



Systèmes d'Information documentaire **Composants et fonctionnalités des offres logicielles**

Production et gestion du document

Odile Giraud, Doc&Co

Accès à l'information

Sylvie Dalbin, ATD - DESYBEL

Diffusion et publication : support, outils, formats

Sylvie Cabral, Ourouk

Mercredi 01 Juin 2005



Systèmes d'Information documentaire **Composants et fonctionnalités des offres logicielles** **2. Accès à l'information**

Sylvie Dalbin

Assistance & Techniques Documentaires - DESYBEL Gie

Mercredi 01 Juin 2005

Sommaire de l'intervention

- 4 Contexte en forte évolution
- 4 Fonctions de recherche
 - Etape 1 - Compréhension du SID
 - Etape 2 - Capture de la requête
 - Etape 3 - Traitements
 - Etape 4 - Présentation des résultats
 - Etape 5 - Aide à l'exploitation et à la lecture
 - Documents multimédia
 - Système question/réponse
 - Navigation hypertexte
- 4 Techniques sous-jacentes
- 4 En conclusion

Contexte en forte évolution

Evolutions « Objets documentaires »

- 4 Objets documentaires sous forme numérique (encodés), manipulables - métadonnées
- 4 Architectures des entrepôts diversifiées, à coordonner
- 4 Donnée, texte, images fixes et animées, son
 - Sources, Formats et filtres des logiciels
 - Documents structurés (sémantiques)
- 4 Multilingue & trans-lingue (multilinguisme croisé)
 - Autodétection de la langue
- 4 Stocks - Flux

Evolution « Utilisateurs »

- 4 Usages et environnements très variés
Société de l'information
- 4 Autonomie - Maturité
- 4 Diversités des besoins
 - ▾ Rechercher/retrouver, trier, classer, ordonner les résultats
 - ▾ Utiliser (lire, comprendre), exploiter (réutiliser), intégrer dans ses données personnelles - valoriser son propre fonds
 - ▾ Exploiter l'informatique pour optimiser ses activités liées à l'information > groupware, wikis...
- 4 Simplicité et performance des interface homme-machine (IHM)

Evolution « Informatique »

- 4 Informatique étendue :
 - en amont du stockage/gestion : acquisition, DSI
 - au cœur : indexation de tout contenu
 - en aval : classement des résultats - fouille - cartographie
- 4 Informatique Web : XML, architecture n-tiers,...
- 4 Informatique ouverte - composants interopérables
Web services
- 4 Architectures d'index performantes
 - fichiers inverses, vecteurs,...
- 4 Modélisation - Informatique "objets"
- 4 de l'Ascii à l'Unicode, ...

Evolution « marché »

- 4 Prolifération d'outils
 - v autonomes et spécialisés versus offres packagées
 - fonction
 - métier
 - intégrée à une offre d'information
- 4 Constitués de briques fonctionnelles
 - v orientation particulière en terme d'usage : Intranet/portail, GED, veille, travail de groupe (groupware), diffusion,...
 - v briques associées à l'accès à l'information

➤ *Techniques ou fonctionnalités adaptées
aux différentes étapes d'une recherche*

Evolution des pratiques de recherche

- 4 Syndrome du « chercheur scientifique »
 - v Recherche précise sur un sujet considéré comme cerné et à préciser, dans un environnement documentaire maîtrisé
- 4 Profils des utilisateurs des SID : révision drastique
 - v Ne souhaite pas chercher, mais être informé
 - v Ne sait pas s'il peut exister une information
 - v Ne pense qu'à trouver : la fouille, de préférence à la requête
 - v Ne connaît rien au sujet
proximité faible avec son domaine d'activité, mais recherche indispensable
 - v Souhaite retrouver quelque chose de connu rapidement
 - v Point d'accès très variés
 - v

Fonctions de recherche

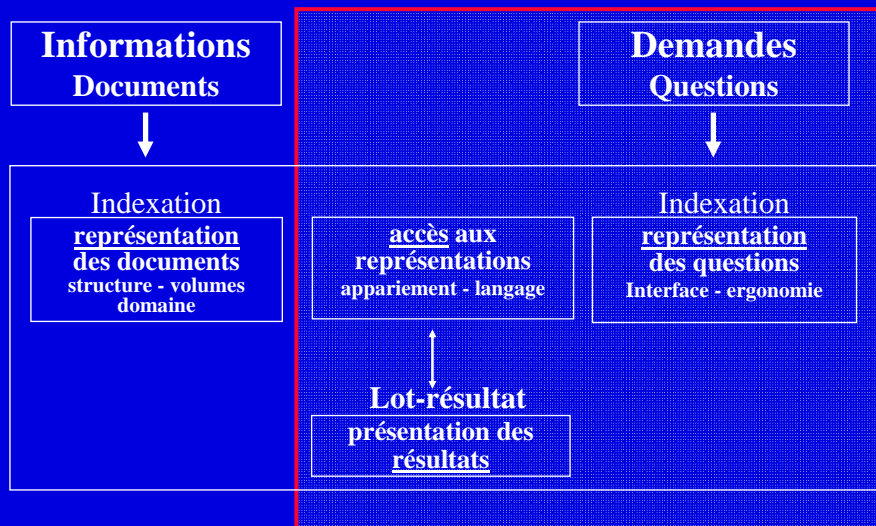
Diversité des types de recherche

- 4 Recherche de données par extraction (date retrieval)
 - v Recherche sur des champs avec ou sans listes contrôlées
 - v Critères précis - Tout ou rien
- 4 Recherche d'information (information retrieval)
 - v Critères plus souples pour palier ambiguïté ou imprécision de la demande - Tolérance aux erreurs
- 4 Question/réponse
 - v Différent de la recherche de données : au sein de fonds de documents
- 4 Navigation

Diversité des types de recherche

- 4 Recherche de données par extraction (date retrieval)
 - v Recherche sur des champs avec ou sans listes contrôlées
 - v Critères précis - Tout ou rien
- 4 **Recherche d'information (information retrieval)**
 - v Critères plus souples pour palier ambiguïté ou imprécision de la demande - Tolérance aux erreurs
- 4 Question/réponse
 - v Différent de la recherche de données : au sein de fonds de documents
- 4 Navigation

Le système documentaire schéma fonctionnel



Etape 1 - Compréhension de l'architecture et de l'organisation du dispositif

*Un utilisateur oriente plus efficacement sa question
s'il a une "bonne idée" de l'espace manipulé*

- 4 Signalétique - Carte - Fil conducteur
- 4 Distinguer espace de production - espace d'utilisation
Structurer l'espace en fonction des pratiques utilisateurs
- 4 Terminologie et ergonomie adaptées aux utilisateurs
- 4 Ergonomie générale
 - v Positionnement à l'écran : recherche rapide - personnalisation
 - v Prise en compte du niveau de l'utilisateur
modes rapides / expert / assisté - modes multiples

➤ **Point de départ : ergonomie**

Etape 2 - Capture de la requête

Diversité des modes d'expression de la requête

- 4 Dialogues de commandes, avec syntaxe
 - v Expression d'une requête à partir de champs structurés
 - v Expression d'une requête sur le contenu des objets
- 4 Dialogue en langage naturel
 - v Indexation de même nature ou différente de celle des documents
- 4 Formulaire
- 4 Navigation
- 4 Sérendipité
art de découvrir sans anticipation, sans hypothèse, par accident et sagacité

➤ *Recherche systématique et finalisée
versus recherche exploratoire*

➤ *Mixité des modes de recherche*

Etape 2 - Capture de la requête

Techniques de reformulation et d'expansion

- 4 Tolérance aux erreurs manipulatoires
 - v Phonétique - orthographique (algorithmes/dictionnaires)
- 4 Multilingue - Traduction des requêtes (translingues)
- 4 Expansion ou restriction de la recherche :
 - v Sur la requête initiale et/ou après un premier niveau de réponse - traitements linguistiques et sémantiques
 - v Traitement linguistique et sémantiques
 - v Interactive à partir d'un document pré-identifié (technique de bouclage ou rétroaction de pertinence - approche similarité)
 - v À partir d'une valeur d'une métadonnée (le nom d'un auteur, d'une société, ...) d'une des réponses fournies
 - v sur une autre base proposée

Etape 3 - Traitements

Techniques de filtrage - Pondération (a)

Avec une représentation enrichie de la requête

- 4 Position des termes dans le document structuré
- 4 Fréquence des termes de la requête dans les documents
 - Densité : nombre de termes pertinents / nombre total de termes du document
- 4 Proximité/ordre des termes de la requête, dans les documents
- 4 Exhaustivité
 - Poids plus important aux documents contenant un plus grand nombre de termes de la question

Etape 3 - Traitements

Techniques de filtrage - Pondération (b)

Avec une représentation enrichie de la requête

- 4 Discriminant : poids importants aux termes rares
- 4 Localisation du terme trouvé (corps du texte, paratexte)
- 4 Distance sémantique des termes (question/document) fréquence sémantique
- 4 Poids assigné par l'utilisateur aux mots de sa question
- 4 Prise en compte des hyperliens (pagerank, Kleinberg)

Etape 3 - Traitements

Techniques de filtrage - Pondération (c)

- 4 Prise en compte de la source
 - v cas de la recherche multisources
 - v pondération préalable des sources
- 4 Prise en compte d'entités nommées
- 4 Par profil utilisateurs (*profiling*)
 - v Structure de données qui décrit les centres d'intérêt de l'utilisateur, en général représentée par des mots clés
extension/filtrage par contenu
 - v Structure utilisée pour filtrer ou pour recommander ce qui a satisfait d'autres utilisateurs, de profil similaire
extension/filtrage collaboratif

➤ **Des bases « Mémoire des utilisateurs », capitalise l'utilisation du système**

Etape 3 - Traitements

Recherche distribuée - fédération de sources

Interrogation unifiée de sources multiples - hétérogènes

- 4 Connecteur (= agent, wrapper, passerelle)
 - v génériques, spécialisés par rapport aux sources
 - v Protocole OAI (Dublin Core ou autres modèles) - ISO 23950 (Z3950 client/serveur) - Profils spécialisés
- 4 Formats des données - structurés, normalisés / plus libres - et syntaxes : adaptation
- 4 Fusion et uniformisation de la présentation des résultats
- 4 Pertinence indépendante de la source
- 4 Administration des sources à interroger

Etape 4 - Présentation des résultats

Présentation du lot-résultat

Aider l'utilisateur à **explorer rapidement les résultats** et à sélectionner les documents les plus pertinents

Tableau de bord d'aide à la sélection

- 4 Tri par ordre de « pertinence » ou autres critères, au choix
- 4 Liste, classification ou cartographie du lot résultat
- 4 Identification de thèmes ou métadonnées connexes
 - v Permet la relance/restriction d'une requête
 - v Thème connexe : identification dans le texte, de mots ou expressions évalués comme importants
 - par catégorie d'information : entités nommées, date,...

Etape 4 - Présentation des résultats

Présentation des informations-résultat

Aider l'utilisateur à **explorer rapidement les résultats** et à sélectionner les documents les plus pertinents

Tableau de bord d'aide à la sélection

- 4 Présentation synthétique du document
 - v Extraction de la phrase dans laquelle se trouve le(s) terme(s) recherché(s)
 - v Résumé automatique - problème de la reformulation
- 4 Positionnement sur le mot (ou groupe de mots) recherché, et son contexte (surbrillance)
- 4 Distinction par "unité d'information"(document, partie de document) et navigation entre elles

Etape 4 - Présentation des résultats

Classification, visualisation graphique

Explorer rapidement un ensemble d'information
Mettre en évidence des relations entre information

4 Mots clés - Listes

∑ Indicateur de contenu

4 Regroupements

- ▾ clusters, classes, thèmes,...

∑ Indicateur de thèmes

4 Cartes

∑ Positionnement de thèmes
dans l'espace de connaissance

Etape 4 - Présentation des résultats

2 méthodes de classification des résultats

4 Classification (clustering)

- ▾ Organiser un ensemble de documents en classes homogènes
- ▾ Caractéristiques
 - non supervisée, calcul dynamique
 - Structure hiérarchique ou à plat - score à l'intérieur de chaque groupe
 - Divers modes de représentation des classes et problème de dénomination de chaque groupe

4 Catégorisation

- ▾ Assigner à chaque document, une catégorie d'un schéma existant
- ▾ Caractéristique
 - supervisée

Company | products | solutions | customers | demos | press

Vivísimo®

problème de dopage dans le sport
the Web

Search Advanced Search Help

Clustered Results

- problème de dopage dans le sport (127)
 - Lutte contre le dopage (25)
 - Conférence (13)
 - Santé (15)
 - La santé dans la pratique (3)
 - Lutte (2)
 - Doctissimo (2)
 - Sport, Dopage Et Conduites Addictives (2)
 - La santé en langue doc-roussillon méditerranée (2)
 - Sénat, Décembre (2)
 - Other Topics (2)
 - Haut niveau (14)
 - Antidopage (11)
 - Président (7)
 - Fédéral, Fifa.Com (6)
 - Peut (6)
 - Clubs (5)

Top 127 results retrieved for the query **problème de dopage dans le sport** (Details)

1. [HARMONISATION DES METHODES ET DES MESURES DE LUTTE CONTRE LE DOPAGE...](#) [new window] [frame] [preview] [clusters]
...du 7-9 mai 1999 - Toulouse4,1752'8&7,21 Il semble que le **problème** du **dopage** se soit posé depuis que le **sport** est connu comme phénomène...
europa.eu.int/comm/research/smt/hardop-fr.pdf - Lycos 1, Ask Jeeves 1, MSN 18
2. [Quelques reflexions sur le dopage dans le sport par le président de...](#) [new window] [frame] [preview] [clusters]
Introduction **Le dopage dans le sport** L'avenir Du "fair-play" à la définition du **dopage** **Problème** de la créatine EPO...
www.bmlweb.org/dopage_ucl.html - Lycos 4, Ask Jeeves 4, MSN 11
3. [Sport, Santé et éthique](#) [new window] [frame] [preview] [clusters]
... épidémie globale selon lui et un **problème de** santé publique encore négligé ... la Convention Internationale sur le **dopage dans le sport** sera ratifiée en Octobre prochain. M ...
Cache: http://cc.msnsocache.com/cache.aspx?q=1874744611583 (=&en-US&FORM=CVRE7
www.un.org/sport2005/calendartunisia_sportde_programme.pdf - MSN Search 7, MSN 14
4. [Lutte contre le dopage](#) [new window] [frame] [preview] [clusters]
... centre fédéral de formation dans le domaine du **sport**, un centre de référence ... en 2001 considèrent le **dopage** comme un **problème** grave, voire très grave pour le **sport** d'élite. Bien que la ...
Cache: http://cc.msnsocache.com/cache.aspx?q=1877391063770 (=&en-US&FORM=CVRE
www.baspo.admin.ch/.../baspo/fr/home/sport00/sport00c.html - MSN Search 1
5. [UNESCO. General Conference: 32nd; Suivi de la Table ronde des ministres et hauts responsables chargés de...](#) [new window] [frame] [preview] [clusters]
... internationale contre le **dopage dans le sport** ci-joint (Annexe II), le Conseil exécutif ... a insisté sur le **problème** du **dopage** qui, à l'heure actuelle, apparaît comme le **péril** le ...
www.cigepe.org/pdf/32c50fr.pdf - MSN 1

Etape 4 - Présentation des résultats

Représentation cartographique des résultats

- 4 Une carte
 - ▾ des nœuds, des liens, une position dans l'espace
- 4 Valeur d'un nœud
 - ▾ un document (Kartoo)
 - ▾ un groupe (cluster) de documents/informations (Mapstan)
- 4 Positionnement dans l'espace (nœuds/liens)
 - ▾ sans signification (Kartoo)
 - ▾ avec signification (>>veille)

Résumé automatique

Recherche Google / Pertinence

« **développement durable** » **batterie**

[Médiaterre actualité scientifique: l'information pour le ...](#)

... mondial francophone pour le **développement durable** Médiaterre
(présentation sur ... ELECTRONIQUE : La premiere **batterie** rechargeable
mobile intelligente ...

www.mediaterre.org/scientifiques/gen.php3/topic/Energie,0,1,0.html - 72k -

[En cache](#) - [Pages similaires](#) - [Résumé](#)

Afficher un résumé d'au plus 20% (547 mots) ou 5.0 % du texte source (2738 mots)

[Copier le résumé affiché](#) Nouveau document à résumer Votre évaluation

Langue sélectionnée : **french** Domaine sélectionné :
133 phrases, 2738 mots, 17369 caractères. Traité en 3s 729ms

Vos mots-clés
batterie (8)
développement durable (4)

Equipe d'une puce électronique, le MPB 4000 peut identifier le type de la batterie et indiquer au chargeur l'algorithme de charge à appliquer, tout en protégeant la batterie d'une surcharge, d'une décharge excessive ou d'un court-circuit.

Ce site de référence a été conçu pour à la fois présenter le progiciel et fournir à ses utilisateurs de nombreuses ressources téléchargeables (environ 45 Mo d'exemples, manuels de référence, notes de prise en mains, notes de réflexion...).

Il constitue une plateforme d'expérimentation virtuelle permettant aux élèves de faire le lien entre la théorie et la pratique en mettant en œuvre les concepts étudiés en cours et de s'initier à la modélisation des systèmes énergétiques.

http://www.pertinence.net - Pertinence Summarizer - Mozilla

cZ Chargé

5. Aide à l'exploitation et à la lecture

Récupérer et réutiliser le lot sélectionné; Lire et exploiter les documents

- 4 Panier ou téléchargement - tout ou sélection - sauvegarde des requêtes - des documents
- 4 Formats permettant une réintégration simple dans d'autres systèmes personnels
- 4 Possibilités d'annotation des documents : un début !
 - v fonctions de résumé et navigation entre la synthèse et les parties de document
 - v colorisation du texte
 - v Annotation de documents HTML en ligne
 - v Un problème : en mode Auteur et non lecteur

Documents multimédia

- 4 Techniques de nature « similaire »
- 4 2 principaux modes d'accès
 - v retrouver des corpus d'images homogènes, par reconnaissance
 - v naviguer dans des corpus d'images hétérogènes
- 4 Attributs exploités
 - v images universels
 - couleur, texture et forme. Extraction de signatures d'images correspondantes pour les bases d'images génériques.
 - v images spécifiques à un domaine
 - dépendent de la connaissance a priori du domaine (ex : excentricité d'un visage, position d'une tumeur)
- 4 Produits
 - v Excalibur, Autonomy ; LTU, iPari, Kinomai (vidéo), Virage (video)

Diversité des types de recherche

- 4 Recherche de données par extraction (date retrieval)
 - v Recherche sur des champs avec ou sans listes contrôlées
 - v Critères précis - Tout ou rien
- 4 Recherche d'information (information retrieval)
 - v Critères plus souples pour palier ambiguïté ou imprécision de la demande - Tolérance aux erreurs
 - v Documents multimedia
- 4 **Question/réponse**
 - v Différent de la recherche de données : au sein de fonds de documents
- 4 **Navigation**

Système de question-réponse

- 4 Recherche d'une donnée, d'un fait, d'un renseignement, au sein d'un corpus textuel
 - v Qui a tué Henri IV ? Ravaillac
- 4 Problématique : recherche de documents et extraction d'information
 - v Analyse de la question : type attendu de la réponse, focus (topic, thème) de la question, type de question; reformulation
 - v Traitement des documents : syntaxique, sémantique
 - v Utilisation de patrons = schémas d'analyse de la réponse
- 4 Efficacité dépendante des sources d'information

➤ **En développement**

Navigation hypertextuelle

- 4 Au sein d'un schéma de départ
 - plan de classification - ontologie, listes de valeurs contrôlées
 - par sélection au sein d'une liste ou par approches successives du thème recherché
- 4 Au sein d'un schéma obtenu par classement des résultats d'une recherche
 - Mode de sélection de documents
 - Clusterisation - cartographie
- 4 Vers d'autres ressources suggérées par l'analyse des résultats (absent de la requête initiale)
 - Extension d'une recherche à d'autres fonds
 - Exemple : Base de dossiers de presse ou documentaires précis après une recherche sur un sujet particulier

Techniques sous-jacentes

Techniques de recherche

- 4 Recherche séquentielle sur chaînes de caractères
- 4 Recherche sur "fichiers inverses"
 - positionnel : n° document, nom du champ, n° de la phrase dans le champ, n° du mot dans la phrase
 - pondérations
 - Traitement d'unitermes / mots composés
- 4 Recherche sur fichier de signatures
 - Un document = vecteur binaire de longueur fixe
 - méthodes de signature diverses, en forte évolution
 - reconstruction d'index limité, encombrement réduit, ajouts de documents rapides
- 4 Recherche sur graphe conceptuel ou arborescence
 - **Enrichissement des représentations des documents et des questions (index)**

Modèles sous-jacents

- 4 Booléen et booléen pondéré
 - Pondéré : des poids préalablement attribués aux mots du corpus et de la question
- 4 Extension : logique floue
 - Pour des données imprécises [à peu près, environ]
- 4 LSI (Latent semantic indexing)
 - Co-occurrence
- 4 Modèle vectoriel (Salton G)
 - représentation d'un document/question (ou de concepts) par un vecteur (matrice) > correspondance partielle
- 4 Modèle probabiliste
 - Réseaux d'inférence bayésiens - calculer la pertinence d'un document en fonction de pertinences connues d'autres documents.
- 4 Réseaux de neurones (apprentissage initial)
- 4 Théorie des possibilités [très, beaucoup, peu, plutôt]

La problématique de l'évaluation

- 4 Pourquoi évaluer ?
 - v Etre pertinent, efficace, efficient par rapport à des finalités (utilisateurs finaux, intermédiaires, gestionnaires)
 - v Etre en conformité avec des procédures, des règlements
- 4 Qu'évalue-t-on ?
 - v Le moteur ; le dispositif d'information ; la réponse au besoin de l'utilisateur ?
- 4 Bruit et silence - mesure de pertinence (relevancy)
 - v Résultats d'une requête sur un système de recherche
 - v Ne mesure pas la performance du système par rapport aux usages, mais par rapport à l'outil technique
 - Prendre en compte : fonds, ergonomie, aides à la sélection des documents au sein du lot, classification, aide à la ré-exploitation

En conclusion

- 4 Interface d'accès orientées usages/utilisateurs
- 4 Exploiter au maximum votre outil documentaire et ses possibilités
 - v Repenser l'architecture des données, des fonctions
 - v Séparer systèmes de production / système de mise à disposition
 - v Exploiter au mieux les fonctionnalités - sortir des sentiers battus
 - v Des traitements et une ergonomie appropriés à la recherche
- 4 Prendre en compte la structure hypertextuelle des documents
 - v Au sein du document ; entre documents ; entre fonds/dossiers
- 4 Etendre à l'image, puis à l'audio

Merci de votre écoute !