Calcul de similarité entre événements sociaux

Armel FOTSOH

Laboratoire LIUPPA, BP-1155, 64013 PAU Université Cedex, France
@ArmelCPU
aftawofaing@univ-pau.fr
armel@cogniteev.com

06 Novembre 2017

Plan

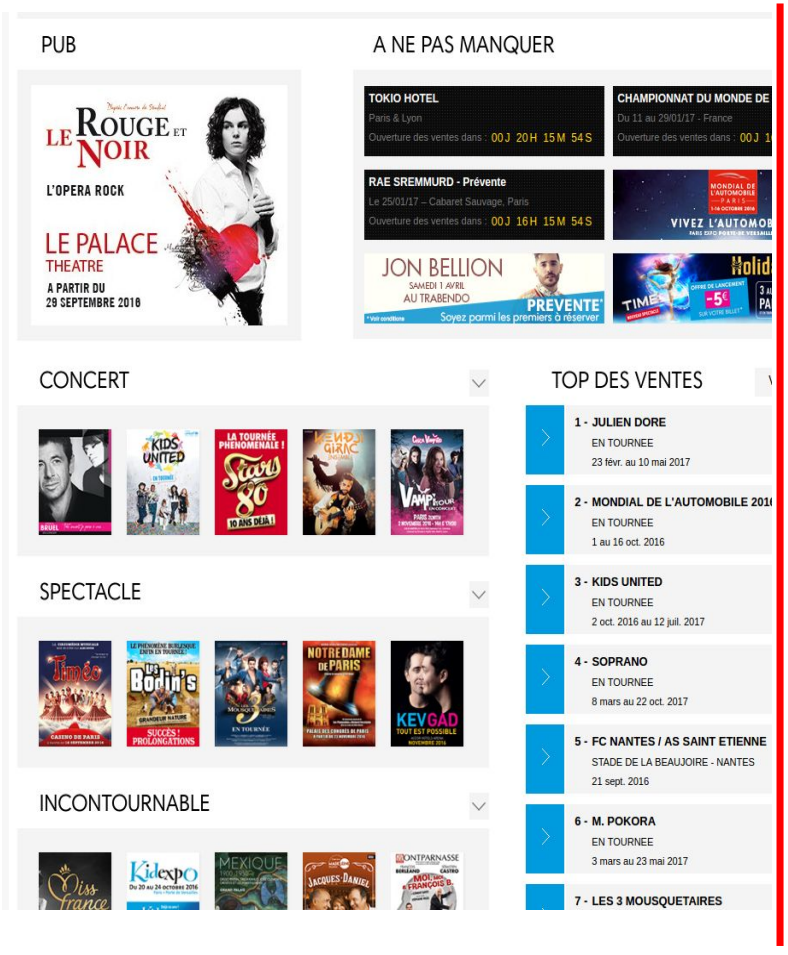
Introduction

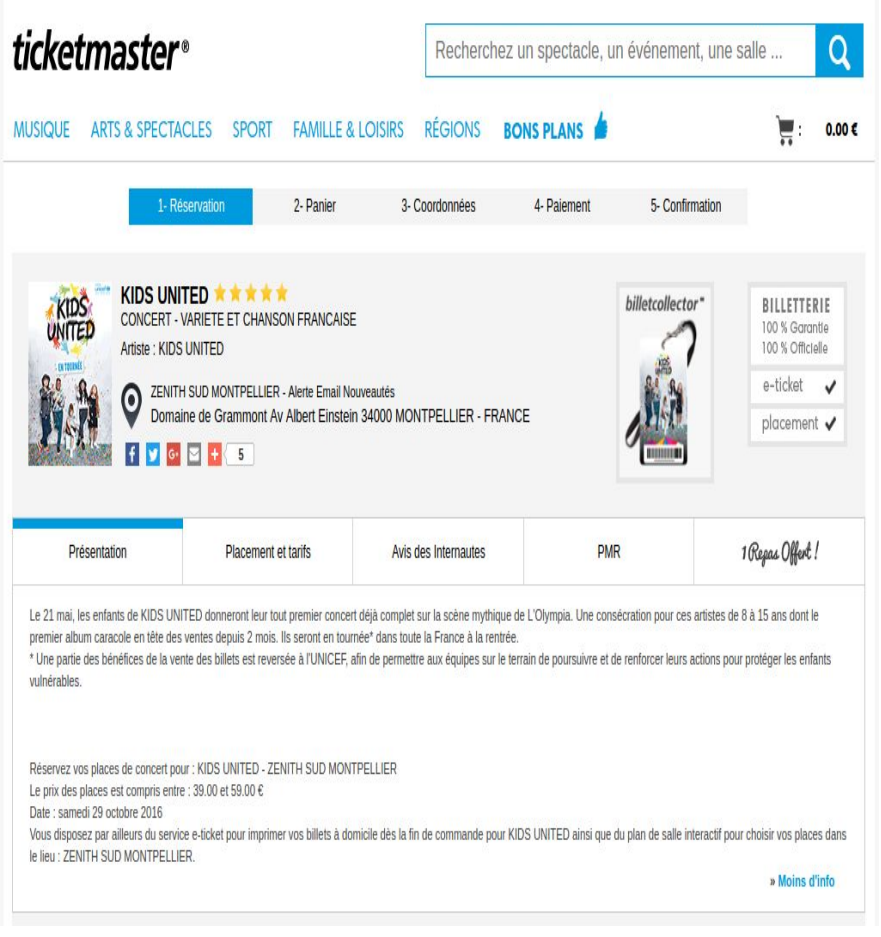
Contexte

- **Cognisearch Event** : Service dédié à l'extraction et à la recherche d'événements sur le web
- **Types d'événements ciblés** : événements socio-culturels (concerts, forums industriels, tournois sportifs, meetings politiques, etc.)
- **Trois dimensions traitées**
  - *Espace* : où ?
  - *Temps* : quand ?
  - *Thème* : qui ? quoi ?

Cogniseach Event

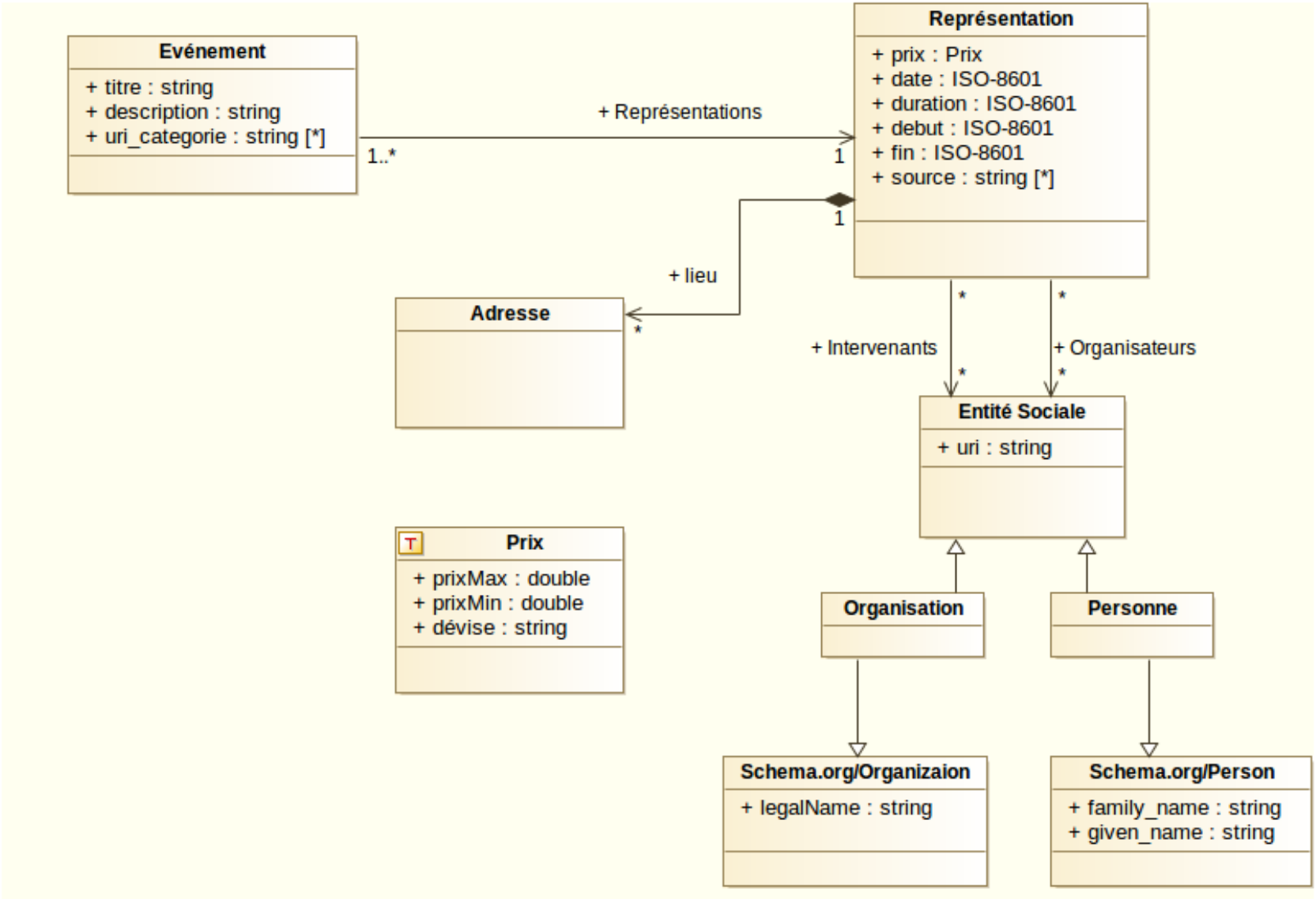
Exemple de pages traitant d'événements



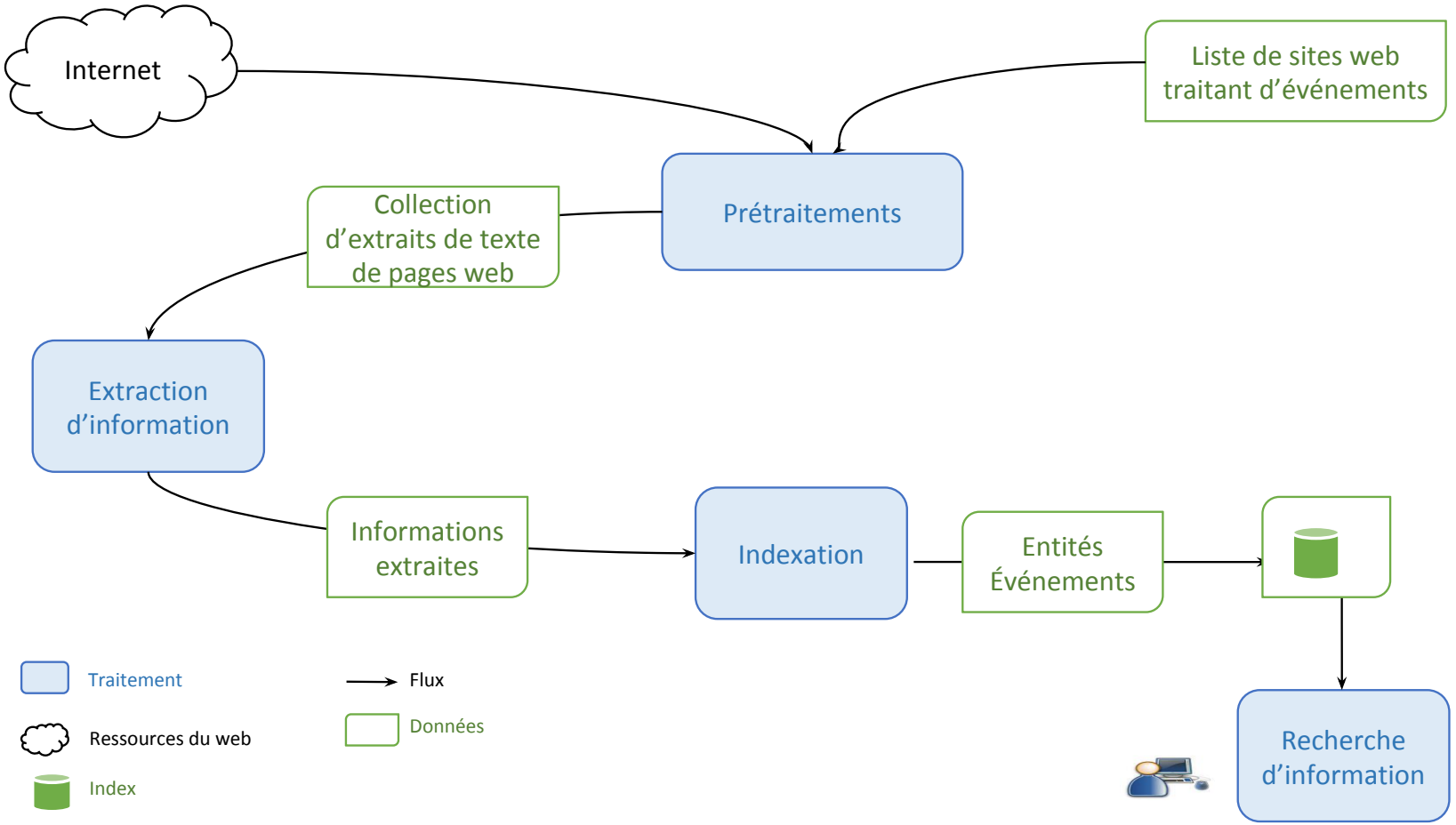


Architecture du service Cogniseach Event

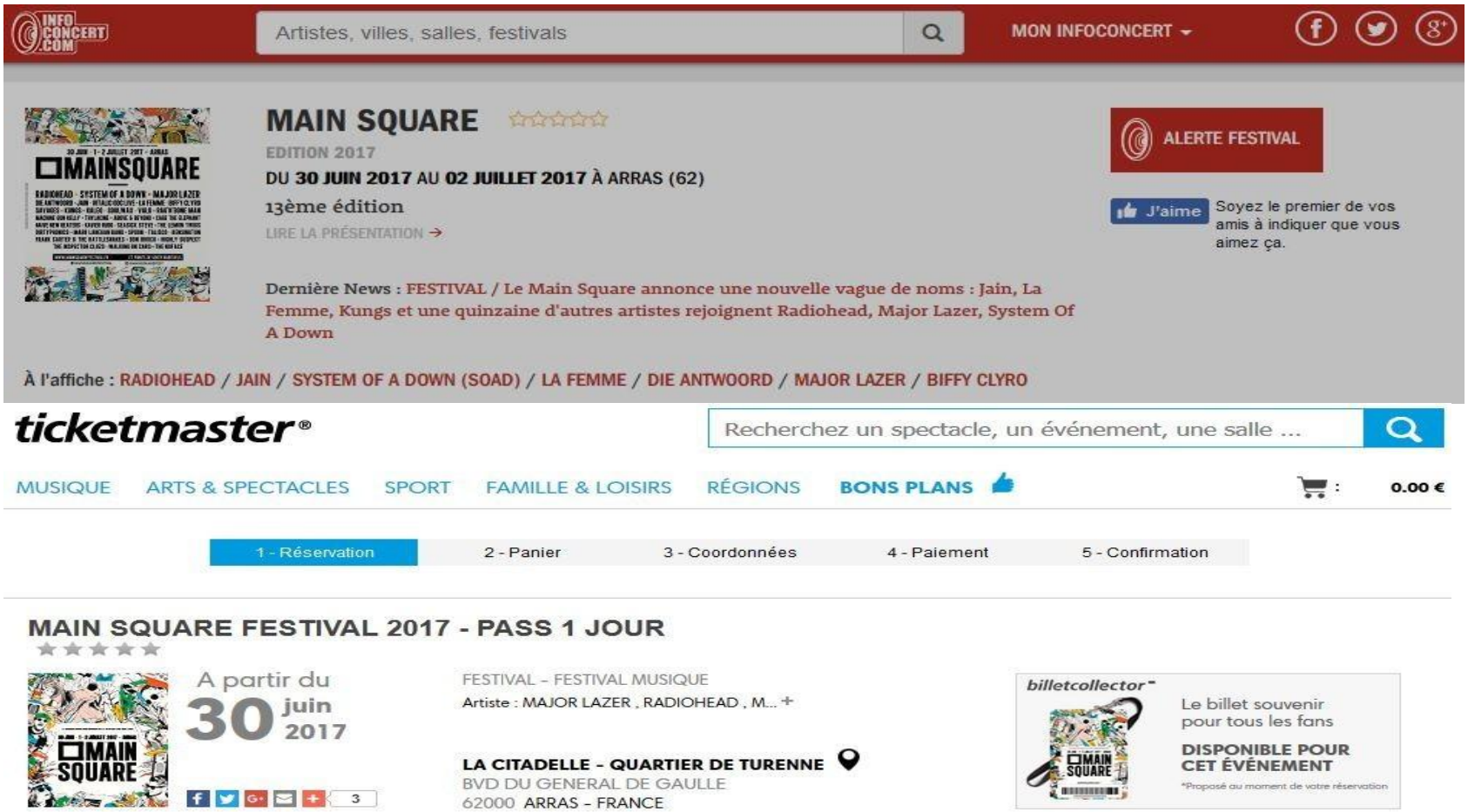
Modèle de représentation d'événement



Enchaînement des modules de l'architecture du service



Problématique : illustration



Problématique : verrous

- *Propriétés exprimées différemment*
  « Arras » vs « La Citadelle - Quartier de Turenne, boulevard du Général de Gaulle 62 000 Arras France »
- *Propriétés non renseignées* « Festival - Festival Musique » vs RIEN
- *Évaluation de la similarité au niveau de l'EN complexe : agrégation de similarités*

Calcul de similarité entre EN complexes

Calcul de similarité entre EN complexes

- Soient o_1 et o_2 deux EN complexes :

$$o_1 =< p_{11}, p_{12}, ..., p_{1n} >$$

$$o_2 =< p_{21}, p_{22}, ..., p_{2n} >$$

- Calcul de similarité en deux temps :
  - Calcul de la similarité entre propriétés de même rang : $s_i(p_{1i}, p_{2i})$
  - Agrégation des s_i

Approches d'agrégation des similarités intermédiaires

- La combinaison linéaire

$$s(o_1, o_2) = \sum_{i=1}^{n} w_i . s_i(p_{1i}, p_{2i}) \text{ avec } \sum_{i=1}^{n} w_i = 1$$

  Plusieurs approches de pondération :
  - Pondération empirique
  - *Prioritized Agregation*
  - Utilisation des méta-heuristiques (Algorithmes génétiques ...)
- Exploitation des fonctions logiques

$$s(o_1, o_2) = (s_2 > \epsilon_2) \quad OU \quad ((s_1 > \epsilon_1) \quad ET \quad (s_3 > \epsilon_3))$$

Calcul des poids par étalonnage

- Elle ré-utilise la combinaison linéaire

$$s(o_1, o_2) = \sum_{i=1}^{n} w_i . s_i(p_{1i}, p_{2i}) \text{ avec } \sum_{i=1}^{n} w_i = 1$$

- Les w_i sont déterminés par étalonnage

0	0	1	0,2	0	0,8	0,4	0	0,6	0,6	0	0,4	0,8	0	0,2	1	0	0
0	0,2	0,8	0,2	0,2	0,6	0,4	0,2	0,4	0,6	0,2	0,2	0,8	0,2	0			
0	0,4	0,6	0,2	0,4	0,4	0,4	0,4	0,2	0,6	0,4	0						
0	0,6	0,4	0,2	0,6	0,2	0,4	0,6	0									
0	0,8	0,2	0,2	0,8	0												
0	1	0															

Agrégation basée sur la régression logistique

- $s(o_1, o_2)$ est la probabilité d'identité étant donnés les $s_i(p_{1i}, p_{2i})$

$$s(o_1, o_2) = P(1|X) \quad \text{avec} \quad X = (s_1(p_{11}, p_{21}), ..., s_n(p_{1n}, p_{2n}))$$

- P(1|X) est évaluée en exploitant la régression logistique

$$s(o_1, o_2) = \frac{1}{1 + e^{-(w_0 + w_1 . s_1(p_{11}, p_{21}) + ... + w_n . s_n(p_{1n}, p_{2n}))}}$$

- Les w_i sont déterminés par descente de gradient

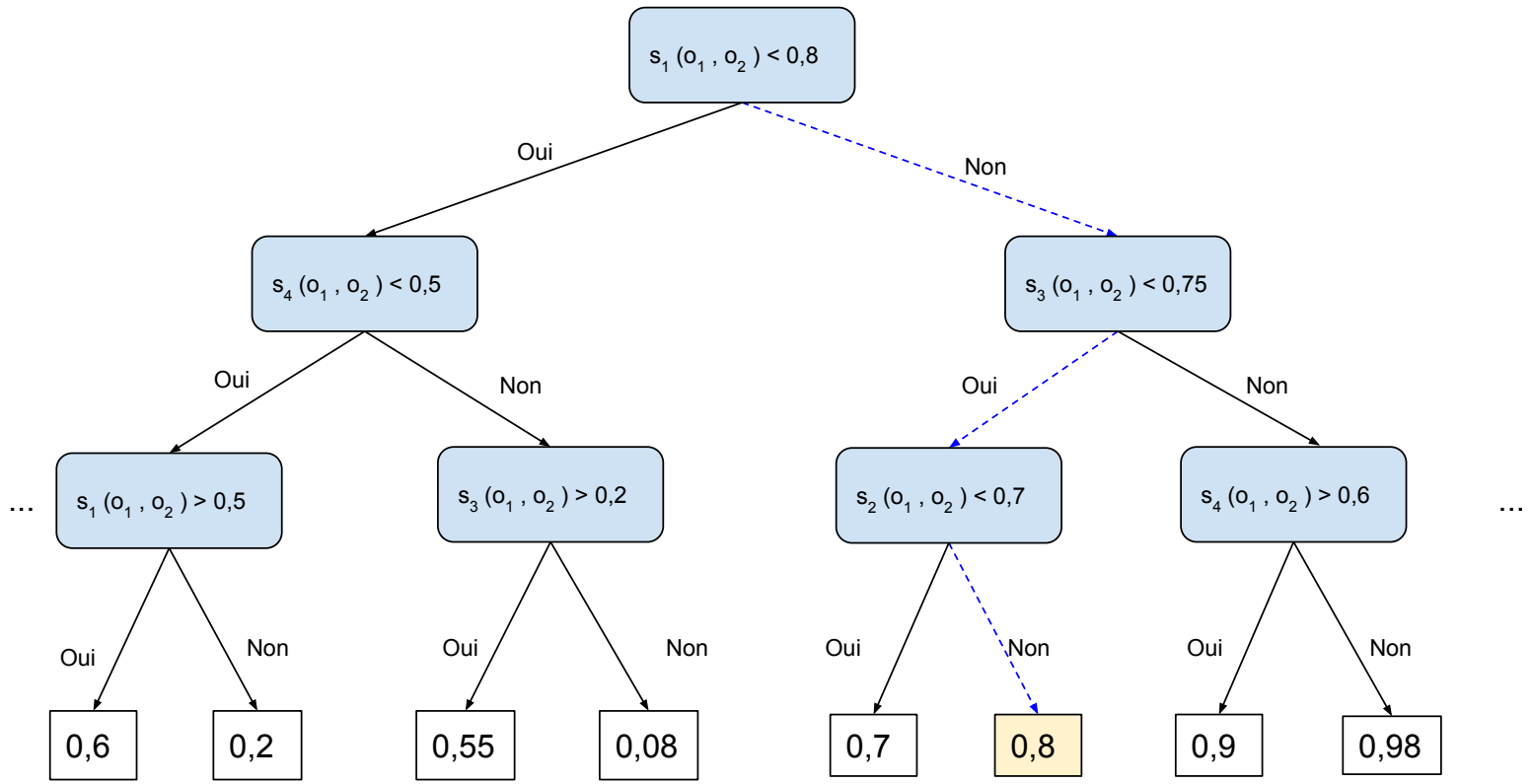
Régression basée sur les arbres de décision : principe

- Segmentation de l'ensemble des couples d'EN en sous-ensembles homogènes : couples d'EN dont les similarités intermédiaires vérifient le même ensemble de conditions

- Les conditions sont définies en fonction des similarités entre propriétés de même rang

Exemple : $s_1(p_{13}, p_{23}) > 0,7$

Illustration de la mise en œuvre des arbres



Application au calcul de similarité entre événements sociaux

Formalisation

- Un événement est défini par :

$$e =< t, Cat_e, A_{P_e} >$$

- Une représentation est définie par :

$$r =< l, d, h > \ , \ r \in R_e$$

- Deux cas à observer :
  - toutes les propriétés renseignées
  - des propriétés sont non renseignées dans au moins une des EN

Expérimentation du calcul de similarité entre parties *event*

- Similarité booléenne : les EN sont similaires ou non (0 ou 1)
- Jeu d'évaluation : 100 couples d'EN *event* (52 - 42)
- Jeu d'entraînement et de validation : 85 couples d'EN (50 - 35)
- Baseline : CombMNZ

  $s(o_1, o_2) =$ somme des s_i non nuls $*$ nombre de s_i non nuls
- Métriques d'évaluation : Précision, Rappel, F_1-mesure et Exactitude

Résultats & Analyses

Résultat sur l'ensemble du jeu d'évaluation

Approches	CombMNZ	Etalonnage	Logistique	Arbres
Précision	79,03	79,03	**89,09**	84,48
Rappel	84,48	**86,21**	84,48	84,48
F_1-mesure	81,67	82,64	**86,73**	84,48
Exactitude	78,00	79,00	**85,00**	82,00

Résultats & Analyses : Scénario 1

Scénario 1 : toutes les propriétés renseignées

Approches	CombMNZ	Etalonnage	Logistique	Arbres
Précision	**92,86**	86,67	100	**92,86**
Rappel	**92,86**	**92,86**	85,71	**92,86**
F_1-mesure	**92,86**	89,66	92,31	**92,86**
Exactitude	91,30	86,96	91,30	91,30

Résultats & Analyses : Scénario 2

Scénario 2 : au moins une propriété manquante dans l'un des EN

Approches	CombMNZ	Etalonnage	Logistique	Arbres
Précision	75,00	77,08	**86,05**	81,82
Rappel	81,82	**84,09**	**84,09**	81,82
F_1-mesure	78,26	80,43	**85,06**	81,82
Exactitude	74,03	76,62	**83,12**	79,22

Conclusion

Conclusion

Synthèse

- Extraction et de recherche d'EN événements sur le web
- Calcul de similarité pour l'intégration des EN et la RI : proposition de nouvelles approches.
- 73% des events et 56% des représentation de l'index obtenu ont des propriétés non renseignées

Perspectives

- Évaluer les approches d'un point de vue temps et ressources de traitement
- Évaluer les approches dans un contexte de RI

Autres travaux

- Développement d'une approche combinant clustering et modèle vectoriel