

Nouveau Modèle de recommandation pour la Classification à facettes

Manel HMIMIDA

Manel.hmimida@cnam.fr

Thèse en cours

Encadrement: Manuel Zacklad & Rushed Kanawati

Journée de recherche sur les moteurs de recommandations

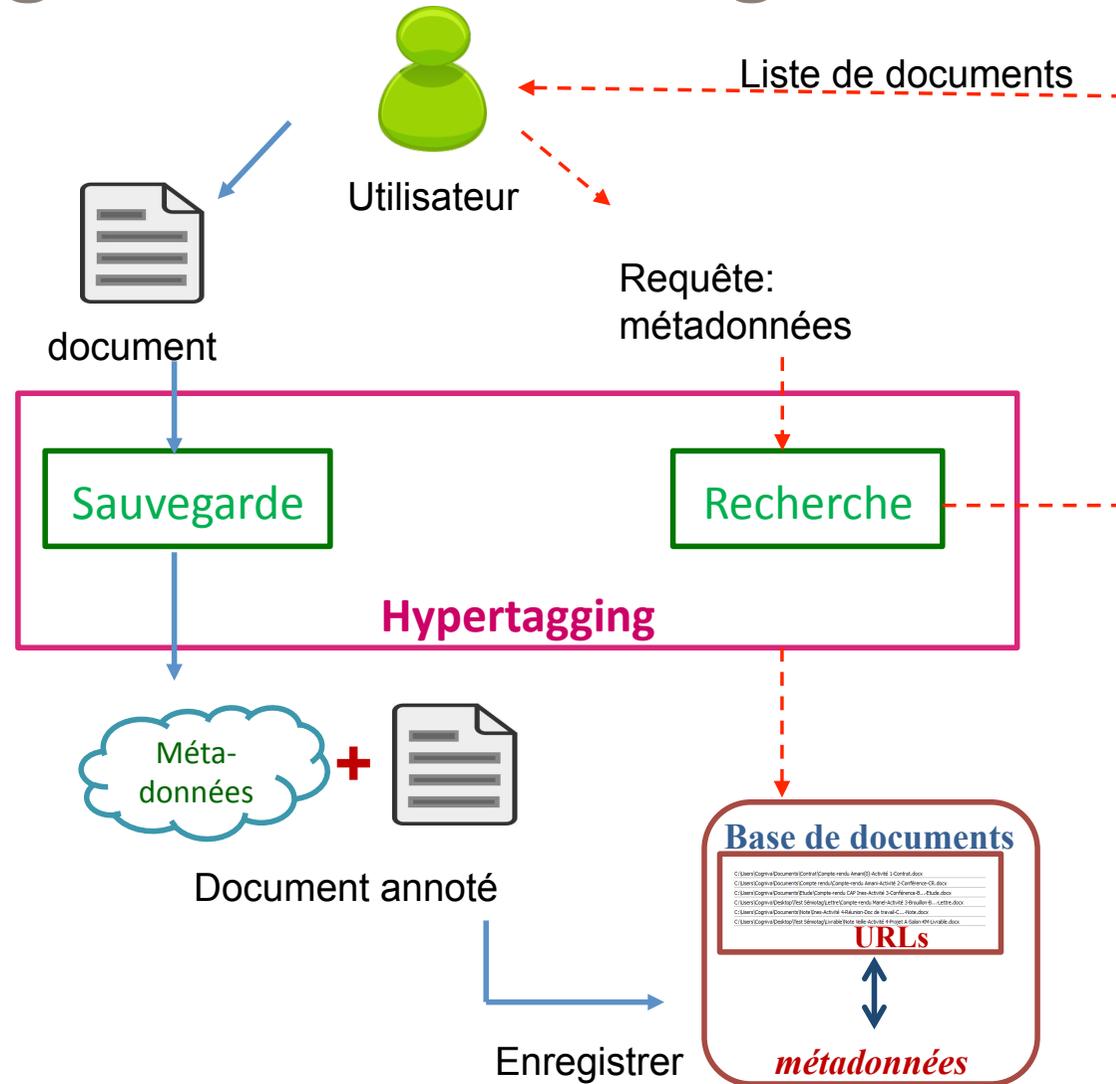
Contexte

- **Projet ANR**: Miipa-Doc (Méthodes et Services Intégrés Institutionnels et Participatifs pour la Classification à Facettes des Contenus Documentaires Complexes)
- **Partenaires** : UTT, EDF, Cogniva, Cnam.
- Faciliter **l'indexation** et l'accès à l'information dans **les organisations**
- Développer un nouveau SOC (Système d'Organisation de connaissances) basé sur **classification à facettes: Hypertagging**

Le Système Hypertagging

- SI basé sur un SOC multidimensionnel et distribué, à base des facettes (Zacklad et. al., 2011)
- Situé dans le domaine de l'organisation des connaissances avec une perspective de classification documentaire participative
- **Principales fonctions**
 - Stockage de documents (nommage, classification)
 - Traitement (recherche)
 - Communication (partage)

Hypertagging: Architecture Logicielle



Les Métadonnées Hiérarchiques

Vues

Les activités métier et les centres d'intérêts utilisateur.

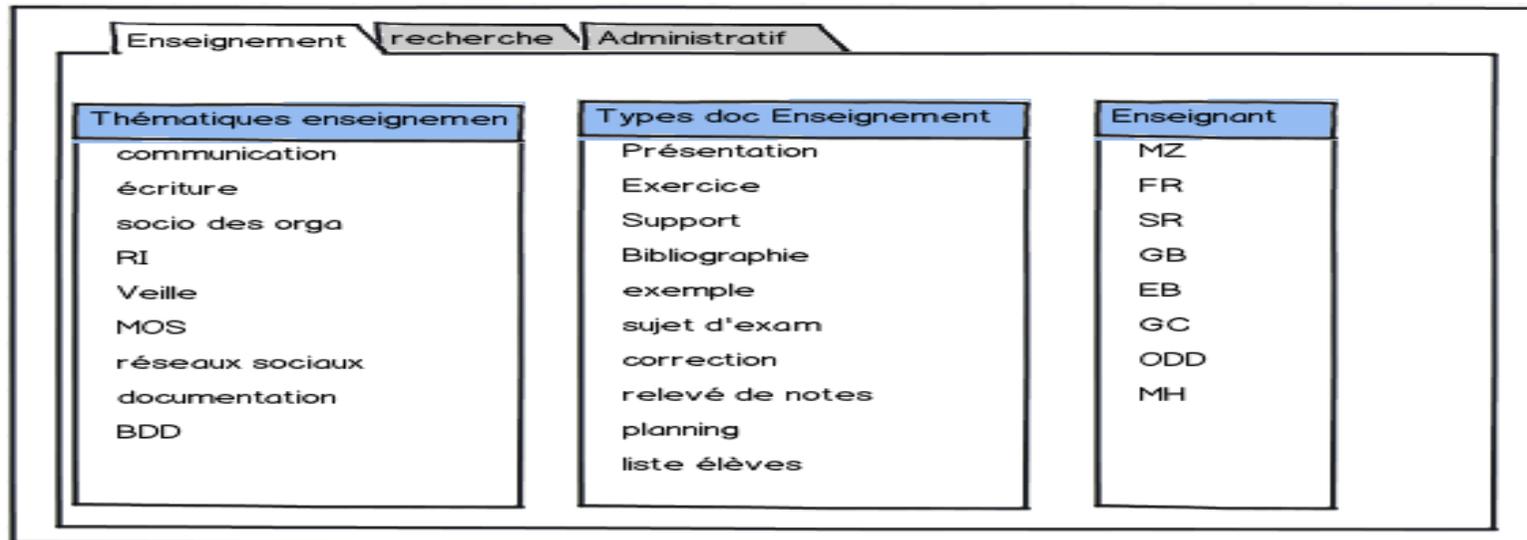
Facettes

Classe ou groupe de tags de nature similaire.

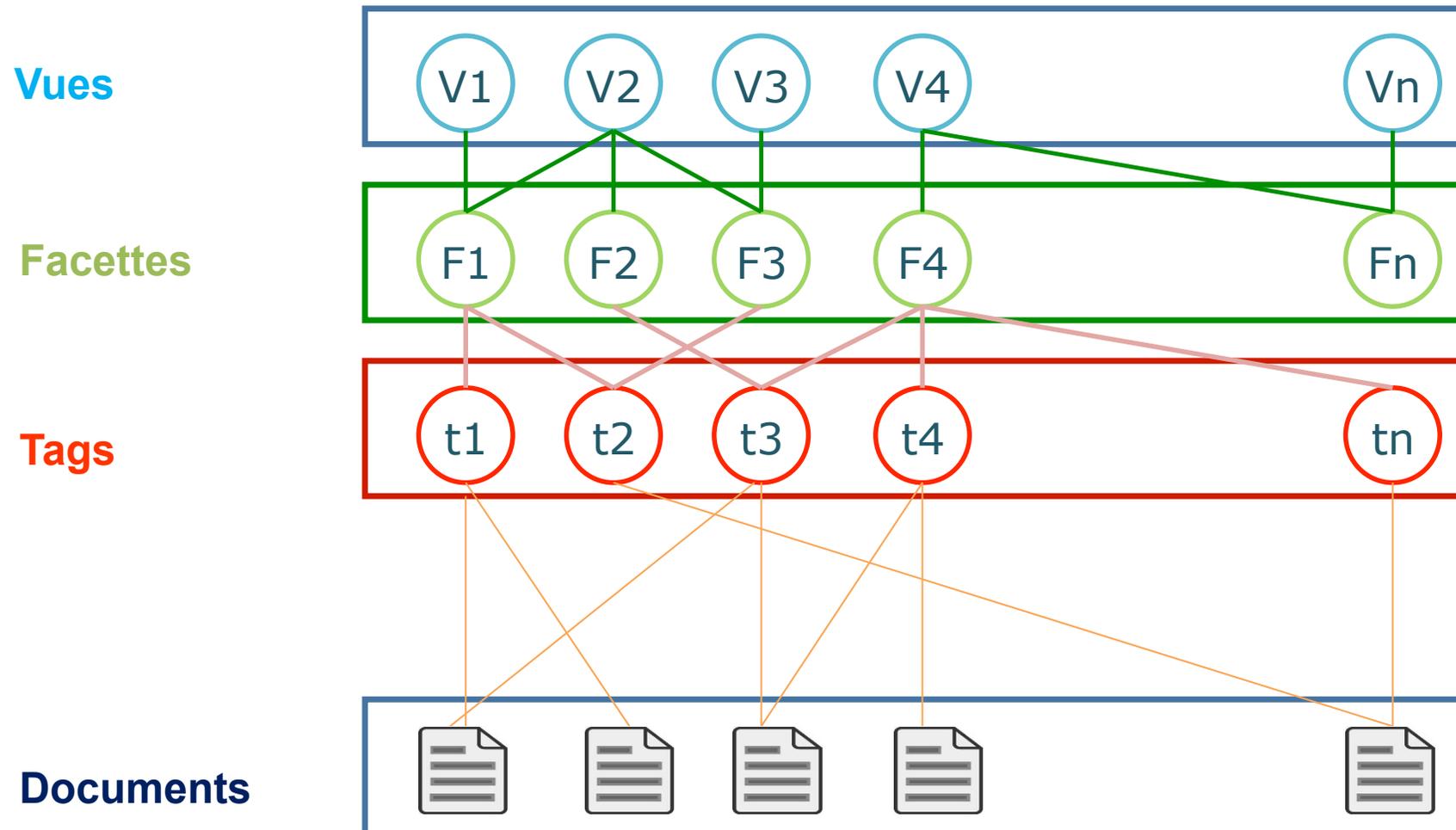
Tags

Ce sont les valeurs de facettes.

Exemple



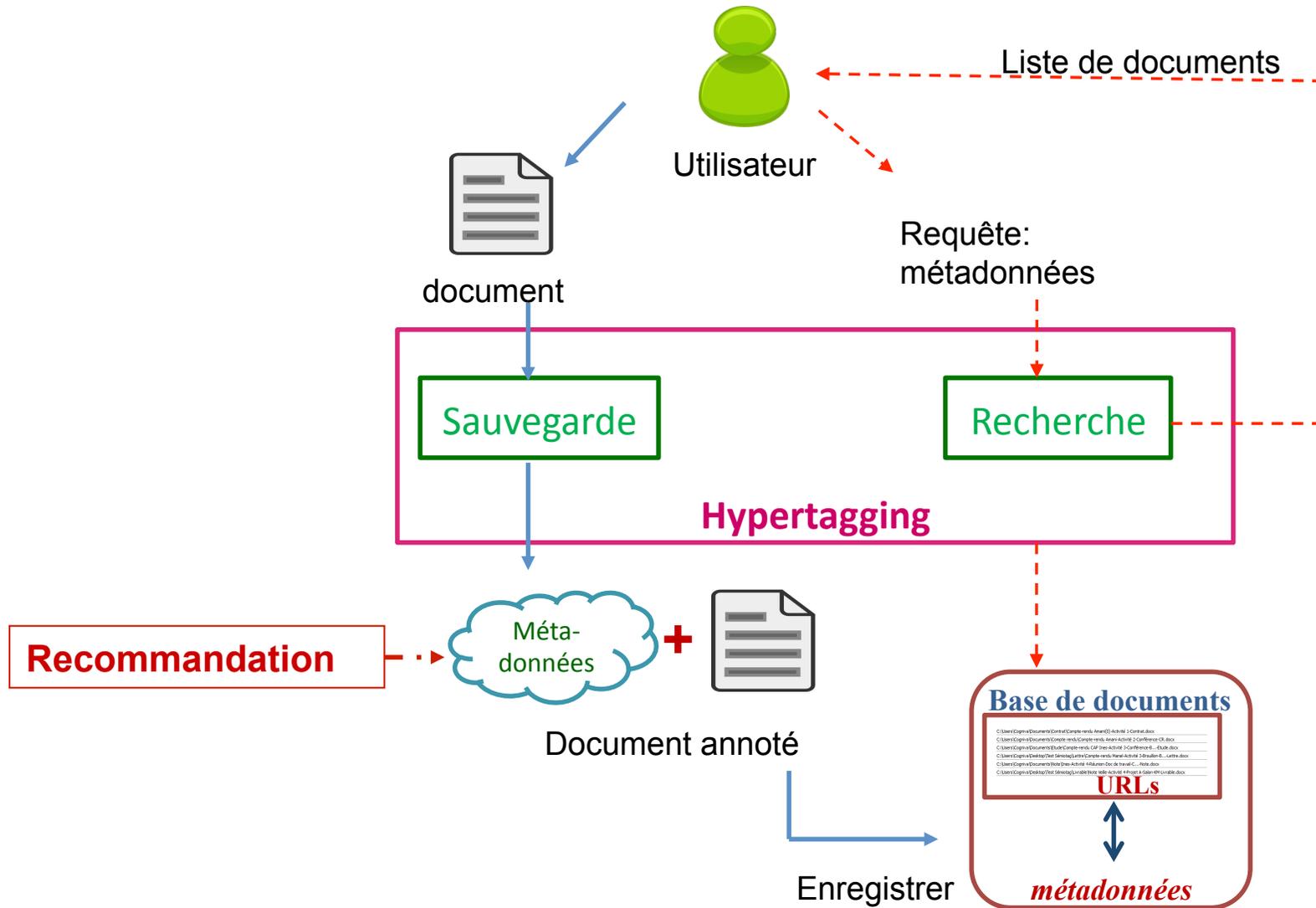
Les Métadonnées Hiérarchiques



Problématique

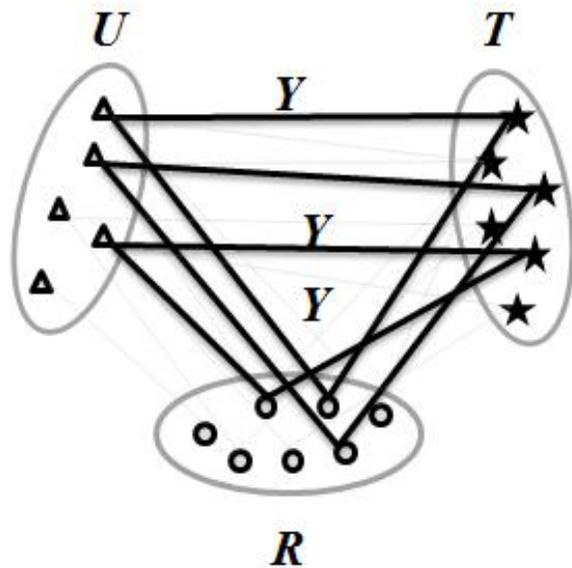
- L'aspect **évolutif** des intérêts utilisateur et des activités métiers.
- **La croissance exponentielle** de la quantité de documents numériques au sein des entreprises.
- **L'évolution du nombre de tags et facettes** engendrent la complexité de la tâche d'indexation des documents.
- **Des conséquences lourdes** sur la performance **individuelle** et **collective** des collaborateurs.

Objectif



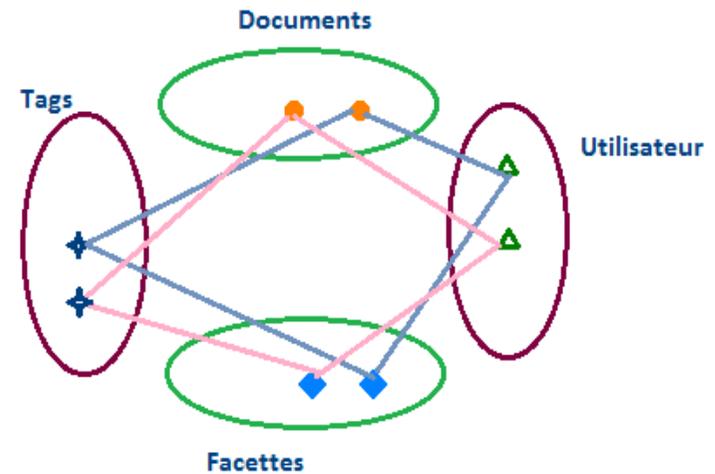
Etat de l'art

Folksonomie



Folksonomie:= (U, T, R)

SOC à facettes



SOC:= (U, T, R, F)

Recommandation de Tags

1. Basée sur le contenu
2. Basée sur le filtrage Collaboratif
3. Approche Topologique

Approche Basée sur le contenu

(Mrosek et. al., 2008), (Lu et. al., 2009)

Principe

- Analyser le contenu de ressources pour proposer des termes/mots les plus fréquents
- Il y a des approches qui filtrent les termes fréquents par les termes déjà utilisés

Avantages

- Pouvoir préconiser de nouveaux tags qui n'ont été jamais utilisés.

Limites

- Ne prends pas en compte le choix de l'utilisateur

Filtrage Collaboratif

(Jäschke et. al., 2007)

Principe

- Recommander des tags utilisés par des utilisateurs similaires

Avantage

- Choisir les tags les plus pertinents

Limite

- Démarrage à froid

Approche Topologique

Principe

- Calculer les tags à partir de l'analyse de graphe tripartite représentant les folksonomies.

- Approche statique (Jäschke et al., 2008)

- Agrégation de l'historique de l'évolution des folksonomies dans un seul graphe.

L'algorithme FolkRank

1. Projeter le graphe de la Folksonomie

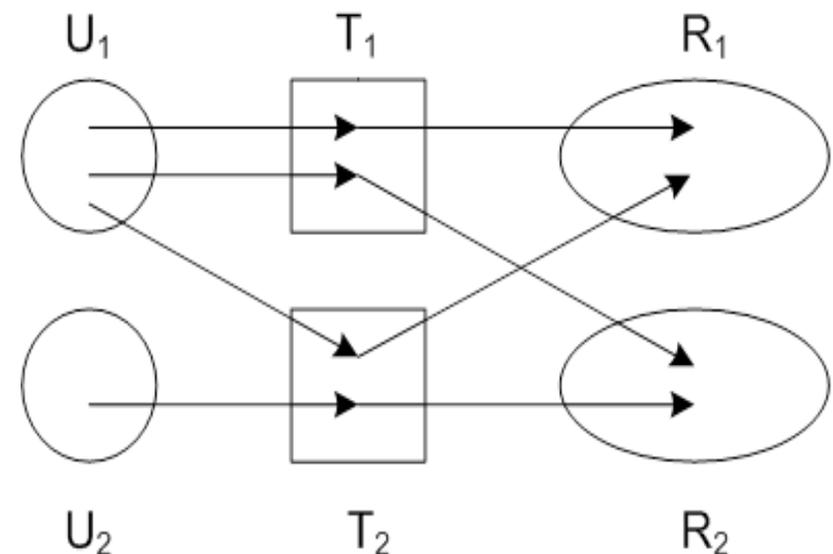
(U, T), (T, R)

1. Appliquer l'algorithme PageRank

2. Une ressource annotée avec des tags importants par des utilisateurs

importants devient elle

même importante



Approche Topologique

- Approche dynamique (Pujari et. Kanawati., 2011)
- Prendre en compte l'aspect dynamique des folksonomies
- Prédiction de liens dans les graphes bi-partites.

L'approche Liptar (Link Prediction based Tag Recommender)

Données d'entrée : utilisateur u , ressource r

1. Déterminer k utilisateurs similaires à u (k -nearest neighbor) dans un graphe bipartite)
2. Associer chaque utilisateur u à une séquence de graphe bi-partite temporel reliant les ressources ajoutées par u et annotées par les tags t
3. Application de l'approche de prédiction de liens dans les graphes bi-partites

Résultat : Liste de tags pour annoter la ressource r

Contribution

- Exploiter **la structure hiérarchique** d'Hypertagging pour générer les recommandations
- Recommandation **par niveau**:
 - Niveau Facette
 - Niveau Tags
- Utilisation de **règles d'association** dans les SOC à facettes.

Utilisation de règles d'association

1. Les Règles d'Association (Agrawal et. al., 1994)

-Approche automatique pour découvrir des relations / corrélations intéressantes entre des objets

-Règle de la forme:

$$X \Rightarrow Y$$

X et Y peuvent être composés de conjonctions

2. Mesures de qualité

$$\text{Support } (X \Rightarrow Y) = P(X \text{ et } Y)$$

$$\text{Confiance } (X \Rightarrow Y) = P(Y | X) = P(X \text{ et } Y)/P(X)$$

Utilisation de règles d'association

Données d'entrées: documents+ Facettes+ Tags

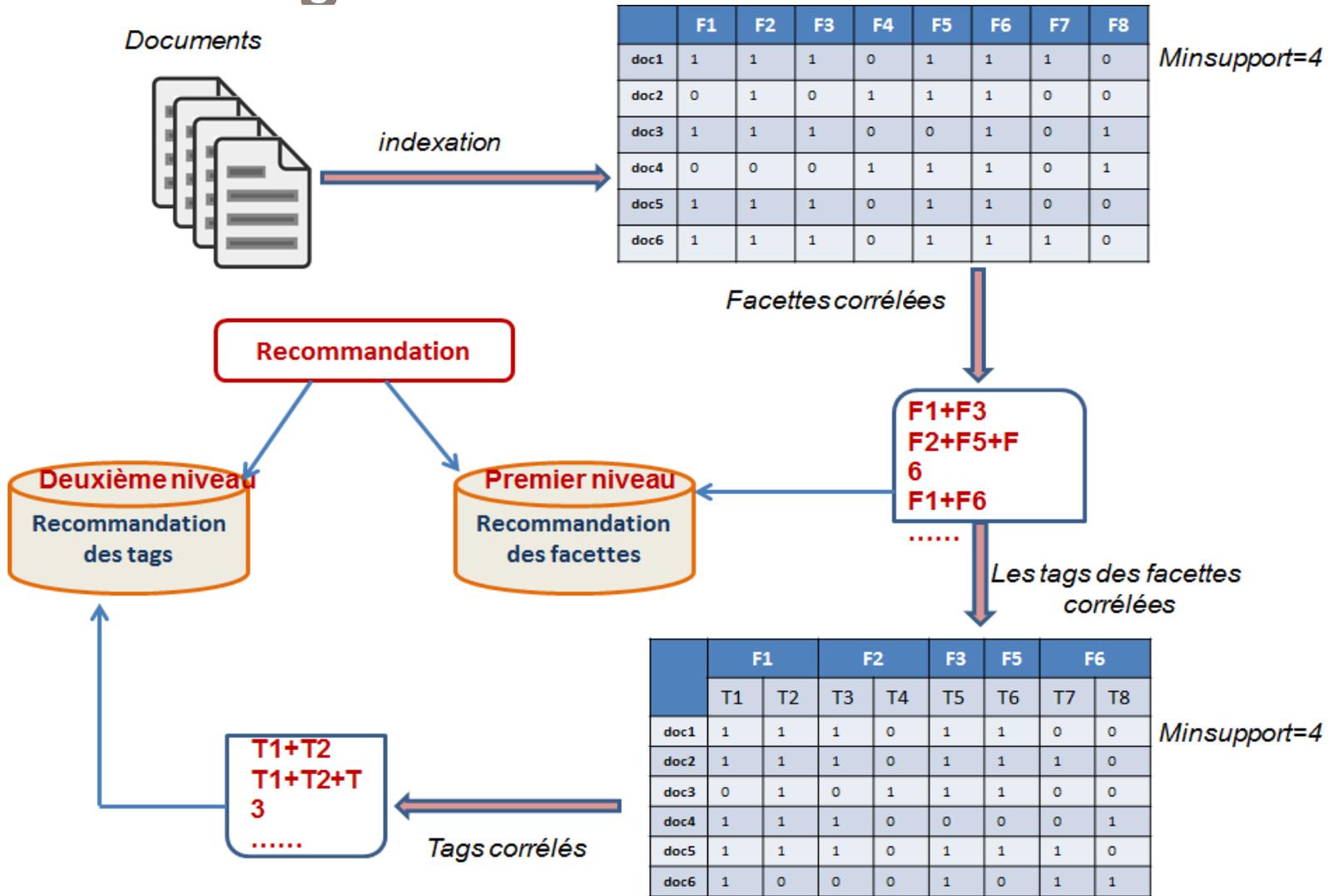
Déterminer les règles d'association entre Facettes

$F_i \Rightarrow F_j$

Déterminer les règles d'association entre Tags

$T_n \Rightarrow T_m$

Méthodologie



Algorithme de recommandation de Facettes/tags

Entrée

FTi : Ensemble de facettes et tags sélectionnés par l'utilisateur

RA : Ensemble de règles d'association telle que chaque règle est représentée par un couple (A, C) où A est l'ensemble des termes antécédent et C l'ensemble de termes conséquent.

Sortie

$F'T'i$: Ensemble de facette et tags de recommandation

Début

$$F'T'i = \emptyset$$

Pour chaque $(A_j, C_j) \in RA$ // Parcours de la base des règles

Si $FTi \subseteq A_j$ alors // Vérification s'il existe une règle associée au choix de l'utilisateur

$$F'T'i = F'T'i \cup C_j$$

Fin Si

Fin Pour

Fin

Protocole d'évaluation

Collection CiteUlike (2005-2007)

○ *Prétraitement:*

- Choisir les tags et les ressources les mieux connectés
- Suppression de tags systèmes
- Suppression de tags non fréquents

○ Taille de la base: 307 users, 1762 documents et 751 tags

- **Basé Utilisateurs** : regrouper les tags dans des classes en fonction de comportements de l'ensemble des utilisateurs.
- **Basé Ressources** : regrouper les tags dans des classes en fonction de leurs utilisation dans l'indexation de ressources.

○ Création de facettes: Appliquer un algorithme de détection de communauté. (les communauté représente les facettes)

○ Appliquer l'algorithme des règles d'association sur les facettes puis les tags.

Conclusion

Nous développons une nouvelle approche de recommandation dites par niveau dans notre SOC à facettes pour le classement documentaire. Notre modèle permet d'apprendre au cours du temps les habitudes et les préférences utilisateurs en exploitant l'historique d'indexation.

Limites

- Démarrage à froid

Perspectives

- Utiliser la technique de Co-clustering pour générer les facettes
- Travailler avec un graphe multipartite en ajoutant les facettes
- Exploiter la dimension Vue

Bibliographie

Agrawal R., Srikant R., «Fast algorithms for mining association rules in large databases» International Conference on Very Large Data Bases, pages 487-499, Santiago, Chile, September 12-15 1994.

Lu, Y.-T., Yu, S.-I., Chang, T.-C., Jen Hsu, J.Y.: A content-based method to enhance tag recommendation. In: Boutilier, C. (ed.) IJCAI, pp. 2064–2069, 2009.

Mrosek, J., Bussmann, S., Albers, H., Posdziech, K., Hengefeld, B., Opperman, N., Robert, S., Spira, G.: *Content-and graph-based tag recommendation: Two variations. In: ECML PKDD Discovery Challenge 2009, CEUR Workshop Proceedings, vol. 497, pp. 189–199 2009.*

Zacklad M., Desfriches-Doria O., Bertin G., Mahe S., Ricard, B., Musnik N., Cahier, J.P, Benel, A., Lewkowicz, E. (2011). *Miipa-Doc : vers une gestion de l'hétérogénéité des classifications documentaires en entreprise*, soumis à la conférence Hypertextes et hypermédias. Produits, Outils et Méthodes (H2PTM 2011), Metz, 2011;

Jäschke, R., Marinho, L.B., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag recommendations in folksonomies. In: Hinneburg, A. (ed.) LWA, pp. 13–20. Martin- Luther-University Halle-Wittenberg (2007)

Jäschke, R., Marinho, L.B., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag recommendations in social bookmarking systems. *AI Commun.* 21(4), 231–247, 2008.

Manisha Pujari, Rushed Kanawati. Supervised machine learning link prediction approach for tag recommendation. *4th International Conference on [Online Communities and Social Computing @ HCI International 2011](#), 9-14 July 2011, Hilton Orlando Bonnet Creek, Orlando, Florida, USA, LNCS Springer.*



**Merci pour votre
Attention**

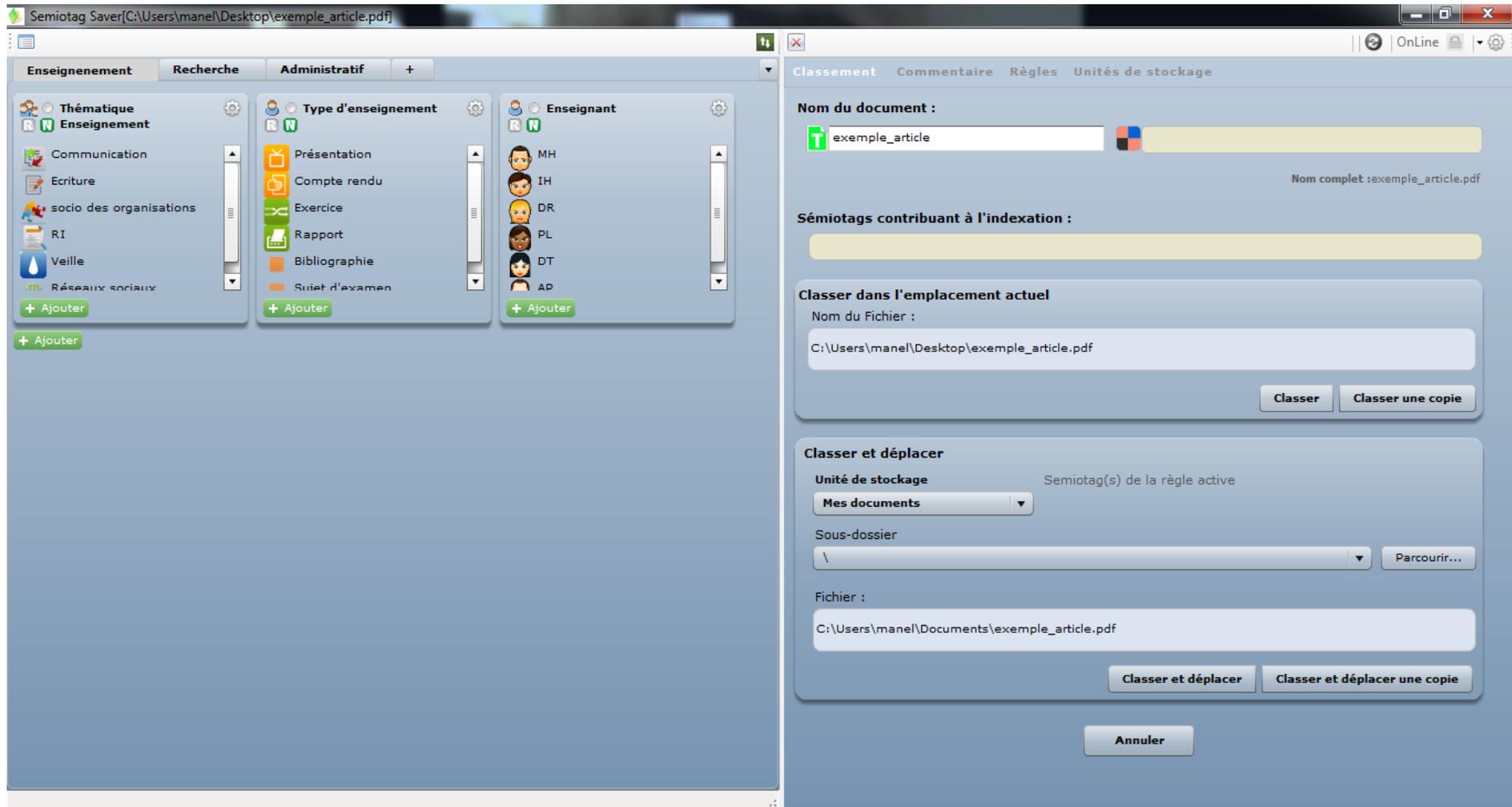


Questions?

Manel.hmimida@cnam.fr

le **cnam**

Interface Graphique d'Hypertagging



Modélisation d'un système de classification à facettes

Définition. Un système (SCF) est composé d'un ensemble d'utilisateurs $u \in U$, tags $t \in T$, facettes $f \in F$, et des documents $d \in D$.

(f_c, t_i, u_k, d_j) : association d'un ensemble de tags et facettes à des documents.

- R_a , un ensemble de quadruplets (f, t, u, d) pour dire que l'utilisateur u annote le document d avec la facette f et le tag t .
- TF_u est l'association de tag T à la facette correspondante F par l'utilisateur $u \in U$.
- TR_u est l'association de tag T à la ressource R par l'utilisateur $u \in U$.
- RF_u est l'association de la ressource R à la facette F par l'utilisateur $u \in U$.
- RT_u est l'association de la ressource R au tag T par l'utilisateur $u \in U$.