Recherche d'Information Contextuelle

Centres d' Intérêt, Localisation, Facteurs (Média-)Sociaux

Lynda Tamine-Lechani
lynda.lechani@irit.fr
http://www.irit.fr/~Lynda.Tamine-Lechani/

Objectifs

- **Définir la recherche d'information contextuelle** et introduire le vocabulaire associé
- **Caractériser le contexte**, ses dimensions et les interactions entre dimensions
- **Présenter** et **illustrer** les catégories de **méthodes de construction** et de **modélisation du contexte** centrés sur les **centres d'intérêt**, la **localisation géographique**, les **signaux et le voisinage sociaux** des utilisateurs
- **Présenter** et **illustrer les catégories de modèles de recherche d'information** qui **utilisent** le **contexte**

Structure du du cours

Recherche d'Information Contextuelle : Quoi, Pourquoi et Comment ?

- *Recherche d'information contextuelle : Quoi et Pourquoi ?*
  - ✓ Motivations
  - ✓ Vocabulaire
  - ✓ Problèmes scientifiques
  - ✓ *Taxonomies du contexte*
- *Recherche d'information contextuelle : Comment ?*
  - ✓ Modélisation du contexte
    - Données observées, facteurs contextuels inférés, outils formels et méthodologiques
  - ✓ *Utilisation du contexte pour la recherche d'information*
    - Techniques de reformulation de requête, de (ré)ordonnancement basées sur le contexte

Cadre général : principaux éléments

- Tâche de recherche d'information
  - **Requête** : besoin en information
  - **Documents** : résultats de la requête
  - *Objectif* : retourner la **réponse pertinente à la requête**

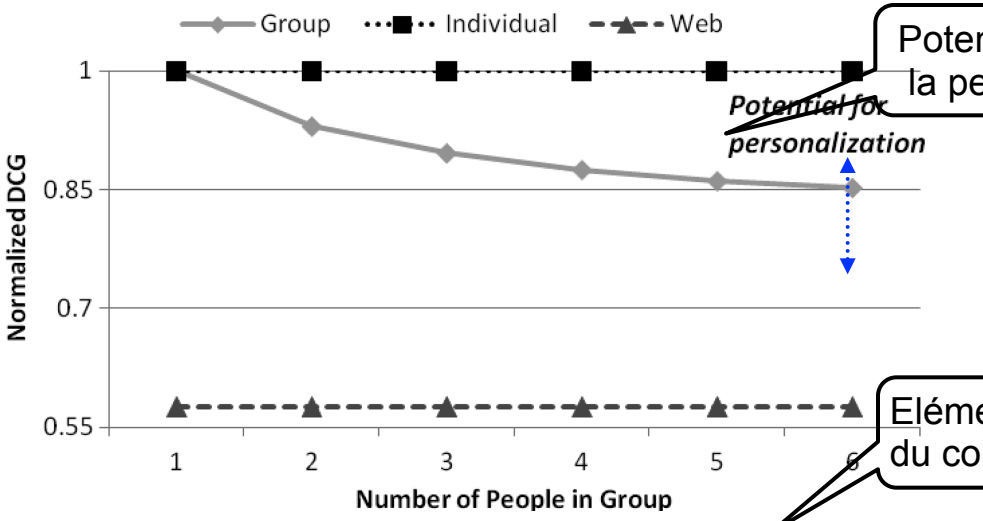
Mais aussi…….

  - **Eléments autres** que la **requête** et les **documents : contexte**
    - *Utilisateur* : préférences, âge, profession, réseau social, localisation
    - *Tâche* : loisir, professionnelle, …
    - ….
  - *Objectif* : retourner la **réponse pertinente à la requête dans son contexte**

Motivations : Résultats d'études empiriques

- Elément de contexte étudié : **centres d'intérêt**
  - Teevan et al. ACM TCHI 2010, Dou et al. WWW 2007
  - Etude Teevan et al., ACM TCHI 2010: *Potential for Personalization*
    - ✓ 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

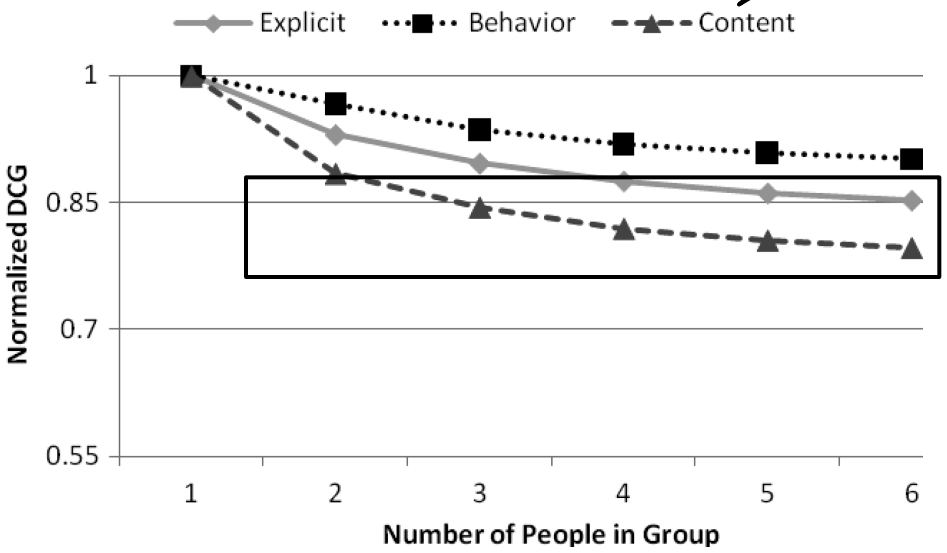
Motivations : Résultats d'études empiriques

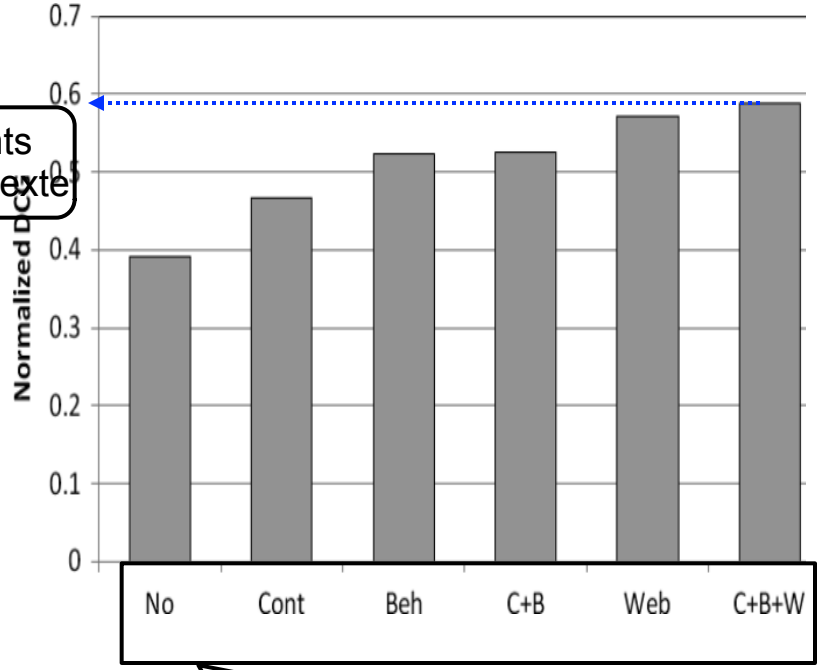


Potentiel de la personnalisation

La prise en compte de l'utilisateur (individu) améliore les performances de recherche

Eléments du contexte

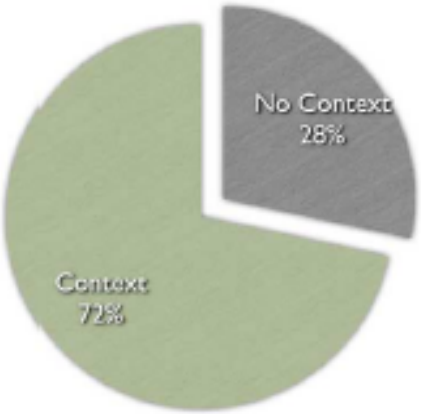




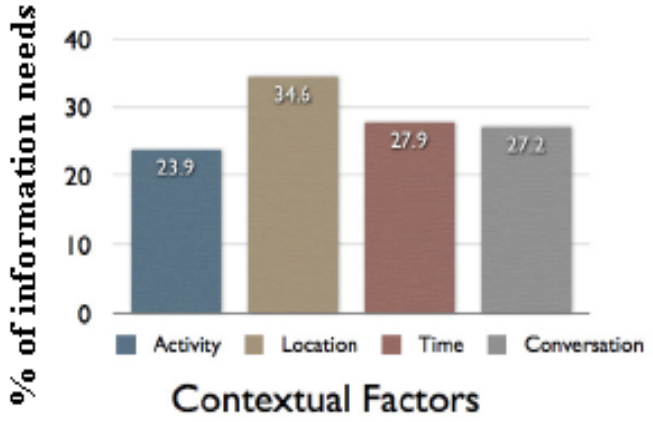
Scénarios de RI contextuelle

Motivations : Résultats d'études empiriques

- Elément de contexte étudié : **localisation**
  - ✓ Kamvar and Baluja CHI 2006, Sohn et al. CHI 2008, Bierig and Göker IIIX 2006



(a)



(b)

- **Principal résultat :** besoin de considérer la localisation de l'utilisateur pour évaluer sa requête
  - ✓ **72% des requêtes** expriment des **besoins sensibles à la localisation**
  - ✓ **Requêtes courtes** (2.3-2.5 termes/requête), **donc** ambigües
  - ✓ **Sessions courtes** (1.6 requêtes/session)

Motivations : Résultats d'études empiriques

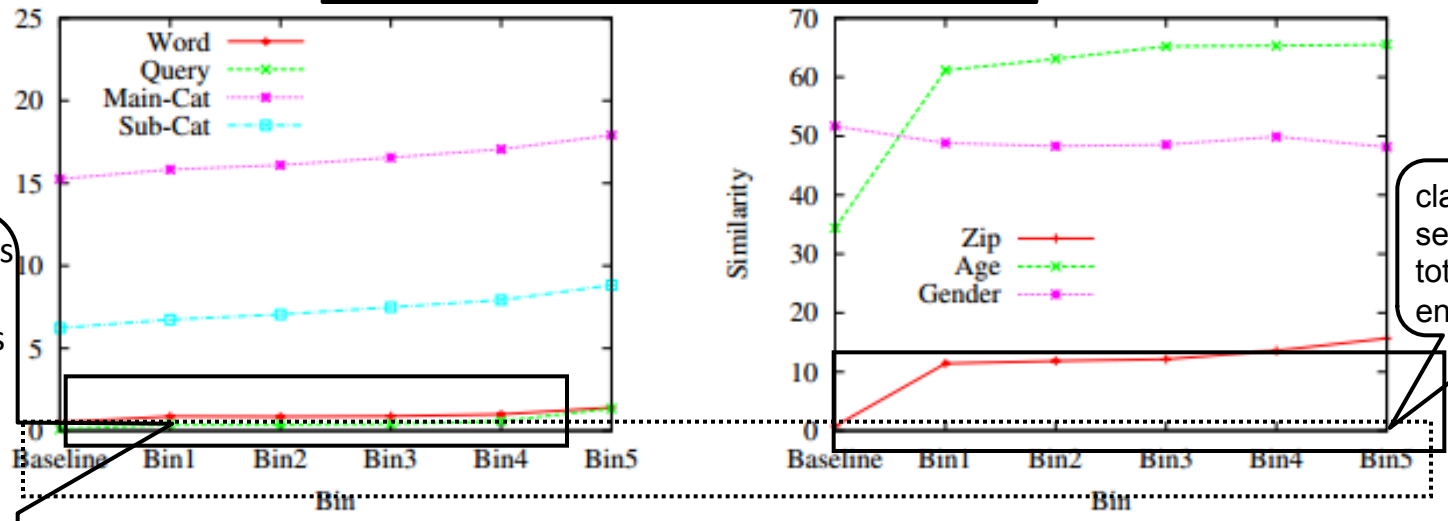
- Elément du contexte étudié : **voisinage social (média sociaux)**
  - Lee et Brusilovsky ACM HT 2010, Singla et Richardson WWW 2008, Chelaru et al. WISE 2012

  - Etude Singla et Richardson WWW 2008: *Yes, There is a Correlation- From Social Networks to Personal Behaviour on the Web*

    Utilisateurs : 25 billions sessions MSN chat, 162 millions utilisateurs, 3.3 billions
    paires d'utilisateurs ayant interagi au moins une fois
    Collecte de données : âge, sexe, département, nombre total de messages chat échangés entre deux utilisateurs, temps total passé en interaction, requêtes, sujets des requêtes (catégories sémantiques)…

    ✓ Question de recherche *:* ***"SI A interagit souvent à B et C est un autre utilisateur choisi aléatoirement, est ce que A ressemble d'avantage à B qu'à C?"***

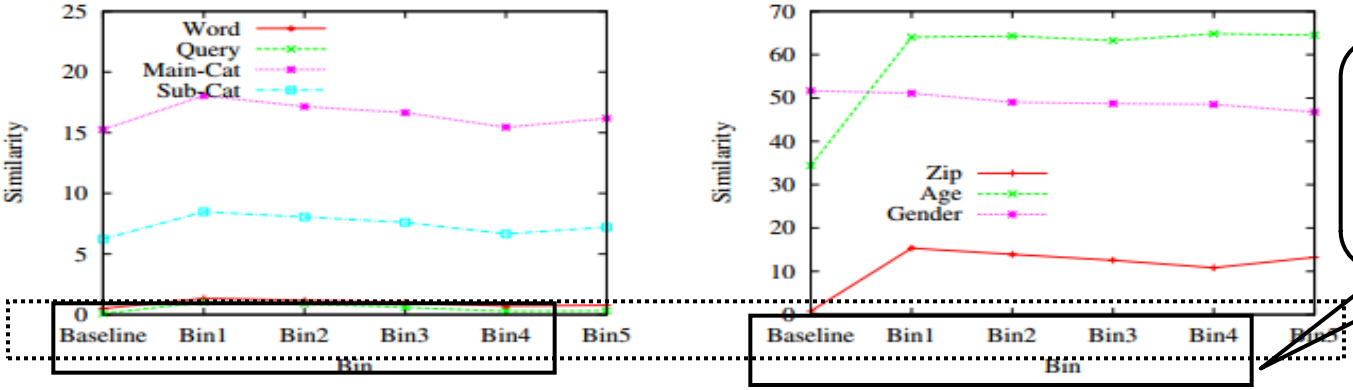
A ressemble d'avantage à B qu'à C



Baseline: temps calculé sur les attributs de toutes les paires d'U possibles

classes d'U selon le temps total passé en interaction

Figure 3: Variation in similarities(%) with total talk duration: query attributes(left) and personal attributes(right)



classes d'U selon le temps moyen passé par message

Figure 5: Variation in similarities(%) with average time spent per message: query attributes(left) and personal attributes(right)

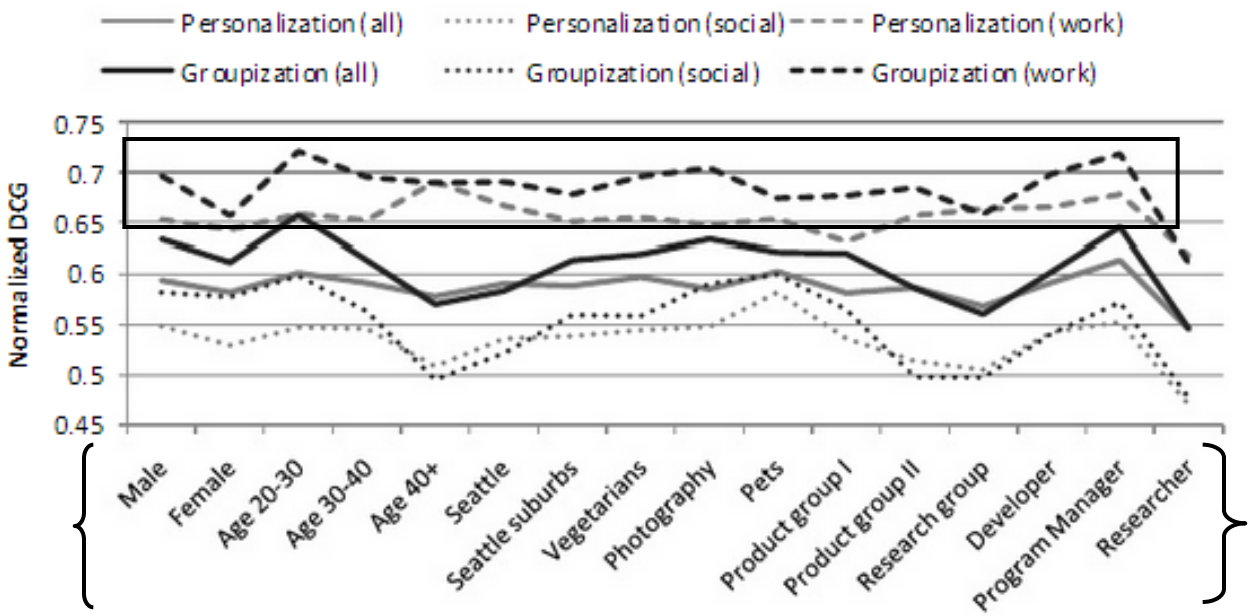
Motivations : Résultats d'études empiriques

- Elément de contexte utilisé : **voisinage social (groupe social)**
  - Teevan et al. WSDM 2009: *Discovering and Using Groups to Improve Personalized Search*
    - ✓ Plus de 160 utilisateurs employés de Microsoft de différents groupes
      - Ayant travaillé ensemble sur les mêmes tâches
      - Ayant la même fonction, rôle dans la même entrprise
      - Travaillant ou vivant dans la même région
    - ✓ Requêtes et jugements de pertinence : requêtes prédéfinies, jugements collectés à trois niveaux (très pertinent, pertinent, peu pertinent)
    - ✓ Collecte de données de profils : e-mails, desktop, pages visitées, documents produits etc.

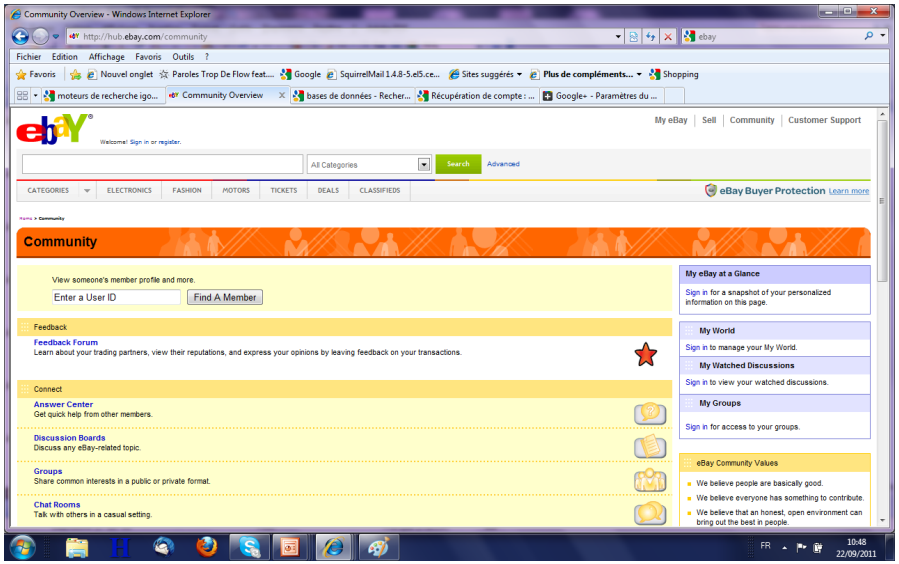
Motivations : Résultats d'études empiriques

La prise en compte du facteur « groupe social » dans le modèle d'appariement améliore la précision de la recherche

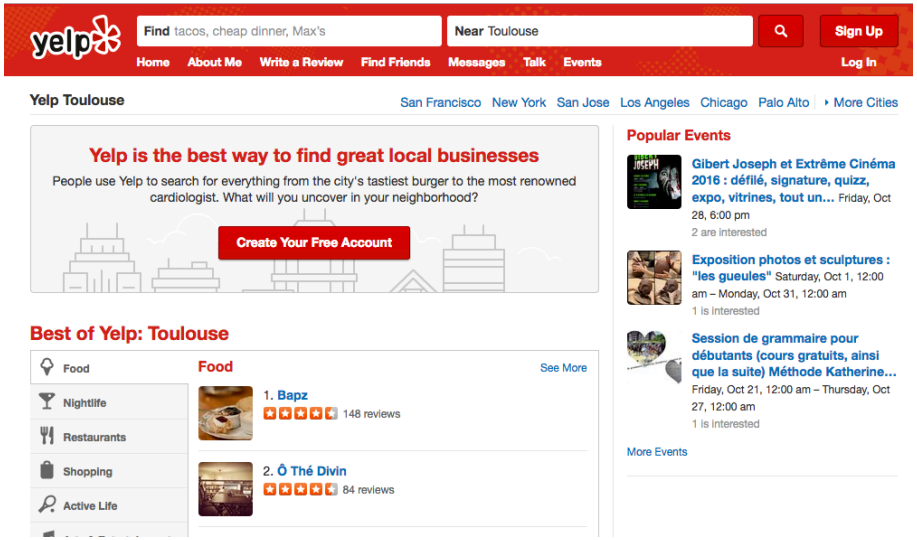
Group		Queries	
		All	Group
Pets	In	0.52	0.40
	Out	0.47	0.31
	Diff.	10%	26%
Photography	In	0.56	0.64
	Out	0.46	0.35
	Diff.	23%	82%
Vegetarianism	In	0.49	0.65
	Out	0.48	0.41
	Diff.	1%	58%
Work groups	In	0.52	0.60
	Out	0.47	0.54
	Diff.	9%	11%
Task groups	In	0.42	0.77
	Out	0.31	0.35
	Diff.	34%	120%

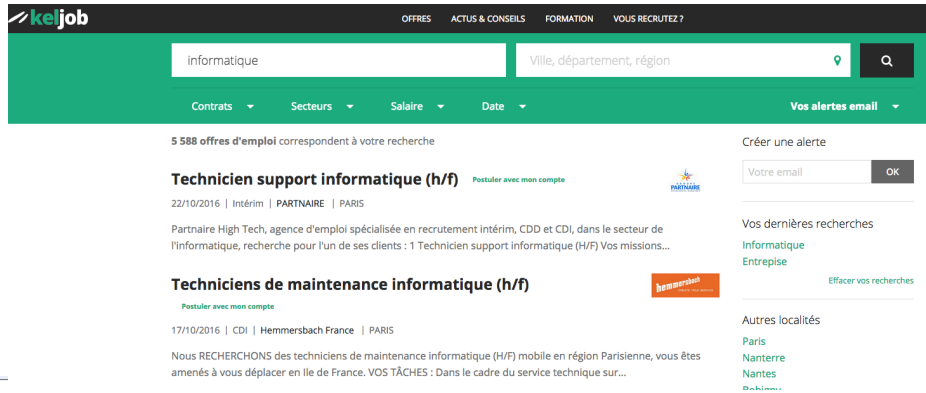


Cadre général : de très…..très nombreuses applications

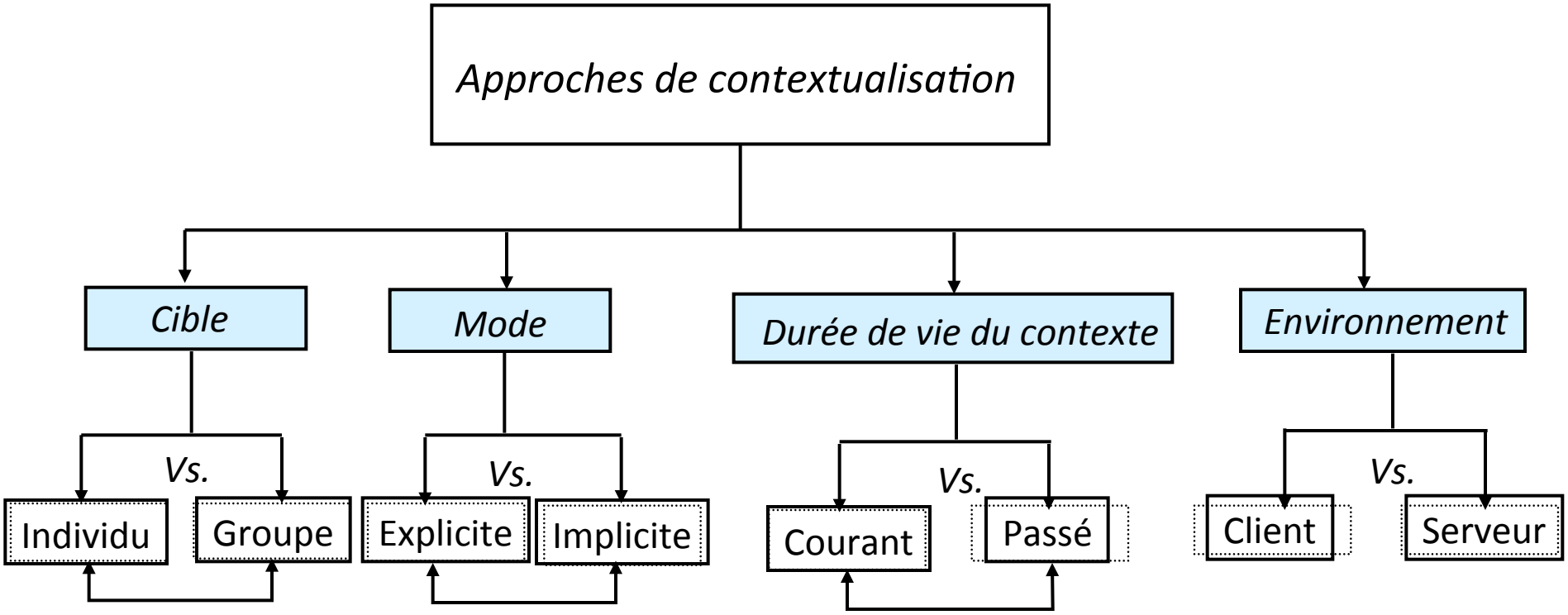


Cadre général : de très…..très nombreuses applications



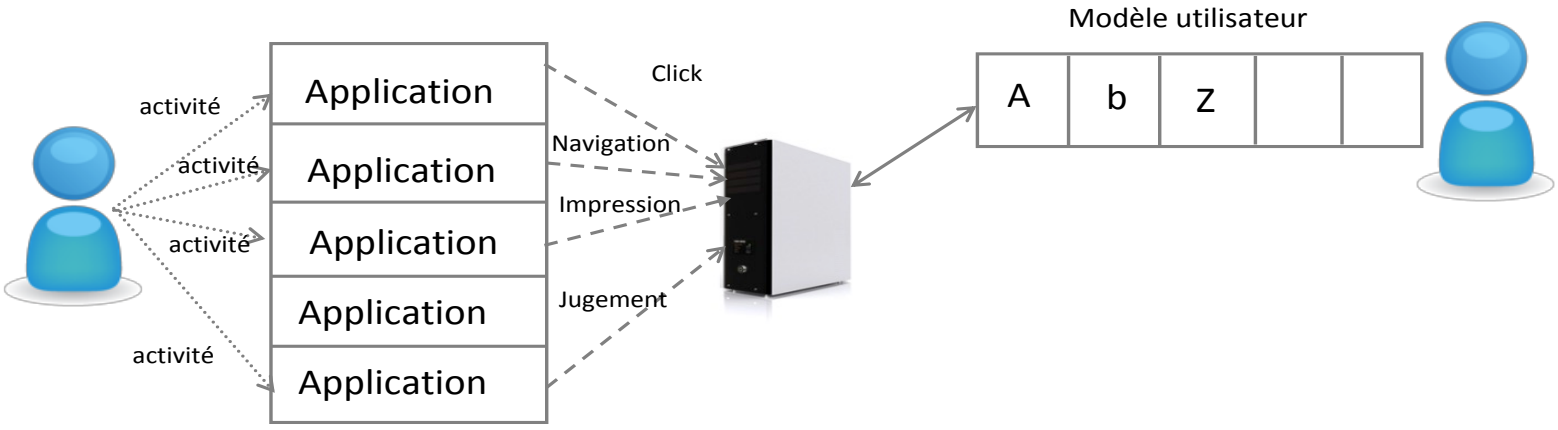


Approches de contextualisation

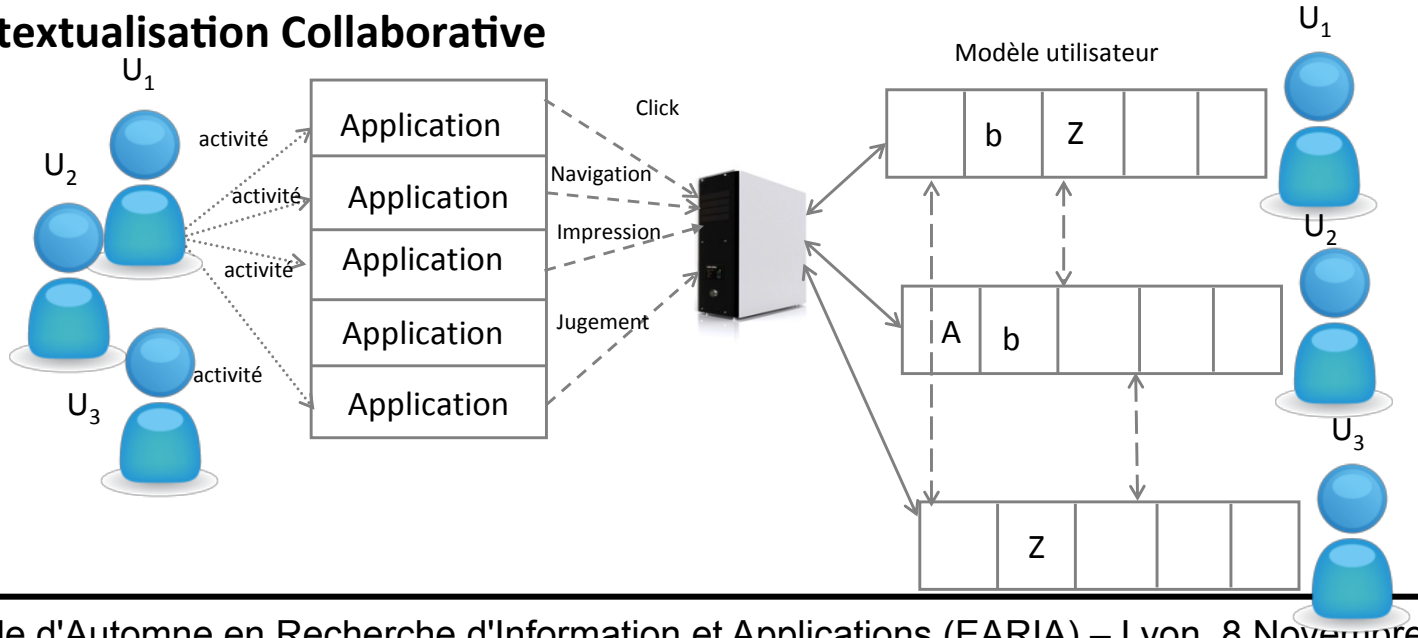


Contextualisation Individuelle Vs. Collaborative

Contextualisation Individuelle



Contextualisation Collaborative



Mode de Contextualisation Explicite Vs. Implicite

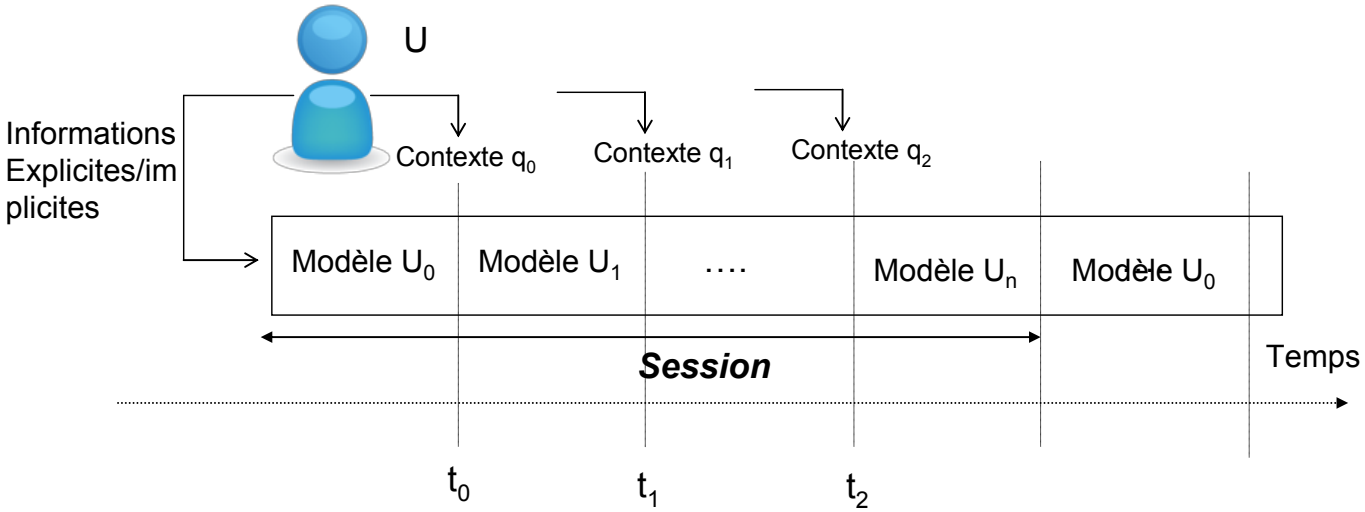
- Mode **explicite** : collecte des données descriptives du contexte **à partir de l'utilisateur lui même**
  - ✓ *Formulaires* : cases à cocher, saisie de mots clés
  - ✓ *Interfaces élabo*rées : expression d'exemples, contre exemples, votes, notes, annotations etc.
  - ✓ *Questionnaires*

- Mode **implicite** : **dérivation automatique du contexte à l'aide d'algorithmes qui utilisent :**
  - ✓ Les *applications utilisées*
  - ✓ *L'historique de localisations*
  - ✓ *Les interactions de l'utilisateur* : Mouvements des yeux, Données de *clicks*, Actions sur les documents (Données de navigation, temps de lecture etc.), Messages (e-mails) envoyés ou reçus, Annotations sociales, *Bookmarks, etc.*

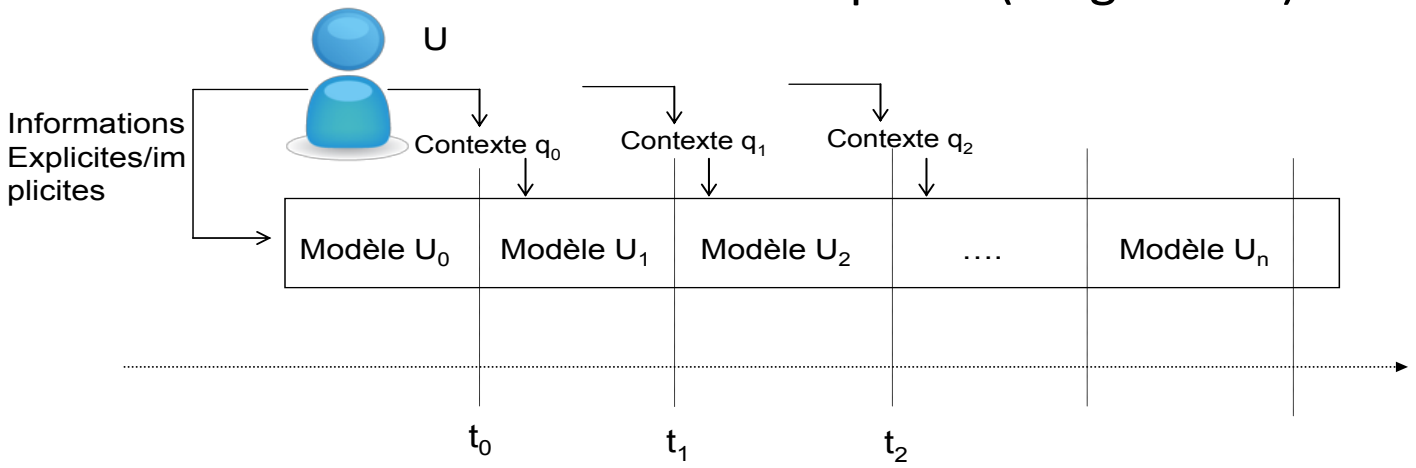
—— Mode de Contextualisation le plus utilisé en Recherche d'Information

Contextualisation à court-terme Vs. Long-terme

- Contextualisation basée sur le contexte courant (court-terme)



- Contextualisation basée sur le contexte passé (long-terme)



Cadre général : Questions Scientifiques Critiques

① Quels sont les éléments du contexte qui impactent un processus de RI ?

② Comment capturer, modéliser le contexte ?

③ Comment utiliser le contexte pour mieux répondre à la requête ?

Cadre général : Questions Scientifiques Critiques

① **Quels sont les éléments du contexte qui impactent un processus de RI ?**

② Comment capturer, modéliser le contexte ?

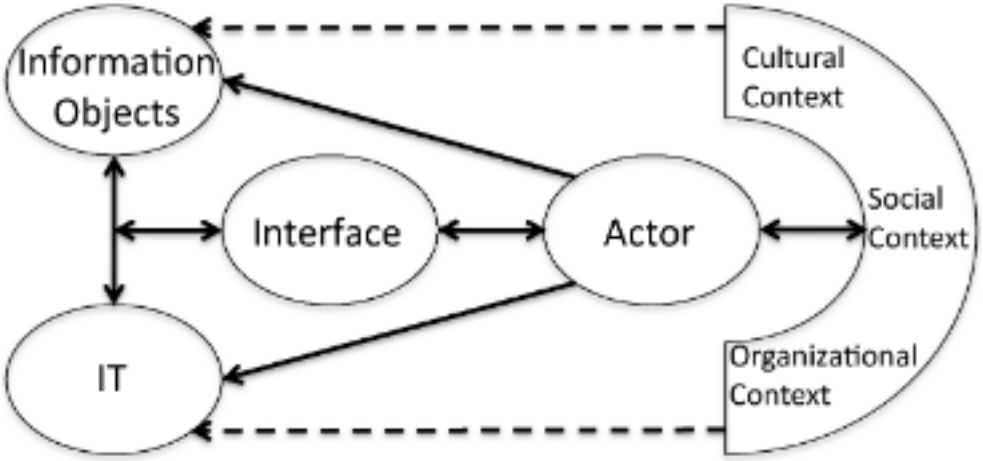
③ Comment utiliser le contexte pour mieux répondre à la requête ?

Plusieurs taxonomies, modèles

- Modèle Ingwersen et Jarvelin (1978)
- Modèles Saracevic (1997)
- Taxonomie Fuhr (2000)
- Taxonomie Dey and Abowd (2000)
- Taxonomie Goker and Maryhuang (2002)
- Modèle Ingwersen et Jarvelin (2005)
- Taxonomie Tamine and al (2010)
- Taxonomie Kaenampornpan and O'Neill (2004)
- Taxonomie Arias and al (2010)
- …

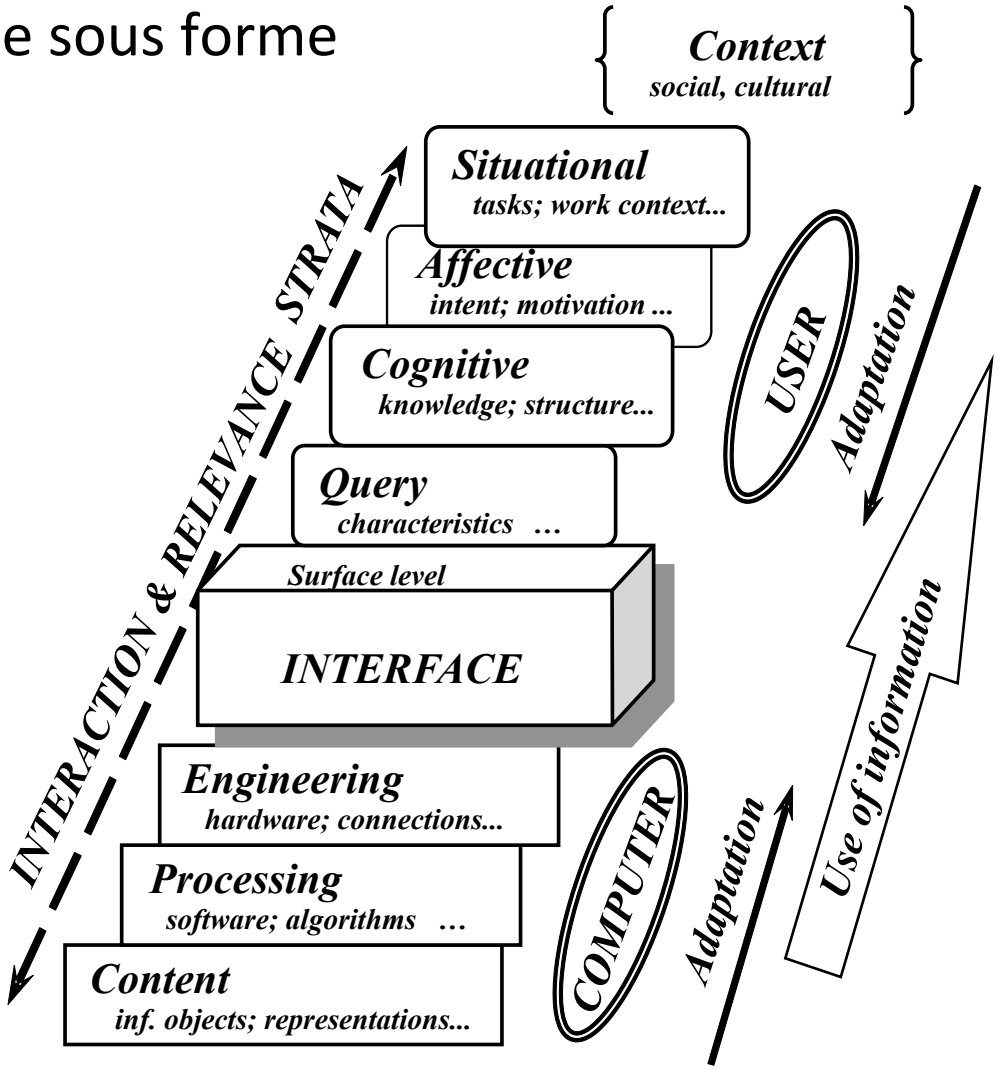
Taxonomie Ingwersen et Jarvelin, The Turn, 2005

Objets d'informations, Interfaces et acteurs



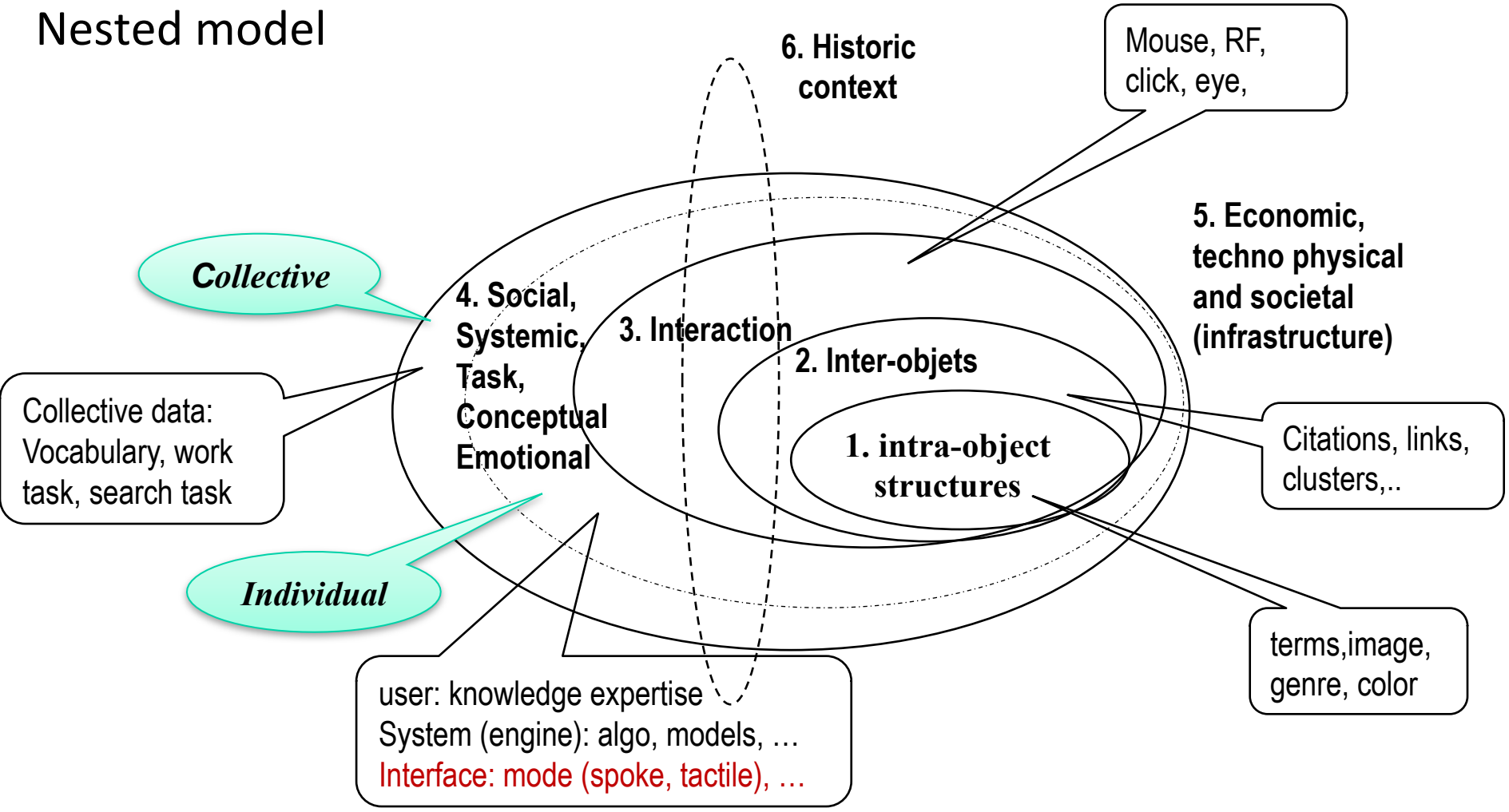
Taxonomie de Saracevic, Saracevic ASSIS 90

Le contexte sous forme de strates



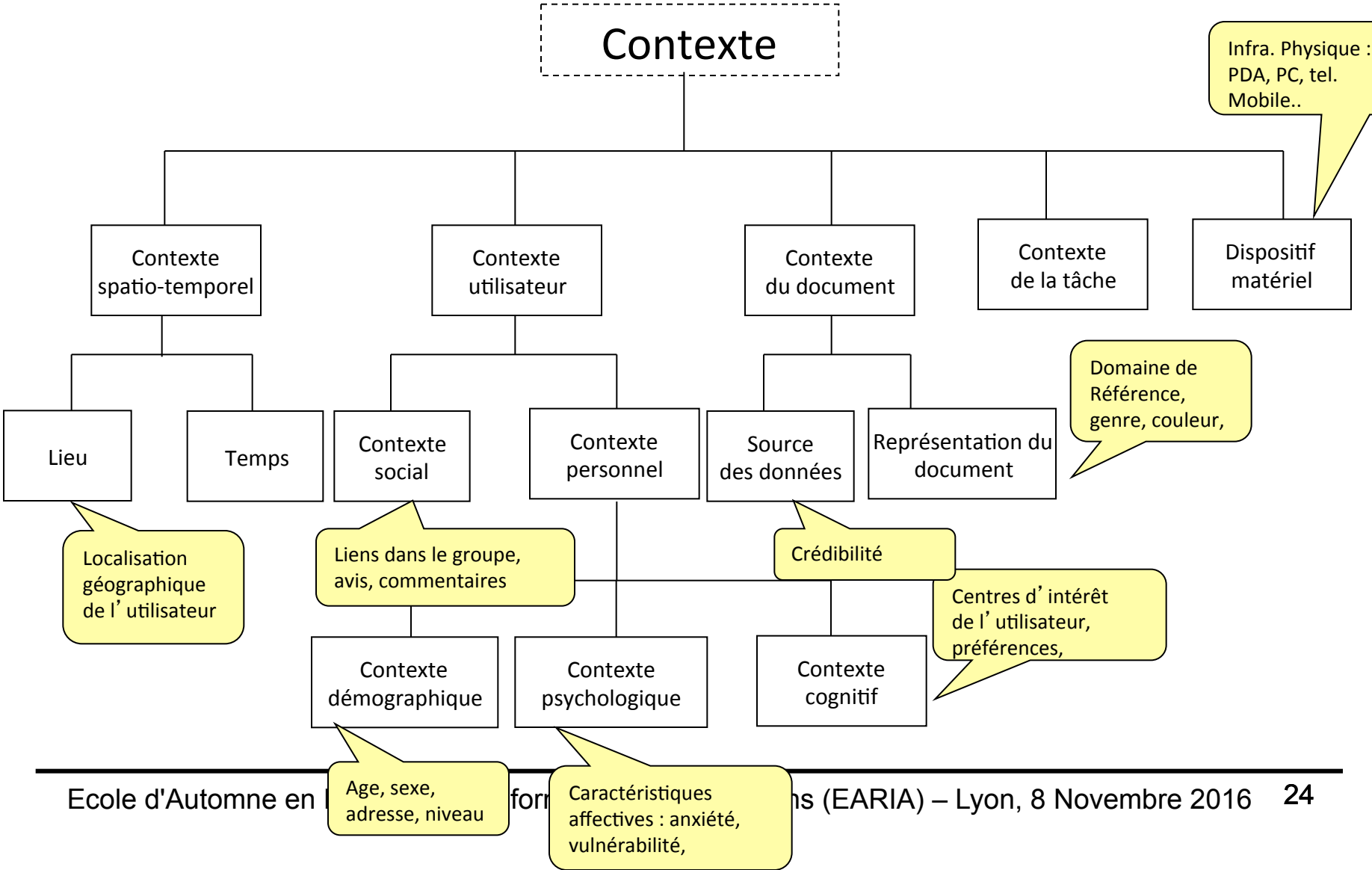
Taxonomie de Ingwersen, The Turn 2005

Nested model



Taxonomie de Tamine et al. , Knowledge and Information Systems 2010

Le contexte comme une hiérarchie de facteurs



Cadre général : principaux contours thématiques

RI orientée tâche

RI sociale

RI contextuelle

RI Mobile

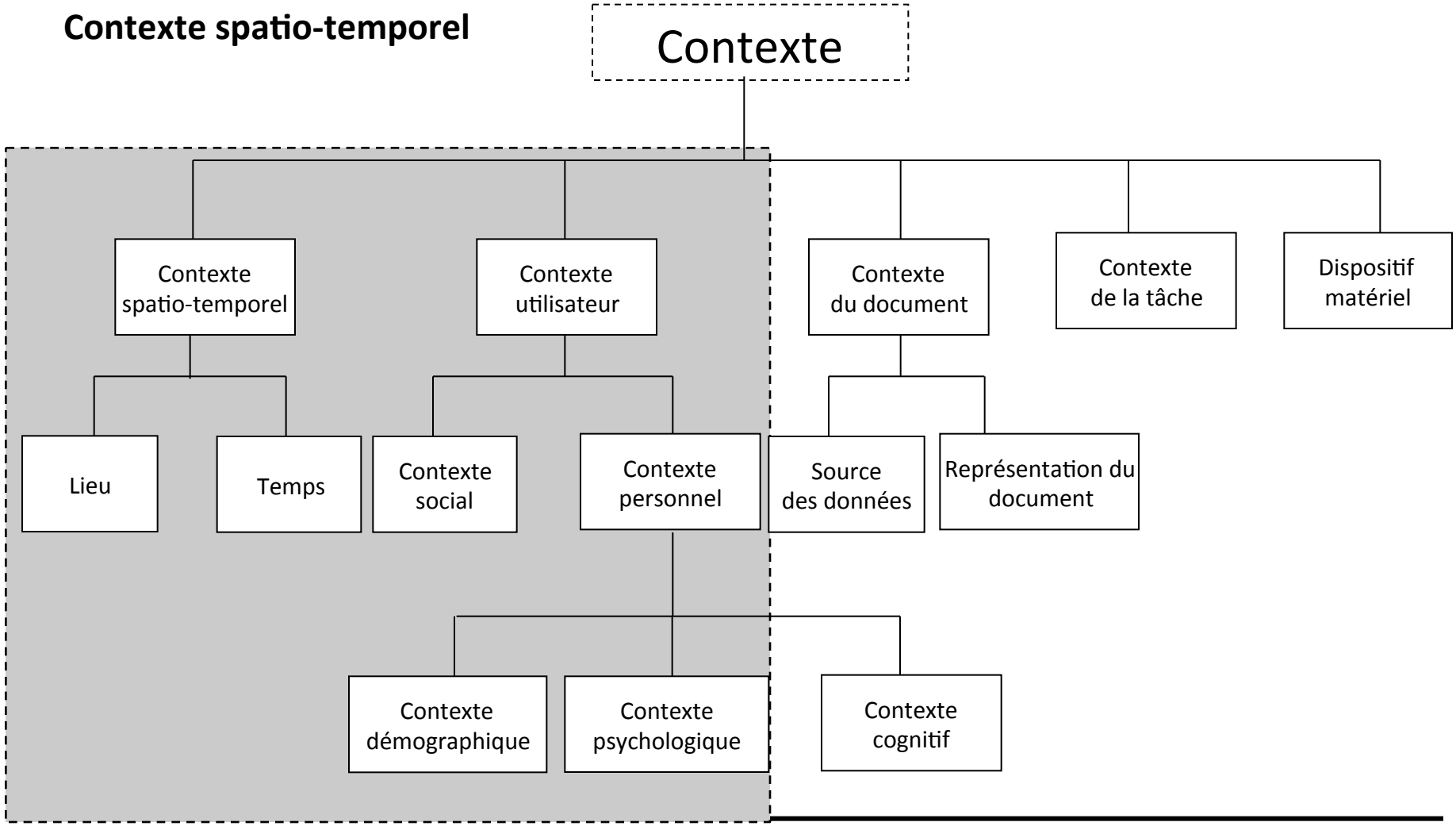
RI Interactive

RI Personnalisée

Quel Type de 'Contexte' dans ce Cours ?

Contexte utilisateur : personnel, social

Contexte spatio-temporel



Cadre général : Questions Scientifiques Critiques

① Quels sont les éléments du contexte qui impactent un processus de RI ?

② Comment capturer, modéliser le contexte ?

③ Comment utiliser le contexte capturé pour mieux répondre à la requête ?

Principe général

- Distinguer entre observations et facteurs dérivés du contexte

Feedback implicite

activité

activité

activité

activité

Application

Application

Application

Application

Application

Click

Navigation

Impression

Jugement

Algorithme de construction du contexte

Feedback explicite

Contexte

Méthodes de l'état de l'art

- Spécifiques au facteur de contexte
  - ✓ **Données démographiques :** âge, sexe,… (Zhang et al. ICWSM 2016, Wang et al. WSDM 2016, Zhong WSDM 2015, Dong et al KDD 2014)
  - ✓ Profession ou secteur d'activités (Hu et al. ICWSM 2016, Preotiuc-Pierto et al. ACL 2015)
  - ✓ **Localisation** : (Kinsella et al. SMUC 2011, Backstorm et al. WWW 2010, Leung et al. ICDE 2010, Bila et al. MDM 2008)
  - ✓ **Centres d'intérêts :** (Xing et al. CIKM 2013, Leung et al. ICDE 2010, Daoud et al. KAIS 2009, Liu et Weng TKDE 2004)
  - ✓ Personnalité (Liu et al. ICWSM 2016, Schwartz et al. PloS ONE 2013)
  - ✓ Orientation politique (Pennacchiotti et Popescu ICWSM 2011)

Facteurs de Contexte les plus utilisés en Recherche d'Information

Méthodes de l'état de l'art : Inférence des données démographiques

Wang, Guo, Lan et al. WSDM 2016. *Your cart tells you: Inferring Demographic Attributes from Purchase Data*

- ***Données de contexte inférées*** : sexe, âge, statut marital, niveau de formation
- ***Données utilisées pour inférer le contexte*** : achats sur le web
- ***Méthode***
  - ✓ Deux (2) problèmes de prédiction :
    1. Etant donné un ensemble incomplet de données démographiques d'utilisateurs, prédire les données manquantes

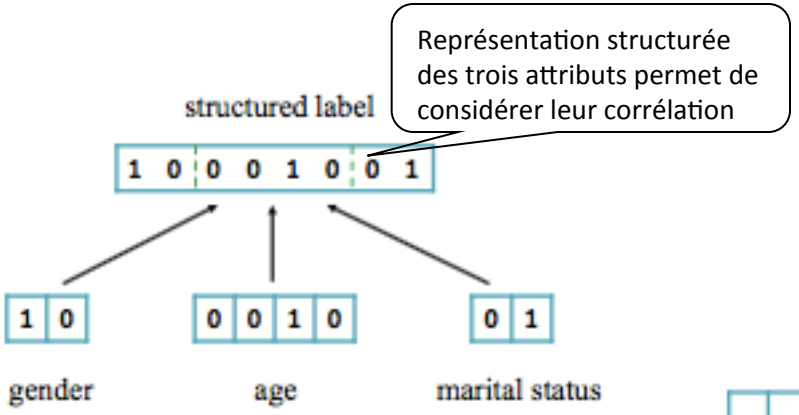
Ensemble d'utilisateur avec données partielles

$$f : X \rightarrow Y^{U}$$

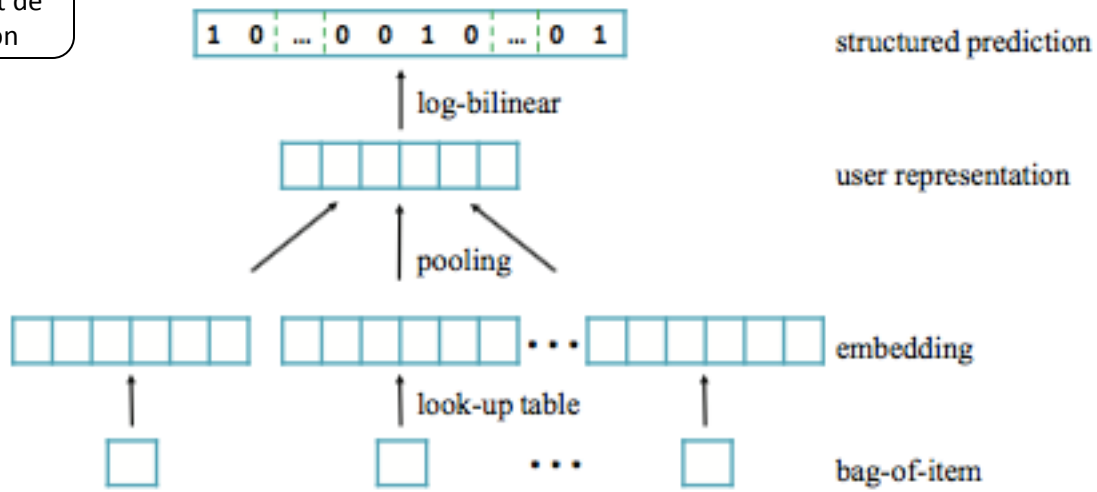
Ensemble d'attributs à inférer sur le même ensemble X d'utilisateurs

    2. Etant donné un ensemble d'utilisateurs avec des données démographiques, inférer les données démographiques pour de nouveaux utilisateurs $f : X^{N} \rightarrow Y^{N}$

Méthodes de l'état de l'art : Inférence des données démographiques



Représentation des attributs



Architecture neuronale du modèle de prédiction

Utilisateur U ($x^{(i)}$, $y^{(i)}$),
$x^{(i)}$: historique des achats
$y^{(i)}$: attributs démographiques

Descripteur agrégé des objets achetés par xi

$$p(y^{(i)} \mid x^{(i)}) = \frac{\sum_{y^{(i)} \in y^{(i)}_{partial}} \exp(v^{(i)} W y^{(i)})}{\sum_{y^{(i)} \in Y} \exp(v^{(i)} W y^{(i)})}$$

Probabilité d'assigner les attributs y sachant l'utilisateur x

Matrice des interactions utilisateurs-objets

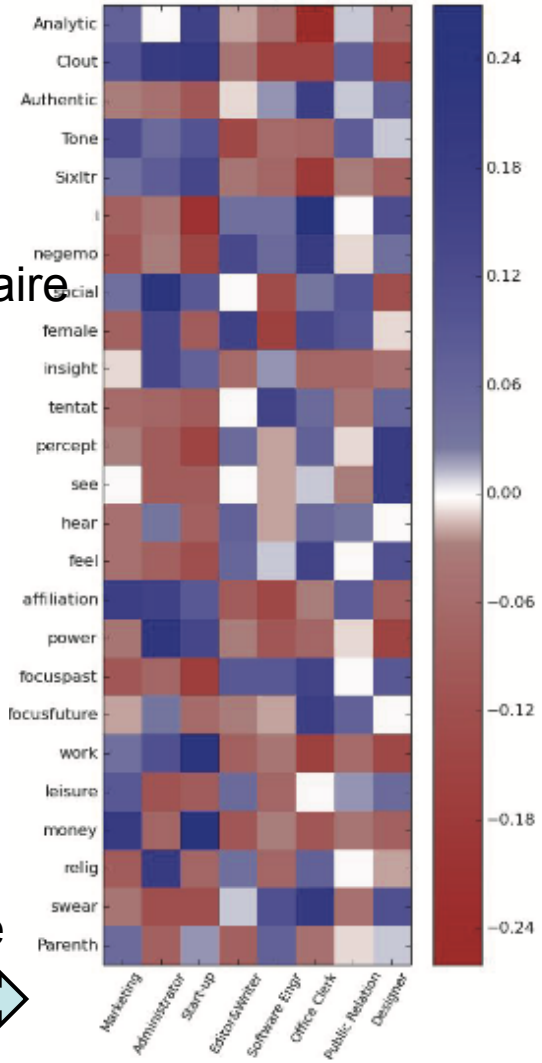
Méthodes de l'état de l'art : Inférence de la profession

Hu, Xiao, Jiebo et al. ICWSM 2016. *What the Language you Tweet Says about your Occupation*

- ***Données de contexte inférées*** : profession
- ***Données utilisées pour inférer le contexte*** : *tweets* des utilisateurs
- ***Méthode***
  - ✓ Deux (2) tâches:
    1. Catégorisation des activités : corrélations avec un vocabulaire fermé, ouvert, ..
    2. Catégorisation des professions selon les caractéristiques pertinentes identifiées en 1

Méthodes de l'état de l'art : Inférence du secteur d'activités

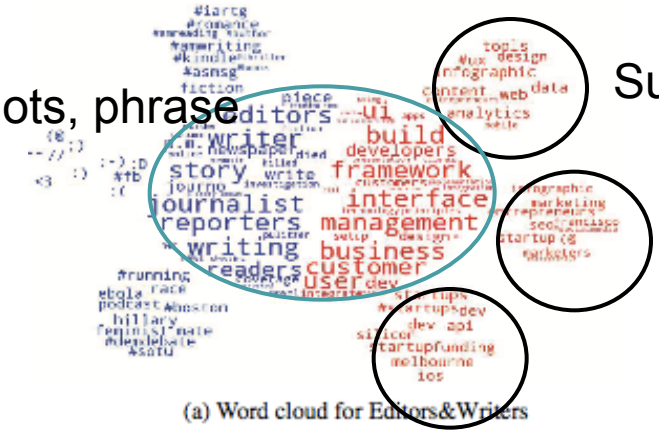
Vocabulaire



Corrrélation

Activité

Mots, phrase



Sujet

(a) Word cloud for Editors&Writers

Classifieur

Profession
…

Caractéristiques linguistiques : poids des mots du vocabulaire, Sujets, fréquence des mots phrases...

Méthodes de l'état de l'art : Inférence de la localisation

Kinsella, Murdock and Ohare, SMUC 2011. *I'm Eating a Sandwich in Glasgow. Modeling Locations with Tweets*

- ***Données de contexte inférées*** : localisation des utilisateurs
- ***Données utilisées pour inférer le contexte*** : *tweets* des utilisateurs
- ***Méthode***
  - ✓ Calculer la vraisemblance de la localisation sachant le *tweet*
  - ✓ Ordonnner les localisations selon cette vraisemblance

Méthodes de l'état de l'art : Inférence de la localisation

1. Calculer la vraisemblance de localisation sachant le tweet

Ensemble des localisations

Tweet

Vraisemblance de la localisation sachant le tweet

$$P(L/T) = \frac{P(T/\theta_L)P(L)}{P(T)}$$

Terme dans le tweet

$$p(T/\theta_L) = \prod_i p(t_i/\theta_L)$$

Fréquence d'un terme dans une localisation

$$p(T/\theta_L) = \frac{c(t,l) + \mu P(t|\theta)}{|L| + \mu}$$

2. Ordonnancement des localisations

$$KL(\theta_T/\theta_L) = \sum_t p(t/\theta_T) \log \frac{p(t/\theta_T)}{\alpha \times p(t/\theta_L)}, \alpha = \frac{\mu}{\mu + |L|}$$

Méthodes de l'état de l'art : Inférence des centres d'intérêt

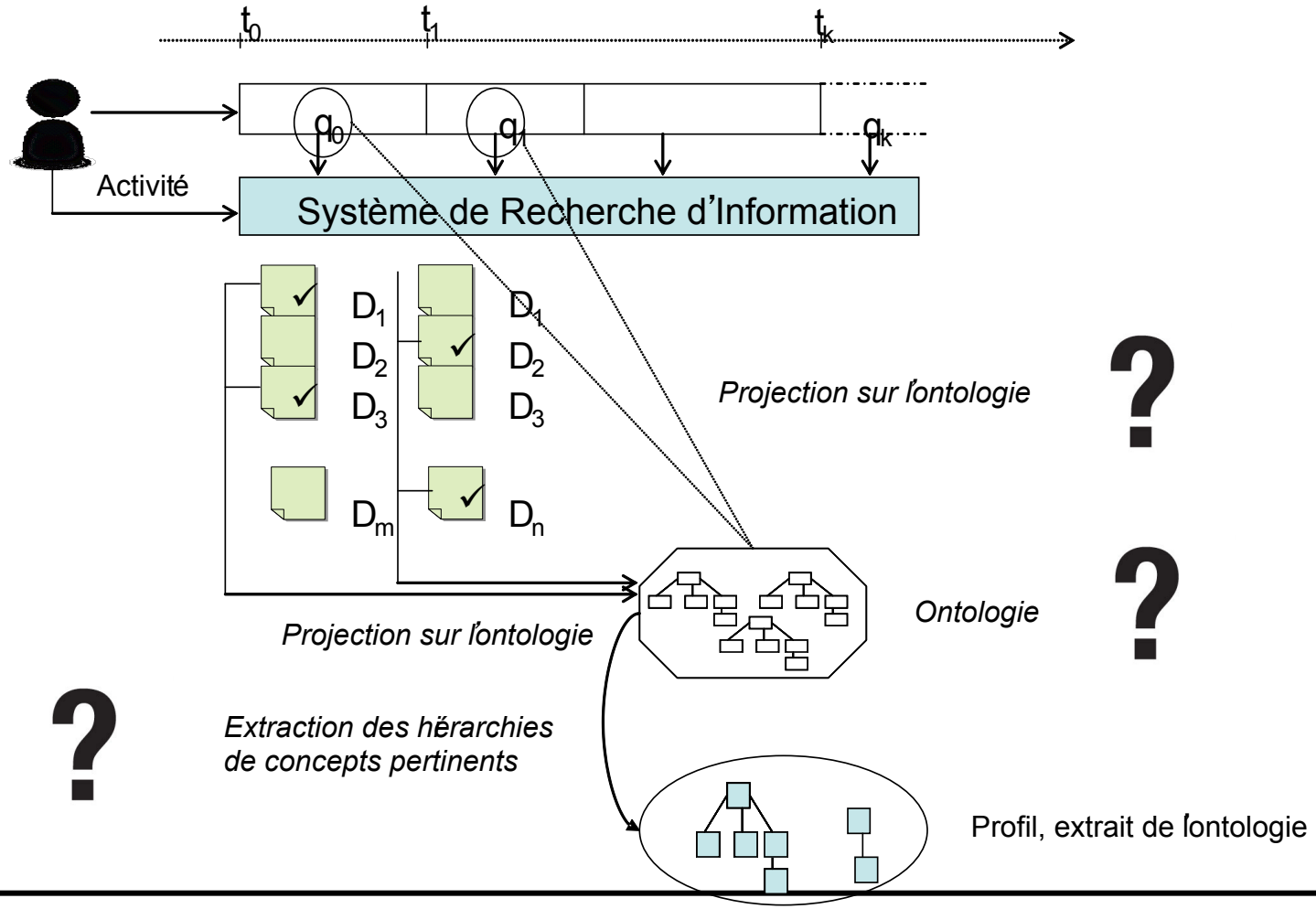
Daoud, Tamine, Boughanem KAIS 2010 *Towards a Graph-Based User Profile Modeling for a Session-Based Personalized Search*

- ***Données de contexte inférées*** : centres d'intérêts à court-terme
- ***Données utilisées pour inférer le contexte*** : Historique de recherche (sessions de recherche : requêtes, résultats des requêtes)

- ***Méthode***
  - ✓ Calculer le profil sémantique du besoin : projeter la requête sur une ontologie de contenu (ODP)-> liste/graphe de concepts
  - ✓ Calculer le profil sémantique des résultats de la requête : projeter les documents résultats sur l'ontologie-> liste/graphe de concepts
  - ✓ Apparier les profils requête et documents
  - ✓ Si corrélation, ie, la requête est dans la même session, alors mettre à jour le profil du besoin

Méthodes de l'état de l'art : Inférence des centres d'intérêt

Cadre général



Méthodes de l'état de l'art : Inférence des centres d'intérêt

Profil utilisateur= liste de concepts

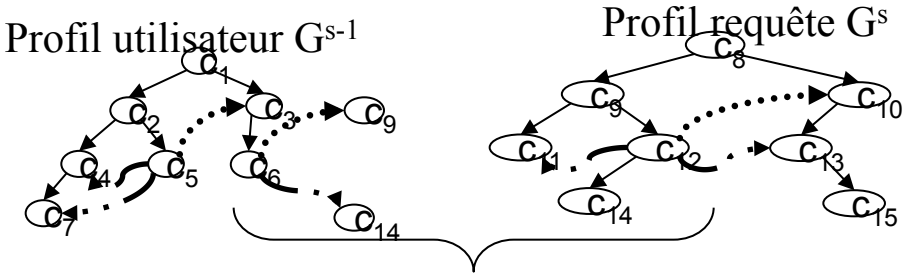
Profil utilisateur C^{s-1} | Profil requête C^s

$$C^{s-1} = \{...,(c^{s-1}{}_j, sw(c^{s-1}{}_j)),...\} \qquad C^s = \{...,(c^s{}_j, p(c^s{}_j)),...\}$$

Mise à jour par une fonction linéaire

$$p_{nouv}(C_j^s) = \begin{cases} \beta * p(C_j^{s-1}) + (1-\beta) * p(C_j^s) \text{ si } C_j \in C^{s-1} \\ \beta * p(C_j^s) \text{ sinon} \end{cases}$$
$$0 < \beta < 1$$

Profil utilisateur= graphe de concepts



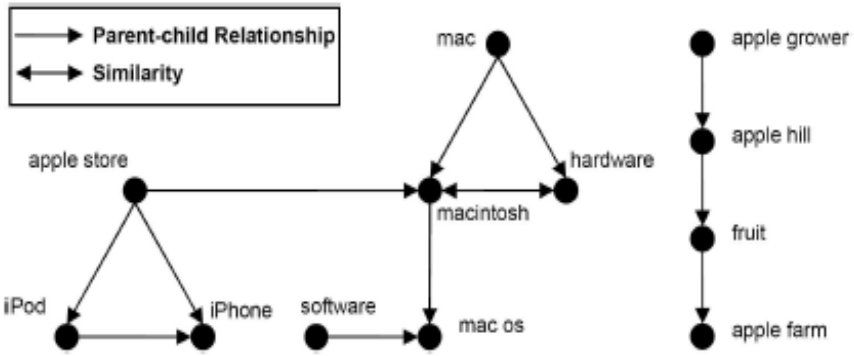
Combinaison de graphes

- accumulation des poids des concepts
- ajout des liens sémantiques

Méthodes de l'état de l'art : Inférence des centres d'intérêt

Leung, Lee and Lee ICDE 2010 *Personalized Web Search with Location Preferences*

- ***Données de contexte inférées*** : centres d'intérêts, préférences lieu
- ***Données utilisées pour inférer le contexte*** : documents visités
- ***Méthode***
  - ✓ Projeter les documents *clickés* sur des ontologies de contenu
  - ✓ Calculer le profil de localisation, le profil du sujet



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

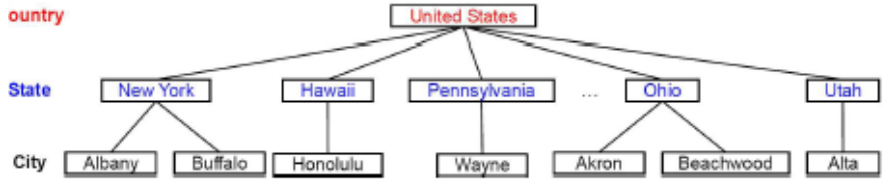


Fig. 3. Example Predefined Hierarchy for "United States".

Méthodes de l'état de l'art : Inférence des centres d'intérêt

- Profil : mesure formelle de calcul d'entropie

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

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

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

$\overline{C_u} = \left\{(\overline{c_{1u}}, \overline{c_{2u}}, \ldots, \overline{c_{tu}}) : |c_{iu}| \right.$ *nombre de documents contenant le concept de contenu* c_i *et clickés par l'utilisateur u*

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

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

$\overline{L_u} = \left\{(\overline{l_{1u}}, \overline{l_{2u}}, \ldots, \overline{l_{vu}}) : |l_{iu}| \right.$ *nombre de documents contenant le concept de localisation* l_i *et clickés par l'utilisateur u*

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

Cadre général : questions scientifiques critiques

① Quels sont les éléments du contexte qui impactent un processus de RI ?

② Comment capturer, modéliser le contexte ?

③ Comment utiliser le contexte pour mieux répondre à la requête ?

Taxonomie des usages du contexte en RI

Ingwersen & Jarvelin, 'The Turn', 2005

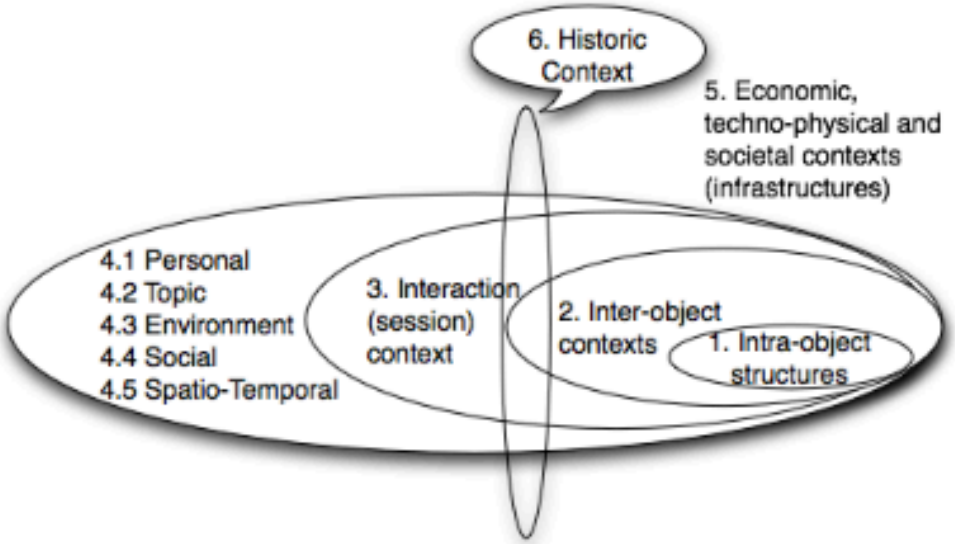


FIGURE 1: Taxonomy for Context Features - variant of (Ingwersen & Järvelin 2005)

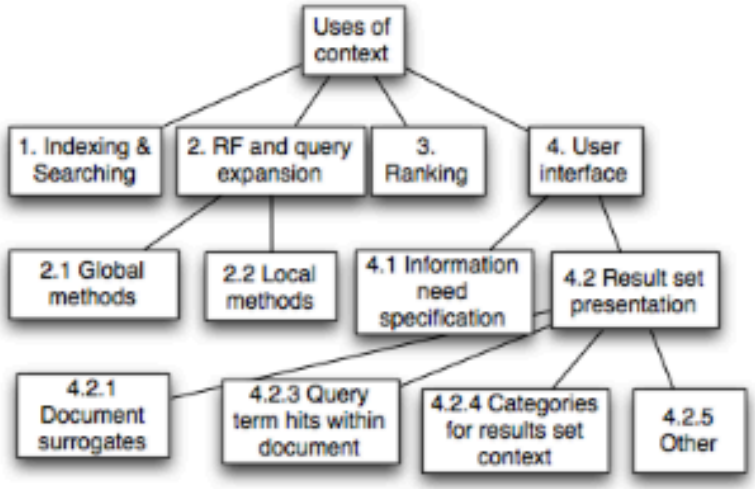


FIGURE 2: Taxonomy for Uses of Context

Principe général

activité

Feedback implicite

activité

activité

activité

Application

Application

Application

Application

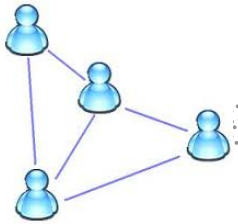
Application

Click

Navigation

Impression

Jugement



Algorithme de construction du contexte

Profil utilisateur

Feedback explicite

Requête *Q*

Traitement de la requête

Contexte

Méthodes de l'état de l'art

- Fondamentalement : techniques de **reformulation de requête** ou de **(ré)ordonnancement des documents** utilisant les **éléments du contexte** (lieu, préférences, …)
  - ✓ **Reformulation (du modèle de langue) de la requête** : Kong et al. SIGIR 2015, Grbovic et al. SIGIR 2015, Tan et al. KDD 2006, Shen et al. 2005, Shen et al. CIKM 2005
  - ✓ **(Ré)ordonnnancement** : Deveau et al. CIKM 2015, Li et al. SIGIR 2015, Zhang et al. SIGIR 2015, White et Awadallah SIGIR 2015, Badache and Boughanem ECIR 2015, Rakesh et al. ICWSM 2014, Chelaru et al. WISE 2012, Karweg et al. CIKM 2011, Bouidghaghène et al. MDM 2011, Qiu and Sho WWW 2006, Speretta and Gauch Web Intell. 2005, Liu and Weng TKDE 2004

Méthodes de l'état de l'art : Reformulation (du modèle de langue) de requête

Tan, Shen and Zhai KDD 2006 *Mining Long-Term Search History to Improve Search Accuracy*

- ***Données de contexte utilisées*** : Historique de recherche
- ***Résultats retournés*** : documents
- ***Méthode***
  - ✓ **Reformuler le modèle de langue de la requête en considérant l'historique de recherche de l'utilisateur**
  - ✓ Calculer le modèle de langue du document
  - ✓ Ordonner les documents par appariement des modèles de langue de la requête et du document

Méthodes de l'état de l'art : Reformulation (du modèle de langue) de requête

- **Estimer le modèle de langue de la requête q_k**

  Calculer pour chaque requête $q_k \in H_k$ un vecteur de termes pondérés qui tient compte de la distribution des termes dans les documents jugés pertinents *C** et documents jugés non pertinents *NC = D-C**

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

❶ *Modèle de distribution du terme t dans les résultats de la requête q_k (Contexte courant)*

❷ *Modèle de distribution du terme t dans l'historique H (Contexte passé)*

Méthodes de l'état de l'art : Reformulation (du modèle de langue) de requête

❶ $$p(t|\theta_k) = \lambda_q p(t|\theta_{q_i}) + (1-\lambda_q)\frac{\sigma_C \sum_{d_j \in C_k} p(t|\theta_{dj}) + \sigma_{NC} \sum_{d_j \in NC_k} p(t|\theta_{dj})}{\sigma_C |C_i| + \sigma_{NC} |C_{NC_i}|}$$

Poids sur le modèle de la requête originale

Modèle de langue classique de la requête

Modèle de langue de la requête à partir des modèles de documents résultats de la requête q

$$p(t|\theta_q) = \frac{TF(t,q)}{|q|}$$

$$p(t|\theta_d) = \frac{TF(t,d) + \mu p(t|\theta_C)}{|d| + \mu}$$

❷ $$p(t|\theta_{H_k}) = \frac{\sum_{q_i \in H_k} \lambda_i p(t|\theta_i)}{\sum_{q_i \in H_k} \lambda_i}$$ → Modèle de pondération simple

$$\lambda_i = 1, \forall q_i \in H_k$$

Méthodes de l'état de l'art : Reformulation de requête

Grbovic, Djuric, Radosavljevic et al. SIGIR 2015, *Context- and Content-Aware Embeddings for Query Rewriting in Sponsored Search*

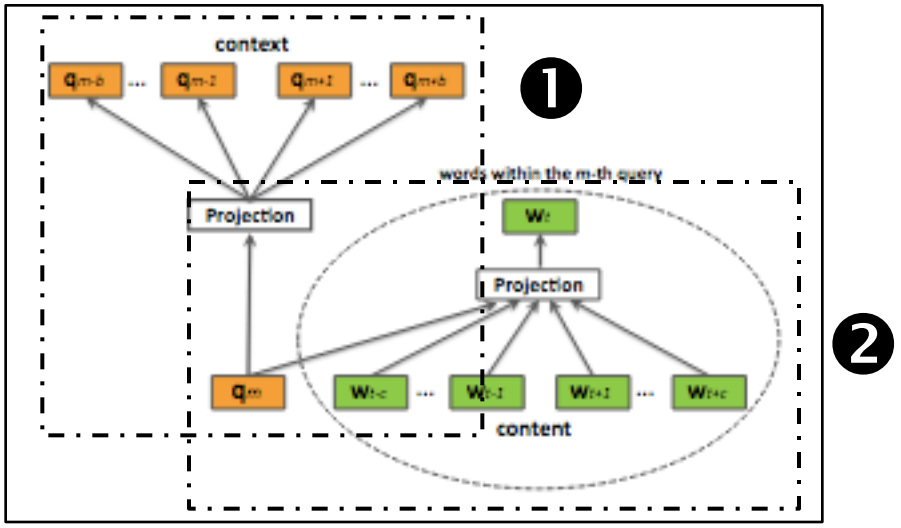
- ***Données de contexte utilisées*** : Historique de recherche
- ***Résultats retournés*** : documents/ressources
- ***Méthode***
  - ✓ **Apprendre la représentation latente et conjointe de la requête et son contexte. Equivalent à une réécriture de requête dans un espace latent**
  - ✓ Apparier les documents avec la requête réécrite dans l'espace latent. Equivalent à une recherche avec les k. plus proches voisins

Méthodes de l'état de l'art : Reformulation de requête

Apprentissage à partir des données
de **contenu** et données
de **contexte** issues de sessions utilisateurs

Maximiser la fonction objectif :

S : sessions de recherche
q_i : ième requête



$$\zeta = \sum_{s \in S} \sum_{q_m \in S} \Big(\sum_{-b \le i \le b, i \neq 0} \log p(q_{m+i} | q_m) + \alpha_m \log p(q_m | w_{m1} : w_{mt_m}) + \sum_{w_{mt} \in q_m} \log p(w_{mt} | w_{m,t-c} : w_{mt+c}, q_m) \Big)$$

❶ **Contexte** : **context2vec utilise le skipgram**
Séquence de requêtes
dans la même session

❷ **Contenu : content2vec utilise le paragraph2vec**
Séquence de mots
dans la même requête

Méthodes de l'état de l'art : (Ré) Ordonnacement des résultats

Karweg, Hutter and BoHm CIKM 2011, *Evolving Social Search Based On Bookmarks and Status Messages from Social Networks*

- ***Données de contexte utilisées*** : Signaux sociaux
- ***Résultats retournés*** : documents/ressources
- ***Méthode***
  - ✓ Calculer le score social d'une ressource sur la base de différents signaux sociaux issus du voisinage : confiance, engagement
  - ✓ Calculer le score de pertinence d'une ressource comme la combinaison linéaire du score social et du score textuel classique

Méthodes de l'état de l'art : (Ré) Ordonnacement des résultats

- Calcul du score social d'une information *i* pour un utilisateur *x*

Score de pertinence sociale

Degré de confiance (trust) de u en x

$$SRS_S(i) = \sum_{x \in E_t} t_s(x) \times e_x(i)$$

Ensemble des utilisateurs avec qui l'utilisateur x a interagi

Intensité de l' engagement

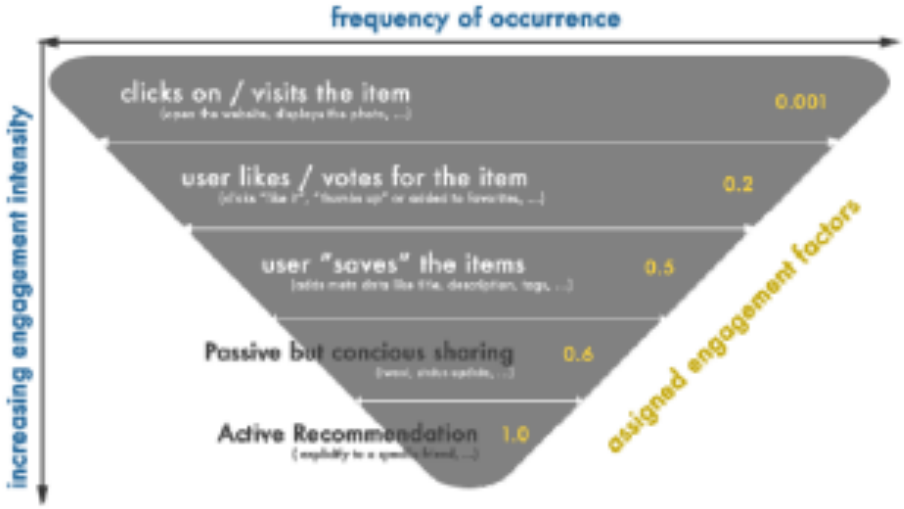


Figure 4. Different Levels of User Engagement

- Calcul du score social d' une information *i* pour un utilisateur *x*

Score total de pertinence

Score de pertinence sociale

Score de pertinence thématique

$$CSR(i) = \alpha * SRS(i) + (1 - \alpha) * FTR(i)$$

Méthodes de l'état de l'art : (Ré) Ordonnacement des résultats

Deveaud, Albakour, MacDonald et al CIKM 2015, *Experiments with a Venue-Centric Model for Personalized and Time-Aware Venue suggestion*

- ***Données de contexte utilisées*** : centres d'intérêts, localisation de l'utilisateur, temps d'émission de la requête
- ***Résultats retournés*** : documents/ressources lis à des localisations vues comme des points d'intérêt à recommander à l'utilisateur
- ***Méthode***
  - ✓ Calculer différents indicateurs : a) popularité de la localisation, b) degré d'intérêt de la localisation
  - ✓ Calculer le score de pertinence d'une ressource comme le produit de ces indicateurs

Méthodes de l'état de l'art : (Ré) Ordonnacement des résultats

- **Estimer la probabilité qu'un point d'intérêt v soit pertinent à un instant *t* pour un utilisateur *u* localisé au lieu *l***

$$p(v|u,l,t) \alpha\ p(v|u) \times p(v|l) \times p(v|t)$$

Pertinence du point d'intérêt v pour l'utilisateur u

Proximité du point d'intérêt v du lieu de l'utilisateur l

Probabilité d'affluence au point d'intérêt v à cet instant t

Popularité du point d'intérêt v

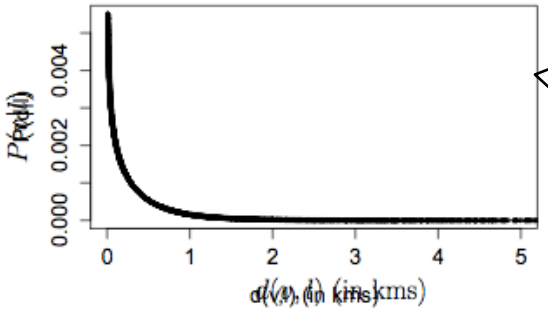
Méthodes de l'état de l'art : (Ré) Ordonnacement des résultats

- **Estimer la pertinence du point d'intérêt *v pour l'utilisateur u***

$$p(v|u) = \sum_{e} \sum_{c \in C} p(v|c)\, p(c|e)\, p(e|u)$$

e sont les localisations entrainées du graphe Facebook de l'utilisateur u et appariées avec une ontologie DMOZ

- ***Estimer la proximité du point d'intérêt v de la localisation de l'utilisateur l***



Estimation de la distribution à partir des lieux consultés par l'utilisateur sur Foursquare

$$p(v|l) = p(d(v,l))$$

Distribution de probabilité fonction de la distance entre le point d'intérêt v et localisation l de l'utilisateur

- ***Estimer la probabilité d'affluence au point d'intérêt v à l'instant t***

$$p(v|t) = \frac{y(v,t) + \mu \times p(v|\mathrm{T})}{\sum_{v' \in V} y(v',t) + \mu}$$

Modèle de série chronoogiques type ARIMA estimé à partir des données de Foursquare sur la localisation v. Lissage Bayesien appliqué pour éviter les probabilités nulles

Méthodes de l'état de l'art : (Ré) Ordonnacement des résultats

Bouidghaghène, Tamine, Pasi et al. AIRS 2011, *Prioritized Aggregation of Multiple Context Dimensions in Mobile IR*

- ***Données de contexte utilisées*** : centres d'intérêts, localisation
- ***Résultats retournés*** : documents
- ***Méthode***
  - ✓ Calculer le score partiel d'une ressource sur la base du sujet et de chaque facteur du contexte (centres d'intérêt, localisation)
  - ✓ Calculer le score de pertinence d'une ressource comme l'agrégation des scores partiels par application d'un opérateur d'agrégation prioritaire

Méthodes de l'état de l'art : (Ré) Ordonnacement des résultats

- Calcul des scores partiels

Score thématique

$$Thème(d,Q) = \sum_{i=1}^{n} IDF(t_i) * \frac{f(t_i,d) * (k_1+1)}{f(t_i,d) + k_1 * \left(1 - b + b * \frac{|d|}{avgdl}\right)}$$

Score « Centres d'intérêt »

$$Intérêts(d,I) = \sum_{c_j \in I \wedge j \in [1,k]} sw(c_j) * \cos(\vec{d}, \vec{c}_j)$$

Score « Localisation »

$$Localisation(d,L) = f(L) + \sum_{L_i \in descendants(L)} f(L_i)$$

- Application d'un opérateur Scoring (F_s)

« plus est le score de satisfaction des critères de plus hautes priorités, moins le score de satisfaction d'un critère de moindre priorité influence le score global d'un document »

$$F_s : [0,1]^n \rightarrow [0,n]$$

$$F_s(C_1(d),\ldots,C_n(d)) = \sum_{i=1}^{n} \lambda_i \cdot C_i(d)$$

$$\lambda_1 = 1$$
$$\lambda_i = \lambda_{i-1} \times C_{i-1}(d), C : critère$$

En résumé

- **Le contexte est un concept multidimensionnel**
  - ✓ Utilisateur (intérêts, localisation…), tâche, temps
  - ✓ Plusieurs sources d'évidence pour inférer le contexte : historique de recherche, signaux sociaux, trajectoires de mobilité, …
- **Différents outils de modélisation**
  - ✓ Classification, apprentissage de représentations, modèles de langue, etc.
- **Contextualisation d'un processus de RI**
  - ✓ Techniques de reformulation de requêtes : réécriture de requête, réécriture du modèle de langue de la requête, apprentissage de représentation de requêtes
  - ✓ Techniques de (ré)ordonnancement des documents : nouvelle variable liée au Contexte en plus du Document et reQuête)

Principaux défis et objets de recherche en lien avec la Recherche d'Information Contextuelle

- **Recherche d'information contextuelle et vie privée :**

  Quel Compromis ?

  - ✓ Personnalisation par pseudonyme, niveau client…
  - ✓ Encryptage des données personnelles

- **Recherche d'information contextuelle et évaluation :**

  Quel Cadre ?

  - ✓ Révision des modèles d'évaluation de tye « Cranfield » : étude de traces, études utilisateurs, nouvelles tâches orientées système (TREC Task, TREC Session, TREC Contextual Suggestion)
  - ✓ Répétabilité et reproductibilité des expérimentations utilisateurs

Recherche d'Information Contextuelle

Centres d' Intérêt, Localisation, Facteurs (Média-)Sociaux

Questions ?