.). - Paris : Hermès, 2000
- Recherche documentaire : du thesaurus au texte intégral / Philippe Lefevre. - Paris : Hermès, 2000
- Comment les logiciels de bases de données bibliographiques et textuelles peuvent-ils répondre aux différents besoins de leur utilisateurs ? Bertrand-Gastaldy Suzanne. [En ligne], [Canada, sans date] [Visité le 28 août 2000]. Disponible sur Internet. http://www.ling.uqam.ca/sato/publications/bibliographie/ind_lang.htm
- Panorama et perspectives des outils de recherche d'information textuelle sur Internet/ François Bourdoncle. - In : IDT 1999 : textes des communications. <http://www.exalead.com/Francois.Bourdoncle/idx99.html>
- La recherche d'information dans les mémoires électroniques. L'enjeu documentaire / Fondin Hubert . Documentaliste - Sciences de l'information, 1999, vol.36, n°4-5, pp. 242-248
- Méthodes de tri des résultats des moteurs de recherche/ Jean-Pierre Lardy. <http://www.adbs.fr/site/repertoires/sites/lardy/risi.htm>

Conception des systèmes d'information

- Les portails d'entreprise : conception et mise en œuvre / Jean-Louis Bénard. Paris, Hermès, 2002. [Caractéristiques du portail d'entreprise, en particulier des technologies mises en œuvre et des principaux acteurs du marché ; démarche de conception]
- L'écrit et l'écran. Captain Doc, mars 2002, n° 6. <<http://www.ftpress-kiosque.com/www/arc/captaindoc-txt/2002-03/thrft1.html#000000>>. [Les rapports de l'écrit et de l'écran ; un entretien avec Brigitte Juanals : "Accès aux savoirs, de la page du livre à la page-écran" ; dossier complet <http://www.captaindoc.com/dossiers/dossier07.html>]
- Michèle Hudon. Structuration du savoir et organisation des collections dans les répertoires du Web.. Bulletin des bibliothèques de France, 2001, t. 46, n° 1. <http://bbf.enssib.fr>

Sylvie Dalbin (ATD/DESYBEL), Rouen le 5 Décembre 2002

57

Annexe 2 : Indexation humaine / assistée par ordinateur

Indexation humaine

- vocabulaire restreint
- utilisation de macro-termes
- variabilité de l'indexeur
- opère sur l'intégralité du document et surtout son "sens" (sémantique)
- effet de généralisation
- amplification de certains éléments par rapport à d'autres (par rapport au fonds documentaire, à ces utilisateurs)
- décalage entre le thesaurus/le contenu des documents
- sélection de termes, voire des concepts

IAO (sans linguistique)

- vocabulaire des auteurs de la langue française, et autres !
- cohésion par rapport à un auteur (pendant une période)
- opère sur l'intégralité du texte du document
- effet de dispersion : absence de terme synthétique
- description spécifique de l'information du texte
- effet d'ambiguïté (mots hors contexte)
- exhaustivité des termes et concepts

Sylvie Dalbin (ATD/DESYBEL), Rouen le 5 Décembre 2002

58

Annexe 3 : Traitement automatique du langage naturel (TALN)

Lire/Ecrire : aide à l'écriture, génération automatique de texte, résumé assisté par ordinateur

Traduire : TAO, bases de données multilingues, terminologies

Décrire, organiser, caractériser : indexation automatique (donner une description, discriminer parmi des fonds importants), classement automatique, structuration des documents

Rechercher, retrouver : interfaces de bases de données factuelles, textuelles, de données mixtes, comparaison de textes

APIL (Ass. Prof. Industries de la langue) ([//www.apil.asso.fr](http://www.apil.asso.fr))

OFIL (Observatoire français des industries de la langue-