



Mesurer les comportements des utilisateurs des sites

Brigitte Trousse
INRIA - Projet AxIS



1. Qui sommes-nous ?

Projet AxIS

*Conception, Analyse et Amélioration
de Systèmes d'Informations
dirigées par l'usage*

*bi-localisation:
Sophia Antipolis et Rocquencourt*

création juillet 2003





L'équipe



4 Chercheurs Inria

- ✓ Brigitte Trousse (CR), resp scientifique
- ✓ Florent Masségia (CR)
- ✓ Yves Lechevallier (DR), resp permanent
- ✓ Thierry Despeyroux (CR)

2 Chercheurs en détachement

- ✓ Eric Guichard (ENS Ulm)
- ✓ Fabrice Rossi (Univ. Paris Dauphine)

2 Assistantes

- ✓ Sophie Honnorat (à 40 %)
- ✓ Stéphanie Aubin (à 25 %)

3 collaborateurs extérieurs/1 post doc

- ✓ Mireille Arnoux (MC, Univ Brest)
- ✓ Patrice Bertrand (MC, ENST Bretagne)
- ✓ Marc Csernel (MC, Univ Paris IX)
- ✓ Brieux Conan-Guez, 1/2 ATER Paris IX

4 doctorants

- ✓ Doru Tanasa (UNSA, 3 ème année, MENRT)
- ✓ Sergiu Chelcea (UNSA, 2ème année, PACA)
- ✓ Aicha El Golli (Paris IX, 3 ème année)
- ✓ Hicham Behja (ENSAM Meknès, 1 ère année)

✓ Mihai Jurca

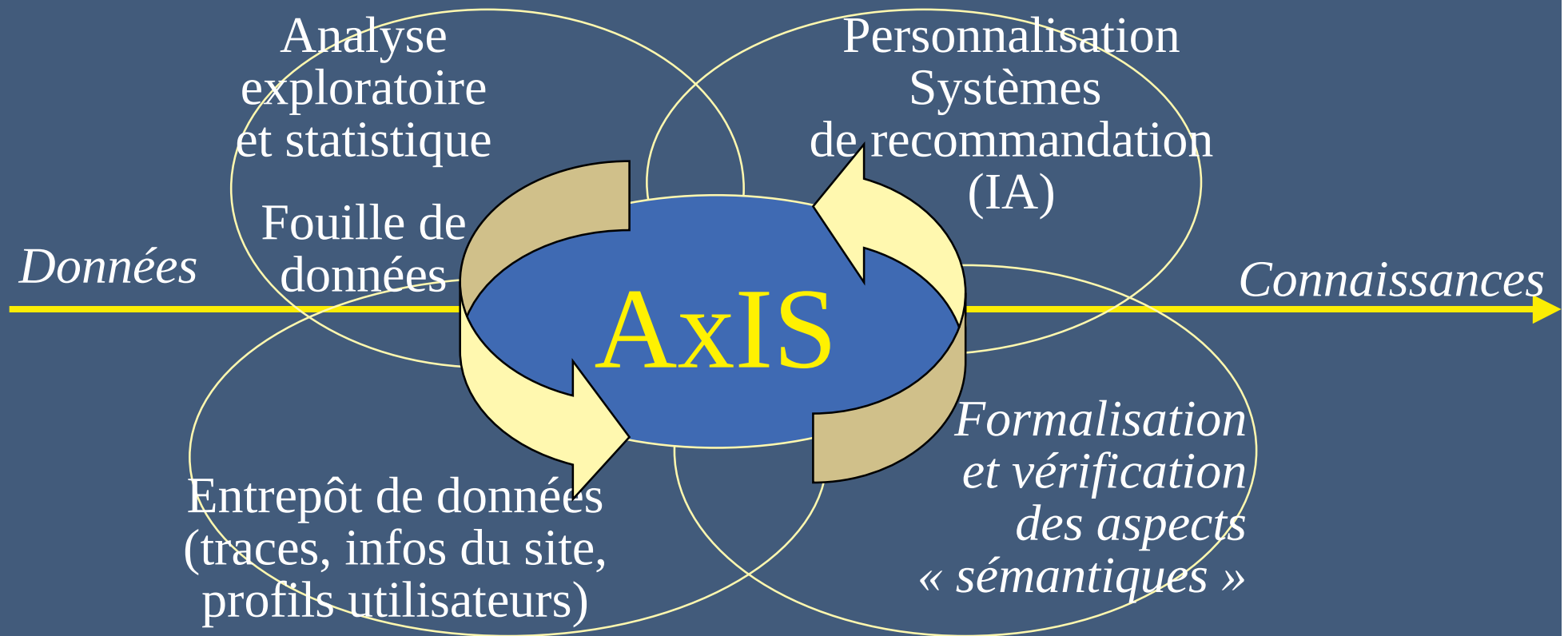
1 ingénieur expert

AS Mesures des flux de l'internet



Brigitte Trousse, 20 novembre 2003

Une démarche pluridisciplinaire



2. Web Usage Mining

- WUM - Web Usage Mining: ou comment mesurer les comportements des utilisateurs de sites?
 - ♦ Quelle démarche?
 - ♦ Quelles motivations de l'analyse des comportements?
 - ♦ Quelles données? Quelles méthodes de recueil?
 - ♦ Quels outils de pré-traitement?
 - ♦ *Quelles méthodes de fouille de données?*
 - ♦ *Quelles méthodes de validation des informations/connaissances extraites?*
 - ♦ *Quels indicateurs de mesure?*

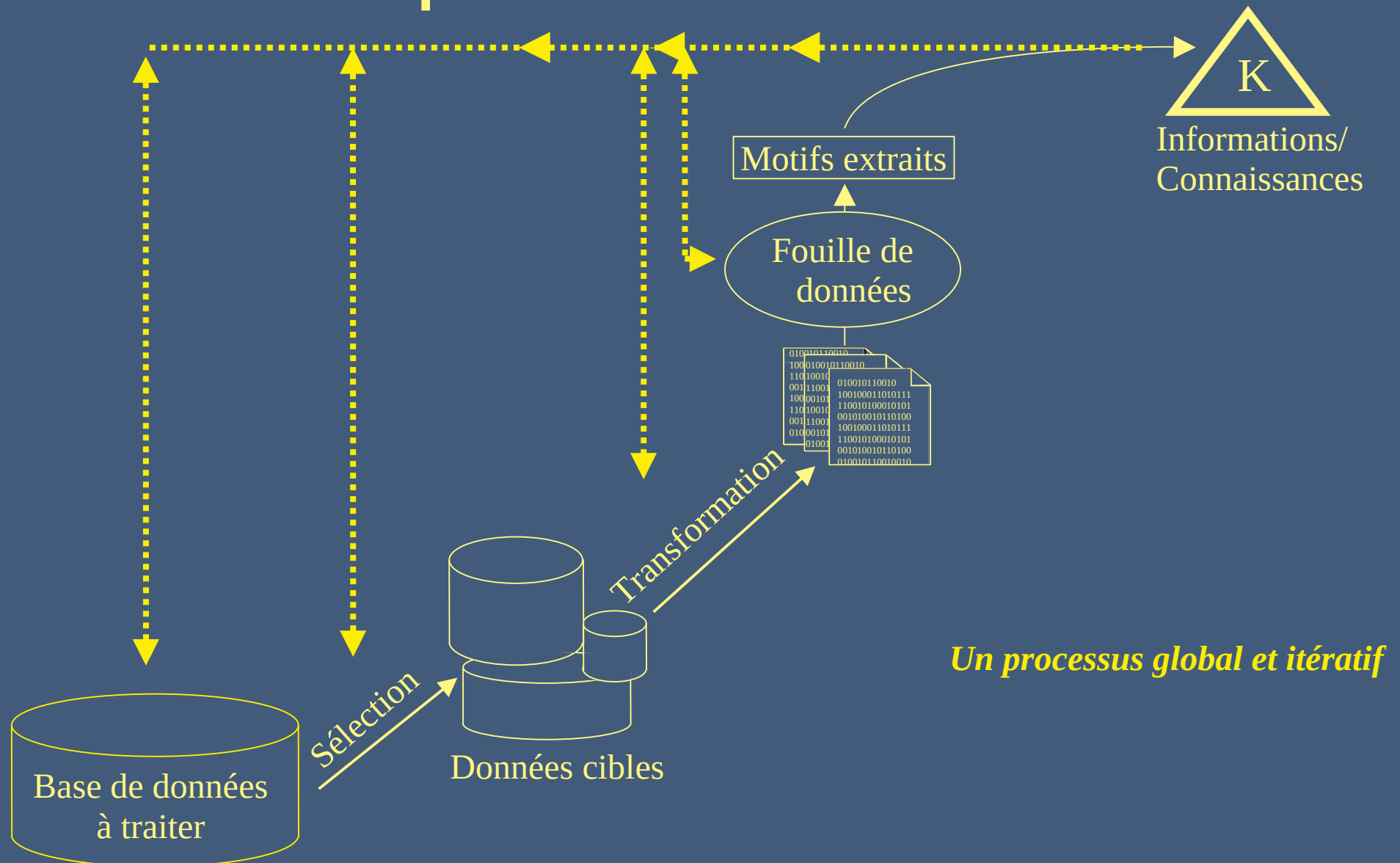
2.1 Quelques mots sur le Web Mining

❖ Application de l'ECD
(Extraction de Connaissances à partir de données)
aux données du Web

❖ Se décline en trois domaines d'application :

- Web Content Mining (contenu des documents)
- Web Structure Mining (structure des documents)
- Web Usage Mining (comportement des utilisateurs)

Le processus d'ECD



Quelques mots sur le Web Mining

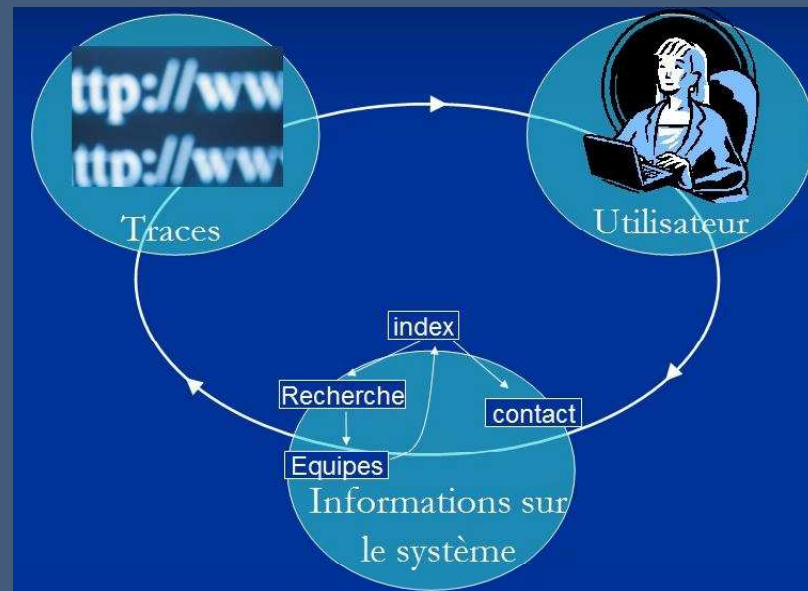
- ❖ **Motivations** pour le Web Usage Mining :
 - « Comprendre » l'internaute
 - Mesure d 'audience
 - Personnalisation des pages
 - Re-conception du site
 - Cache « intelligent »
 - ☆ Application d 'algorithmes de fouille!

2.2 Quel entrepôt de données pour une véritable analyse des comportements?

(Vers une combinaison des approches coté utilisateur et site)

- Pour un site WEB, trois sources d'informations peuvent être utilisées afin de prétendre à une analyse dédiée à ce contexte :

Logs HTTP (issus éventuellement de divers serveurs)
Traces structurées
Traces issues de panels



Profils utilisateurs

Données subjectives:
Enquêtes,
Questionnaires

Indicateurs des analystes

Caractérisation du contenu
Structure, carte du site
Informations syntaxiques
Informations sémantiques

2.3 Illustrations

- Traces structurées (visites, utilisateurs)

a) Broadway-Web (1997) : proxy (client), tag « image invisible » (site), applet (client), identification user, satisfaction explicite

b) Educaid (1998-2000 FT-Inria) : servlet (site), notion de visite, satisfaction vue par le concepteur

c) e-behaviour (2000) : servlet (site), système eye-tracking (client), notion de visite, satisfaction vue par le concepteur

- Logs HTTP (visites)

- pré-traitement des logs

a) Broadway-Web : aide à la navigation sur le Web

The image shows a Netscape browser window and a Broadway Manager Java applet. The browser window is titled "Netscape: AID (Action)" and displays the URL "http://www.inria.fr/Equipes/AID-eng.html". The browser's toolbar includes a "Broadway toolbar" with various icons. The main content area of the browser shows the title "AID (Action)" and the subtitle "Knowledge Management For Cooperative Design". Below this, there is a paragraph of text describing the AID research group at INRIA Sophia Antipolis. To the left of the text is a small image of four people. The Broadway Manager applet is a Java Applet Window with a menu bar (File, Navigations, Options, Help) and buttons for "Start navigation", "Stop navigation", "Advice upon request", "Continuous advice", "Request", "GO", "Explain", "Comments", and "Stats". The "Advice list" section shows a list of items: "77% ACTION AID - SOFTWARE", "73% INRIA - DEVELOPMENT AND INDUSTRIAL", and "38% ACTION AID - PUBLICATIONS".

Broadway Manager

File Navigations Options Help

Start navigation Stop navigation Advice upon request Continuous advice Request

Advice list:

- 77% ACTION AID - SOFTWARE
- 73% INRIA - DEVELOPMENT AND INDUSTRIAL
- 38% ACTION AID - PUBLICATIONS

GO Explain Comments Stats

Java Applet Window

Netscape: AID (Action)

File Edit View Go Communicator Help

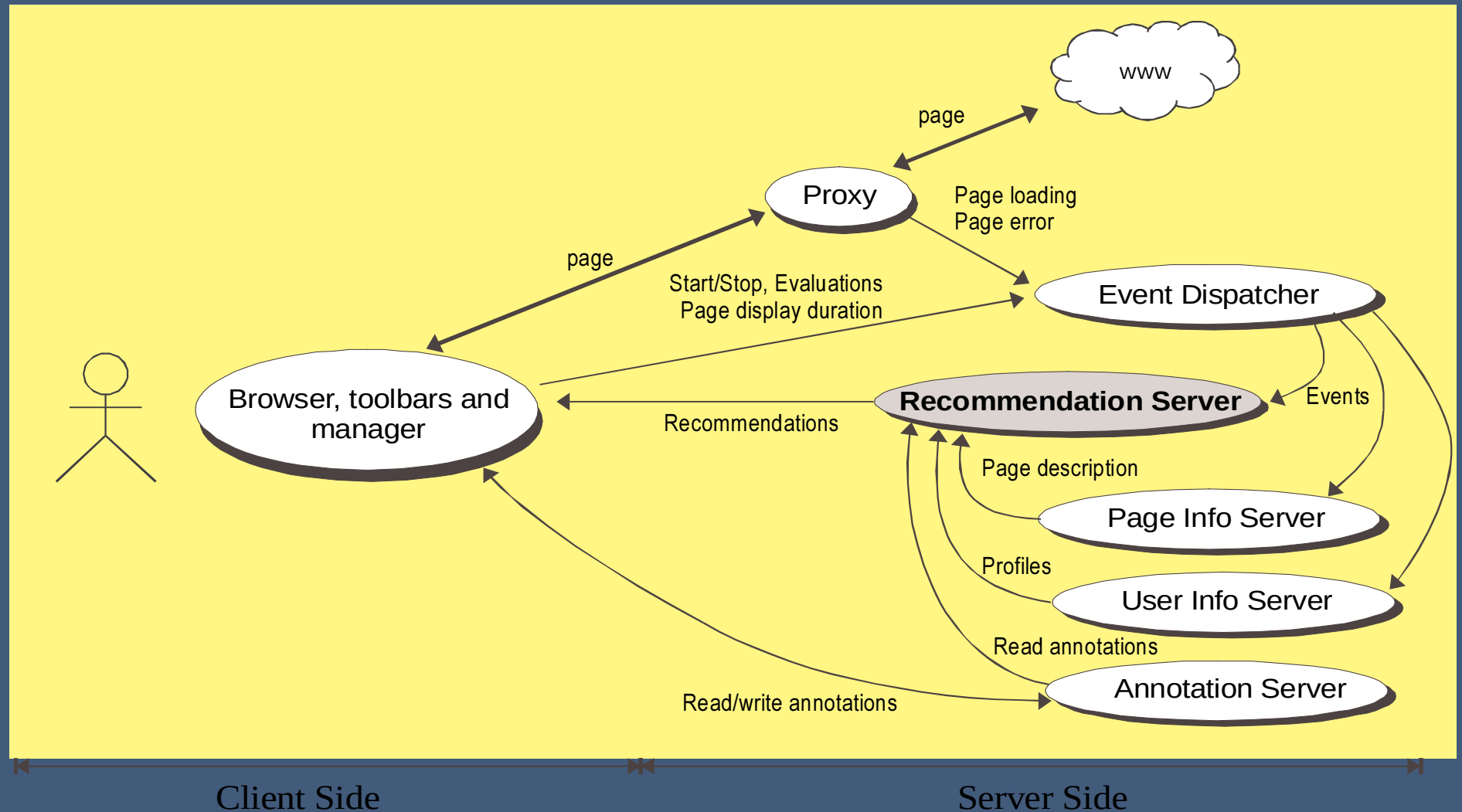
Bookmarks Location: <http://www.inria.fr/Equipes/AID-eng.html>

AID (Action)
Knowledge Management For Cooperative Design

Work in the AID research group of INRIA Sophia Antipolis focuses on reasoning techniques in Artificial Intelligence (AI), the aim being to develop models, techniques and tools for supporting Knowledge & Information Management for Computer-Supported Cooperative Work (CSCW) in Design applications on Internet/Intranet. Moreover, to construct an appropriate interpretation context of the information that is embedded in these processes or in past experiences, in order to make it more intelligible and increase its reusability.

To know

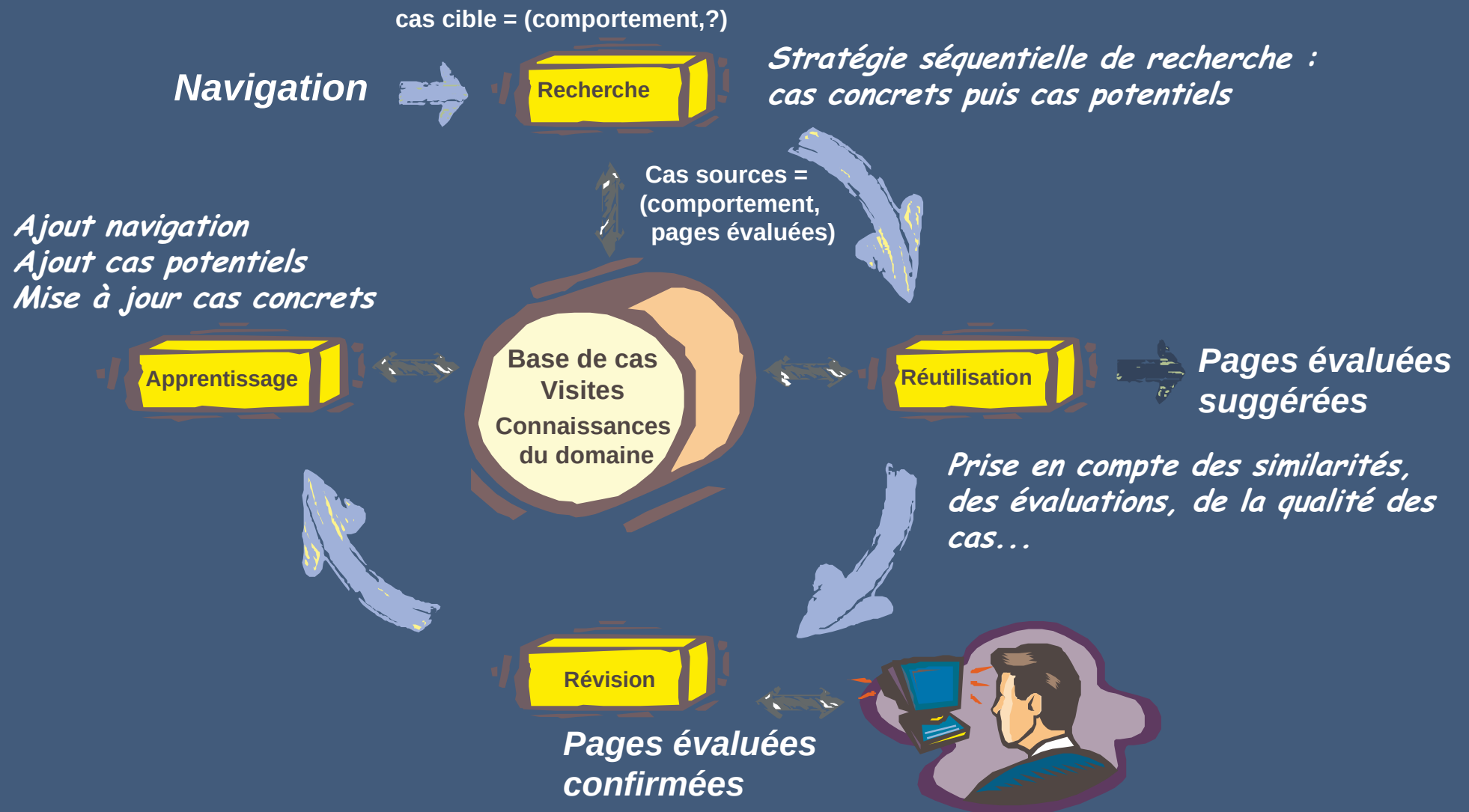
a) Broadway-Web : une approche « user & site centric » (1997)



a) Broadway-Web : représentation des visites

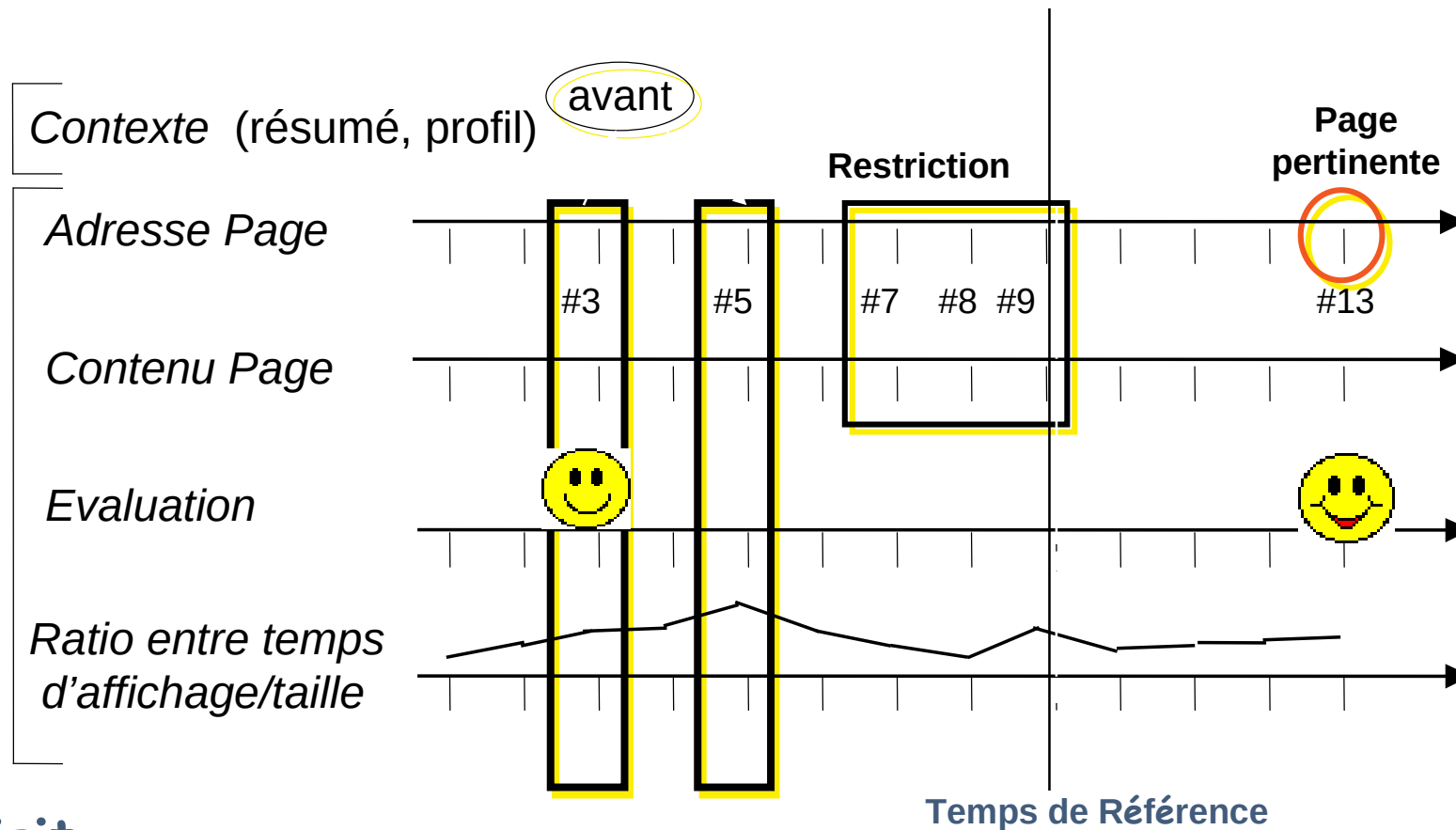
- Observation de 4 variables
 - ◆ Adresse (URL)
 - ◆ Contenu (racines des mots clés extraits)
 - ◆ Evaluation explicite (via les boutons d'évaluation)
 - ◆ Evaluation implicite (ratio temps affichage/ taille du document)
- Navigation d'un utilisateur
 - ◆ identification de l'utilisateur
 - ◆ Start/stop décidé par l'utilisateur
 - ◆ Evolutions des variables sauvées dans des séries temporelles
 - ◆ Résumé (union des mots clés et des sites visités)

a) Broadway-Web: Raisonnement dans Broadway-Web



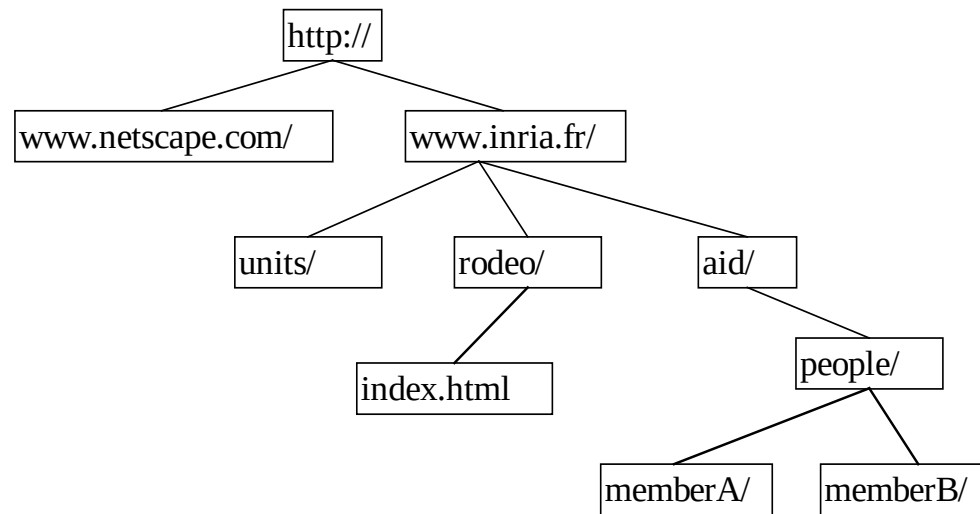
a) Broadway-Web: structure de “cas”

Cas = situation comportementale + liste de pages pertinentes



Visite

a) Broadway-Web : similarité entre adresses



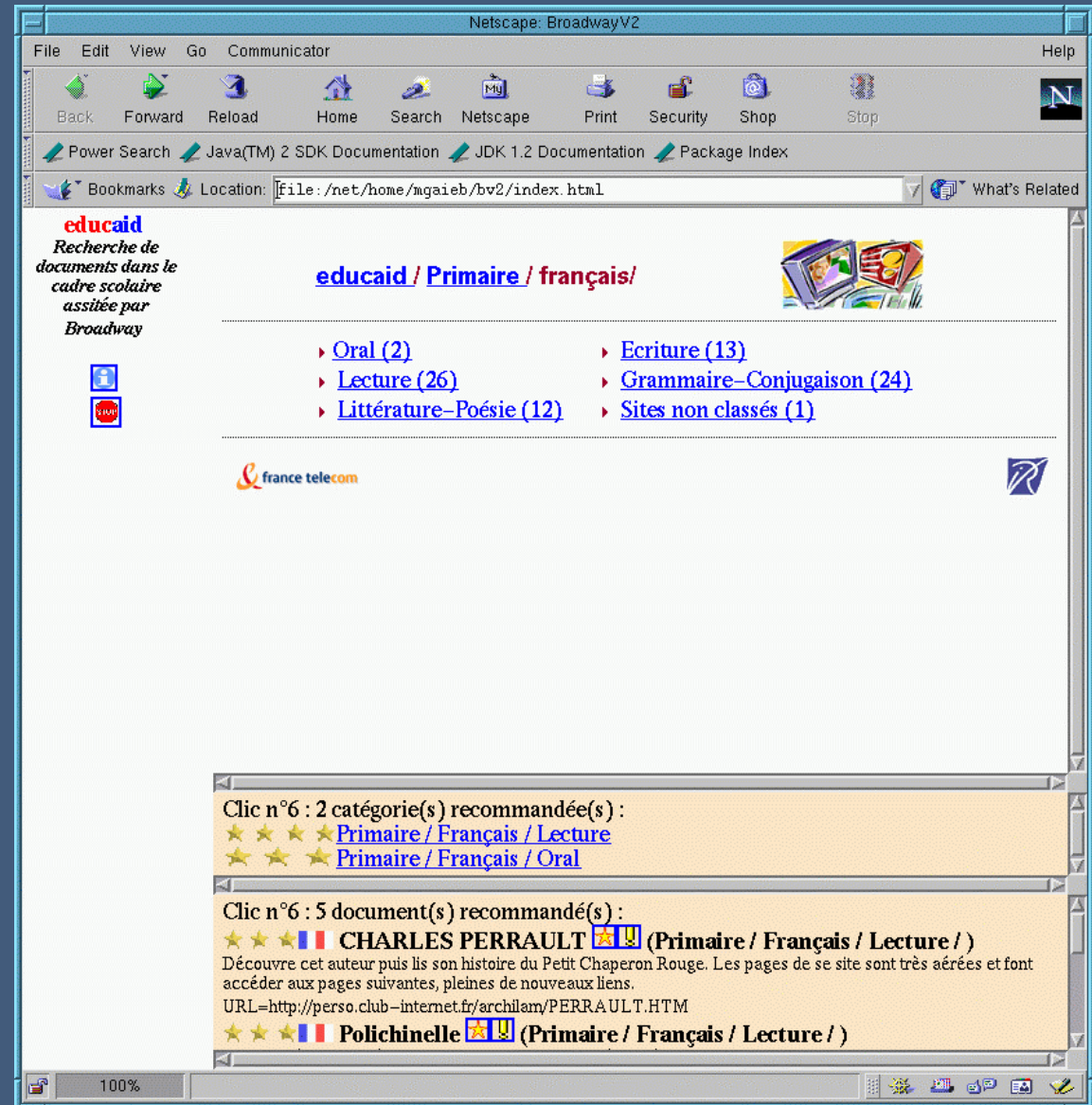
$$S_{ad}(p, q) = 1 - \frac{h(p, MSCA(p, q)) + h(q, MSCA(p, q))}{h(p, \text{root}) + h(q, \text{root})}$$

$S_{ad}(\text{http://www.inria.fr/aid/people/memberA}, \text{http://www.inria.fr/aid/people/memberB}) = 0,75$

$S_{ad}(\text{http://www.inria.fr/aid/people/memberA}, \text{http://www.inria.fr/rodeo/index.html}) = 0,28$

$S_{ad}(\text{http://www.inria.fr/rodeo}, \text{http://www.inria.fr/aid}) = 0,5$

**b) Educaid (FT-Inria
1998-2000):
aide à la navigation
dans un portail de
ressources éducatives**



c) Projet e-behaviour (Inria-UNSA)

Tâche 3 : Vous cherchez toutes les équipes du thème 3A ayant des activités d'enseignement ou de formation en ergonomie. Trouvez au moins un document par équipe présentant ces activités.

RA	 ra / 2001 / Theme 3A / UR Sophia Antipolis / acacia / Diffusion
<i>Recherche de documents dans le site des rapports annuels de l'INRIA</i>	Il y a 5 document(s)
Document trouvé:	▶ Animation de la Communauté scientifique ▶ Organisation de colloques et de cours ▶ Divers ▶ Enseignement ▶ Participation à des colloques, séminaires, invitations
Sortir :	
	
	
	Clic n°8 : 3 document(s) recommandé(s) :
	★★★★★ Enseignement (2001/Theme 3A/UR Sophia Antipolis/acacia/Diffusion/)
	★★★★★ Enseignement universitaire (2001/Theme 3A/UR Rocquencourt/eiffel/Diffusion/)
	★★★★★ Enseignement universitaire (2001/Theme 3A/UR Rocquencourt/merlin/Diffusion/)

c) Projet e-behaviour: données de comportement visuel

Projet e-behaviour - Représentation des données visuelles

Fichier Visualisation Analyse

Tache 3 : Vous cherchez toutes les équipes du thème 3A ayant des activités d'enseignement ou de formation en ergonomie. Trouvez au moins un document par équipe présentant ces activités.

RA

Recherche de documents dans le site des rapports annuels de l'INRIA

Document trouve: 

Sortir: 

[ra / 2001 / Theme 3A / UR Sophia Antipolis / aid / Diffusion](#)
Il y a 3 document(s)

- ▶ [Animation de la communauté scientifique](#)
- ▶ [Formation](#)
- ▶ [Participation à des colloques, séminaires](#)

Clic no 11 : 3 document(s) recommande(s) :

- ★ [Enseignement universitaire](#) (/ra/2001/Theme 3A/UR Rocquencourt/merlin/Diffusion/Enseignement universitaire)
- ★ [Aide à la modélisation et construction de mémoire d'entreprise](#) (/ra/2001/Theme 3A/UR Sophia Antipolis/acacia/Résultats nouveaux/Aide à la modélisation et construction de mémoire d'entreprise)
- ★ [Enseignement universitaire](#) (/ra/2001/Theme 3A/UR Rocquencourt/eiffel/Diffusion/Enseignement universitaire)

Suivant >> 12 sur 16 << Antérieur

The GIMP
File Xtns Help

Prétraitement des logs HTTP

- Introduction

- ♦ Exemple d'un fichier log Web
- ♦ Motivations
- ♦ WUM multi-sites

- Le prétraitement des données

- ♦ Nettoyage des données

- Suppression des requêtes pour les ressources Web non-analysées (N1)
- Suppression des requêtes provenant des robots Web (N2a,b,c)

- ♦ Transformation des données

- Fusionner les fichiers log ensemble (T1)
- Rendre anonymes les IP des utilisateurs (T2)
- Identification des utilisateurs (T3)
- Identification des vues de page (T4)
- Identification des visites (T5)
- Identification des épisodes (T6)

- ♦ Ajout d'indicateurs de mesure (variables agrégées)

Exemple de fichier log Web

Fragment d'un fichier log Web contenant 7 requêtes HTTP (en format ECLF)

194.78.232.8 - - [10/Sep/2001:15:33:43 +0200] "GET /orion/liens.htm HTTP/1.1" 200 1893 "http://www-sop.inria.fr/orion/index.html" "Mozilla/4.0 (compatible; MSIE 5.0b1; Mac_PowerPC)"

lucy.ins.cwi.nl - - [10/Sep/2001:15:34:07 +0200] "GET /stacs2002/ HTTP/1.0" 200 1012 "[unknown origin]" "Mozilla/4.74 [en] (WinNT; U)"

lucy.ins.cwi.nl - - [10/Sep/2001:15:34:07 +0200] "GET /stacs2002/home.html HTTP/1.0" 200 483 "[unknown origin]" "Mozilla/4.74 [en] (WinNT; U)"

lucy.ins.cwi.nl - - [10/Sep/2001:15:34:09 +0200] "GET /stacs2002/Images/affiche_vierge.jpg HTTP/1.0" 200 281281 "http://www-sop.inria.fr/stacs2002/home.html" "Mozilla/4.74 [en] (WinNT; U)"

194.78.232.8 - - [10/Sep/2001:15:34:09 +0200] "GET /orion/Telescope/Telescope.html HTTP/1.1" 200 4433 "http://www-sop.inria.fr/orion/liens.htm" "Mozilla/4.0 (compatible; MSIE 5.0b1; Mac_PowerPC)"

lucy.ins.cwi.nl - - [10/Sep/2001:15:34:10 +0200] "GET /stacs2002/cfp.html HTTP/1.0" 200 10334 "http://www-sop.inria.fr/stacs2002/home.html" "Mozilla/4.74 [en] (WinNT; U)"

194.78.232.8 - - [10/Sep/2001:15:34:23 +0200] "GET /orion/Telescope/Videosurveillance.html HTTP/1.1" 200 2979 "http://www-sop.inria.fr/orion/Telescope/Telescope.html" "Mozilla/4.0 (compatible; MSIE 5.0b1; Mac_PowerPC)"

Construction de deux sessions en considérant le même (IP, User Agent)

L'utilisateur provenant de 194.78.232.8 avec l'agent Mozilla/4.0 (compatible; MSIE 5.0b1; Mac_PowerPC)

/orion/liens.htm /orion/Telescope/Telescope.html /orion/Telescope/Videosurveillance.html

L'utilisateur provenant de lucy.ins.cwi.nl avec l'agent Mozilla/4.74 [en] (WinNT; U)

/stacs2002/ /stacs2002/home.html /stacs2002/cfp.html

Format standard d'un fichier log

[ip] [name] [login] [date] [url] [status] [size] [referrer] [agent]

- **ip** *adresse électronique de l'utilisateur*, cette adresse correspondant souvent au nom de domaine d'un serveur si l'utilisateur est connecté à internet via un fournisseur d'accès ou une entreprise
- **name/login** supposent que *l'utilisateur se soit lui-même identifié*
- **date** *date et heure* précises de réception de la requête.
- **URL** adresse de la page visitée sur le *site* (www.<...>)
- **statut** *code retour* qui indique si l'action s'est bien déroulée.
- **size** indique la *taille du fichier* retourné.
- **referrer** signale *l'adresse de laquelle a effectué la requête*, la page de provenance
- **agent** le *navigateur et le type de système d'exploitation* de l'utilisateur

Le WUM Multi-sites

- Une vision globale des actions faites par les utilisateurs sur les sites Web d'une même organisation
- Plus de 70% de visites contiennent des pages de deux sites Web (www.inria.fr & www-sop.inria.fr)
- Nombreuses rubriques transversales sur les deux sites (Recherche, Formation, Publications, Documentation, etc.)
- **Objectifs:**
 - ♦ Trouver les rubriques/pages où il manque des liens
 - ♦ Trouver les pages mal-placées dans l'hierarchie des sites Web

Le prétraitement des données

Les étapes de prétraitement

Unité de traitement	requête				session		visite		
Nettoyage des données		N1	N2a, N2b				N2c		
Transformation des données	T1			T2	T3	(T4)	T5		(T6)

Nettoyage des données

- N1** Suppression des requêtes pour les ressources Web non analysées
- N2** Suppression des requêtes/visites provenant des robots Web

Transformation des données

- T1** Fusionner les fichiers log
- T2** Rendre anonymes les IP des utilisateurs
- T3** Identification des utilisateurs
- T4** Identification des vues de page
- T5** Identification des visites
- T6** Identification des épisodes

Nettoyage des données (N1)

Suppression des requêtes pour les ressources Web non-analysées

- **Les ressources Web non-analysées, par exemple:**
 - ♦ Les fichiers images: "*.gif", "*.jpg", etc.
 - ♦ Autres fichiers qui «composent» une page (fichiers de style)
- Toutefois, il existe des images qu'on peut "cliquer"
- Utiliser la carte du site pour être sûrs

Nettoyage des données (N1)

Exemple INRIA

Trois mois de logs Web de quatre serveurs Web

Log \ Mois	Nov 2001	Dec 2001	Jan 2002	Total
Logs www.inria.fr	897	896	1069	2862
Logs www-sop.inria.fr	507	449	598	1554
Logs moteurs de recherche	3	7	6	16
Total	1407	1352	1673	4432
Nettoyé	602	680	760	2042
% dimension initiale	42.8	50.3	45.4	46.1

- Dimensions des fichiers logs (en MB) -

Nettoyage des données (N2)

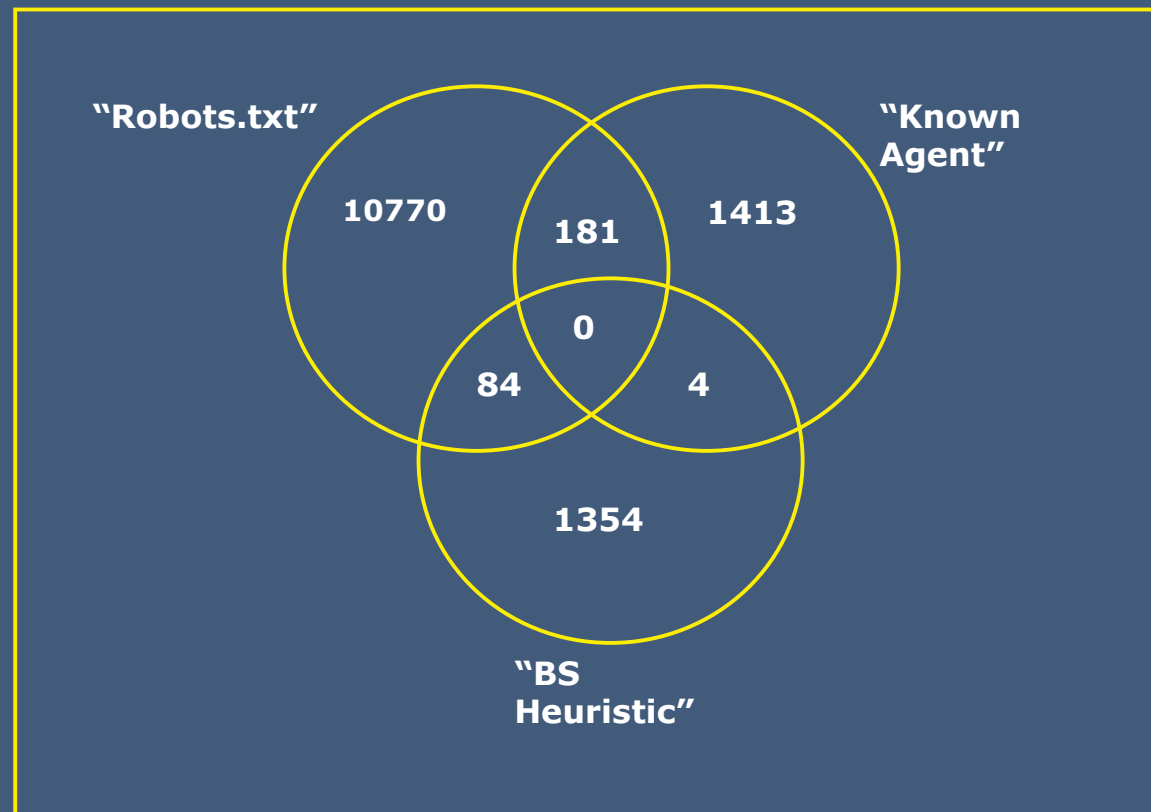
Suppression des requêtes/visites provenant des robots Web
(N2a,b,c)

- Les requêtes des robots Web représentent du «bruit»
- On réduit la dimension du fichier log de 50%
- Trois méthodes pour détecter les robots Web (WR):
 - ♦ Identifier les couples (IP, Agent) depuis lesquels il y a eu une requête pour "robots.txt" (N2a)
 - ♦ Utiliser une liste des agents connus comme WR pour identifier les couples (IP, Agent) dont l'agent fait partie de la liste (N2b)
 - ♦ Utiliser un seuil pour la vitesse de navigation (BS - « Browsing Speed »), qui est égale (N2c) :

$$BS = \frac{\text{Nombre des pages visitées}}{\text{Durée de la visite (secondes)}}$$

Nettoyage des données (N2)

Suppression des requêtes provenant des robots Web (N2a,b,c)



Total : 13268 hosts

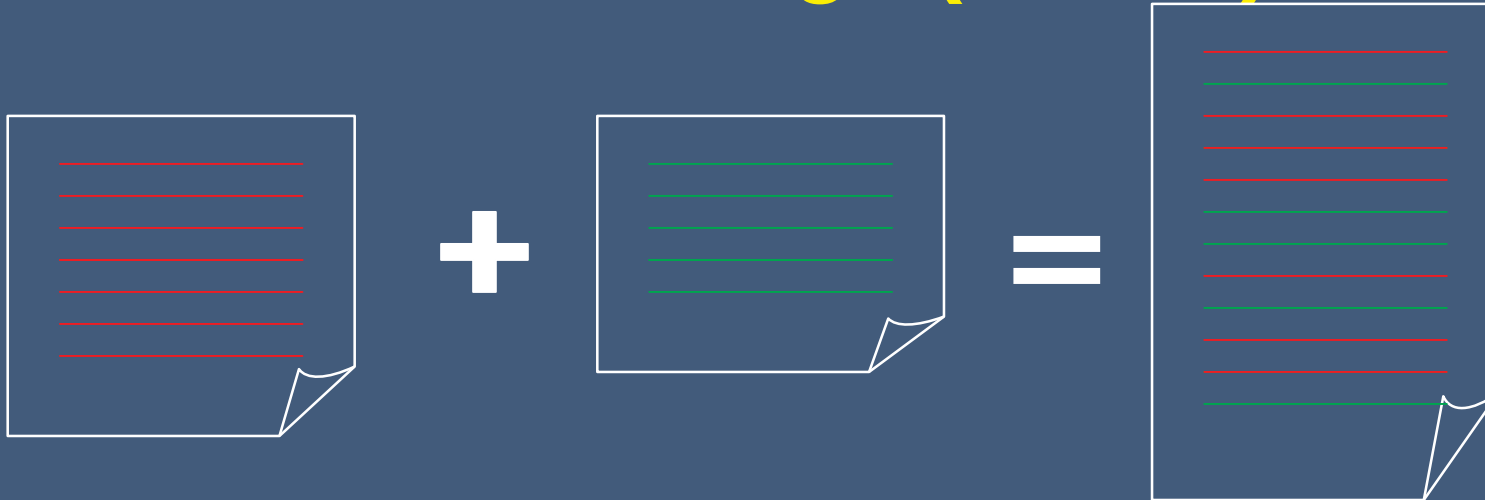
Nettoyage des données (N2)

Exemple INRIA

Méthode \ Mois	Nov 2001	Déc 2001	Jan 2002	Total	Total(%)
<i>Dimension initiale</i>	602	680	760	2402	100
N2a - robots.txt	416	430	518	1364	66.9
N2b - "User agent" connu	535	590	691	1816	88.9
N2c - BS	529	605	691	1825	89.4
Les trois	314	339	440	1093	53.5

Dimensions des fichiers logs Web après la suppression des WRs (MB)

Fusionner et rendre anonymes les fichiers logs (T1, T2)



- Les fichiers logs (ordonnés par la date de la requête) sont mis ensemble
- Chaque requête est modifiée pour:
 - ♦ inclure le "ID" du serveur Web dans la requête ("ID" = nom du serveur Web)
 - ♦ Rendre anonyme le IP/nom de la machine qui a effectué la requête

Fusionner et rendre anonymes les fichiers logs (T1, T2)

Exemple pour l'INRIA

www.inria.fr

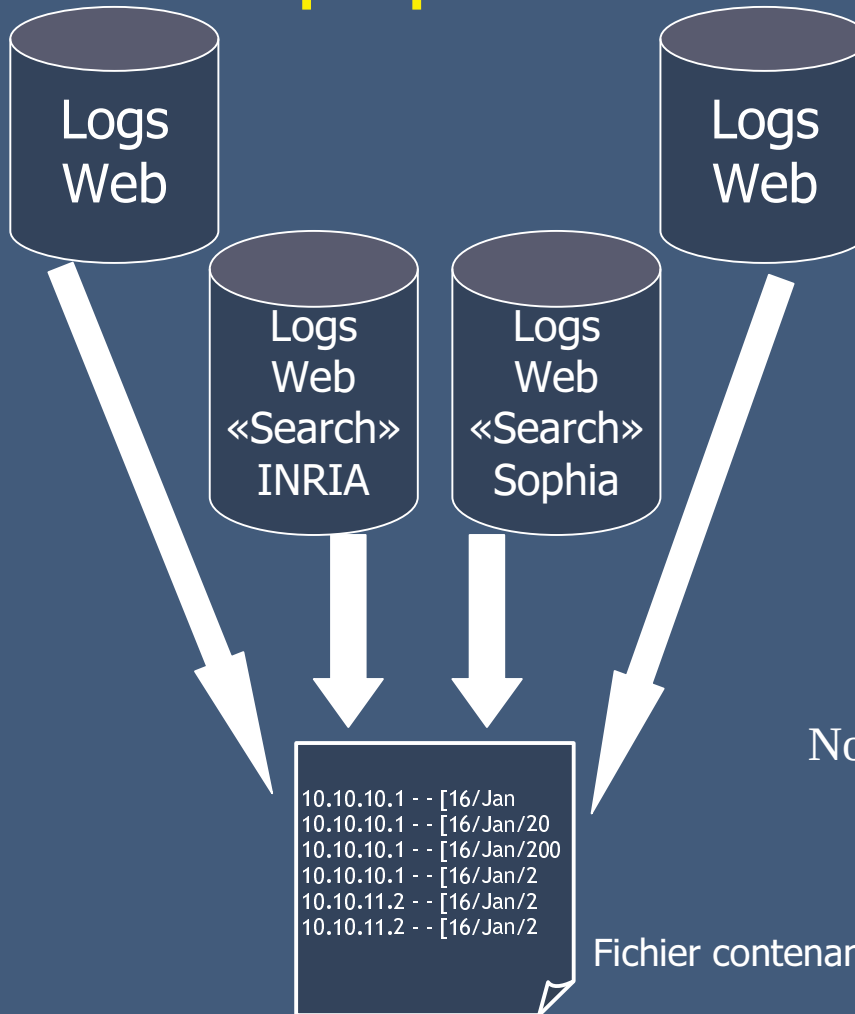


Serveur Web de l'INRIA national

www-sop.inria.fr



Serveur Web de l'INRIA Sophia Antipolis



10.10.10.1 - - [16/Jan
10.10.10.1 - - [16/Jan/20
10.10.10.1 - - [16/Jan/200
10.10.10.1 - - [16/Jan/2
10.10.11.2 - - [16/Jan/2
10.10.11.2 - - [16/Jan/2

Fichier contenant toutes les requêtes

Notion de visite multi-site

Transformation des données (T3)

Identification de l'utilisateur

Une tâche compliquée - raisons:

- ♦ Serveurs proxy
- ♦ Adresses dynamiques
- ♦ Cas d'utilisateurs utilisant le même ordinateur
- ♦ Mêmes utilisateurs utilisant plus d'un navigateur Web ou plus d'un ordinateur

Solutions possibles:

- ♦ Cookies
- ♦ Pages Web dynamiques (IDSession)
- ♦ Utilisateurs enregistrés
- ♦ Navigateur modifié
- ♦ Utilisation de la carte du site + le référeur

- Nous avons utilisé le couple (IP, Agent) ou/et name/login

Transformation des données (T4)

Identification des vues de page

- Grouper les requêtes générées par un même clic dans une seule requête !
- La carte du site est importante ici
- Plusieurs algorithmes proposés par Cooley (2000)

Transformation des données (T5, T6)

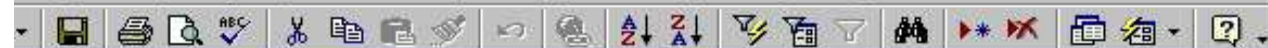
Identification des visites et épisodes

Identification des visites :

- Couper la session utilisateur si la distance entre 2 clics > 30 minutes (standard)
- Couper la session utilisateur si sa longueur dépasse 100 clics (un robot Web?)
- Un processus direct

Identification des épisodes :

- **Diviser la visite en épisodes sémantiques**
- On peut utiliser une carte du site Web amélioré qui contient une **description sémantique des pages Web**
- Un autre chantier en cours pour nous ...



NAVIGATION : Table

	IDNavigation	OldIDNavigation	NbRequests	NbRequestsDif	NbInstance	DureeTotale	Repetition
+	73	103	1	1	1	0	0
+	74	104	1	1	1	0	0
+	75	105	1	1	1	0	0
+	76	106	2	1	2	12	0,5
+	77	107	1	1	1	0	0
+	78	108	1	1	1	0	0
+	79	109	2	2	2	3	0
+	80	110	1	1	1	0	0
+	81	111	4	4	4	117	0
+	82	112	2	2	2	787	0
+	83	113	1	1	1	0	0
+	84	114	1	1	1	0	0
+	85	115	7	7	5	28	0
-	86	116	27	24	27	1440	0,1111111111

	IDRequest	IDSession	IDIP	IDUserAgent	IDDateTime	Duration	FileSize	IDStatus	IDURL	IDReferer	▲
▶	179	82	56887	26	15423	29	14204	1	37327	26984	
	180	82	56887	26	15432	27	10450	1	85604	121709	
	181	82	56887	26	15436	9	12221	1	85501	121709	
	182	82	56887	26	15439	24	10296	1	85518	139582	
	183	82	56887	26	15444	12	9882	1	85516	139582	
	184	82	56887	26	15446	1	5173	9	79117	139592	
	185	82	56887	26	15447	21	7616	1	79215	139592	
	186	82	56887	26	15450	42	22117	1	79202	138273	
	187	82	56887	26	15454	9	10453	1	79249	1	
	188	82	56887	26	15458	54	10453	1	79249	26984	
	189	82	56887	26	15465	342	299905	1	85511	139592	
	190	82	56887	26	15498	55	31282	1	85529	139592	
	191	82	56887	26	15502	108	35178	1	85524	139592	
	192	82	56887	26	15515	278	224477	1	85523	139592	
	193	82	56887	26	15550	22	10315	1	79243	138292	
	194	82	56887	26	15552	83	8636	1	79093	138286	

Conclusions

- Les données du Web se prêtent bien à la fouille de données
- Beaucoup de contributions en WUM consistent simplement à appliquer les algorithmes sur les logs (sans utiliser d'autres sources d'informations)
- Trop peu de réflexions sur les démarches de pré-traitement (combinaison approches site et user) et de validation de la qualité et de l'utilité des résultats
- Besoin d'une approche globale pluridisciplinaire pour répondre à ces problèmes



**Vers de véritables
« approches centrées utilisateurs »
(site, user) plus poussées...**

**Vers la gestion des points de vues
dans le processus d 'analyse des
comportements des utilisateurs**

Quelques références

URL: <http://www-sop.inria.fr/axis/>

Publications en 2003 (à partir de logs HTTP + infos sur site)

- Pré-traitement des logs HTTP

JFT 2003

- Classification automatique appliquée aux données Web

SFC 2003

- Deux méthodes d'extraction de tous les comportements utilisateurs d'un site supportées chacune par une plateforme

« Divide and Discover » : BDA 2003

« Cluster and Discover » : EGC 2004