

ACI Masse de données - Projet Micromégas
approches multi-échelles pour la navigation dans les
masses de données familières

Pointage dans les mondes d'information (LMP)

In Situ (PCRI)

MErLIn (INRIA)

Pôle Informatique (Institut Pasteur)

Étude de la mémoire des documents chez l'utilisateur pour la conception des systèmes d'aide à leur récupération

Tristan Blanc-Brude & Dominique Scapin,
MErLin, projet commun INRIA Rocquencourt & LORIA
(Université Henri Poincaré, Nancy).



Le problème de la gestion de nos propres données numériques

- La quantité de données numériques produites, acquises, reçu et stockées dans nos ordinateurs durant nos activités de travail ou dans notre vie privée est devenue considérable
 - ➡ problème de la gestion de ces informations

- Limites de la métaphore du bureau physique et de l'organisation hiérarchique des fichiers électroniques (Jones et Dumais, 1986; Marsden et Cairns, 2003)
 - ➡ besoin de nouveaux outils



Personal Information Management

(PIM)

- Domaine de recherche qui étudie la manière dont les utilisateurs gèrent leurs documents physiques (livre, cahier, feuille etc.) ou numériques (fichier, email, page web etc.) dans le but de concevoir des outils informatiques d'aide à la gestion des documents numériques (*PIM tools*).
- Activité de gestion des documents comprend 4 sous-activités :
l'acquisition, l'organisation, la maintenance et la récupération des documents.
(Boardman, 2004)

Le problème de la récupération de nos documents électroniques

- Leur nombre ne cesse de croître dans nos machines
 - ➡ les documents sont souvent difficiles à retrouver "dans la masse".

- But principal de l'activité de gestion des documents = permettre leur récupération ultérieure.
(Lansdale, 1988; Barreau et Nardi, 1995; Jones, Dumais et Bruce, 2002; Bellotti et al., 2002)

- Les utilisateurs ressentent moins le besoin de maintenir une hiérarchie complexe de leurs documents lorsqu'ils disposent d'un outil de recherche performant
(Dumais et al., 2003)

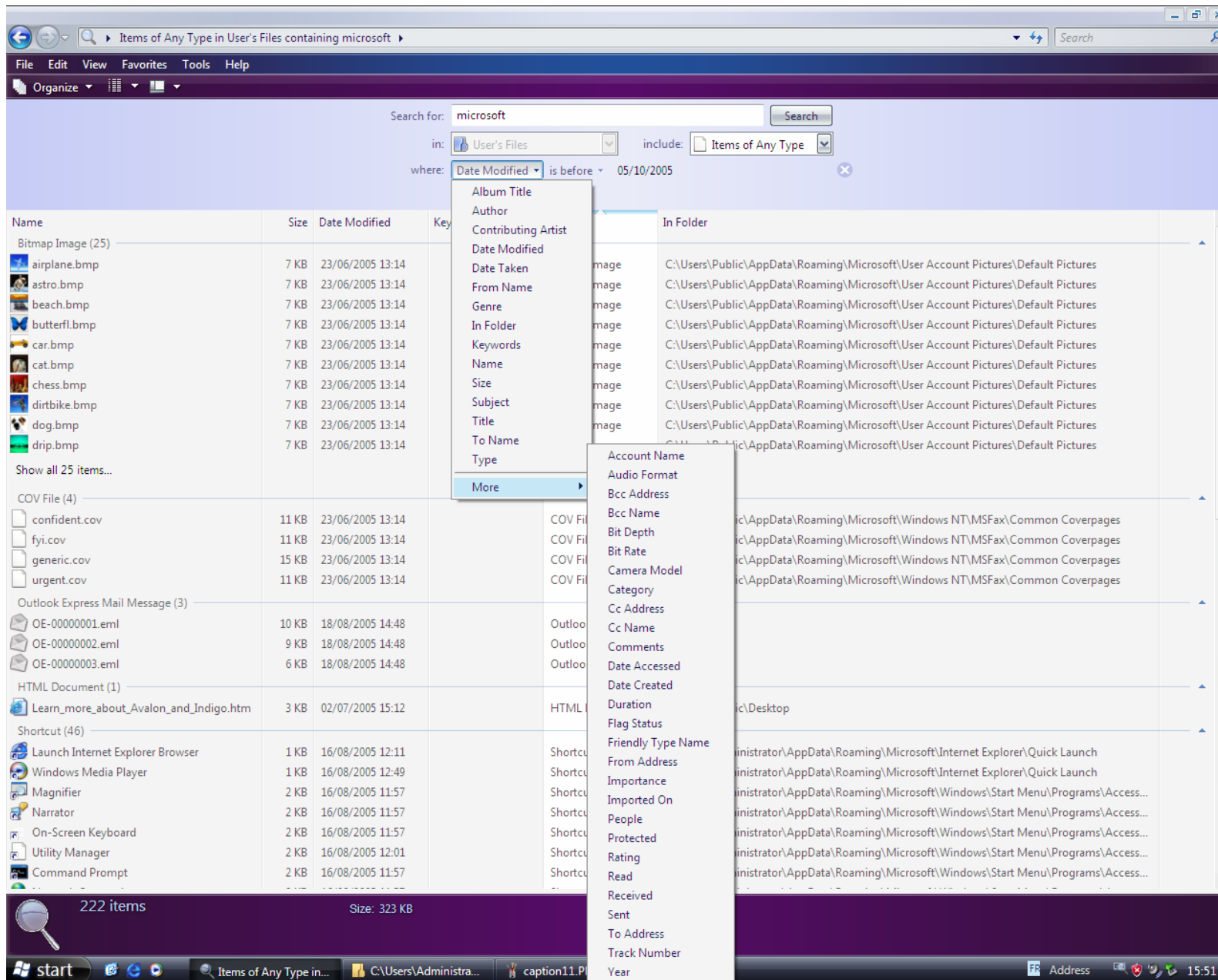


Le rôle des attributs des documents dans leur récupération

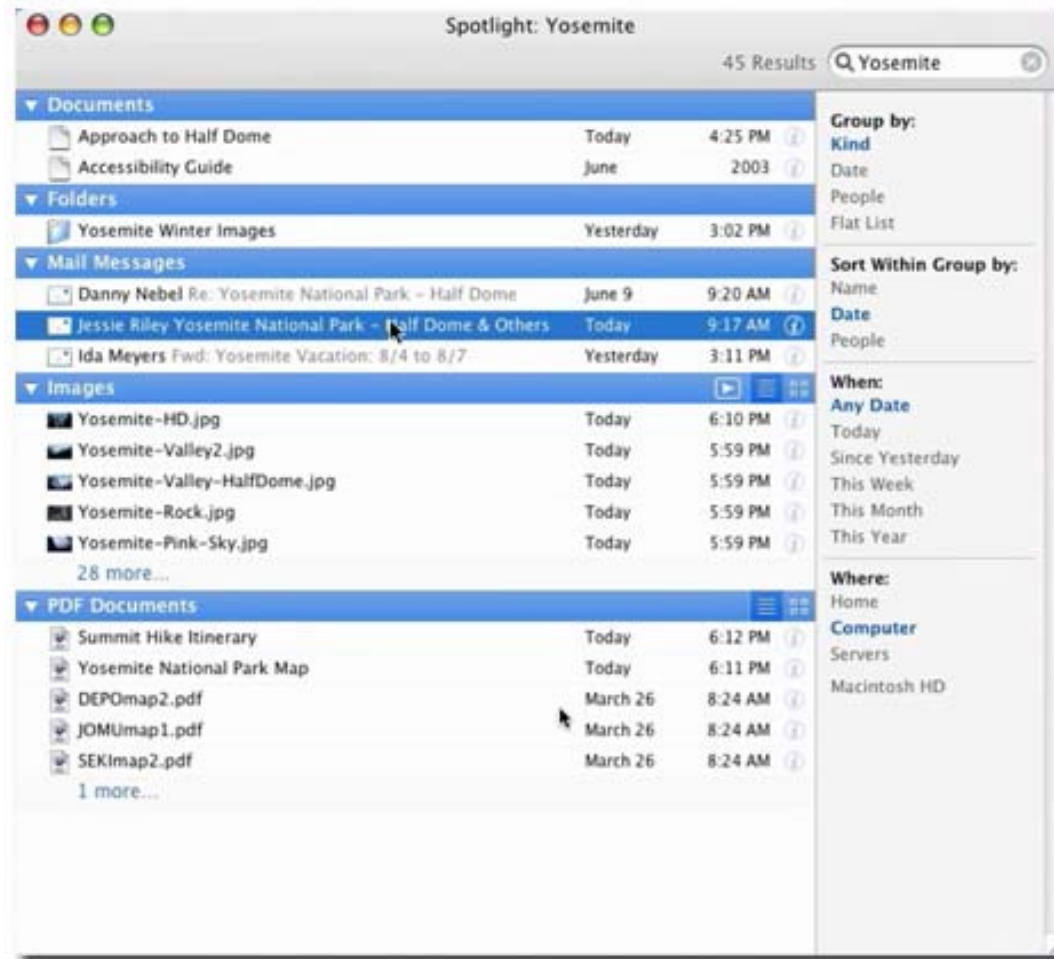
- Attributs d'un document : les différentes dimensions pouvant le caractériser comme sa taille, son titre, sa couleur etc.

- Les *utilisateurs* exploitent différents types d'*attributs* des documents
« où l'ai-je rangé ? » (localisation) « a quoi ressemble t-il ? » (apparence) etc.

- Les *outils de gestion et de recherche* reposent sur différents types d'*attributs* des documents
 - Systèmes d'exploitation classiques : localisation spatiale et nom
 - *Lifestreams* (Freeman et Gelernter, 1996) : temps lié à l'utilisation des documents
 - *UMEA* (Kaptelinin, 2003) : projet d'appartenance des documents
 - Outils de recherche des systèmes d'exploitation de nos ordinateurs : nom, contenu, date, taille etc.



Outil de recherche de MS Windows Vista (Beta 1)



Spotlight dans Mac OS X Tiger

Limite des outils actuels d'aide à la récupération des documents

- Les outils de recherche des systèmes d'exploitation, ne sont utilisés qu'en dernier recours, les utilisateurs préférant parcourir les répertoires (Barreau et Nardi, 1995, Ravasio et al. 2004).

- Les outils de recherche sont
 - Trop lents
 - Pas adaptés aux utilisateurs

" Existing search tools have never been designed for average users but rather for experts" (Ravasio et al., 2004)



Le choix des attributs et la manière dont ils sont exploités doivent être justifiées par des données empiriques sur l'utilisateur.

Etudes de la mémorisation des attributs des documents

- La conception des *outils de recherche* doit reposer sur l'investigation des mécanismes cognitifs sous-jacents à l'activité de *récupération*.
 - ➔ Investigation de l'activité cognitive de mémorisation et de rappel des attributs des documents
(Dumais & Jones, 1985; Jones & Dumais, 1986; Lansdale, 1988; Lansdale, 1991; Lansdale, Simpson & Stroud, 1990; Czerwinski et al. 1999)
- Limites des études précédentes: manque de validité écologique du matériel utilisé



Objectifs de notre étude

- De quels attributs les utilisateurs se rappellent-ils à propos de leurs "propre" documents personnels ?
De quelles informations précises relatives à un attribut donné les utilisateurs se souviennent-ils ?
- Quelle est la fiabilité du rappel des utilisateurs ?
Cela dépend-t-il du type d'attributs ?
- Comment les utilisateurs expriment-ils leur rappel ?
(sous quelle forme ? avec quel degré de précision ?)



Méthodologie

- Participants

 - 14 membres du personnel de L'INRIA :

 - 7 chercheurs / 7 administratifs

- Procédure

 - Phase de rappel des attributs

 - Spontané

 - Indiqué

 - 11 attributs : **localisation, type ou format, nom de fichier, titre, taille, date de dernière utilisation, mots-clefs, éléments visuels, évènements associés, liens et opérations**

 - Phase de récupération des documents



Résultats : Phase de “Rappel spontané”

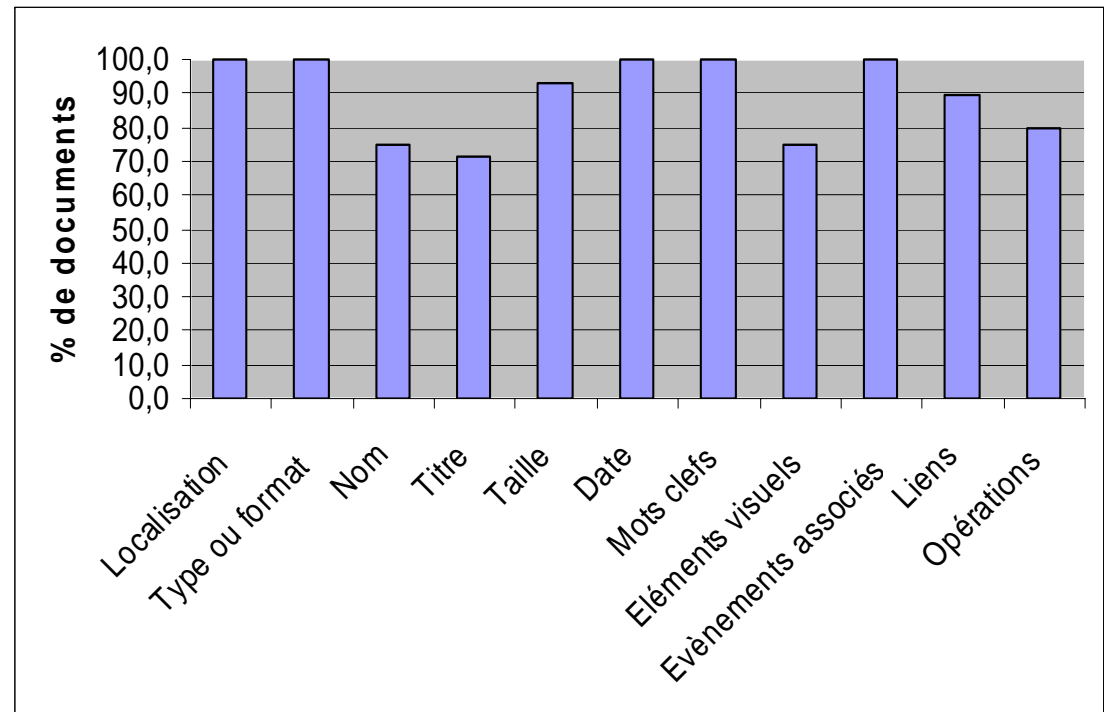
- Types d'attributs rappelés
 - **contenu textuel du document** (résumés du contenu, titre ou autres parties spécifiques du texte)
(71.4 %)
 - **éléments visuels** (couleurs, images)
(25%)
 - **type ou format** (e.g., "livre au format A5", "tableau Excel")
(21,4%)
 - + **taille, documents liés et auteur(s).**

- Moyenne du nombre d'attributs différents rappelés par utilisateur =1.46

Résultats : Phase de “Rappel Indicé”

□ Fréquence des attributs rappelés

- 5 attributs ont été rappelés dans tous les cas : **localisation, type ou format, date, mots clefs et évènements associés.**
- Pourcentage de rappel élevé pour tous les attributs



Pourcentages de documents pour lesquels chaque attribut a été rappelé

Résultats : Phase de “Rappel Indiqué”

□ Exactitude des rappels

- Les attributs les mieux rappelés sont **type ou format** et **éléments visuels**
- Le rappel des attributs est souvent **partiellement exact**

Pourcentages des documents pour lesquels le rappel concernant chaque attribut était totalement exact, partiellement exact, ou totalement faux.

	Rappel totalement exact	Rappel partiellement exact	Rappel totalement faux
Localisation	36%	32%	32%
Type ou format	93,3%	3,3%	3,3%
Nom	25%	66,7%	8,3%
Titre	33,3%	46,7%	20%
Taille	15,4%	30,8%	53,8%
Date	4,8%	47,6%	47,6%
Mots-clefs	32%	68%	0%
Éléments visuels	76,2%	23,8%	0%



Résultats : Phase de “Rappel Indicé”

- **Caractéristiques des rappels partiels**
 - **mots clefs, titre ou nom**
une partie exact mais l'autre non
 - **localisation**
exact jusqu'à un certain point (dossiers parents rappelés mais pas les derniers sous-dossiers)
 - **taille et date**
approximations (e.g., "entre 40 et 60 pages ", "en février 2005")

- **Caractéristiques des rappels faux**
 - **taille et date**
étendues des erreurs proportionnelles aux grandeurs rappelées



Résultats : Phase de “Rappel Indicé”

□ Moyens d’expressions utilisés

■ **date:**

- spécification d’une période (un mois dans la plupart des cas, e.g., “en février 2004”)
 - information sur le temps écoulé (e.g., “il y a deux mois”)
- etc.

■ **taille:**

- adverbes d’approximation (e.g., “environ 500 pages”)
 - intervalles (e.g., “entre 10 et 20 pages”)
- etc.



Résultats : Phase de “Rappel Indicé”

- **Caractéristiques des informations rappelées**
 - **éléments visuels:**

pas seulement dans la première page du document mais aussi (ou seulement) dans les autres pages
 - **opérations:**

modification;
impression, envoi et déplacement
 - **évènements associés:**

un peu plus de la moitié ont produit une "trace" électronique
= mails
 - **liens entre documents:**

pas seulement "abstrait" (i.e. de contenu ou de fonction) mais également "concret" (i.e. d'action comme le copier/coller ou de localisation)



Résultats : Phase de récupération

- La méthode la plus utilisée =
parcours dans la hiérarchie des répertoires + puis recherche visuelle du document cible
- Seul un utilisateur a utilisé la fonction "rechercher" de son OS parce que la stratégie classique avait échoué.
- 4 utilisateurs n'ont pas réussi à retrouver leurs documents (dont 3 documents papiers et 1 document électronique).
- Sur les 26 documents retrouvés, 8 n'ont pas été retrouvés facilement dont 2 papiers et 6 électroniques.
Dans la plupart des cas (5 sur 8) c'est la première phase qui posait problème



Implications pour la conception

- Favoriser les attributs appropriés
- Prévoir de la flexibilité dans les requêtes
e.g., mots-clefs
- Favoriser les expressions appropriées
e.g., temps, taille
- Permettre l'extensibilité des résultats
e.g., localisation, temps, taille
- Fournir des visualisations du contenu des documents
- Exploiter les relations explicites entre les documents
- Fournir les logs de certaines opérations effectuées sur les documents
- Permettre la combinaison semi-automatique d'attributs



Recherches futures

Tester l'utilisabilité et l'utilité de nouveaux systèmes dont la conception sera basée sur ces recommandations, par rapport aux systèmes actuels.