



Université
Paul Sabatier
TOULOUSE III



Recherche d'information contextuelle

Recherche d'information personnalisée, mobile

Lynda Tamine Lechani

lechani@irit.fr

<http://www.irit.fr/~Lynda.Tamine-Lechani/>

Partie 1 RI contextuelle : notions et périmètres

- Introduction au domaine
- Taxonomies du contexte en RI

Partie 2 RI personnalisée, RI mobile

- Approches de RI
- Modèles de contextes
- Modèles de *ranking*
- Questions ouvertes



Partie 1 :

RI contextuelle : notions et périmètres



- La Recherche d'information aujourd'hui face :
 - à la diversité (média, sources, utilisateurs)
 - au volume (sources, utilisateurs, contenus générés)
 - ambigüité des requêtes
 - degré de subjectivité de la pertinence
 - ...

■ Diversité des services et profils d'accès à l'information

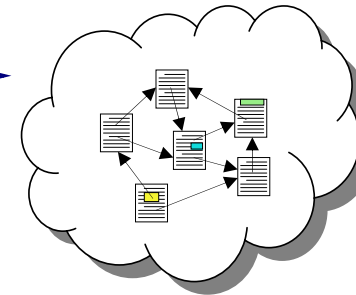
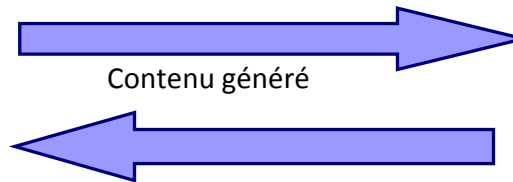




■ Diversité des sources d'information, des profils des utilisateurs

250, 000 sites

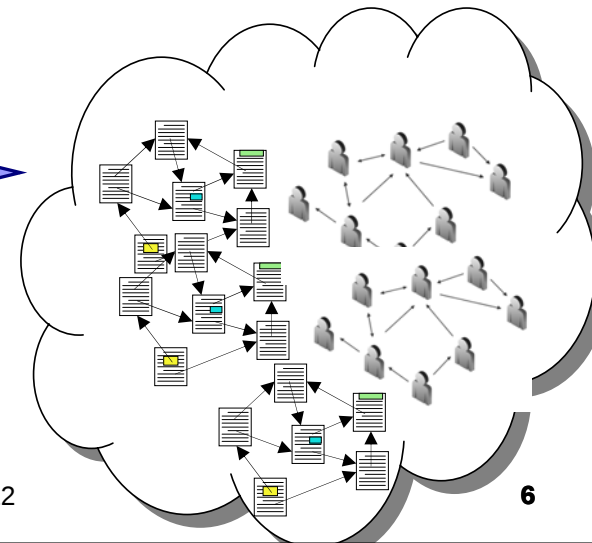
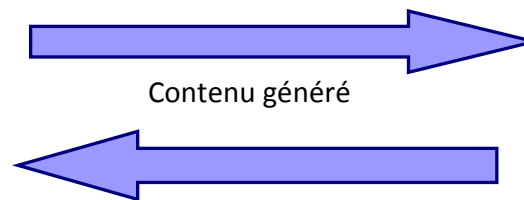
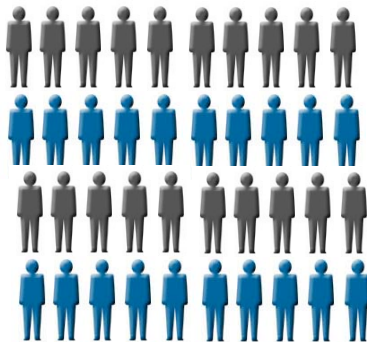
Web 1.0 en 1996



45 millions utilisateurs

80,000,000 sites

Web 2.0 en 2009



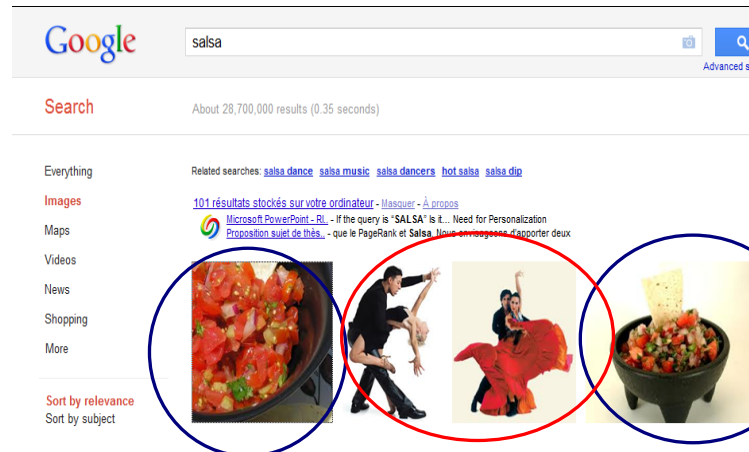
1 billion utilisateurs



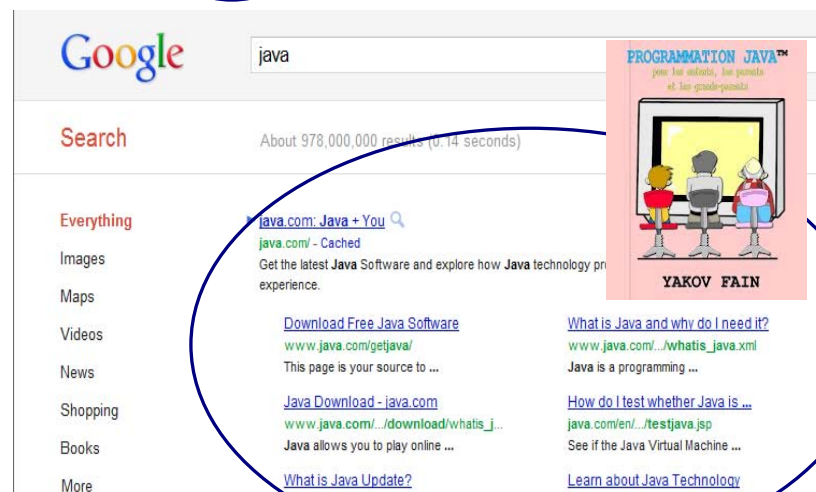
■ Le web en 2009 : quelques chiffres (source : rankstat.com)

- 21.96 billions de pages
- Un taux de croissance de ~ 8 millions de pages/jour
- Volumes indexés en croissance
 - *Google* : 15.5 billions de pages
 - *MSN Live search* : 3.6 billions de pages
 - *Yahoo* : 50 billions de pages
- Volumes des résultats d'une requête importants
 - La requête « JO » retourne 185,000,000 résultats
 - La requête « Elections France 2012 » retourne 159,000,000 résultats
- Requêtes courtes **donc ambiguës** ?
 - Plus d'1/4 (~28%) des requêtes soumises aux moteurs ont 2 termes
 - Environ 1/6 (~16%) des requêtes soumises aux moteurs ont 4 termes

■ Ambiguïté des besoins à l'information (système d'accès à l'information)

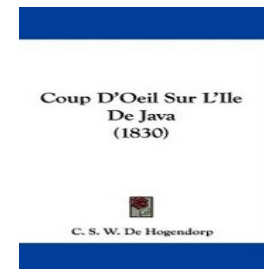


Ambiguïté sémantique



?

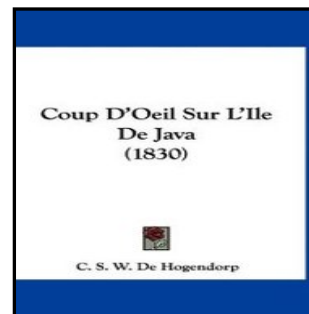
?





■ Ambiguïté des besoins à l'information (système d'accès à l'information)

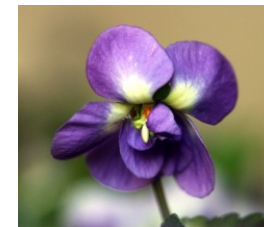
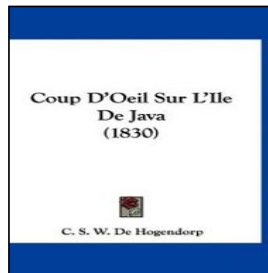
Ambiguïté liée au point de vue de l'utilisateur





■ Ambiguïté des besoins à l'information (système d'accès à l'information)

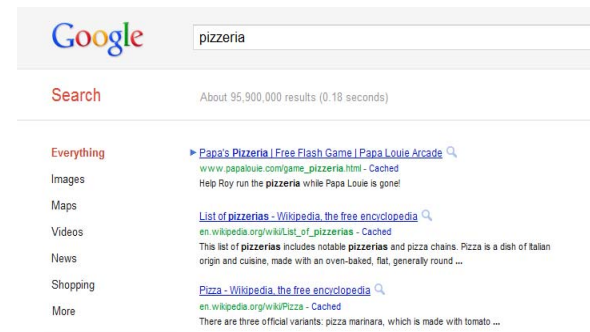
Ambiguïté liée à la langue, à la culture



■ Ambiguïté des besoins à l'information (système d'accès à l'information)



Ambiguïté liée au lieu





- ...Résultats d'études empiriques
 - Autour des utilisateurs et de leur expression des besoins
 - Les utilisateurs ne font pas d'efforts pour affiner leurs besoins en information (Jansen & al, 2000)
 - Même en faisant des efforts, l'expression des besoins n'est pas toujours claire (Spink & Jansen, 2004)
 - Tâche d'expression des besoins difficile même pour des utilisateurs expérimentés (Leake & Scherle, 2001)
 - Autour de la pertinence
 - Différents types de pertinence (Saracevic 2007, Borlund 2003, Mizzaro, 1997)
 - Pour un même type de pertinence, différents facteurs (Nielson 1998, Saracevic, 2007)

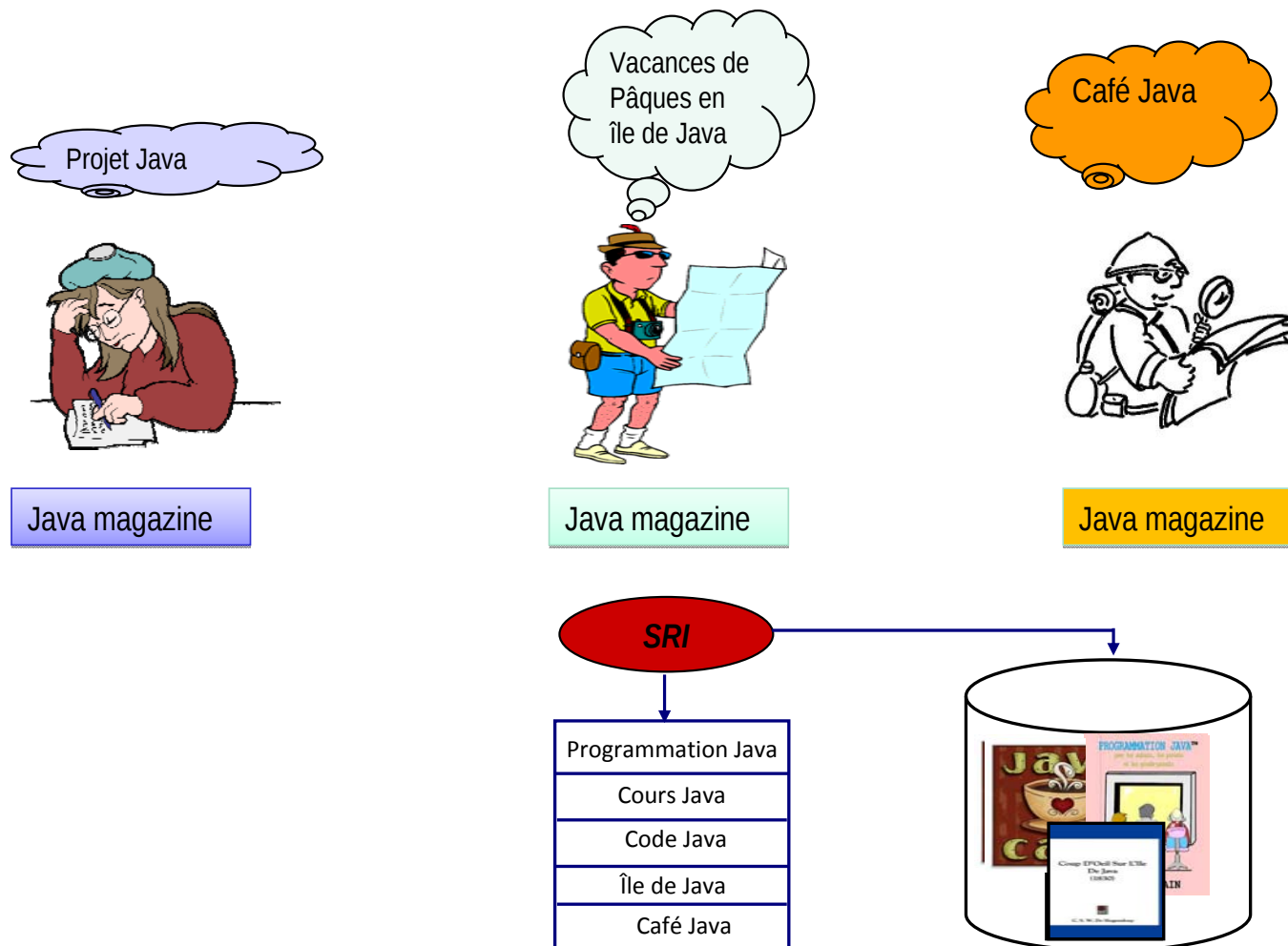


■ Problèmes de la recherche d'information centrée requête

- Résultats impersonnels : approche « *One size fits all* » (Laurence 2000)
 - Dégradation de la précision
- Informations disponibles mais pas accessibles ?
 - Paradoxe avec les volumes d'information qui croissent
- Résultat pertinent pour la requête mais pas pour l'utilisateur, pas pour sa tâche, pas pour son lieu, ...

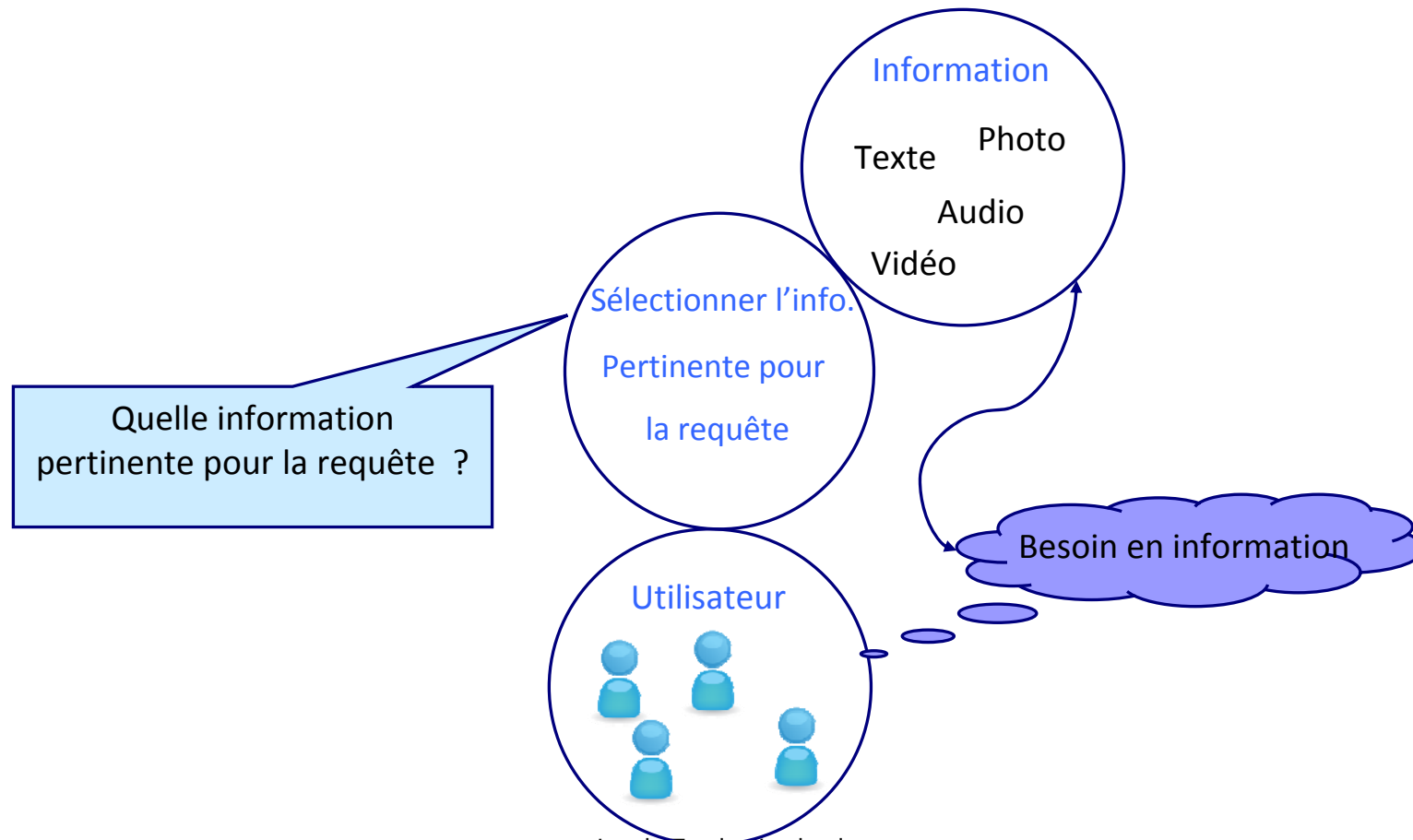
■ Défis à relever

- Fournir des résultats « adaptés » à l'utilisateur (Individu)
 - Point de vue de l'utilisateur : son expérience (niveau de connaissances, tâche etc.), ses préférences (langue, type de l'information etc.), sa localisation, etc.
 - Son infrastructure matérielle : mobile, pc portable, pc fixe etc.
 - Son groupe
 - Etc.

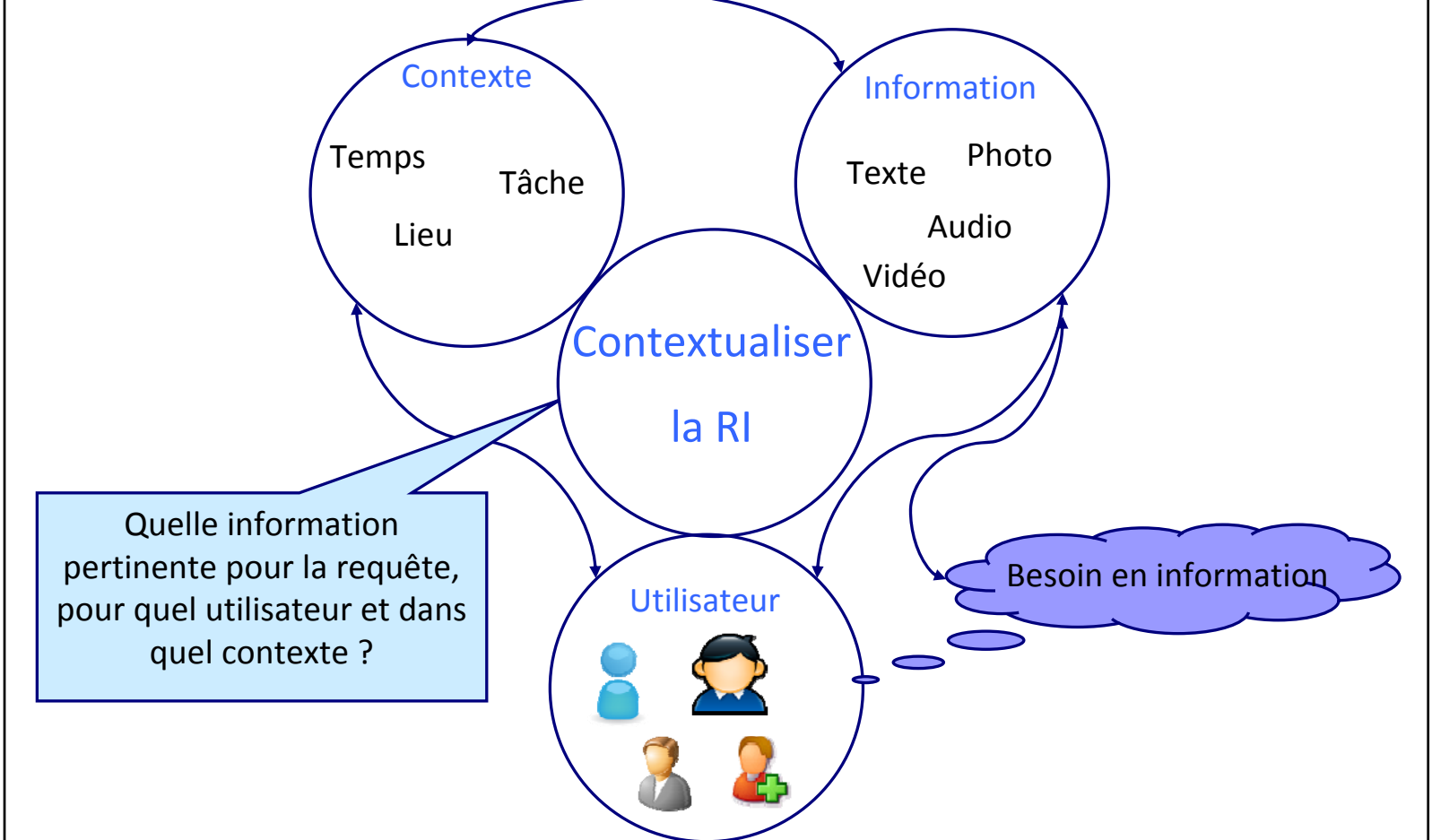




De la RI centrée requête...



À la RI centrée contexte...





■ Analyser le triplet {requête, utilisateur, contexte}

1. Interpréter le sujet de la requête

- Selon son **contenu** (termes) → approche classique
- L'**utilisateur** qui a exprimé cette requête
 - ✓ Analyser son historique de recherche (à quoi s'intéresse-t-il généralement?)
 - ✓ Analyser son intention de recherche (que cherche-t-il précisément ?)
 - ✓ Analyser ses préférences (langue, genre etc.)
- Le contexte de la situation : temps, lieu, matériel, météo, etc.

2. Estimer la pertinence du résultat selon cette interprétation

- Combiner, intégrer, ...
 - ✓ Pertinence liée au sujet de la requête
 - ✓ Pertinence liée à l'utilisateur
 - ✓ Pertinence liée au contexte de la recherche
 - ✓ ...

Niveaux de Pertinence
(Borlund, 2003)



■ A propos du contexte: **pas de définition consensuelle**

➤ *Webster's New Twentieth Century Dictionary (1980)*

« whole situation, background or environment relevant to some happening or personality »

➤ *Dey and Abowd (1999)*

« where the user is, who the user is with, and what resources are nearby »

➤ *Dey and Abowd (2000)*

« any information that can be used to characterize the situation of entities (i.e., whether a person, place, or object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves. Context is typically the location, identity, and state of people, groups, and computational and physical objects »



■ A propos du contexte en RI

➤ Jarvelin (2004)

« Whatever has important relationships with the application of IR, should be »

➤ Kelly (2006)

« Context is one of the most important concepts in information seeking and retrieval research. However, the challenges of studying context are great; thus, it is more common for researchers to use context as a post hoc explanatory factor, rather than as a concept that drives inquiry »



■ A propos de la RI contextuelle

➤ Allan (2003)

«Combine search technologies and knowledge about query and user context into a single framework in order to provide the most 'appropriate' answer for a user's information needs »

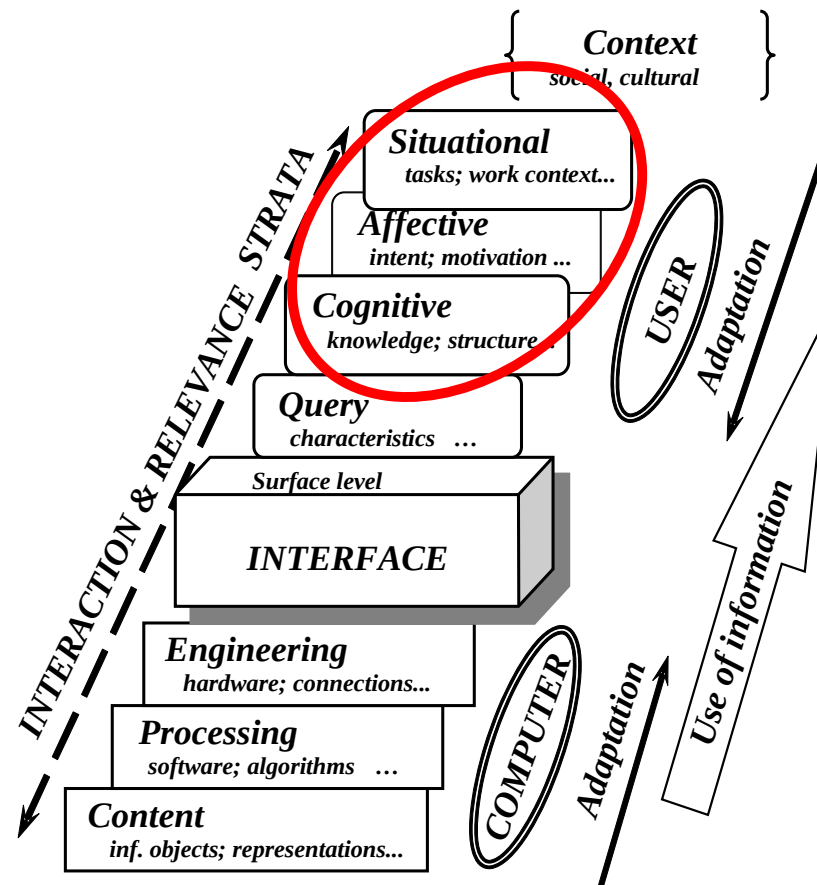
➤ Ingwersen & Belkin (2004) suggested IR in context means one of two things:

- a) *the design of IR methods that are context-aware, that is, able to elicit the implicit knowledge of the participatory searcher environment related to the context into an explicit representation*
- b) *the specification models for context and its features that can be used to focus an IR task according to such a context.*

■ Plusieurs taxonomies, modèles

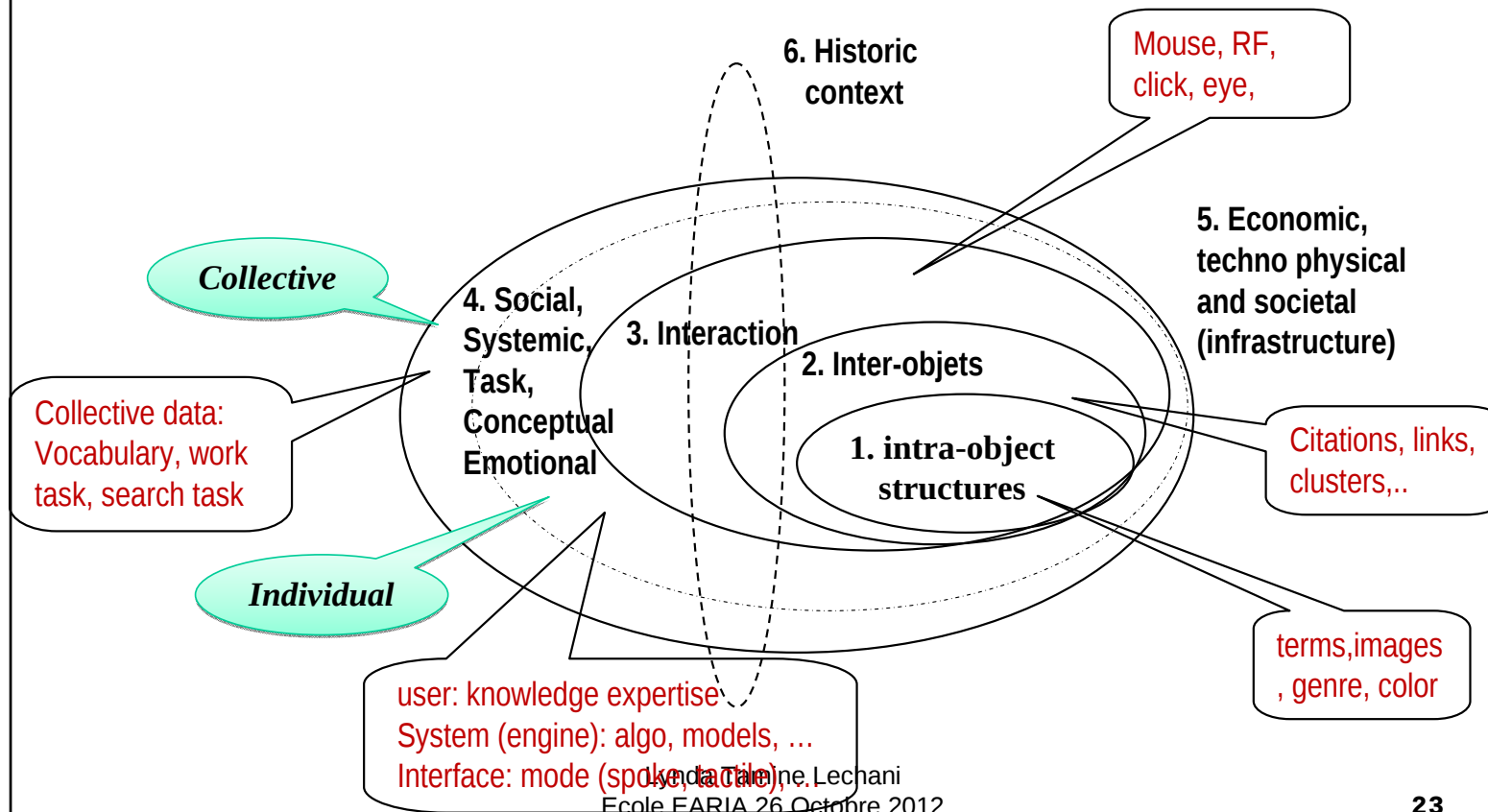
- Taxonomie Fuhr (2000)
- Taxonomie Dey & Abowd (2000)
- Taxonomie Saracevic (1997)
- Taxonomie Ingwersen and Jarvelin (2005)
- Taxonomie Tamine and al (2010)
- Taxonomie Goker and Maryhuang (2002)
- Taxonomie Kaenampornpan and O'Neill (2004)
- Taxonomie Arias and al (2010)
- ...

Modèle stratifié de Saracevic 1997

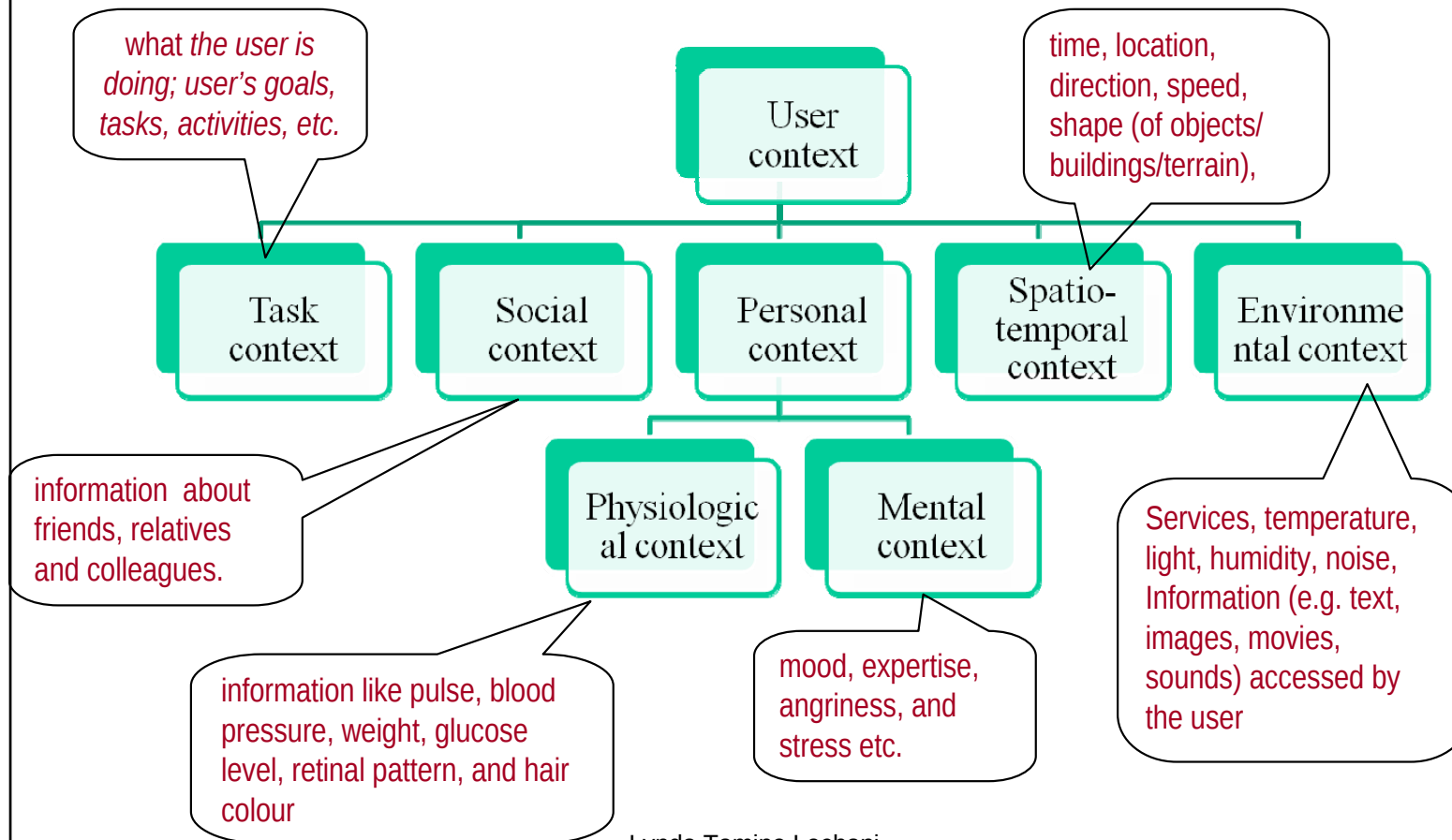




Modèle (nested model) de Ingwersen et Jarvelin 2005

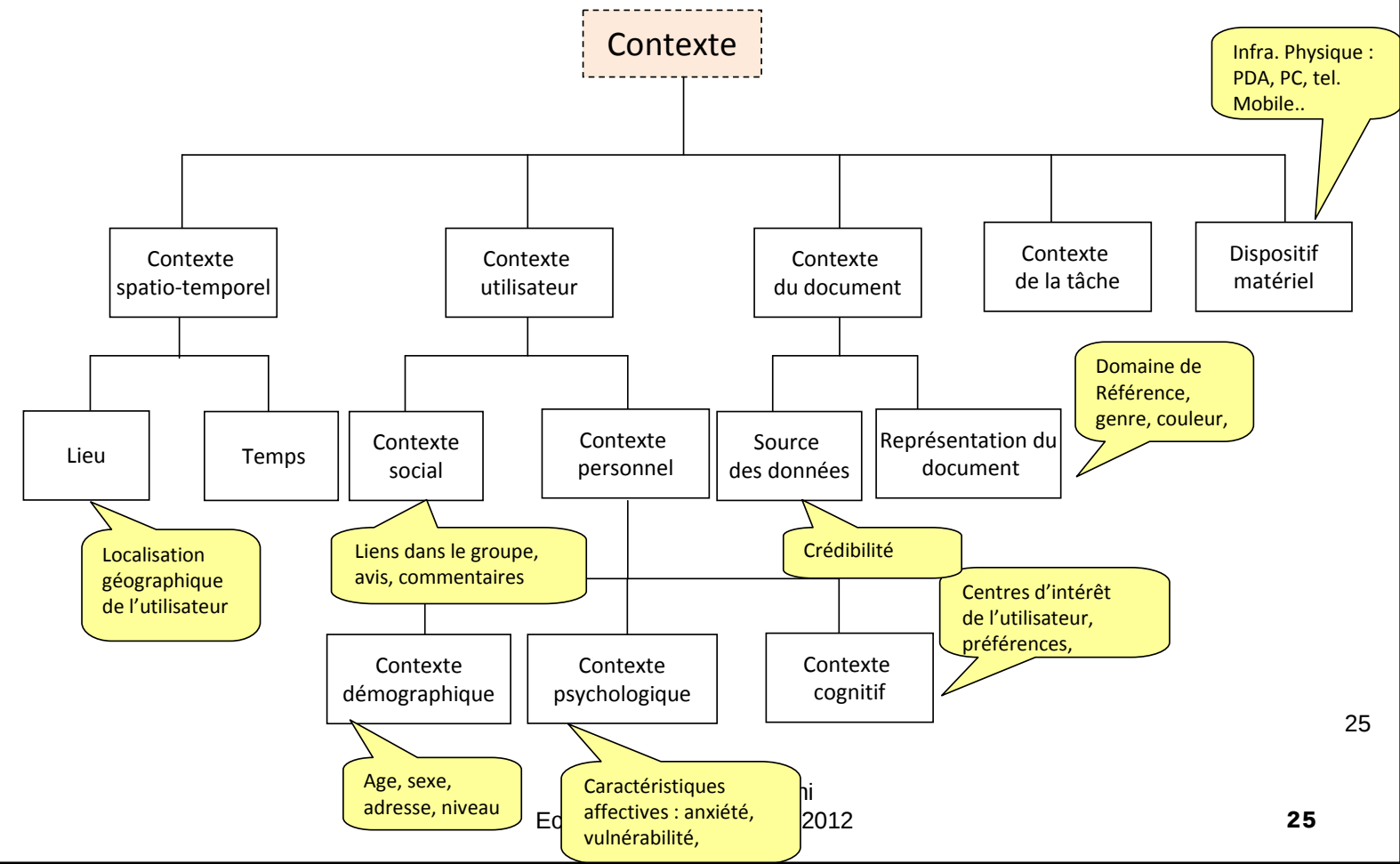


Modèle de Goker et Myraugh 2002





Modèle de Tamine et al. 2010



■ En résumé...

➤ De nombreuses dimensions liées à

- o L'utilisateur
 - o son expertise, ses centres d'intérêts, son affectivité, ...
- o la tâche
- o la localisation
- o le groupe social
- o ...

➤ Quelles dimensions exploiter ?

- o Une seule ?
- o Plusieurs : comment les combiner ?

➤ Pour quelle tâche ?

- o Recherche d'information
- o Recommandation

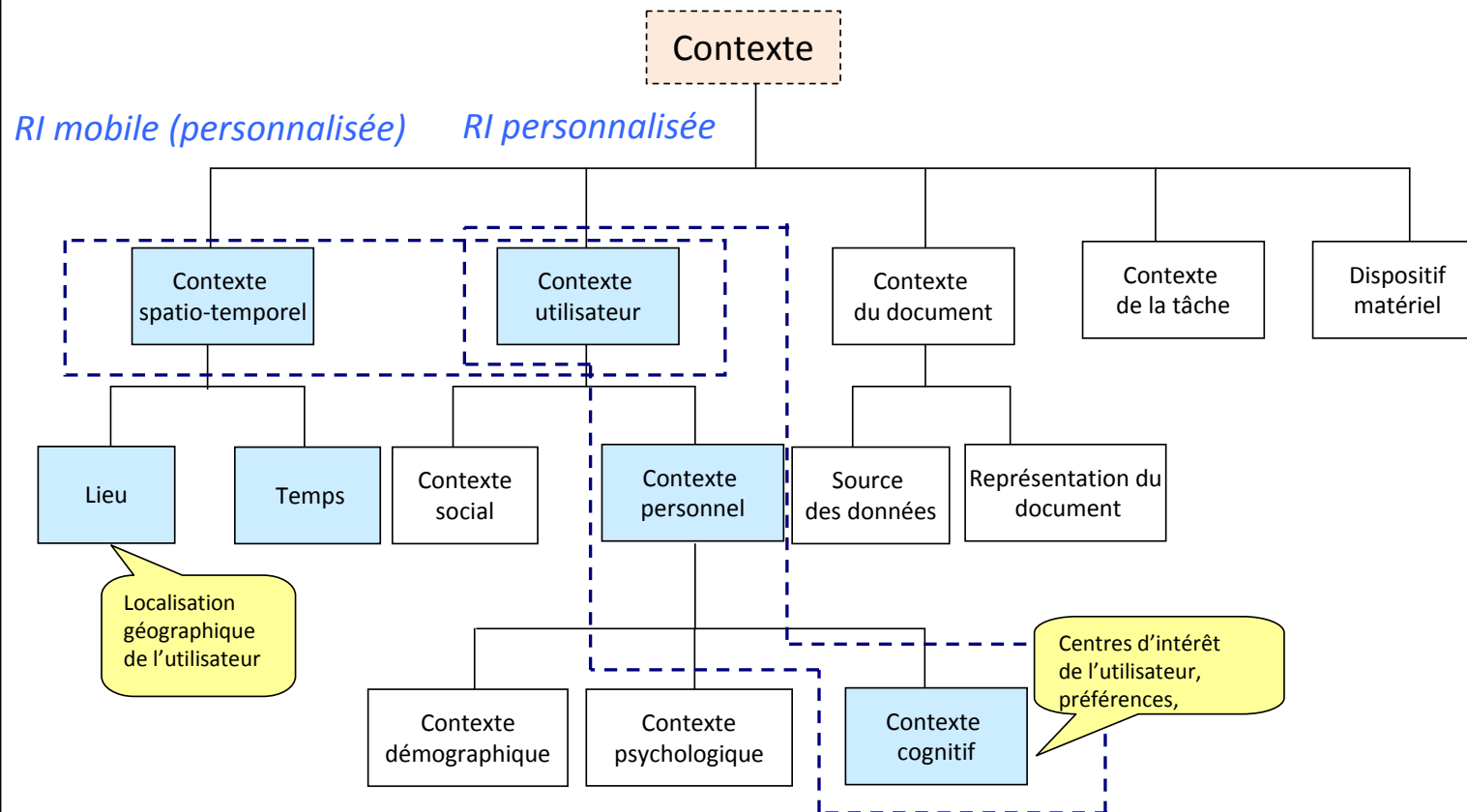
RI collaborative,
filtrage collaboratif,
RI affective,
RI mobile,
RI orientée tâche,
RI sociale,
RI affective,
RI personnalisée,

..

26

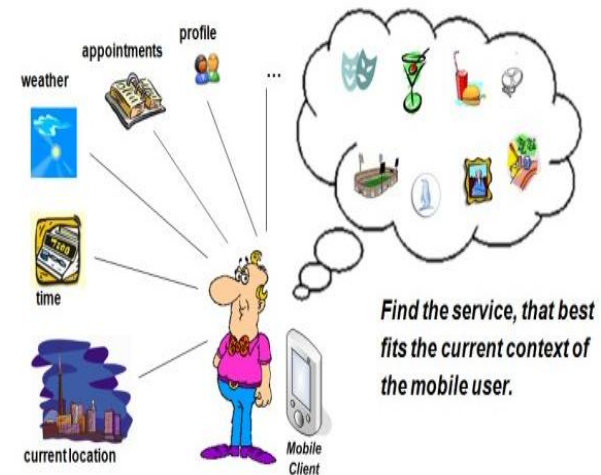


Périmètre du cours



■ Défis de la RI mobile

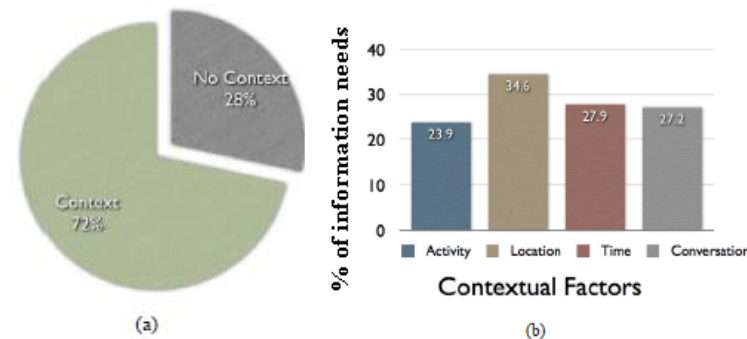
- Adaptation à l'appareil mobile
 - Caractéristiques d'affichage
 - Types de capteurs
 - Capacité mémoire
 - Bande de transmission
 - ..
- Adaptation aux facteurs contextuels
 - Localisation géographique
 - Centres d'intérêts
 - Temps
 - Objets environnants
 - ..



■ Quel intérêt à personnaliser la recherche pour l'utilisateur mobile ?

➤ Quelques études phares

- ✓ Kamvar and Baluja 2006, Church and Smith 2009, Sohn et al. 2008, Bierig and Göker 2006



➤ Principaux résultats

- ✓ 72% des requêtes expriment des besoins sensibles à la localisation
- ✓ Requêtes courtes (2.3-2.5 termes/requête), donc ambiguës
pas de différence significative avec les *desktop* queries
- ✓ Sessions courtes (1.6 requêtes/session)
- ✓ 28,7% de requêtes successives sont considérées comme des reformulations des requêtes précédentes



■ Quel intérêt à personnaliser / activités, centres d'intérêt ?

➤ Quelques études phares

Teevan et al., 2010, Dou et al. 2007, Teevan et al. 2008

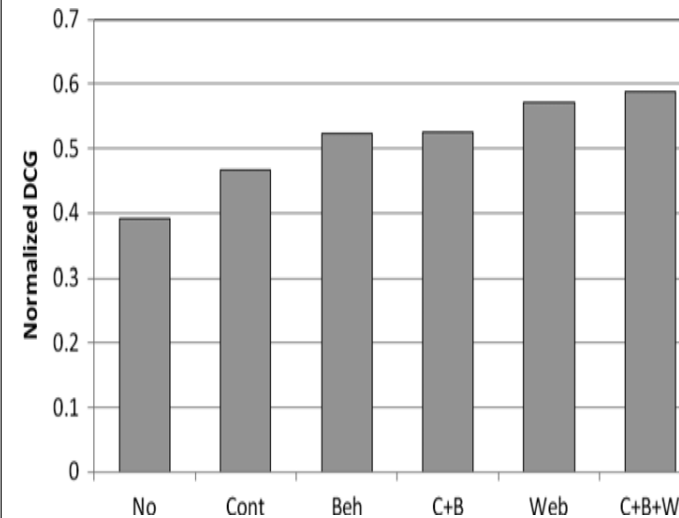
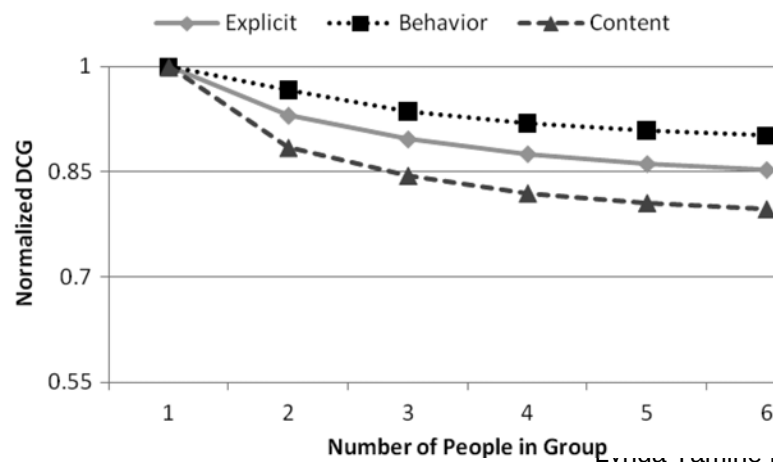
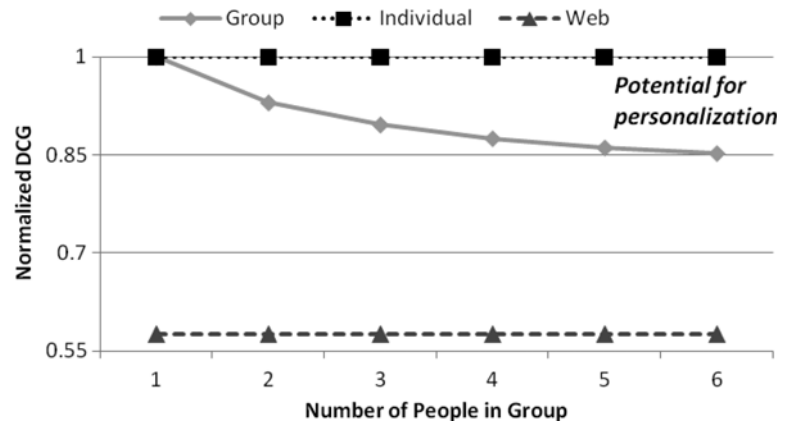
➤ Etude Teevan et al., 2010

- ✓ 125 utilisateurs employés de Microsoft, différents secteurs d'activité (administration, juridique, ventes, recherche etc.)
- ✓ Collecte de Jugements de pertinence : jugement explicite pour 699 requêtes, jugement basé sur le contenu pour 822 requêtes, jugement basé sur l'activité de l'utilisateur pour 2,400,654 requêtes
- ✓ *Pool* de requêtes communes pour différents utilisateurs : jugement explicite pour 11 requêtes, jugements basés sur le contenu pour 24 requêtes, jugements basés sur l'activité pour 44,002 requêtes



■ Quel intérêt à personnaliser / activités, centres d'intérêt ?

➤ Principaux résultats (Teevan et al, 2010)





RI contextuelle
RI en contexte
RI contextualisée
RI personnalisée
RI mobile
RI spatio-temporelle



Contextual Information Retrieval

Contextual retrieval

IR in context

Context-sensitive IR

Context-aware IR

Context-based IR

Personalized IR

Personalized search

Spatio-temporal search

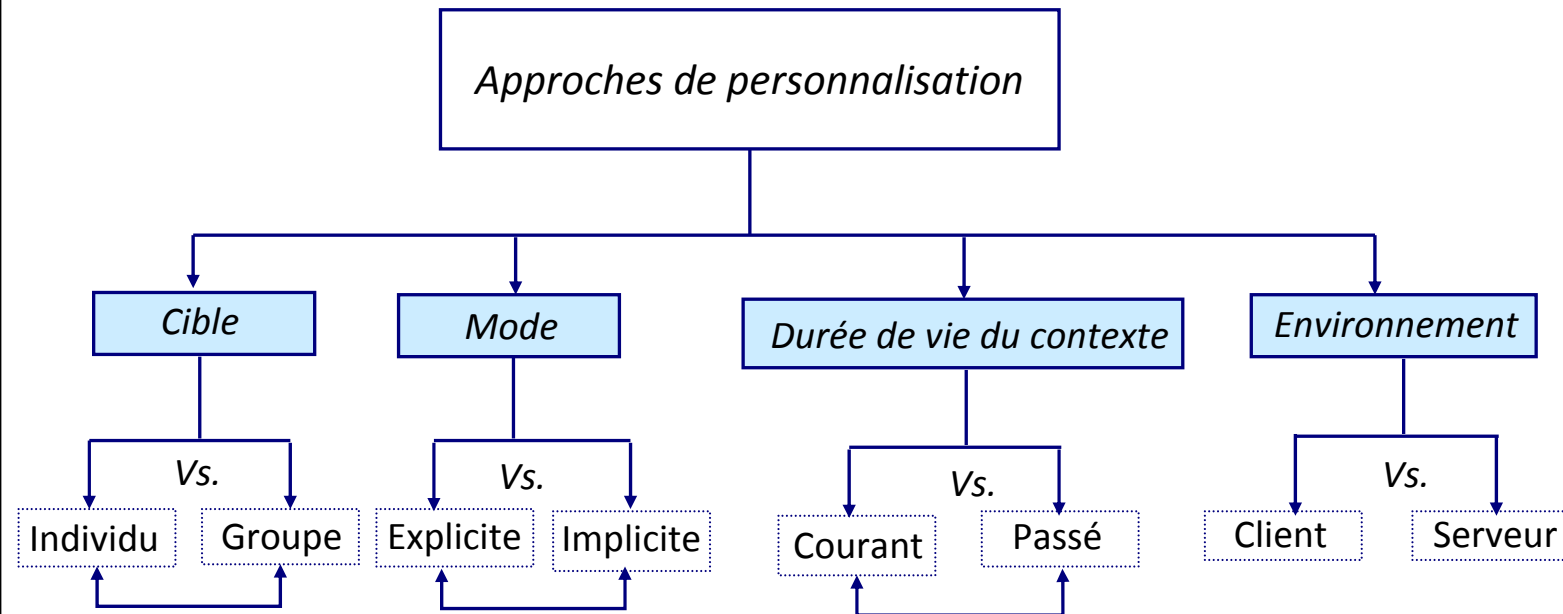
Mobile search

Mobile IR



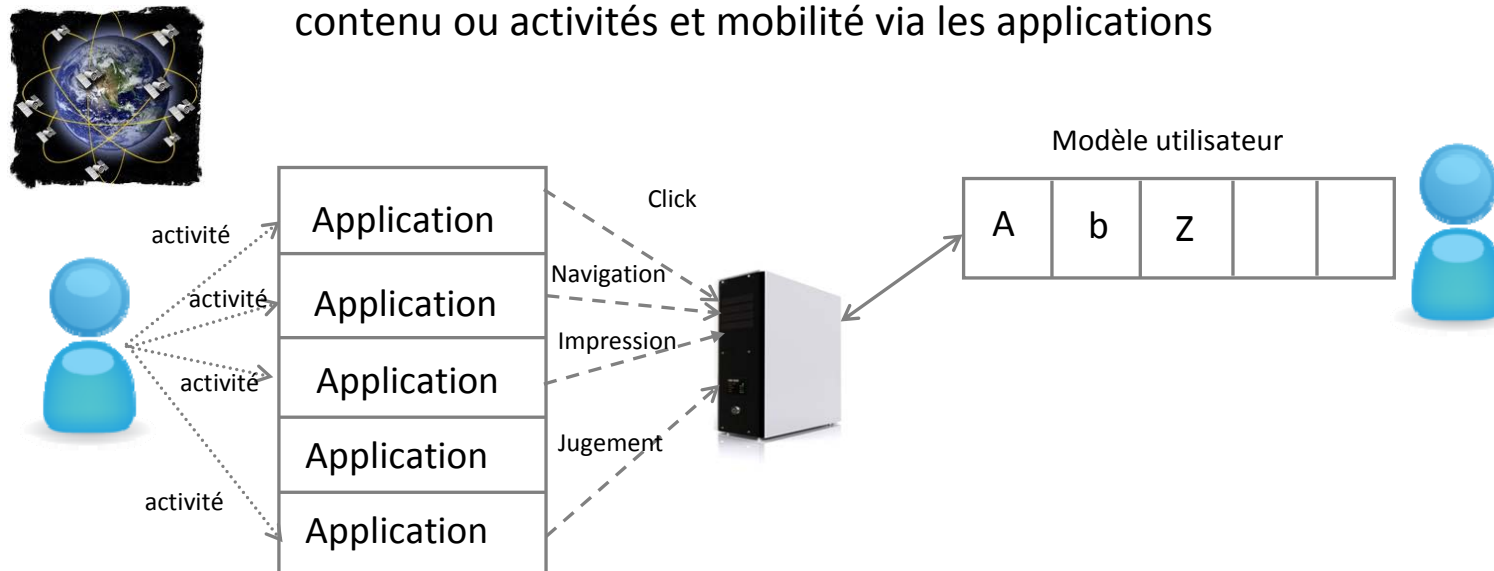
Partie 2 :

RI personnalisée, RI mobile



■ **Personnalisation individuelle Vs. Collaborative** (*Content Vs Collaborative-based personalisation*)

- Personnalisation individuelle : *principe (Liu et. al 2004, Sieg et. al, 2007, ...)*
 - ✓ Utilisateur considéré individuellement
 - ✓ Modèle utilisateur construit sur la base de ses préférences de contenu ou activités et mobilité via les applications



■ Personnalisation individuelle Vs. Collaborative (*Content Vs Collaborative-based personalisation*)

➤ Personnalisation individuelle : *objectifs*

- ✓ Modéliser l'utilisateur à travers ses activités, préférences
- ✓ Fournir un accès à l'information adapté à son modèle
 - o Recommander des objets selon son modèle (recommandation)
 - o L'aider à accéder à l'information pertinente (recherche)

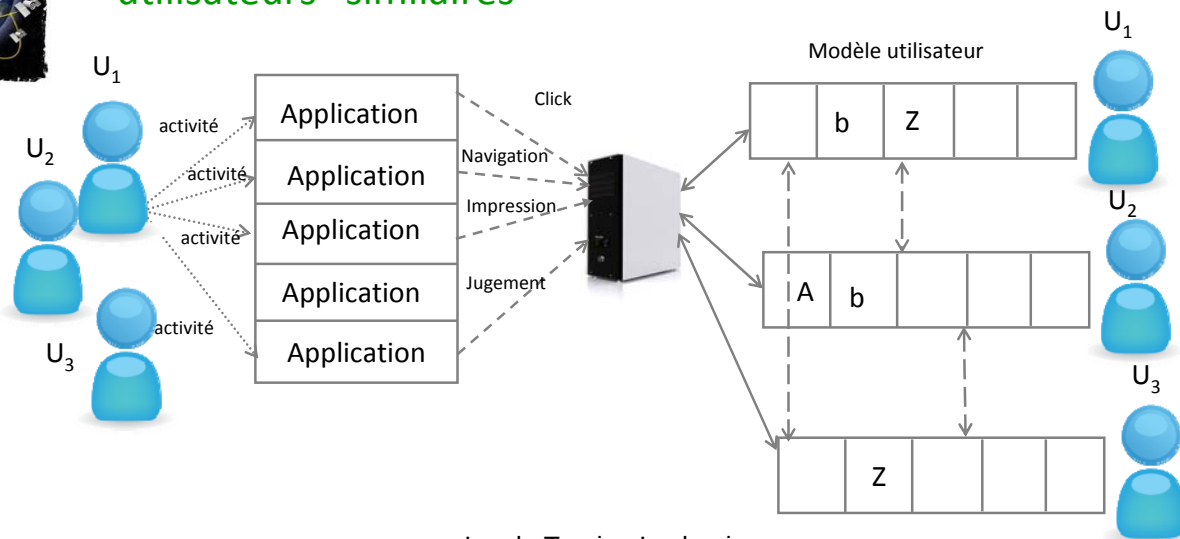
➤ De nombreux systèmes font évoluer la version personnalisée



■ **Personnalisation individuelle Vs. Collaborative (*Content Vs Collaborative-based personalisation*)**

➤ Personnalisation collaborative : *principe (Teevan et. al 2009, Morris et. al 2008, ...)*

- ✓ Utilisateur **considéré comme membre d'un groupe d'utilisateurs**
- ✓ Modèle utilisateur construit sur la base de ses préférences de contenu ou activités via les applications + **modèles des autres utilisateurs "similaires"**

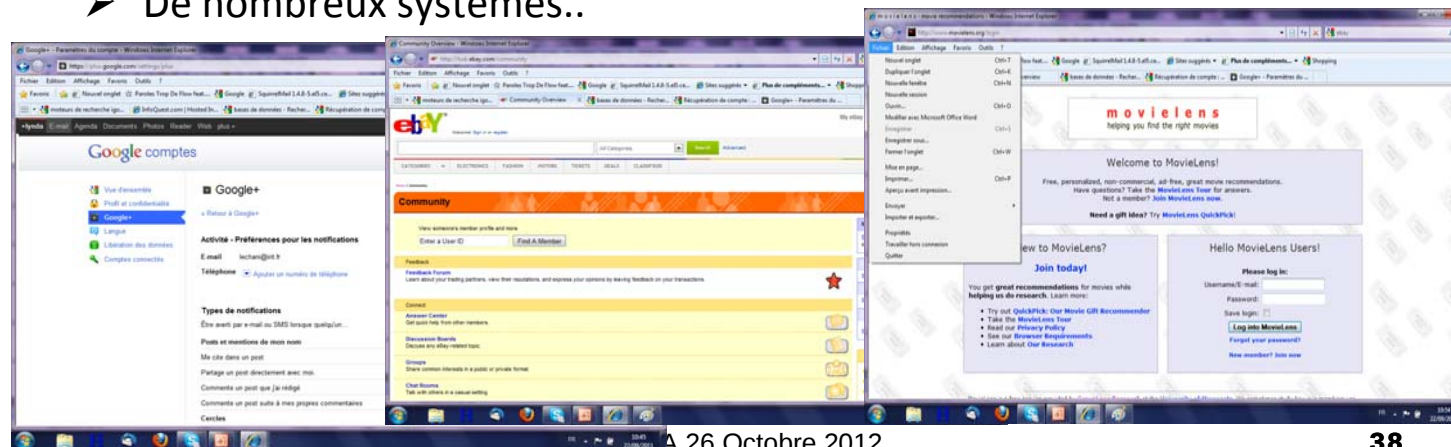


■ Personnalisation individuelle Vs. Collaborative (*Content Vs Collaborative-based personalisation*)

➤ Personnalisation collaborative : *objectifs*

- ✓ Modéliser l'utilisateur à travers ses activités, ses **préférences et celles des autres utilisateurs du groupe**
- ✓ Fournir un accès à l'information adapté à son modèle
 - o Recommander des objets selon son modèle (recommandation)
 - o L'aider à accéder à l'information pertinente (recherche)

➤ De nombreux systèmes..



26 Octobre 2012



■ Personnalisation Explicite Vs. Implicite (*Explicit Vs Implicit*)

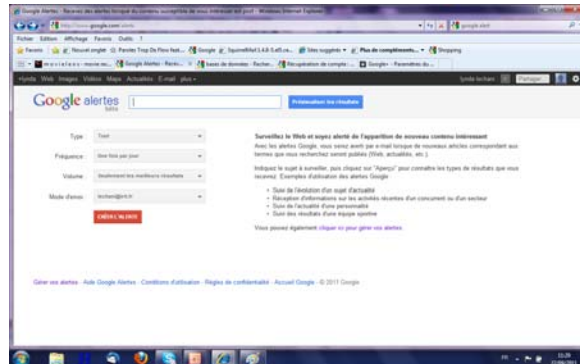
- Personnalisation explicite : *données*
 - ✓ Collecte des données descriptives de l'utilisateur à partir de l'utilisateur lui même
 - o Données démographiques (âge, sexe, langue maternelle etc...) (Hupfer & Detlor, 2006; Weber & Castillo 2010)
 - o Choix des préférences (langue, genre de documents (Martin & Jose 2004; Sieg et al 2007), sujets d'intérêt (Cheverest et al 2000, Panayioutou et al 2005)
 - o Jugements qualitatifs sur l'information (j'aime, j'aime pas) (Chen and Syscara 1998, Liu et al 2004)
- Personnalisation explicite : sources et moyens de collecte des données
 - ✓ Formulaire : cases à cocher, saisie de mots clés
 - ✓ Interfaces élaborées : expression d'exemples, contre exemples, votes, notes, annotations etc.
 - ✓ Questionnaires

■ Personnalisation Explicite Vs. Implicite (*Explicit Vs Implicit*)

➤ Personnalisation explicite : *systèmes*

✓ Commercialisés

Google alert



citeulike

Web page

Primary	Mathematical and Computer Sciences	Information Systems
Secondary		
Other [?]		

Research Fields

About me

[Preview About Me](#) [Editing help](#)

My Interests

[Preview My Interests](#) [Editing help](#)

Email Contact ☒ From time to time we would like to send you emails with the latest news about CiteULike. These will be very infre

✓ Non commercialisés

- NewT's (Sheth, 1993), ifWeb (Asnicar, 1997), WebWather (Armstrong, 1995), Expertclerk (Shimazu, 2002), ...

-



■ Personnalisation Explicite Vs. Implicite (*Explicit Vs Implicit*)

➤ Personnalisation implicite : *données*

- ✓ Données démographiques : peuvent être dérivées
 - o des interactions et activités des utilisateurs (Hu et al. 2007)
 - o des styles d'écriture (Argamon et al. 2003)
 - o des requêtes (Jones et al, 2007)

- ✓ Données situationnelles
 - o Lieu et temps d'émission de la requête (Cheverst 2000, Setten and al 2004, Abowd and al 1997)
 - o Environnement social : amis (Church et al 2010, Hwang et al 2007) ou personnes physiquement proches (Rhodes 2003)
 - o Agenda électronique (Belloti and al., 2008)

- ✓ Données sociales
 - o Annotations, posts, blogs etc. (Cai and Li 2010 ; Carmel et al. 2009; Vallet et al. 2010)



■ Personnalisation Explicite Vs. Implicite (*Explicit Vs Implicit*)

➤ Activités de l'utilisateur

- ✓ Pages visitées, requêtes passées (Shen et al.2005, Liu et al 2004)
- ✓ Données de navigation (Yau et al 2003, Speretta and Gauch 2005, Sugiyama et al. 2004)
- ✓ Applications utilisées (Yau et al 2003, Dumais et al, 2003, Chirita et al 2006,)
- ✓ Favoris (Dumais et al 2003, Jeh and Widom 2003)
- ✓ Historique de localisations (Bila et al 2008, Ying et al 2010)
- ✓ Interactions de l'utilisateur
 - Mouvements des yeux (Sun et al. 2005; Puolamaki and al, 2005)
 - Données de clicks (Sun et al. 2005; Tan et al. 2006; Joachims 2002)
 - Actions sur les documents (ouverture, fermeture, impression, temps de lecture etc.) (Pretschner & Gauch 1999; Chen et al. 2002; Morita and Shinoda, 1994)
 - Messages (e-mails) envoyés ou reçus (Belloti and al. 2008)



■ Personnalisation Explicite Vs. Implicite (*Explicit Vs Implicit*)

➤ Personnalisation implicite : *sources et moyens de collecte des données*

- ✓ Collecte des données descriptives de l'utilisateur à partir de ses activités, à travers :

- o Le navigateur ou proxy

- ☺ Partage des données sur plusieurs sites (appli Web)

- ☹ Données récupérées que d'une seule machine

- ☹ Partage de données machine-serveur non facile

- o Applications clientes (barre de bureau *toolbar*)

- ☺ Partage des données sur plusieurs sites (appli Web)

- ☺ Accès au bureau, fichier système etc.

- ☹ Nécessité de l'installation de l'agent (*toolbar*)

- ☹ Vulnérabilité des accès agent-spiware



■ Personnalisation Explicite Vs. Implicite (*Explicit Vs Implicit*)

➤ Personnalisation implicite : *sources et moyens de collecte des données*

- o Compte utilisateur (cookies, données de *log*)

☺ Historique de l'ensemble des activités

☹ Nécessité de créer le compte

✓ Collecte des données descriptives du contexte spatio-temporel de l'utilisateur à travers son infrastructure matérielle :

- o Global Positionning System (GPS)

☺ Données précises sur le lieu, instant, orientation

☹ Fiable pour le positionnement général, pas local

- o Systèmes de positionnement infrarouge

- o Réseaux sans fil

- o Technologie *Bluetooth*

- o *Horloge interne*



■ Personnalisation hybride (Explicite, Implicite)

➤ Combinaison des principes de collecte explicite, implicite

- o Collecte explicite puis implicite (système WAIR, Seo & Zhang 2001)

1. Get the initial profile from the user. Set $t = 0$. (Explicite)
2. (Retrieval) Generate a query from the profile to retrieve N URLs.
3. (Filtering) Evaluate the relevance of documents.
Rank the N documents and present M of them to the user.
4. (Interface) Get the feedback by observing user behavior. (Implicite)
5. (Learning) Update the user profile.
6. Set $t = t + 1$. Go to step 2.

- o Collecte implicite puis explicite (Liu and Weng 2004) : construction du contexte à partir des interactions utilisateur puis validation par ce dernier

■ **Personnalisation Explicite Vs. Implicite (*Explicit Vs Implicit*)**

➤ Personnalisation implicite : *systèmes*

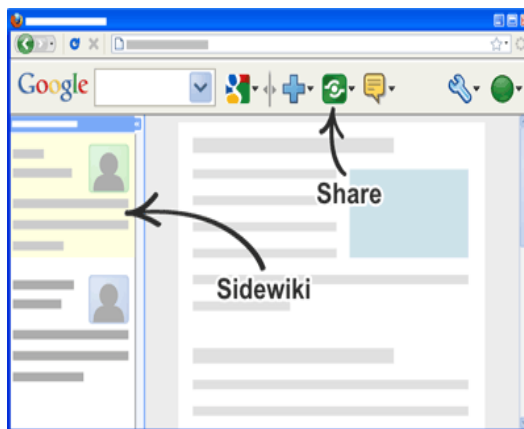
✓ Commercialisés



■ Personnalisation Explicite Vs. Implicite (*Explicit Vs Implicit*)

➤ Personnalisation implicite : systèmes

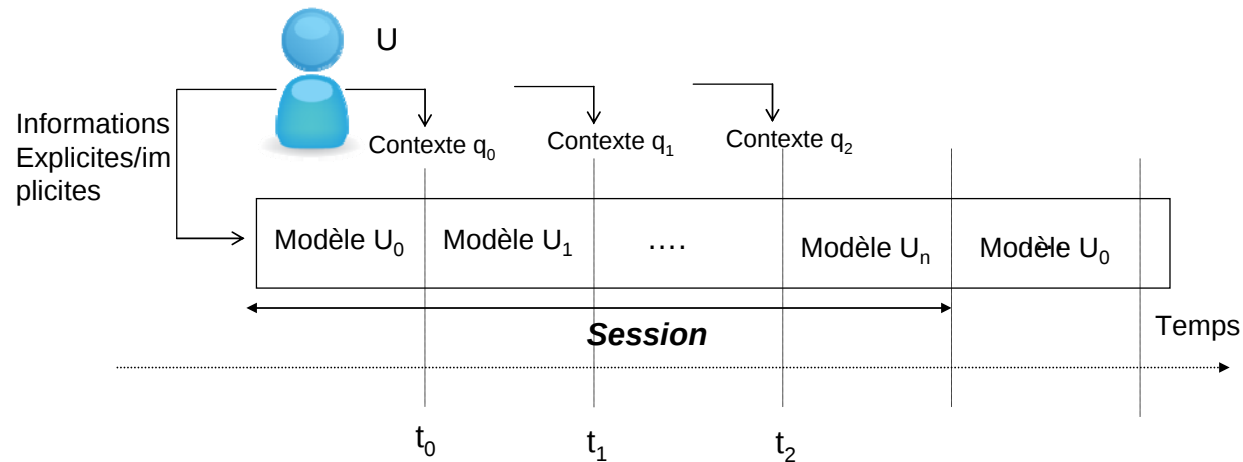
✓ Commercialisés



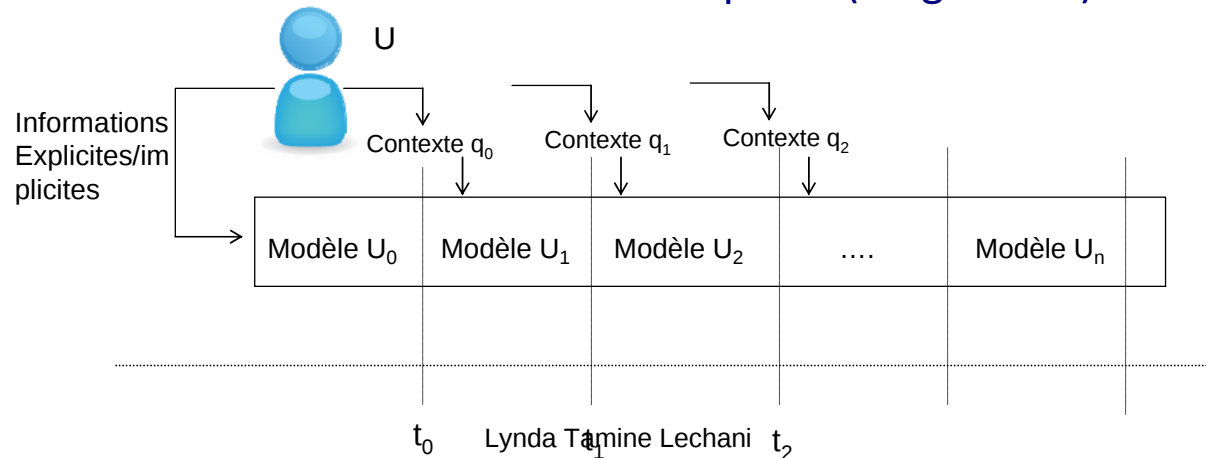
✓ Non commercialisés

- Letizia (Lieberman, 1997), WebMate (Chen, 1998), Personal Web Watcher (Mladenic, 1998), Cyberguide (Abowd and al, 1997)

■ **Personnalisation basée contexte courant (court-terme)**



■ **Personnalisation basé sur le contexte passé (long-terme)**





- Personnalisation basée contexte courant Vs. contexte passé (*Current Vs Past*)
 - Personnalisation basée sur le contexte courant (court terme)
Yahoo Contextual Search (Y!Q), Letizia (Lieberman 1995) WebPersonae (Rhodes 2000), OBIWAN (Gauch and al, 2003)
 - ✓ Mémoire de personnalisation = durée de la session
 - ✓ Problème de délimitation de la session de recherche
 - o Délimitation basée sur le temps (30 min.) (Silverstein et al. 1999)
 - o Délimitation basée sur la « similitude » de requêtes successives (He et al 2002), (Jones and al. 2008)
 - o Délimitation basée sur la « similitude » des résultats de recherche (Radlinski and Joachims 2005; Daoud et al., 2009)



- Personnalisation basée contexte courant Vs. contexte passé (*Current Vs Past*)
 - Personnalisation basée sur le contexte passé (long terme)

Google's Alert, Google's personalized search 1.1, Yahoo My Web, (Gowan 2003, Sieg et al. 2004, Tan et. al 2006, Daoud and al. 2011)

 - ✓ Mémoire de personnalisation = durée du compte utilisateur
 - ✓ Principe d'ajout (accumulation des données), suppression des centres d'intérêts
 - ✓ Délimitation des sessions de recherche (Gauch 2003, Sieg et al. 2007, Sieg et al. 2004) ou sans délimitation (Liu and Weng. 2004)



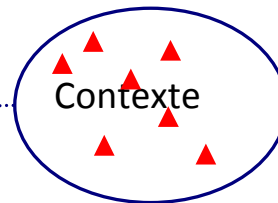
■ Personnalisation niveau client Vs. niveau serveur

	<i>Pers. Niveau client</i>	<i>Pers. Niveau serveur</i>
Données accessibles	- Données utilisateur de la machine cliente - Partage des données sur différentes applications	- Données de tous les visiteurs du site hébergé par le serveur - Données non partagées sur différentes applications
Cible	- Personnalisation individuelle - Personnalisation collaborative non applicable	- Personnalisation individuelle et personnalisation collaborative applicables
Mode	- Implicite - Explicite	- Implicite - Explicite
Durée de vie	- Contexte courant - Contexte passé	- Contexte courant - Contexte passé

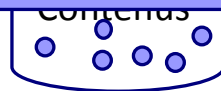
■ Problématiques scientifiques

①

Comment construire le contexte,
le modèle utilisateur ?



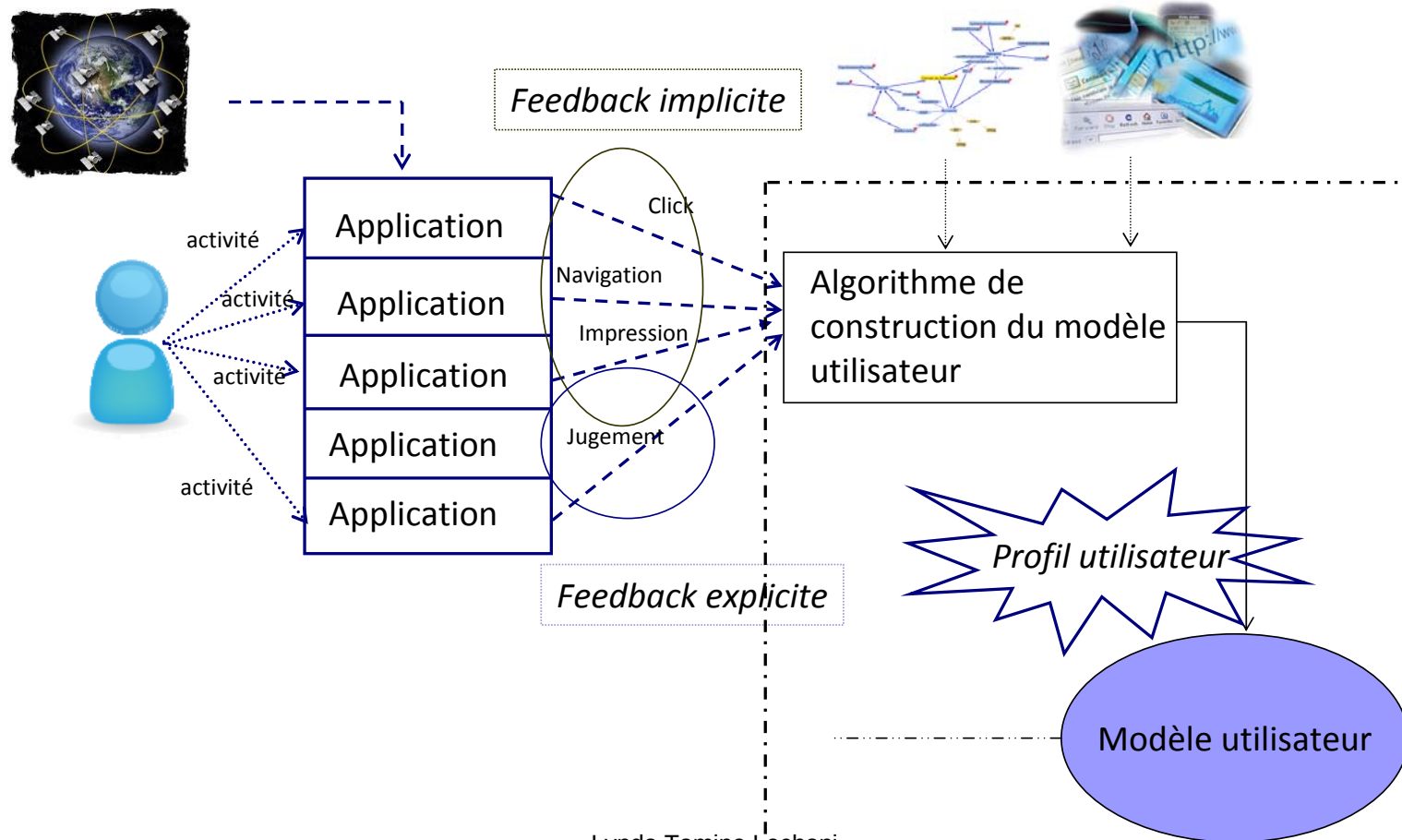
Réponses à ces questions en RI mobile, RI personnalisée



②

Comment exploiter le contexte
pour sélectionner les contenus ?

■ Principe de modélisation du contexte



■ **Aperçu des catégories de modèles en RI personnalisée, mobile**

- Modèles non structurés
 - ✓ Historique de recherche
- Modèles structurés
 - ✓ Modèles basés attribut-valeur
{<élément contextuel= valeur>}
 - ✓ Modèles basés termes
 - Vecteur de termes
 - (Hiérarchie de) classes de termes
 - Réseau de termes
 - ✓ Modèles basés objet, concept
 - Modèles orientés objet
 - Modèles conceptuels

■ Modèles non structurés

Usage en RI personnalisée (Agichtein et al. 2006) et en RI mobile (Purnima, 2004) pour prédire les centres d'intérêt de l'utilisateur

- Trace (représentation) des données brutes dans les *logs*, *proxy*
- Ce qui est généralement représenté
 - ✓ Caractéristiques des requêtes (*Query text features*)
 - taux de couverture du titre, taux de couverture du résumé, taux de couverture de l'URL, longueur de requête
 - ✓ Caractéristiques de navigation (*Browsing features*)
 - temps passé sur une page, temps cumulé passé sur l'ensemble des pages visitées, temps passé sur le domaine de la requête
 - ✓ Caractéristiques des clicks (*Clickthrough features*)
 - nombre de clicks sur les résultats de requête, position de l'URL *clickée*

■ Modèles structurés : modèle attribut valeur

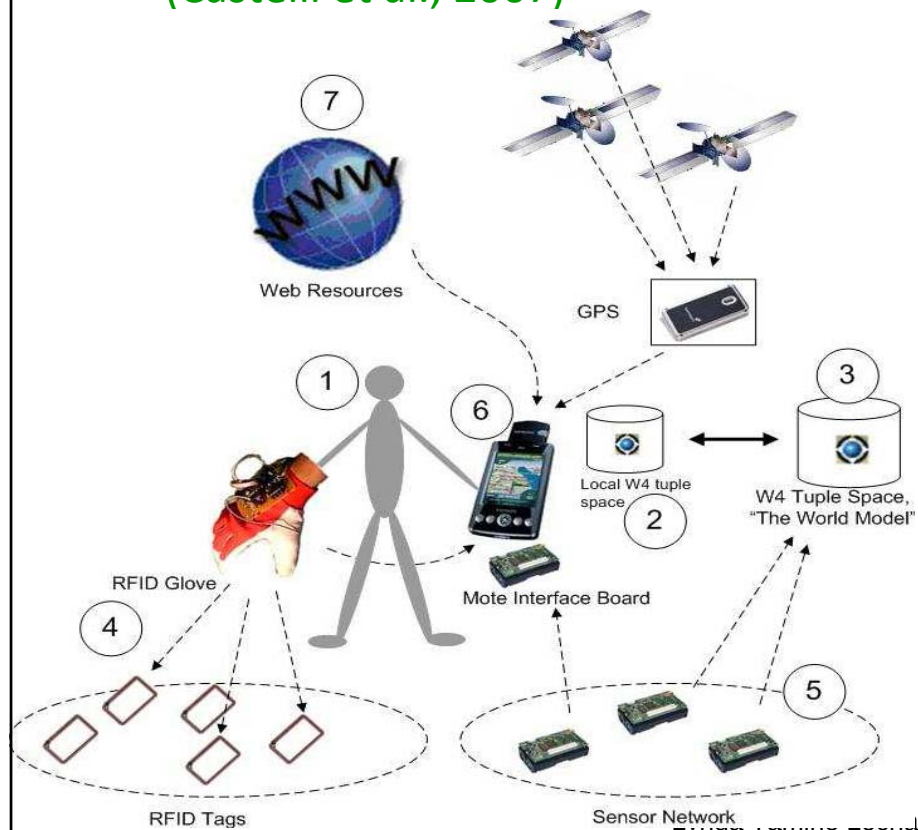
Usage en RI mobile (Rhodes 2003, Yau et. al, 2003, Castelli et al; 2007)
pour instancier des attributs contextuels

➤ Principaux attributs

- ✓ Localisation physique de l'utilisateur
 - Lieu
 - Coordonnées polaires
 - Expression de positionnement : ici, à 500 m, etc.
- ✓ Personnes à côté (*near persons*)
- ✓ Temps
 - Instant
 - Intervalle de temps
 - Expression temporelle : maintenant, aujourd'hui
- ✓ Activité
 - Lecture, travail, etc..

■ Modèles structurés : modèle attribut valeur

Infrastructure nécessaire pour capturer puis instancier les attributs
(Castelli et al., 2007)



■ Modèles structurés : modèles basés termes

Usage en RI personnalisée, RI mobile pour avoir une « image » des centres d'intérêt de l'utilisateur

➤ Modèles basés sur la construction de vecteurs de termes

Pazzani 1996; Lieberman 1997; Chen 1998; Tan 2006; Sun 2008, Shen 2005

➤ Modèles basés sur la dérivation de (hiérarchies) de classes de termes

Gowan 2003; Chen 2005; Widyantoro 1999

➤ Modèles basés sur la construction de réseaux de termes

Begg 1993; Micarrelì 2004

■ Modèles structurés basés sur la construction de vecteurs de termes : *WebMate (Chen, 1998)*

➤ Profil utilisateur : *représentation et mode de construction*

- Mode explicite (*I like, I dislike*), contexte passé, niveau client
- $P = \{V_1 \dots V_N\}$, N est le nombre de centre d'intérêts fixé a priori
 $V_i \langle t_{i1}, \dots, t_{iT} \rangle$ vecteur de termes, T taille de l'index

➤ Profil utilisateur : *algorithme de construction*

1. Calculer un vecteur V_i pour chaque document d_i jugé pertinent $V_i (w_{i1}, w_{i2}, \dots, w_{iT})$
2. SI ($i < N$) ALORS $P = P \cup V_i$
3. SINON - calculer $V_m = \text{Argmax}_{k=1..N} (\text{SIM}(V_i, V_k))$,
- $V_m = V_m + V_i$
SIM est une fonction de similitude vectorielle (Cosinus, Jacquard...)

■ Modèles structurés basés sur la construction de réseaux de termes (Micarelli, 2004), système Wifs

➤ Profil : représentation et mode de construction

- ✓ Profil $P = \langle \{ \text{Stéréotype actif}, \text{Domaine}, \text{Sujet}, \text{Liens sémantiques}, \text{Liens de justification} \} \rangle$
 - *Stéréotype actif* : stéréotype qui correspond au profil P :
 - *Domaine (Domain)* : domaine du stéréotype
 - *Sujet (Topic)* : sujet d'intérêt, représenté par un terme
 - *Poids (Weight)* : degré d'importance du stéréotype pour l'utilisateur, valeur dans $[-10...+10]$
 - *Liens sémantiques (Semantic links)* : termes qui occurrent dans les documents, liés au sujet de chaque slot du stéréotype
 - *Liens de justification (justification links)* : liens vers les formulaires saisis ou interview (feedback explicite), documents jugés (feedback implicite)

Modèles de contexte ▶ Construction et modélisation

Données personnelles : -----

Stéréotypes actifs

Researcher

Slot 1

Domain : Artificial Intelligence

Topic : Learning

Weight : 9

Semantic links → co-keywords

Justification links → Interview

Slot 2

Domain : Object Oriented Language

Topic : C++

Weight : 7

Semantic links → co-keywords

Justification links → Interview

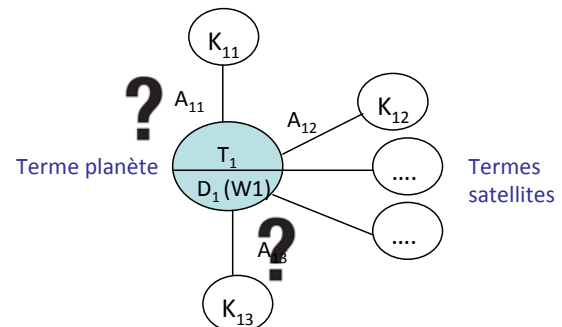
.....

Lynda Tamine Lechani

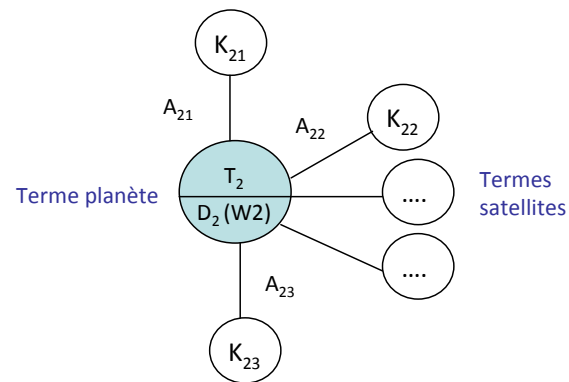
Ecole EARIA 26 Octobre 2012

Exemple de profil utilisateur (Micarelli, 04)

62



Slot 1



Slot 1

Stéréotype actif ?

.....

Stéréotype actif 2

- Modèles structurés basés sur la construction de réseaux de termes (Micarelli, 2004), système Wifs

- Profil : algorithme de construction

- ✓ Attribution de stéréotype
 - Interview
 - Dérivation automatique par réseau de neurones

- ✓ Adaptation du modèle de profil

Trois (3) principales étapes : Soit le modèle P de l'utilisateur

- **1** Calcul de la pertinence de chaque document D pour la requête Q en entrée (Q = domaine dans le cas du filtrage !)
- **2** Mise à jour des slots : adaptation des slots
- **3** Mise à jour du profil P : ajout de nouveaux slots

- Modèles structurés basés sur la construction de réseaux de termes (Micarelli, 2004), système Wifs
 - Profil : algorithme de construction
 - ❶ Calcul de la pertinence de chaque document D en entrée pour le profil P

Développée dans la partie suivante du cours

■ Modèles structurés basés sur la construction de réseaux de termes (Micarelli, 2004), système Wifs

➤ Profil : algorithme de construction

② Mise à jour des slots : *adaptation des slots*

Mise à jour des
poids des slots

$$\begin{cases} t \in P, t \in D & W_{slot} = W_{slot} + \frac{k_1 \times (f - |w_{slot}|) \times occ(t)}{occ(t) + 1} \\ t \in P \wedge t \in Q & W_{slot} = W_{slot} + (f - |w_{slot}|) \\ t \in P \wedge t \notin Q & W_{slot} = W_{slot} - k_3 \times |f| \end{cases}$$

f pertinence utilisateur pour D

Termes du slot
(Domain)

Mise à jour des poids
des termes du slot

Mise à jour des liens
sémantiques (co-Keywords)

$$\begin{cases} t \in P & A = A + \frac{k \times f}{|occ(t) - occ_{ref}(t)| + 1} \quad \forall co-keyword \in D, \\ A = A - k \times f & \forall co-keyword \notin D \\ t \in P \wedge t \in Q & A = A + k \times f \quad \forall co-keyword \in D \\ A = A - k \times f & \forall co-keyword \notin D \end{cases}$$

■ Modèles structurés basés sur la construction de réseaux de termes (Micarelli, 2004), système Wifs

➤ Profil : algorithme de construction

③ Mise à jour du profil : *ajout de nouveaux slots*

Terme devient *Topic*
d'un nouveau slot

Termes du document
(non présents dans le
profil)

Calcul des co-keywords
du slot

$$\left\{ \begin{array}{l} t \notin P \wedge t \in TDB \quad W_{slot} = W_{slot} + \frac{k_1 \times |f| \times occ(t)}{occ(t) + 1} \\ t \notin P \wedge t \in TDB \wedge t \in Q \quad W_{slot} = k_2 \times |f| \\ t \notin P \wedge t \notin TDB \wedge t \in Q \quad W_{slot} = k_3 \times |f| \end{array} \right.$$

f pertinence utilisateur pour D

$$\left\{ \begin{array}{l} t \notin TDB \wedge t \notin Q \quad A = \frac{k_1 \times |f|}{|occ(t) - occ(co(t))| + 1} \quad co(t) : \text{terme cooccurent avec } t \text{ dans } D \end{array} \right.$$

■ Modèles basés objet

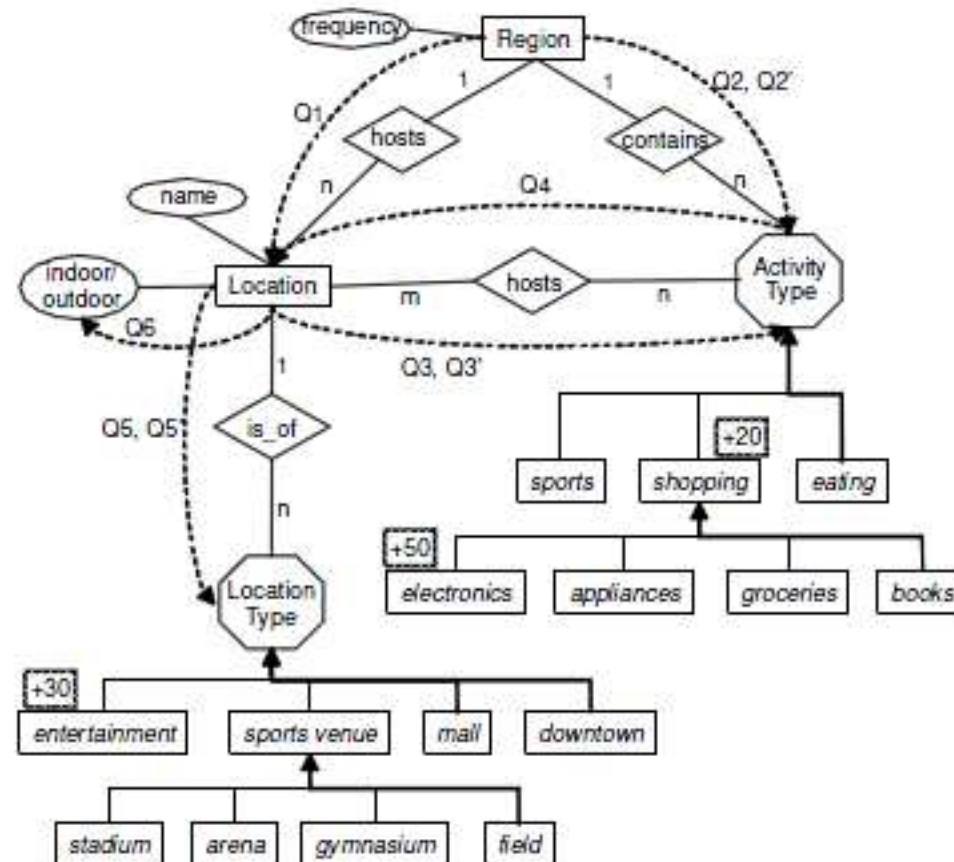
Usage en RI mobile pour avoir une « image » du profil « situationnel » de l'utilisateur (Cheverst 2000, Bila et al. 2000, Yoon et al 2007)

- Quels objets ?
 - ✓ Objets de l'espace : personnes, localisations, etc.

- Représentation des propriétés des objets
 - ✓ Données démographiques (nom, âge, sexe, adresse, etc.)
 - ✓ Activités de l'utilisateur (que fait il ?)
 - ✓ Données de localisation (type, coordonnées polaires)
 - ✓ Données temporelles (instant)
 - ✓ Infrastructure matérielle

- Représentation des liens entre objets

■ Modèles basés objet (Bila et. al 2008)



- Modèles basés concept

Usage en RI mobile et RI personnalisée pour avoir une description
« normalisée » du contexte

- Usages d'ontologies générales

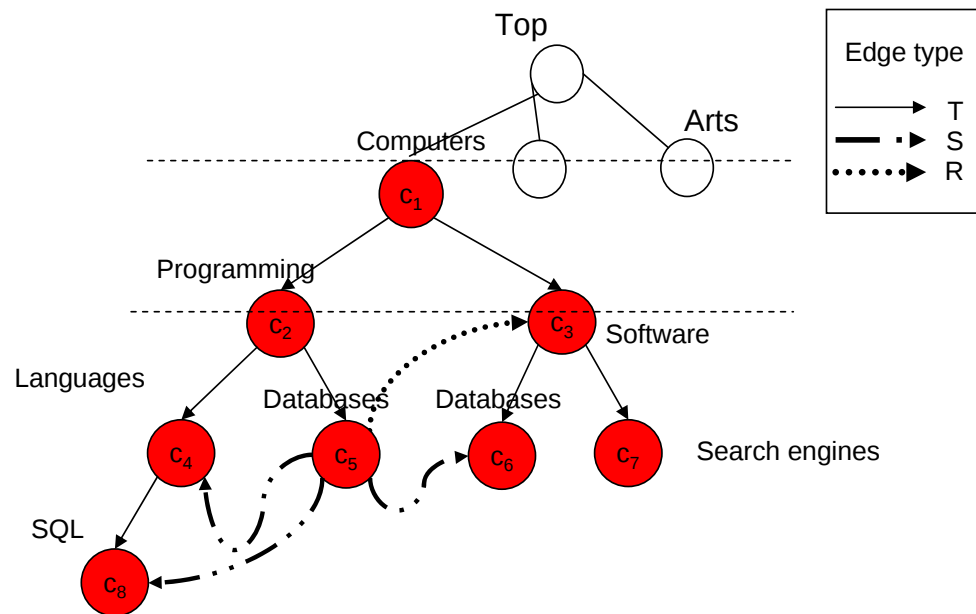
ODP (Liu 2004, Alexandru et al. , Chirita et al. 2006, Mylonas et al. 2004, Doud 2009) , Yahoo (Sieg et al. 2007) , YAGO (Celegari and Pasi 2011)

- Usages d'ontologies de domaine, propriétaires

Santé (Jahnke and al. 2005) , géographique (Leung et. al, 2010)

- Open Directory Project : la plus largement utilisée en RI personnalisée

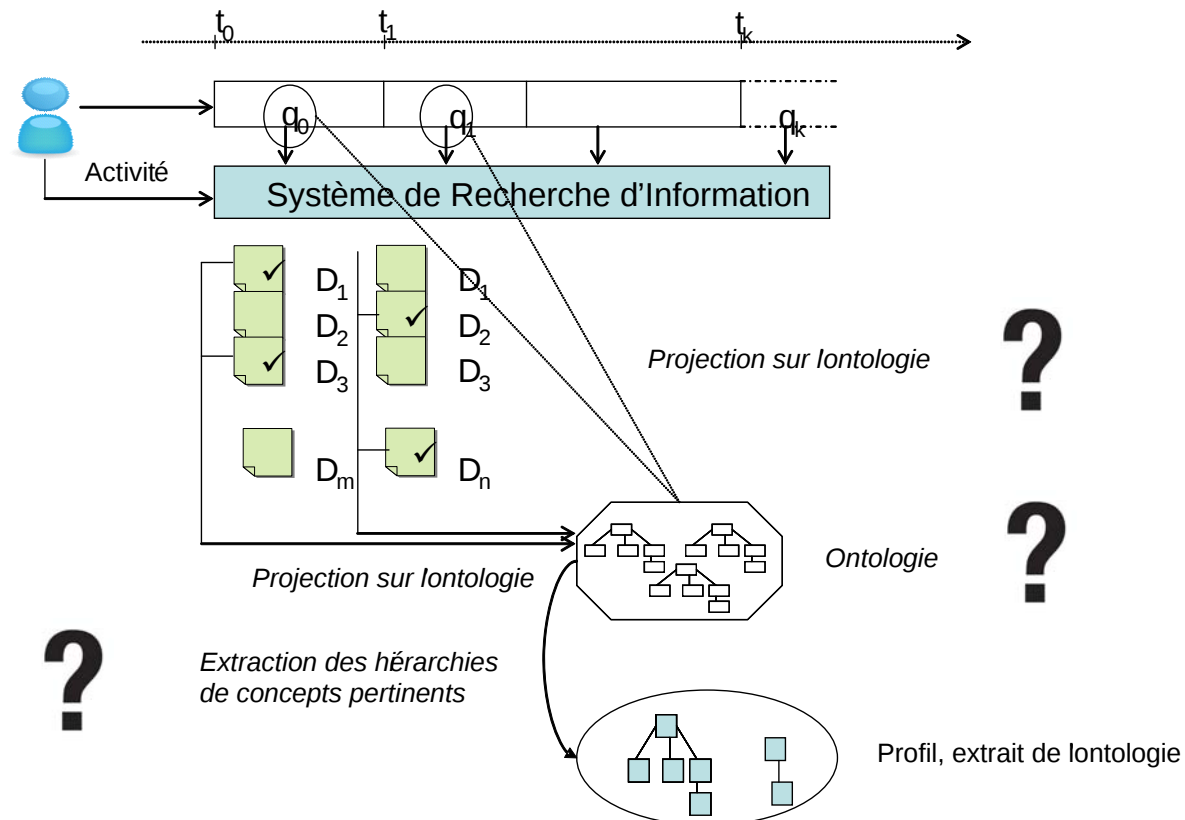




$$\vec{C} = \frac{\left[\sum_{d \in Docs(C)} \vec{d} \right]}{|Docs(C)|}$$

■ En RI personnalisée

➤ Principe (général) de construction du profil

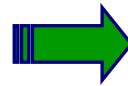


■ En RI personnalisée : (Liu and Weng 2004)

➤ Profil : représentation et mode de construction

- ✓ Mode implicite/ explicite (historique de recherches, feedback), contexte courant/passé, niveau client/serveur
- ✓ Profil P : triplet $\langle I, P_u, P_g \rangle$
 - I : intention de recherche dérivé de la requête $I=\{c_i, w_i\}$, liste de concepts
 - P_g : profil général représenté par l'ontologie générale
 - P_u : Profil utilisateur dérivé de son historique de recherche, hiérarchie de concepts $\{ \langle C_i \rangle \}$, $C_i = \{ \langle t_{ij}, w_{ij} \rangle \}$

Concept/ Document	C_1	C_2	..	C_m
D_1	1/0	1/0	1/0	1/0
D_2	1/0	1/0	1/0	1/0
..				
D_n	1/0	1/0	1/0	1/0

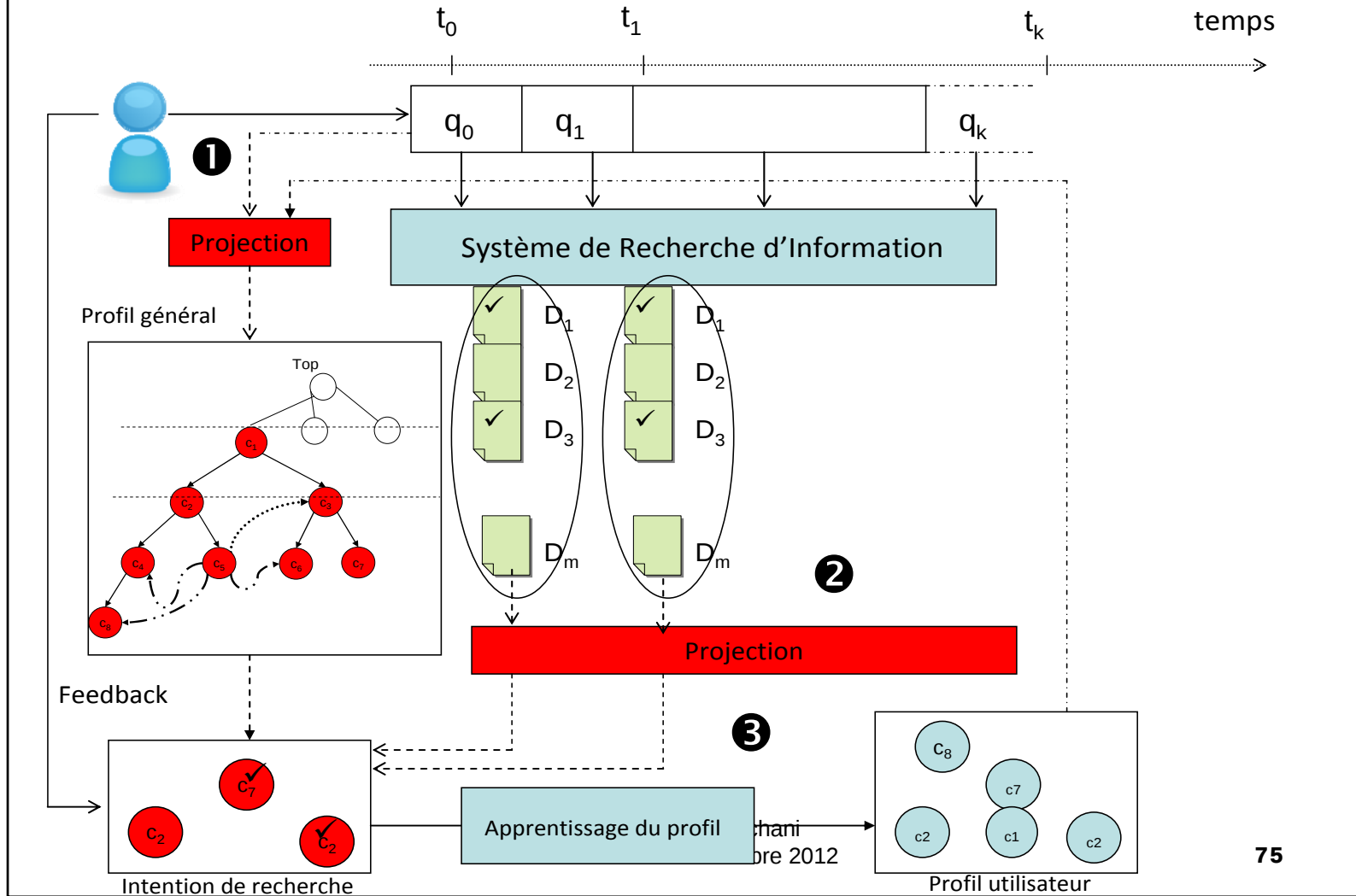


Terme/ Concept	t_1	t_2	..	t_T
C_1	w_{11}	w_{12}		w_{1T}
C_2	w_{21}	w_{22}		w_{2T}
..				
C_n	w_{n1}	w_{n2}		w_{nT}

DC: Matrice de l'historique de recherche

M : Matrice du profil utilisateur

■ En RI personnalisée : (Liu and Weng 2004)



■ En RI personnalisée : (Liu and Weng 2004)

➤ Profil : algorithme de construction

- ✓ *Projection* : sur le profil général, profil utilisateur, sur les deux

Principe : **a-** calculer une similarité vectorielles des descripteurs requête et profil

① ②

$$Sim(q, c) = (Sim(q, c^u) + Sim(q, c^g)) / 2$$

c^u : concept du profil utilisateur

c^g : concept du profil général

b- retenir les n concepts les plus similaires ($n=3$)

- ③ ○ *Apprentissage du profil utilisateur* : plusieurs méthodes basées sur la réduction de la dimension de la matrice M . Principe de la LSI pour l'approximation de la matrice DC .

$$M = DC^T * U * \Sigma^+ * V^T$$

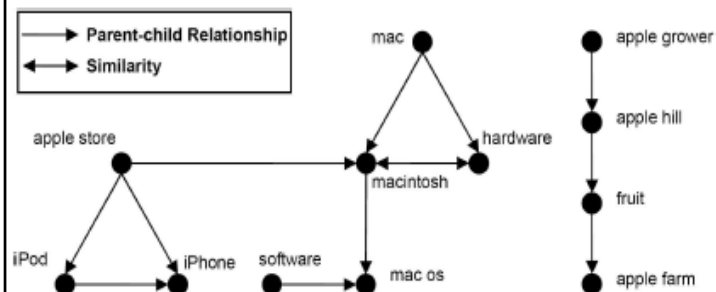
$$DT = U * \Sigma * V^T \quad DT \text{ matrice Document - Terme}$$

U, V matrices orthogonales, Σ^+ matrice inverse de Σ

■ En RI mobile : (Leung et al. 2010)

➤ Profil : représentation et mode de construction

- ✓ Mode implicite (clicks, feedback), contexte courant, niveau client/serveur
- ✓ Profil P : paire $\langle P_c, P_l \rangle$
 - P_c : profil de contenu
 - P_l : profil de localisation
- ✓ Projection des documents cliqués sur des ontologies de contenu, ontologie de localisation, sélection des concepts



Example Content Ontology Extracted for the Query “apple”.

STATISTICS OF THE LOCATION ONTOLOGY

No. of Countries	7	Total No. of Nodes	16899
No. of Regions	190	Country-Region Edges	190
No. of Provinces	6699	Region-Province Edges	1959
No. of Towns	10003	Province-City Edges	14897

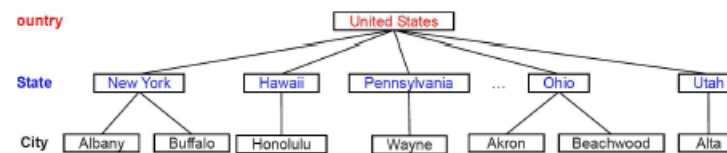


Fig. 3. Example Predefined Hierarchy for “United States”.

■ En RI mobile (Leung et al. 2010)

➤ Profil : mesure formelle de calcul d'entropie

$$H_{\overline{C}}(q, u) = - \sum_{i=1}^t p(\overline{c_{iu}}) \log p(\overline{c_{iu}})$$

$$H_{\overline{L}}(q, u) = - \sum_{i=1}^v p(\overline{l_{iu}}) \log p(\overline{l_{iu}})$$

t : nombre de concepts présents dans les documents cliqués par l'utilisateur u

$\overline{C_u} = \{\overline{c_{1u}}, \overline{c_{2u}}, \dots, \overline{c_{tu}}\}$: $|c_{iu}|$ nombre de documents contenant le concept de contenu c_i et cliqués par l'utilisateur u

$$|\overline{C_u}| = |\overline{c_{1u}}| + |\overline{c_{2u}}| + \dots + |\overline{c_{tu}}|, p(c_{iu}) = \frac{|\overline{c_{iu}}|}{|\overline{C_u}|}$$

v : nombre de concepts de localisation présents dans les documents cliqués par l'utilisateur u

$\overline{L_u} = \{\overline{l_{1u}}, \overline{l_{2u}}, \dots, \overline{l_{vu}}\}$: $|l_{iu}|$ nombre de documents contenant le concept de localisation c_i et cliqués par l'utilisateur u

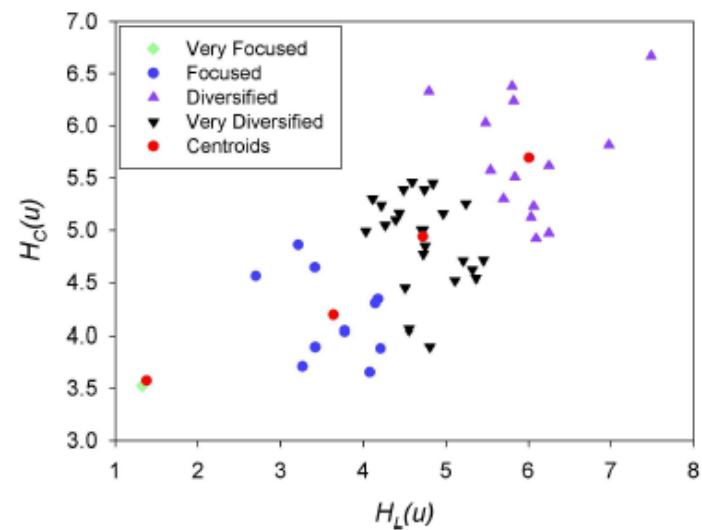
$$|\overline{L_u}| = |\overline{l_{1u}}| + |\overline{l_{2u}}| + \dots + |\overline{l_{vu}}|, p(l_{iu}) = \frac{|\overline{l_{iu}}|}{|\overline{L_u}|}$$

■ En RI mobile (Leung et al. 2010)

- Profil : requête personnalisable ou pas ?

$$e_{C(q,u)} = \frac{H_C(q)}{H_{\bar{C}}(q,u)}$$

$$e_{L(q,u)} = \frac{H_L(q)}{H_{\bar{L}}(q,u)}$$

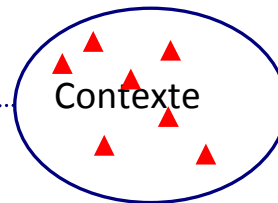


(c) $H_L(u)$ vs $H_C(u)$

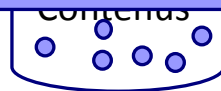
■ **Problématiques scientifiques**

①

Comment construire le contexte,
le modèle utilisateur ?



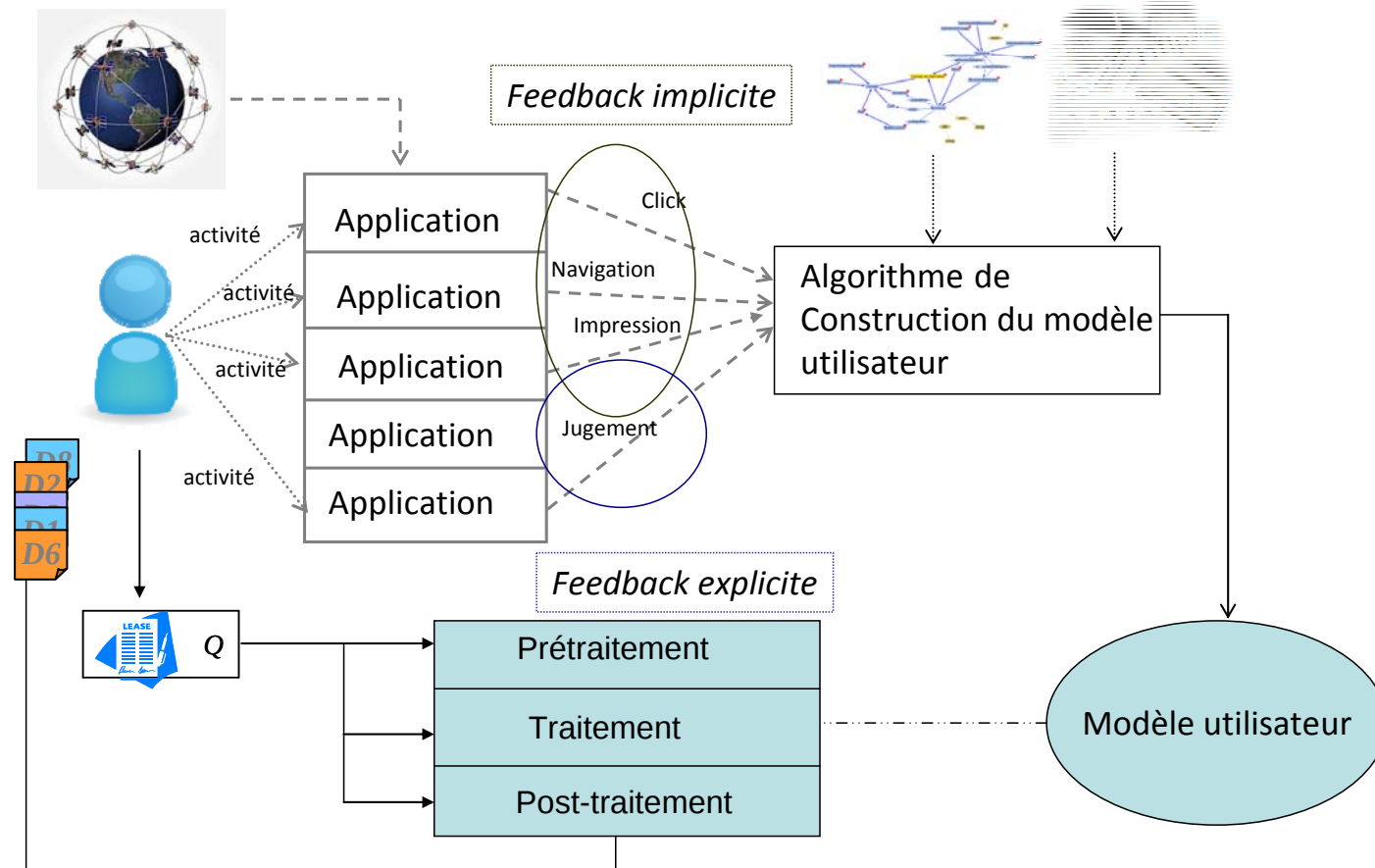
Réponses à ces questions en RI mobile, RI personnalisée



②

Comment exploiter le contexte
pour sélectionner les contenus ?

■ Approches d'appariement personnalisé



■ Prétraitement : reformulation de la requête

- Principe de réécriture de la requête (reformulation et/ou expansion) en utilisant les éléments du profil
 - Ajout ou repondération de termes :
 - Extraits du profil lié aux centres d'intérêts (Chen & Sycara 1998; Chen & Kuo 2000, Koutrika & Ioannidis 2005, Pitkow et al. 2002, Chirita et al. 2006, Cui & al 2003)
 - Extraits du profil lié à la situation (Yau et. al 2003)
 - Concepts pertinents de la requête et profil (Liu et al 2004, Sieg et. al 2004, Daoud 2009)
 - Distribution des termes sur l'historique de recherche via un modèle de langue (Shen et al. 2005, Tan et al. 2006)

■ Reformulation basée sur la technique de Rocchio

➤ Système UCAIR (Shen et al. 05) : reformulation basée sur le contexte courant, représentation vectorielle

✓ Soit q_t requête courante, D_t résultats associés; q_{t-1} requête précédente, D_{t-1} résultats associés (**pré-calculés, non présentés à l'utilisateur**)

✓ Détecter si q_t et q_{t-1} sont de la même session : $Sim(q_t, q_{t-1}) < Seuil$

$$Sim(q_t, q_{t-1}) = Cos(\overline{D_t}, \overline{D_{t-1}})$$

$$\overline{D_t} = \frac{1}{|D_t|} \sum_{d \in D_t} d$$

✓ Si q_t et q_{t-1} sont de la même session alors reformuler q_t

$$q_t' = \alpha * q_t + (1 - \alpha) * \frac{1}{|D_t|} \sum_{i=1}^{|D_t|} d_i$$

■ Reformulation basée sur une combinaison des modèles de requête

- Combinaison du modèle de contexte courant et contexte passé (Tan, 2006)

$$p(t|\theta_{q_k}, H_k) = \lambda_{q_k} p(t|\theta_{q_k}) + (1 - \lambda_{q_k}) p(t|\theta_{H_k})$$

Modèle de distribution du terme t dans les résultats de la requête q_k

Modèle de distribution du terme t dans l'historique H

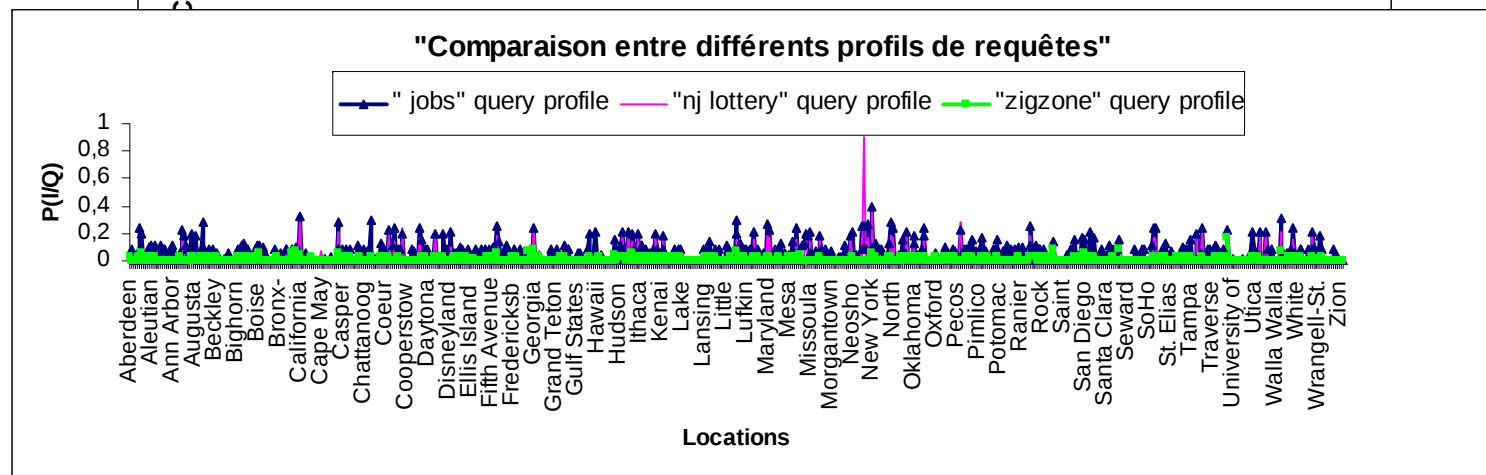
■ **Postraitement : réordonnement des résultats de recherche**

- Principe de combinaison (linéaire) du score initial (non personnalisé) et d'un score personnalisé
- Sources d'évidence pour le calcul du score personnalisé
 - ✓ Localisation géographique de l'utilisateur (Varma et al. 2006, Goenka et al., 2010, Bierig 2008)
 - ✓ Temps, météo, environnement social (Lan et al. 2010)
 - ✓ Termes (pertinents) du profil (Shen et. al 2005, Chen et al. 2002, Teevan et al. 2005)
 - ✓ Concepts (pertinents, sélectionnés) du profil (Gauch et. al 2003, Sieg et al. 2007, Daoud et al. 2010, Liu et. al 2004)

■ Personnalisation de requêtes sensibles à la localisation (Bouidghaghen et al. 2011)

- Etape 1 : calculer le profil de la requête pour détecter sa sensibilité pour la localisation

$$\hat{P}(I/Q) = \sum_{D \in P} \hat{P}(I/D) \cdot \frac{P(Q/D)}{\sum P(Q/D')}$$



$$P(I/Q) = \lambda \cdot \hat{P}(I/Q) + (1 - \lambda) \cdot \hat{P}(I/C)$$

- Personnalisation de requêtes sensibles à la localisation (Boudghaghen et. al 2011)

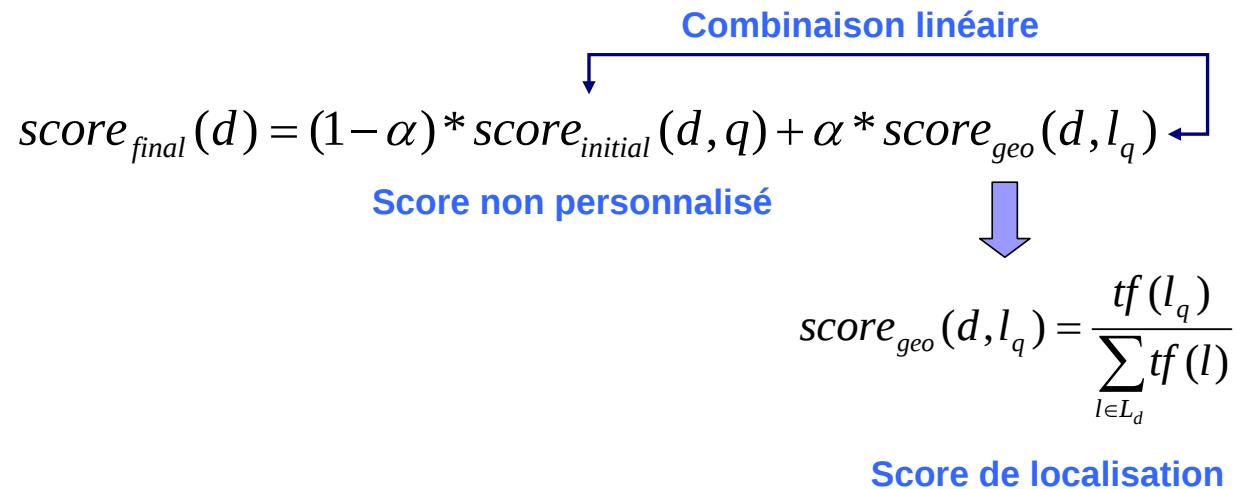
- Etape 2 : Personnaliser la recherche

Combinaison linéaire

$$score_{final}(d) = (1 - \alpha) * score_{initial}(d, q) + \alpha * score_{geo}(d, l_q)$$

Score non personnalisé

Score de localisation

$$score_{geo}(d, l_q) = \frac{tf(l_q)}{\sum_{l \in L_d} tf(l)}$$


■ Traitement : modèle d'appariement intégré

- Le profil utilisateur devient variable du modèle d'appariement /filtrage
- Sources d'évidence pour la définition de la variable « profil »
 - ✓ Tâche de filtrage (en RI mobile)
 - Catégories, sujet du profil (Belloti et al. 2008, Papadogorgaki et al. 2008, Malo et Siitari, 2010)
 - ✓ Tâche de recherche d'information (en RI personnalisée)
 - Termes du profil (Lin et al. 2005, Tamine et al. 2010)
 - Liens de navigation (Haveliwala et al. 2002, Qui et al. 2006)

- Quand le profil devient une variable d'appariement
 - Révision des modèles d'appariement vectoriel (Micarelli 2004)

$$Score(D) = \sum_{t \in D} \underline{Occ(t)} \times \underline{Rel(t)}$$

$$Occ(t) = k_1 \times freq_{body}(t) + k_2 \times freq_{title}(t)$$

$$\begin{cases} t \in P & Rel(t) = Rel(t) + k_1 \times \sum_j w_j \quad \forall w_j : t \in slot_j \text{ sin on} \\ t \in P \wedge t \in Q & Rel(t) = Rel(t) + w_{slot} \text{ sin on} \\ t \notin P \wedge t \in TDB \wedge t \in Q & Rel(t) = Occ(t) + k_2 \text{ sin on} \\ t \in P & Rel(t) = Rel(t) + w_j \sum_i A_i \quad \forall co\text{-keyword}_i \text{ of } slot_j : co\text{-keyword}_i \in D; \forall slot_j : topic_j \in D \end{cases}$$

■ Page Rank Personnalisé (PPR) (Qui, 2006)

➤ Extension du TSPR

✓ **Profil utilisateur :**

- L'utilisateur est une variable du modèle de ranking
- Un utilisateur est représenté par un vecteur de préférences

$$T = [T(1), \dots, T(m)], \sum_{i=1..m} T(i) = 1$$

m : nombre de sujets

$T(i)$: degré de préférence de l'utilisateur pour le sujet i
 $E_j(v) =$

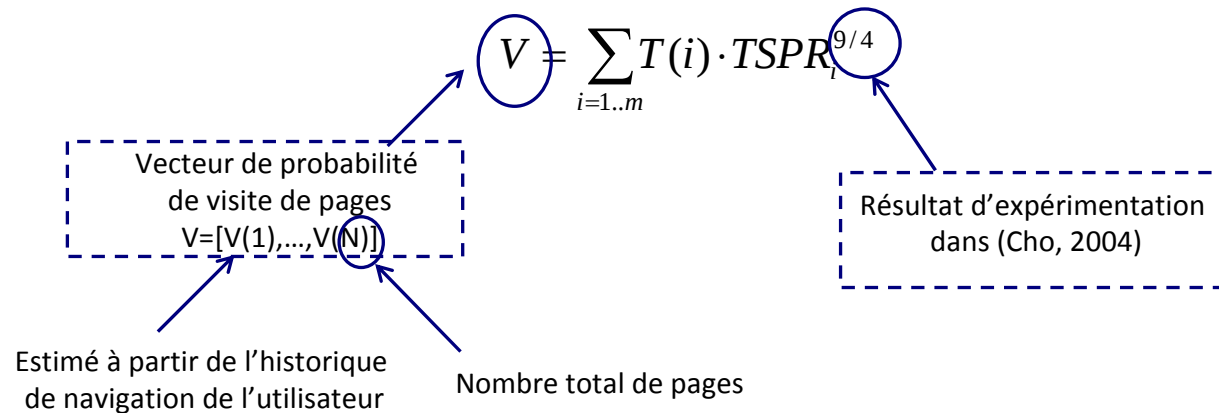
✓ **Modèle de navigation aléatoire orienté par les préférences thématiques (topical preferences)**

- L'utilisateur considère un sujet t avec une probabilité $T(t)$
- Il atteint une page v traitant du sujet de façon équiprobable relativement aux pages qui traitent du sujet ($E_t(v) \neq 0$)
A partir de la page v , avec une probabilité d , il atteint une autre page traitant du même sujet t , avec une probabilité $(1-d)$, il atteint une page qui traite d'un nouveau sujet

■ Page Rank Personnalisé (PPR) (Qui, 2006)

➤ Algorithme de construction du vecteur de préférences

- ✓ **Objectif** : Apprendre le vecteur $T = [T(1), \dots, T(m)]$
- ✓ **Problème** : Trouver le vecteur T qui satisfait



✓ **Solution**

- ✓ Régression linéaire

$$T = (TSPR^t TSPR)^{-1} TSPR^t V$$

- ✓ Maximum de vraisemblance $T = \arg \max_T \left(\prod_{i=1}^k V_T(p_i) \right)$

■ Page Rank Personnalisé (PPR) (Qui, 2006)

➤ Ordonnancement personnalisé

$$PPR_T(p) = \sum_{i=1}^m T(i) * P(Q/T(i)) * TSPR_i(p)$$

PageRank personnalisé (U->T)

w_i

TSPR

■ Questions ouvertes

1. Personnalisation systématique ou sélective ?

To personalize or not to personalize? (Teevan et al. 2008)

- Potentiel de personnalisation diffère d'une requête à l'autre : requête informationnelle, ambiguë, multi-domaines, tâche, etc.

■ Questions ouvertes

2. Personnalisation et vie privée : quel compromis ?

- Appréhension, polémiques : « Beacon advertising », « Instant personalization »,...
- Quelques pistes qui réduisent les risques (Toch et al, 2012) :
 - Personnalisation via pseudonymes
 - Personnalisation au niveau client
 - Encryptage (bruitage) des données de personnalisation

■ Questions ouvertes

3. Comment évaluer un SRI personnalisé ?

- Limites de l'évaluation à la « Cranfield »
- Quelques pistes :
 - Analyse des logs utilisateurs (*log analysis*)
 - Etudes basées utilisateurs (*User study*)
 - Etudes journalières (*Diary study*)
 - ...TREC depuis 2010 : Session track, Context suggestion track