Approches topologiques pour la prévision de liens dans des grands graphes de terrain: application au calcul de recommandation

N. Benchettara, **R. Kanawati**, M. Pujari, C. Rouveirol, G. Santini

prénom.nom@lipn.univ-paris13.fr

Université Paris 13, Sorbonne Paris Cité, LIPN, CNRS UMR 7030, F-93430, Villetaneuse, France.

11 juin 2012

Journée moteurs de recommandation - CNAM Paris

Plan

1. Contexte & problématique
2. Approches de prévision de liens
3. Recommendation de collaborations
4. Recommendation de tags
5. Conclusion

Motivations

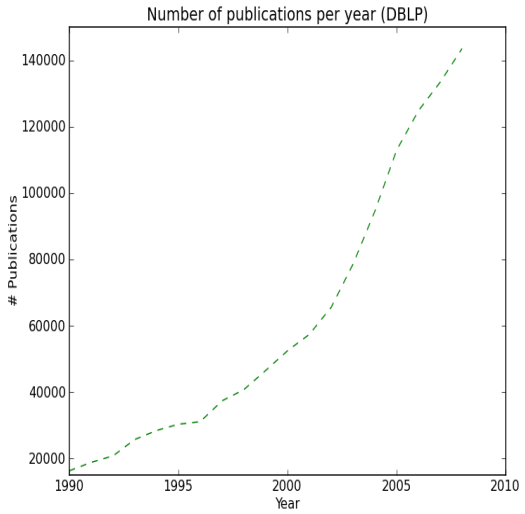
Domaine d'application cible :

Recommandation pour la recherche bibliographique

Contexte : Mutation de la publication académique

- Croissance accélérée de nombre de publications, de colloques, de revues, . . . , etc.
- Généralisation du modèle de l'accès libre.
- Mise en lignes de grandes Bibliothèques numériques (Springer, ACM, DBLP, etc.)
- Sites de partage social : *Les bibsonomies* (CiteULike, Bibsonomy, etc.).
- Outils collaboratifs de gestion de bibliographie : Mendeley, Zotero, Wizfolio, Connotea, etc.

Publications / année (DBLP)



Problème

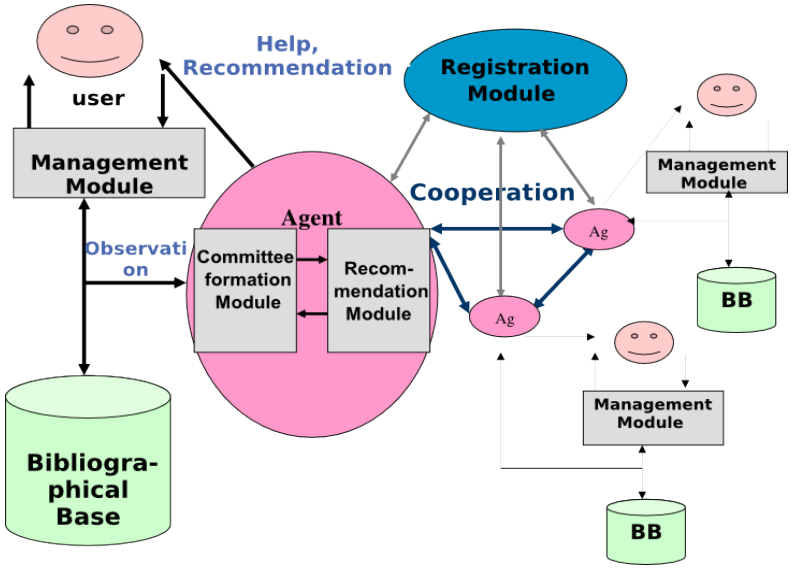
Perdu dans le WEB Bibliographique ! → **Systèmes de recommandation**

Travaux antérieurs : Le Système COBRAS

Système P2P pour la recommandation de références bibliographiques

- Utilisation **individuelle** d'une taxonomie **commune**.
- Un système de raisonnement à partir de cas (RàPC) pour l'**Alignement des taxonomies**
- Un autre moteur RàPC pour apprendre à construire des **communauté d'intérêts**.
- (Kanawati & Karoui, 2009) *A P2P collaborative bibliography recommender system*. **Fourth International Conference on Internet and Web Applications and Services, ICIW 2009**, 24-28 May 2009, Venice, Italy. IEEE Computer Society 2009. pp. 90-96

Architecture

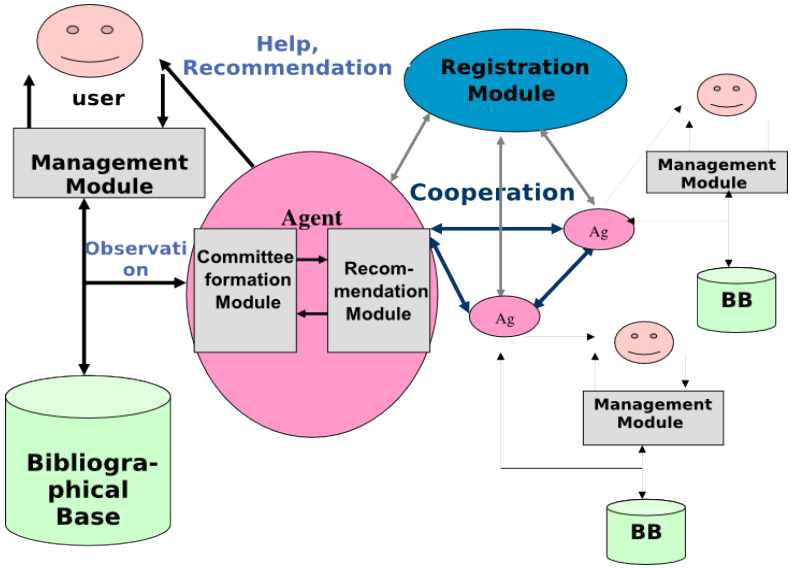


Travaux antérieurs : Le Système COBRAS

Système P2P pour la recommandation de références bibliographiques

- Utilisation **individuelle** d'une taxonomie **commune**.
- Un système de raisonnement à partir de cas (RàPC) pour l'**Alignement des taxonomies**
- Un autre moteur RàPC pour apprendre à construire des **communauté d'intérêts**.
- (Kanawati & Karoui, 2009) *A P2P collaborative bibliography recommender system*. **Fourth International Conference on Internet and Web Applications and Services, ICIW 2009**, 24-28 May 2009, Venice, Italy. IEEE Computer Society 2009. pp. 90-96

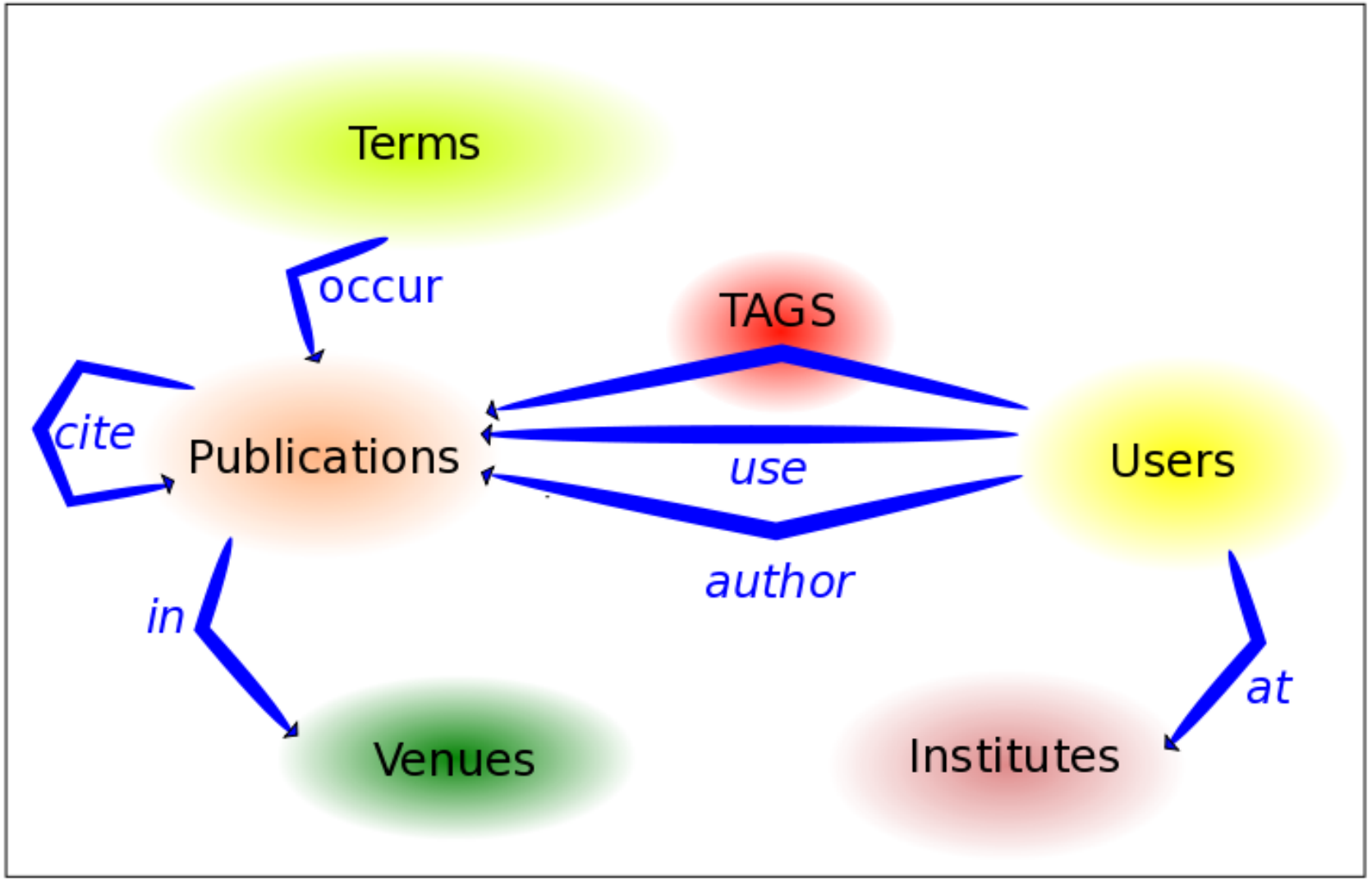
Architecture



Enseignements

- Pertinence de la détection de *communautés*.
- Similarité de références bibliographiques : approches basées sur l'usage.
- Fouille de réseaux bibliographiques

Réseaux Bibliographiques



By Date

Tâches de recommendation

Exemples de tâches :

- Recommendation d'utilisateurs
  - Recommendation de collaborations académiques [Benchattara et.al. 2010]
  - Recommandation de contacts, de personnes à suivre
- Recommendation de références.
- Recommendation de colloques
- Recommendation de tags
- Recommendation de citations *manquantes*
- ...

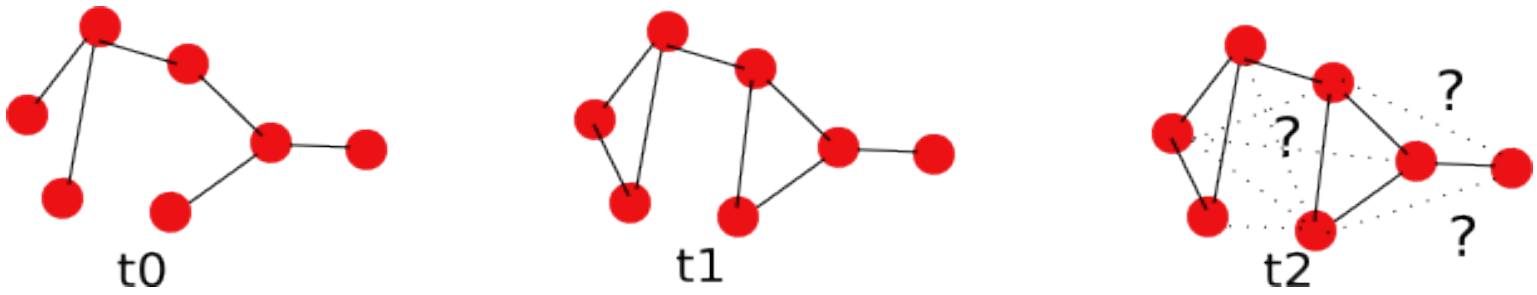
Approche :

Techniques de prévision de liens dans des grands graphes.

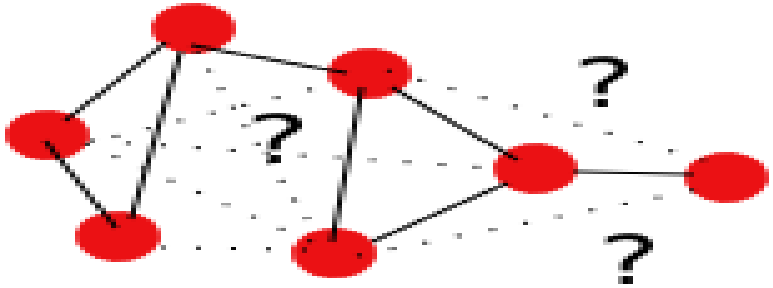
Tâches de prévision de liens

Trois différentes tâches :

- Prévision temporelle : Prévoir l'apparition de liens à un temps $t+1$ à partir de l'observation du graphe jusque l'instant t.



- Prévision structurelle : Calcul de liens cachés, complétion de lien



- Prévision d'attributs de lies : Calcul de labels inconnus à partir de label existants (e.g. réseaux multi-relationnels).

Classification des approches de prévision de liens

- **Approche : Dyadiques / Structurelles.**
  - Dyadique : Evaluer le *score* d'un lien entre deux nœuds v_i, v_j
  - Structurelle : prévoir l'évolution de la matrice d'adjacence du graphe ou prévoir l'évolution de sous-graphes (prédiction de plusieurs liens en même temps)
- **Type d'attributs : topologiques / caractéristiques des nœuds.**
  - Approche topologique : utiliser seulement le graphe du réseau.
  - L'emploi des approches fondées sur l'analyse du contenu des nœuds nécessite une expertise dans le domaine de l'application.

Recommendation de collaborations

Approche dyadique topologique non-supervisé

- Scores topologiques simples
- Combinaison de scores topologiques : Vote non supervisé

Approche dyadique topologique supervisé

- Apprentissage supervisé (semi-supervisé) : Prévision de liens = classification binaire.
- Vote supervisé

Approche structurelle topologique supervisé

- Recherche de motifs fréquents

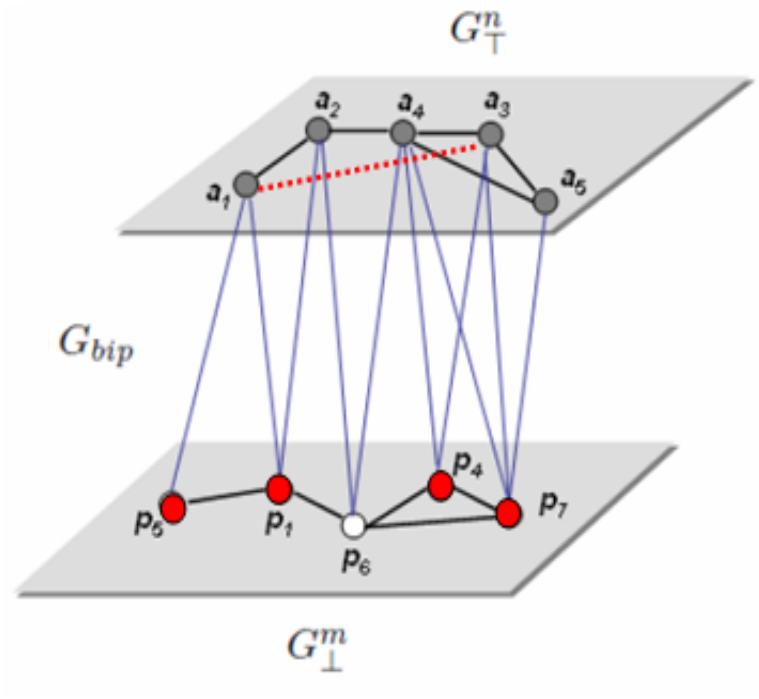
Approche dyadique non-supervisé

- Score d'un lien (u, v) est calculé comme une fonction des *caractéristiques* topologiques de u et v
- Trois familles de mesures de similarités topologiques :
  - Produits des caractéristiques des nœuds : Attachement préférentiel, produit des centralités, pageRank, etc.
  - Mesures basées sur le voisinage des nœuds : Voisins communs, Jaccard, Adamic-Adar.
  - Mesures basées sur les distances entre les nœuds : plus-court-chemin, Katz, Marche aléatoire (MFA, CT, ..).
- *Prédicteur de lien* = un unique attribut *topologique*
- Tous les exemples (paires de nœuds) sont ordonnés en fonction de la valeur de la mesure employé.
- Les k top liens sont retournés.
- Deux types de scores à calculer :
  - directes : calculé dans le graphe de co-publication
  - indirectes : calculés entre des nœdus d'autres types liées aux auteurs (ex. publications)

Mesure indirecte pour les réseaux de collaborations

Une mesure indirecte dans un **réseau de collaborations** est calculée à partir du graphe $G_{\perp}^{m}$:

$$A_{\perp}^{m}(x,y) = \Phi_{u \in \Gamma_{G_{bip}}(x), v \in \Gamma_{G_{bip}}(y)}(A(u,v))$$



$x \in \top$ et $y \in \top$
$A_{\perp}^{m}$ est l'attribut indirect du point de vue du graphe $G_{\perp}^{m}$
Φ est la fonction d'agrégation
$\Gamma_{G_{\perp}^{m}}(y)$ est l'ensemble de tous les voisins du nœud y dans le graphe $G_{\perp}^{m}$

Exemple

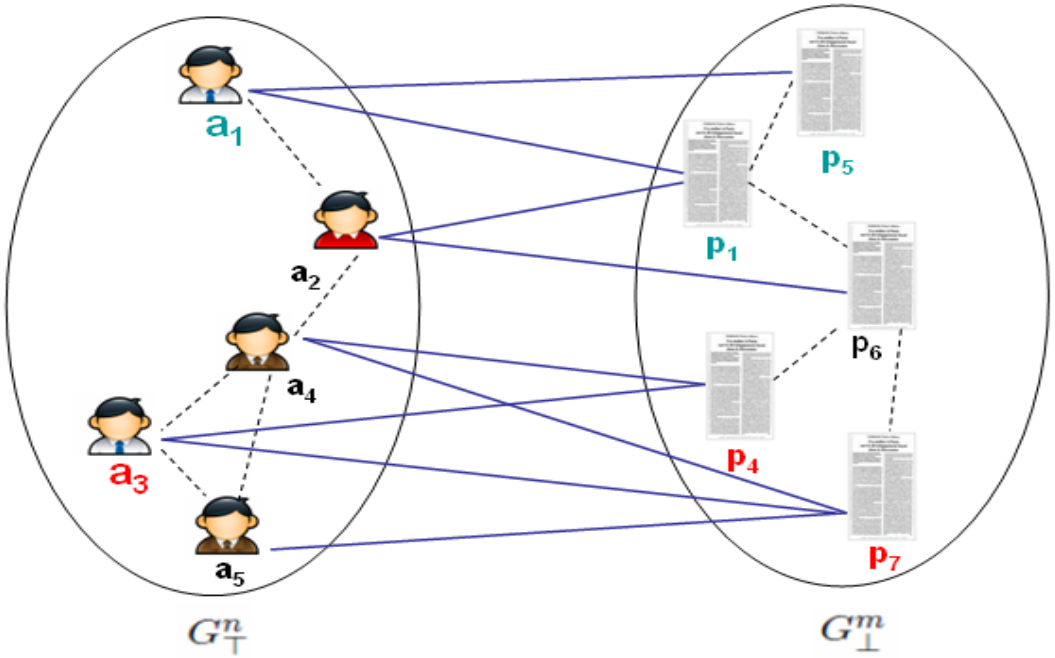
Voisins communs direct :

- $vc(a_1, a_3) = |\Gamma_{G^1_\top}(a_1) \cap \Gamma_{G^1_\top}(a_3)| = |\{a_2\} \cap \{a_4, a_5\}| = 0$

Voisins communs indirect :

- $vc^1_\perp(a_1, a_3) = Max_{u \in \Gamma_{G_{bip}}(a_1), v \in \Gamma_{G_{bip}}(a_3)} vc(u, v) = 1$

$vc(p_1, p_4) = |\Gamma_{G^1_\perp}(p_1) \cap \Gamma_{G^1_\perp}(p_4)|$
$vc(p_1, p_7) = |\Gamma_{G^1_\perp}(p_1) \cap \Gamma_{G^1_\perp}(p_7)|$
$vc(p_5, p_4) = |\Gamma_{G^1_\perp}(p_5) \cap \Gamma_{G^1_\perp}(p_4)|$
$vc(p_5, p_7) = |\Gamma_{G^1_\perp}(p_5) \cap \Gamma_{G^1_\perp}(p_7)|$

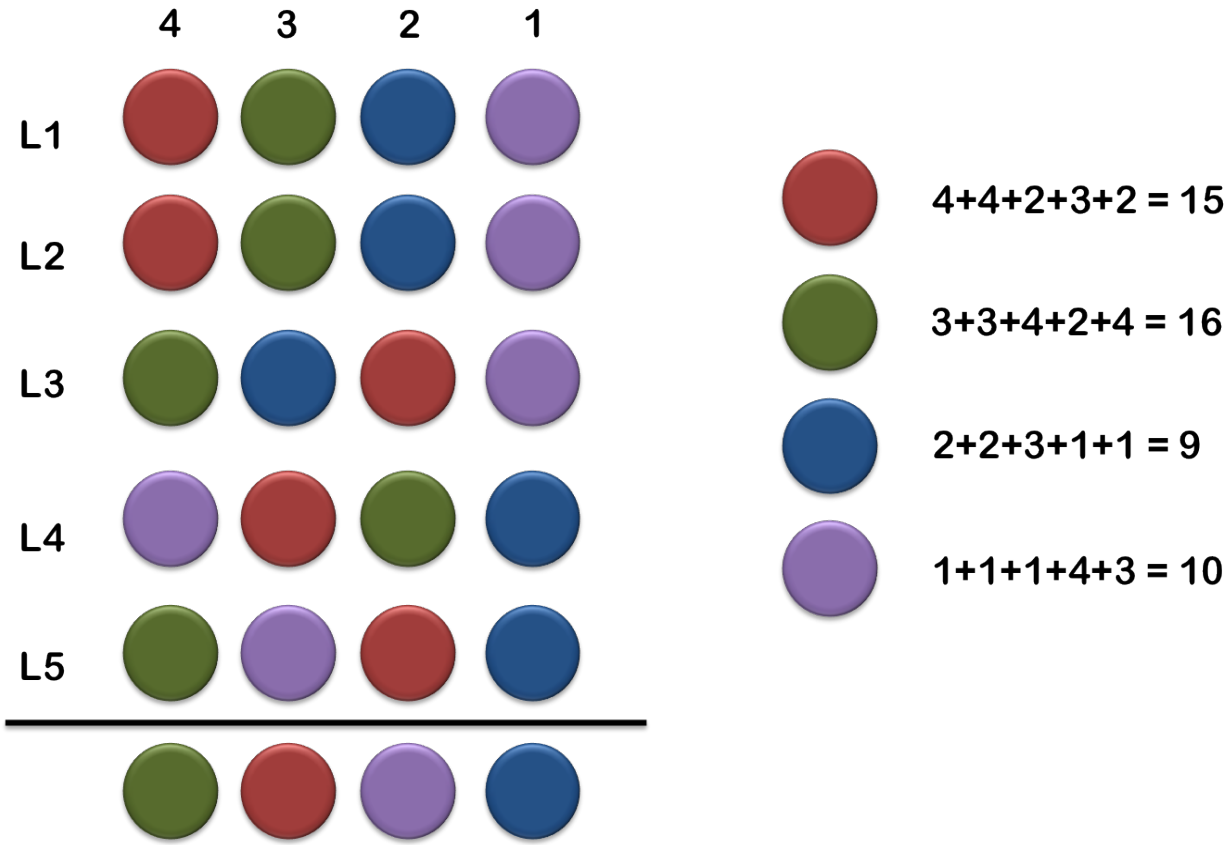


Combinaison non-supervisée des attributs

- Chaque mesures topologique trie la liste de liens candidats selon les valeurs obtenues
- Les différents tris peuvent être combinés en utilisant des techniques de vote (non-supervisé)
- Technique de vote :
  - Borda : approche basée sur les positions des candidats
  - Kemeney : Election de vainqueur de Condorcet : un candidat est placé au rang i s'il a la majorité contre tous les candidats de rang $j > i$

Approche de Borda

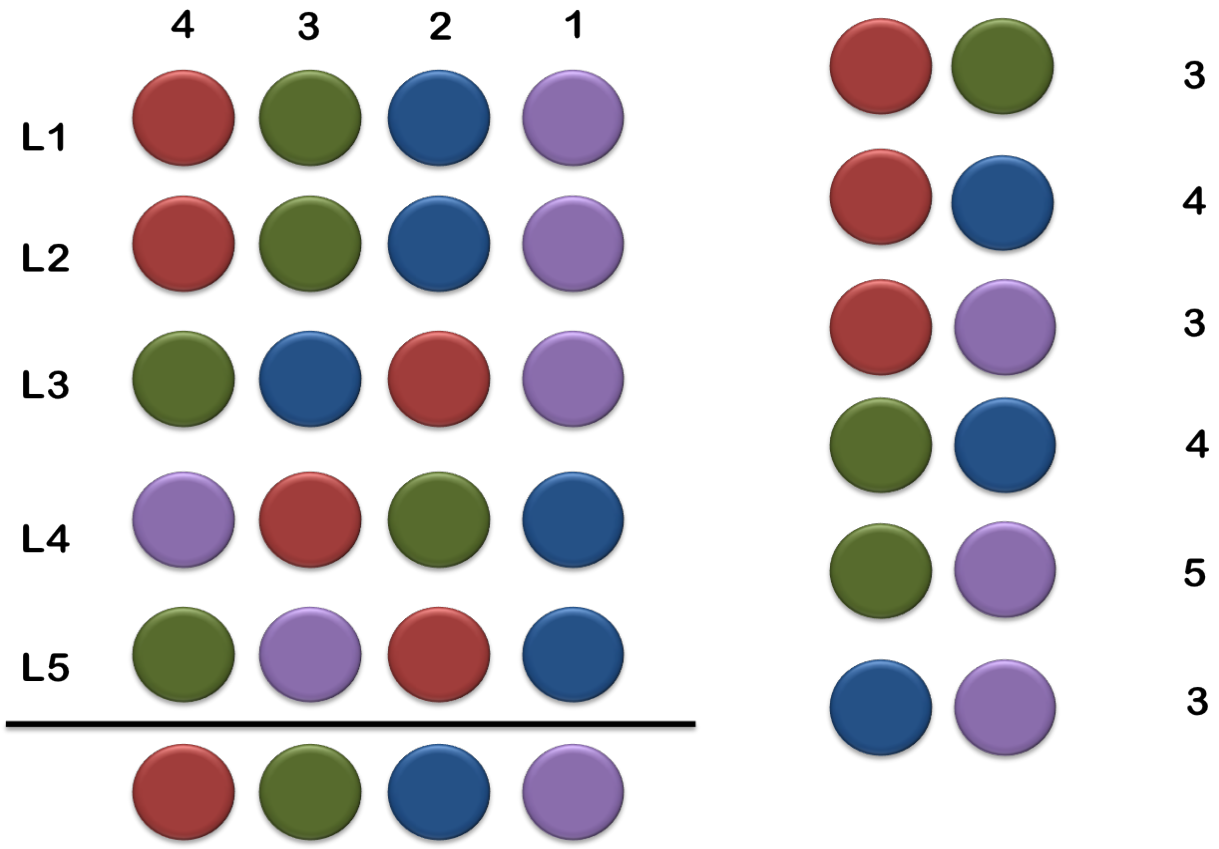
$$B_{L_k}(i) = \{count(j) | L_k(j) < L_k(i) \& j \in L_k\} \quad ; \quad B(i) = \sum_{t=1}^{k} B_{L_t}(i) \tag{1}$$



Approche de Kemeny

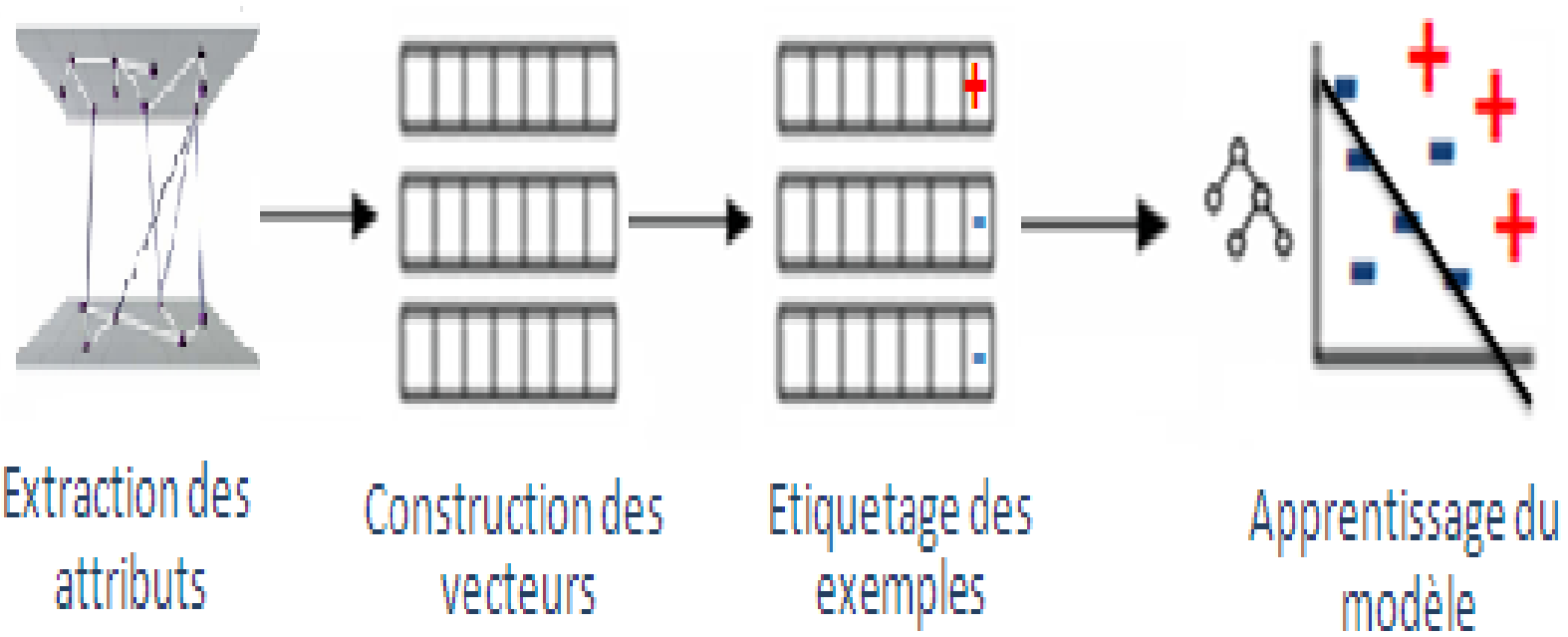
$$SK(\pi, L_1, L_2, L_3, ..., L_n) = \sum_{i \in [1,n]} K(\pi, L_i) \quad (2)$$

- Kendall Tau Distance : $K(L_1, L_2) = \mid (i,j) \ s.t. \ L_1(i) < L_2(j) \ \& \ L_1(i) > L_2(j) \mid$



Approche basée sur l'apprentissage automatique

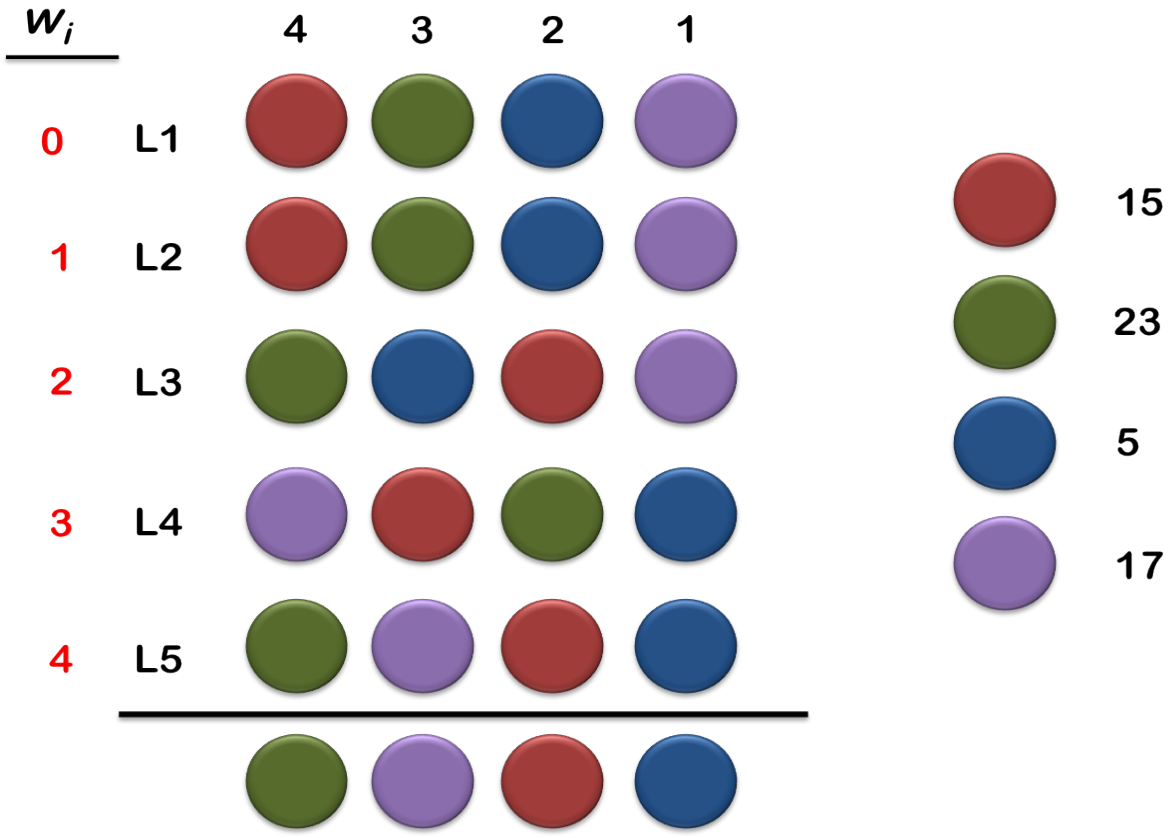
- Les exemples d'apprentissage sont ensuite étiquetés :
  - Classe **positive** : les paires de nœuds qui sont liées dans G_E.
  - Classe **négative** : les paires de nœuds qui ne sont pas connectées dans G_E, et dont l'un des nœuds participe à la construction d'un exemple positif.
- On applique enfin un algorithme d'apprentissage supervisé.



Borda supervisé

Borda score

$$B(i) = \sum_{t=1}^{n} w_i * B_{L_t}(i) \quad ; \quad where \quad t \in [1, k] \tag{3}$$



Kemeny supervisé

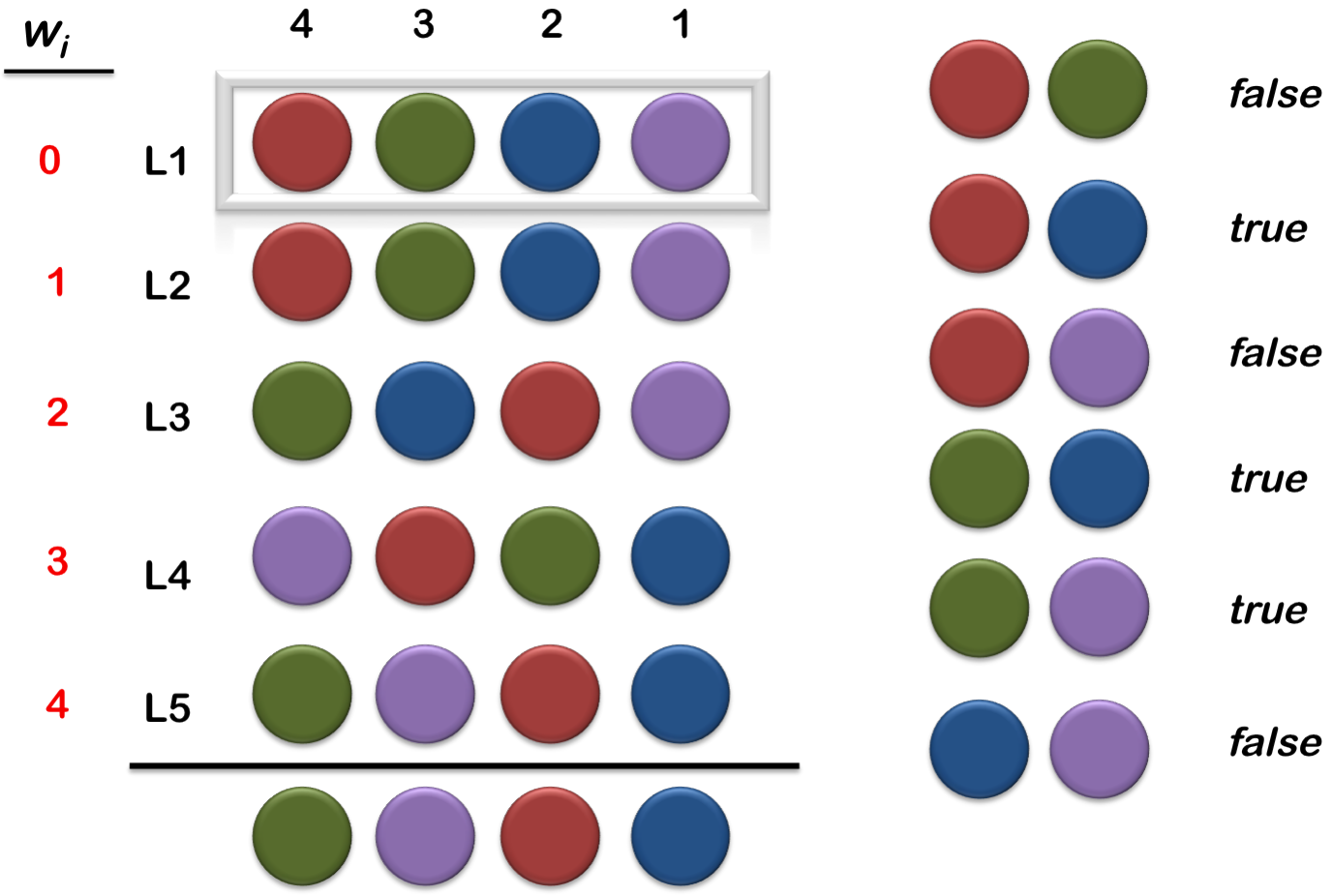
Steps :

1. $L = [L_1, L_2, \ldots, L_n]$, $[w_1, w_2, \ldots, w_n]$, m elements(U)
2. *Initialize* $m \times m$ matrix M with $M(x, y) = 0$
3. $\forall (x, y) \in U$, *Compute*

$$score(x, y) = \sum_{i=1}^{n} (w_i * (x \succ y)) \text{ where } x \succ y = \begin{cases} 0 & \text{if } L_i(x) < L_i(y) \\ 1 & \text{if } L_i(x) > L_i(y) \end{cases}$$

4. *If* $score(x, y) > 0.5 * \sum_{i=1}^{n} w_i$, *Insert* $M(x, y) = true$ and $M(y, x) = false$
5. Initial aggregation $R = L_1$
6. For $x, y \in R$, $Swap(x, y)$ if $M(x, y) = false$
7. R is the final aggregation.

Supervised Local Kemeny Aggregation



Calcul de poids

par Maximisation de la précision :

$$W_{a_i} = n * Precision_{a_i} \tag{4}$$

par Minimisation de taux de faux positives

$$W_{a_i} = \frac{n}{FPR_{a_i}} \tag{5}$$

- $|\top|$ est le nombre des auteurs, $|\perp|$ est le nombre de publications.
- E est le nombre de liens dans le graphe biparti.
- cc est le coefficient de clustering, et $Avg(dis)$ est la distance moyenne.
- d_s est la densité globale.

- Les données de la base bibliographique **DBLP** (http ://www.informatik.uni-trier.de/ley/db/)
- Le graphe biparti est composé de deux types de nœuds : *auteurs* et *publications*.

G_o	$\lvert\top\rvert$	$\lvert\perp\rvert$	$\lvert E\rvert$	cc	d_s	$Avg(dis)$
[1970,1973]	2661	1487	6634	0.88	6.10^{-4}	10
[1972,1975]	4536	2542	10855	0.87	3.10^{-4}	18
[1974,1977]	6844	4128	16281	0.84	7.10^{-5}	4
[1980,1983]	13345	9547	52720	0.5	7.10^{-5}	5.6
[1990,1993]	50560	65247	244146	0.44	2.10^{-5}	10

Précision des attributs simples

1***Attributes**	1* **Dataset1**	1* **Dataset2**	1* **Dataset3**
Katz	0	0	0.0061
MFA	0.0244	0.0732	0.0488
Produit PageRank	0.0244	0.0244	0
Voisins communs	0.5122	0.4268	0.1829
Jaccard	0.2195	0.1707	0.0488
Adamic-Adar	0.1463	0.1463	0.1463
Attachement pref.	0.0488	0	0
Distance	0	0.0122	0.0244
Indirect Distance	0.6098	0.5366	0.8171
Indirect Adamic-Adar	0.0976	0.0488	0.0366
Indirect Attachement pref.	0.0244	0	0
Indirect MFA	0.0488	0.0732	0.0427
Indirect Jaccard	0.0488	0.1098	0.0549
Indirect Prod. PageRank	0.0488	0	0
Indirect Voisins communs	0.0488	0.0488	0.0183
Indirect Katz	0.1220	0.1098	0.0488

Expérimentation : Approches supervisées

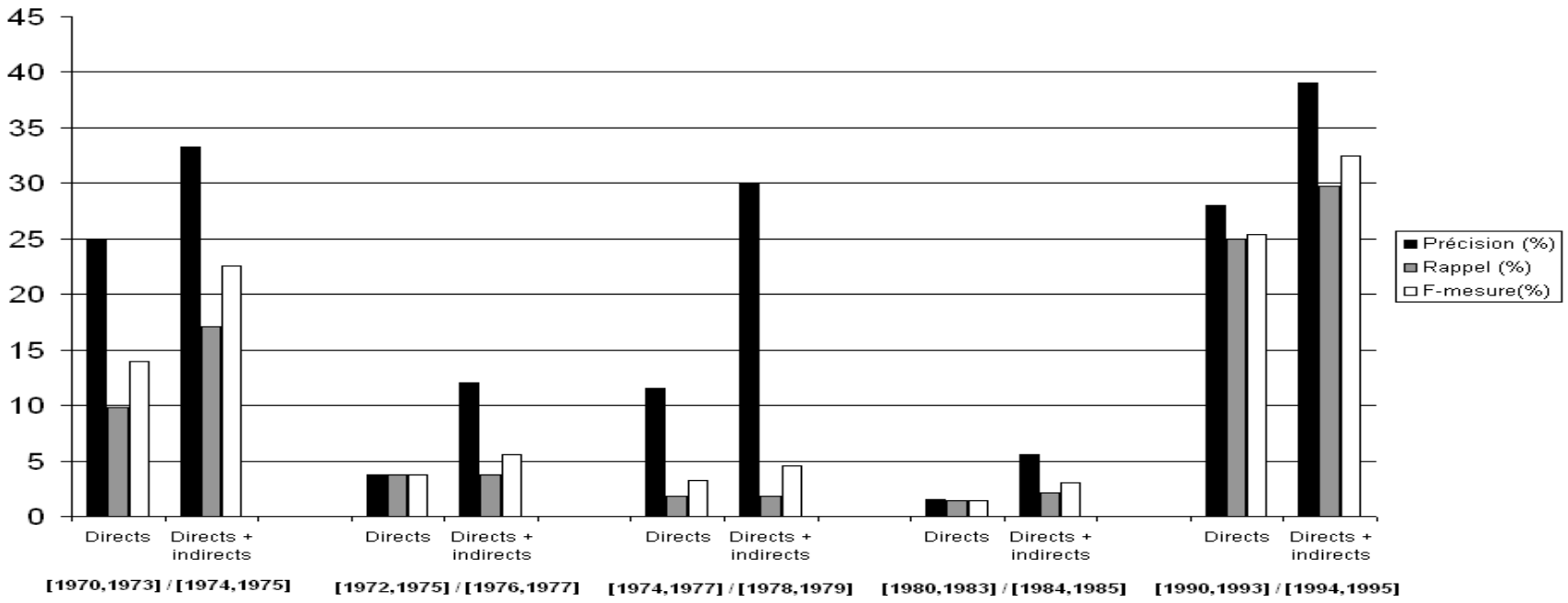
Distribution des exemples

2*Datasets	1*Training Time	1*Validation 1*Time	Training examples		Test examples	
			Positive	Total	Positive	Total
Dataset 1	[1970-1975]	[1971-1976]	30	1693	41	3471
Dataset 2	[1972-1977]	[1973-1978]	87	19332	82	18757
Dataset 3	[1974-1979]	[1975-1980]	102	35190	164	60046

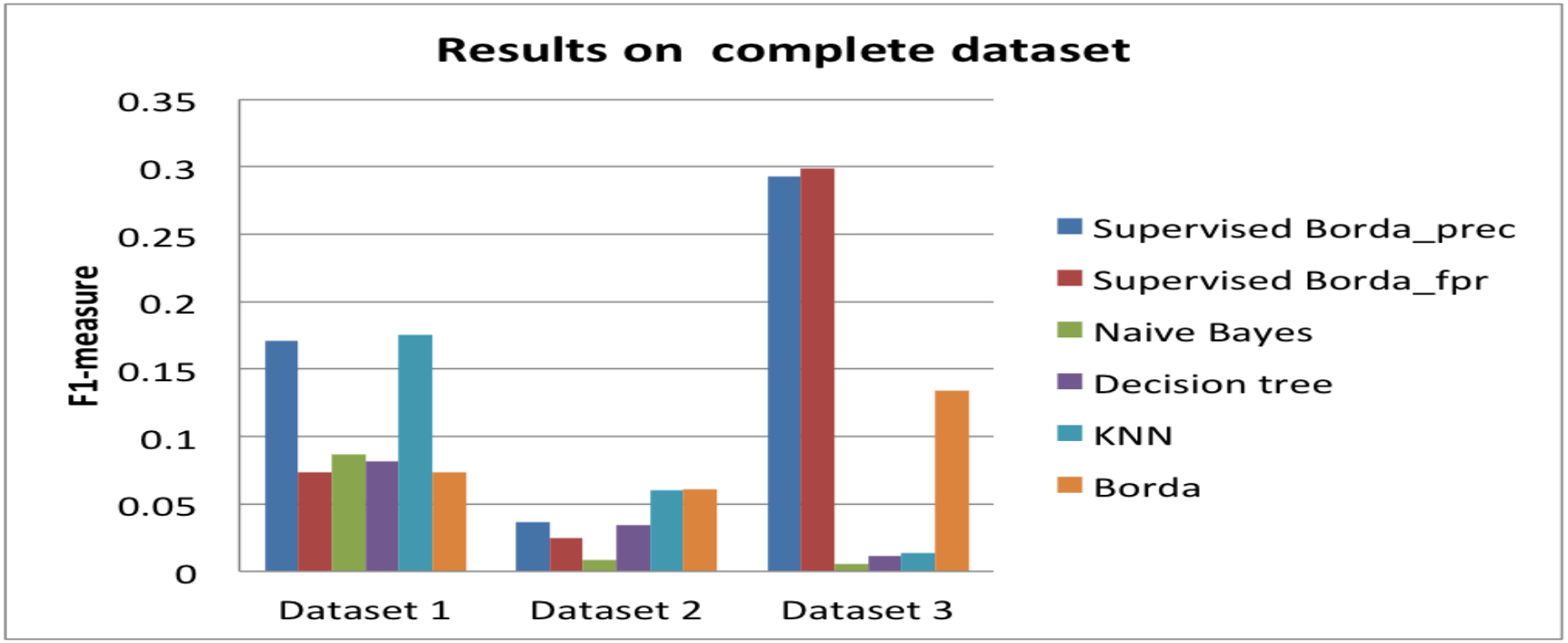
Critère d'évaluation

$$F = 2 \times \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

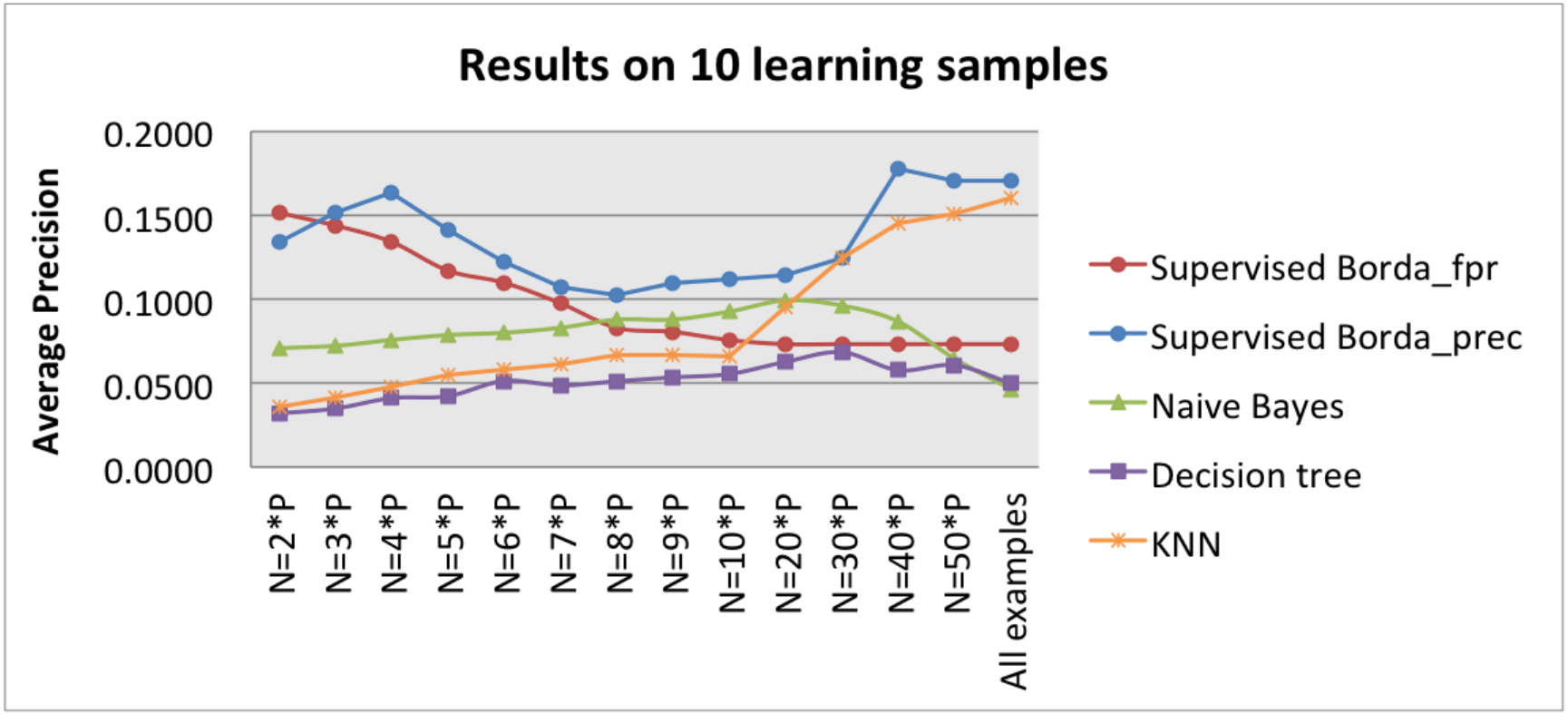
Résultats : Apprentissage supervisé



Résultats : approches supervisées

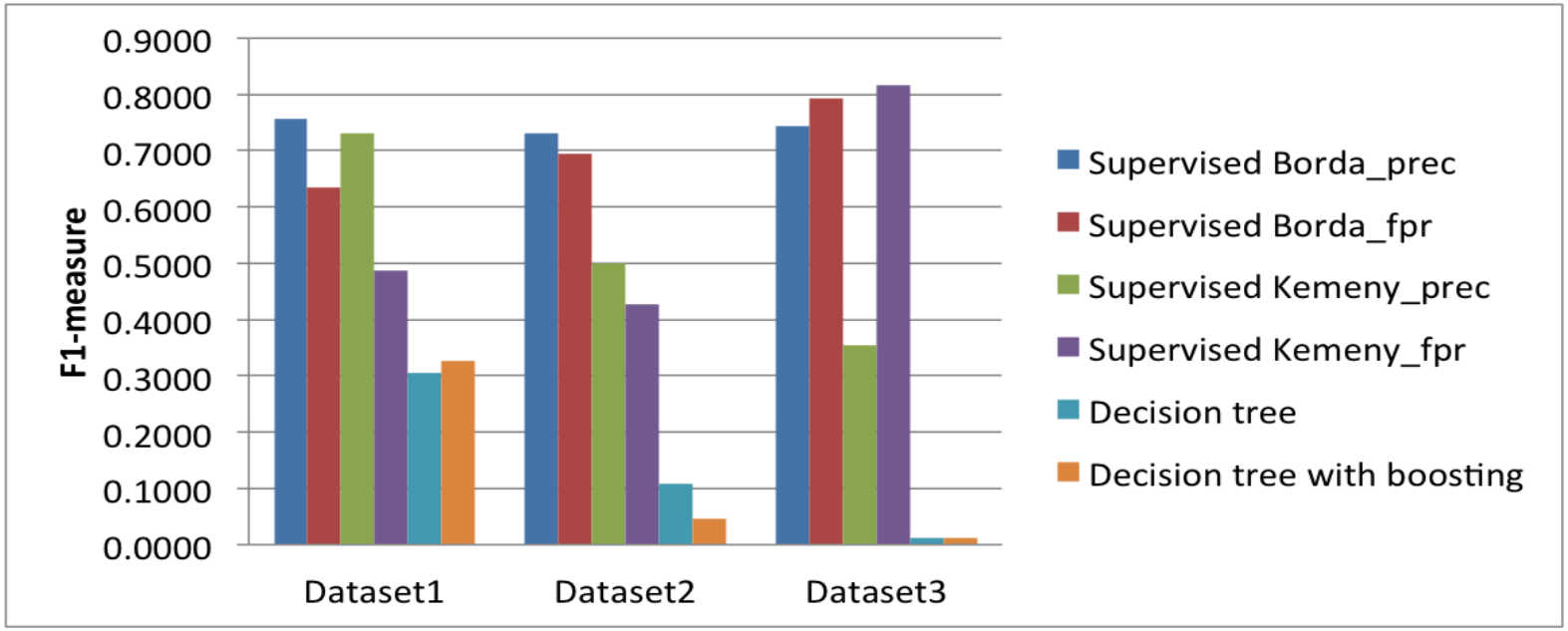


Résultats sur ensembles avec échantillonage $|N| = m \times |P|$

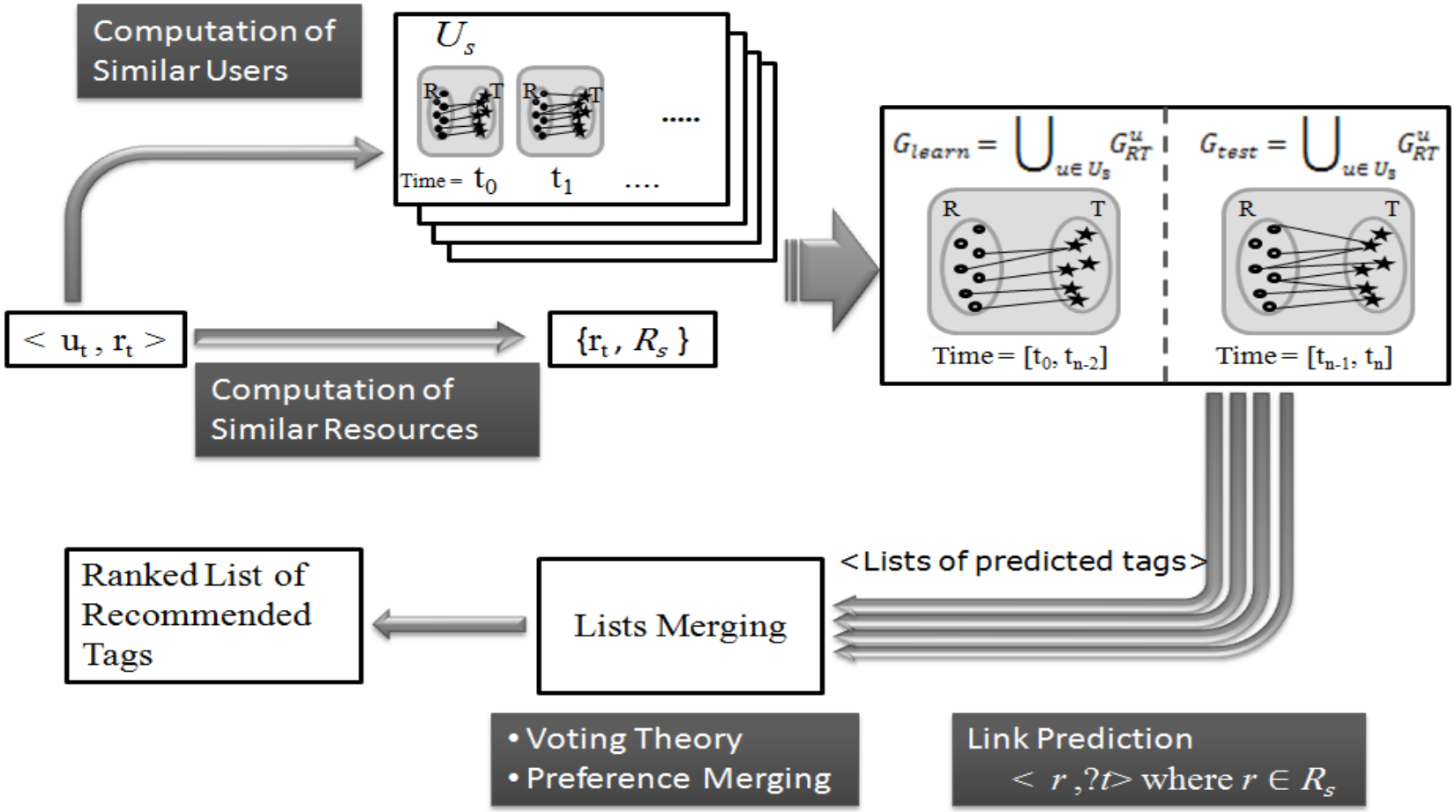


Résultats : $|N| = 5 \times |P|$

2*Datasets	Training examples		Test examples	
	Positive	Total	Positive	Total
Dataset 1	30	1693	41	246
Dataset 2	87	19332	82	492
Dataset 3	102	35190	164	984



LipTAR : Approche de recommendation de tags



R_s : Set of resources similar to target resource
G_{learn} : Learning graph
U_s : Set of users similar to target user
G_{test} : Test graph

LipTAR : Résultats préliminaires

2*Approche	**Average Precision**	**Average Recall**	**Average F-measure**
LiPTaR	0.0136	0.3648	0.0200
CF	0.0134	0.011	0.0111

TABLE: Résultats pour les utilisateurs les plus actifs dans CitULike

Recherche de motifs fréquents au voisinage des nœuds exemples (pivots)

Génération d'une base de graphes pour l'apprentissage

- $\mathbb{EX}^+ = \{e_i^+\}_i = \{(x_i, y_i)\}_i$ l'ensemble des couples de nœuds pivots formant les exemples positifs, et
- $\mathcal{N} : (e_i^+, g_o, d) \rightarrow g_{e_i^+} \subseteq g_o$ une fonction de sélection d'un sous-graphe de voisinage **connexe** de g_t centrée sur le couple de nœuds d'un exemple $e_i^+ = \{x_i, y_i\}$; par exemple, le sous graphe de g_o induit par l'union
  - des nœuds des plus courts chemins entre les nœuds de e_i^+ et
  - des nœuds distants d'une marche de d arêtes des nœuds de e_i^+.
- $\mathbb{D} = \mathcal{N}(\mathbb{EX}^+) = \{g_{e_i^+}\}_i$: la base de graphes pour l'apprentissage.
- Σ : Étiquetage des nœuds et des arêtes.
  - $\Sigma(x_i) = \begin{cases} (deg_{g_o}(x_i),' p') & \text{si } x_i \text{ est un pivot} \\ (deg_{g_o}(x_i),' o') & \text{sinon} \end{cases}$
  - $\Sigma(x, y) = card(\{(x, u)\} \cap \{(y, v)\})$, le nombre de publication en comunes à x et y.

Apprentissage : Recherche de motifs fréquent $\mathbb{S}$ dans $\mathbb{D}$: gSpan [X. Yan and J. Han. (2002)]

- utilise une représentation canonique des graphes : Code DFS minimal,
- explore en profondeur les motifs fréquents $\mathbb{S}$ de la base $\mathbb{D}$.
- $\mathbb{S} = \{s_i\}/support_{\mathbb{D}}(s_i) \geq minsup$

Recherche de motifs fréquents au voisinage des nœuds exemples (pivots)

Algorithm 1: Main

1: $\mathbb{S} \leftarrow \emptyset$
2: $gSpan(root, \mathbb{D}, \mathbb{S})$
3: $\mathbb{P} \leftarrow \emptyset$
4: **for all** $s \in \mathbb{S}$ **do**
5: **if** si s présente 2 nœuds pivots **then**
6: $\mathbb{P} \leftarrow \mathbb{P} \cup s$
7: **end if**
8: **end for**

Résultats

- DBLP jusqu'aux années 1990, pour $minsup = 0.2$,
- 40-80 règles,
- $Rappel \simeq 0,8$.

Perspectives : Optimisation

- Élagage de l'arbre de recherche si un motif ne comporte pas 2 pivots (1 a 2 ordres de grandeur),
- Restriction de la recherche aux maximaux (MARGIN ou SPIN) (>1 ordre de grandeur).

Perspectives : Exploration

- Choix de fonctions d'étiquetage,
- Selection de règles maximisant la precision
- ...

Conclusion

- Pertinence des approches topologiques (dyadiques)
- Problème de déséquilibre de classes.
- Première validation de la fouille des réseaux hétérogènes.
- Travaux en cours :
  - Prise en compte du niveau mésoscopique des communauté pour la prévision de liens : apprentissage des modèles inter et intra communautés.
  - Nouvelle approche de détection de communautés
  - Généralisation de la fouille des réseaux hétérogènes (prise en compte d'autres types de nœdus acteurs dans les réseaux bibliographiques.

Pub : MARAMI/FGG 2012, 17-19 Octobre 2012

3ième conférence sur les modèles et l ?analyse des réseaux : Approches mathématiques et informatiques / Journée de foule de Grands Graphes (FGG)

pour plus d'informations :

http ://lipn.fr/MARAMI12, http ://lipn.fr/fgg12

Deadline **29 Juin 2012**