

## 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



**COGNITEEV**

# Plan

## Introduction

## Architecture du service Cogniseach Event

- Modèle de représentation d'événement

- Chaîne de traitement

## Approches d'agrégation de similarités entre propriétés

- État de l'art

- Approche par étalonnage

- Approche par régression logistique

- Approche basée sur les arbres de décision

## Application au calcul de similarité entre événements sociaux

## Conclusion

# 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

**PUB**

**A NE PAS MANQUER**

**LE ROUGE ET LE NOIR**  
L'OPERA ROCK  
LE PALACE THEATRE  
A PARTIR DU 28 SEPTEMBRE 2018

**TOKIO HOTEL**  
Paris 6ème  
Quintessence des années 80  
002 20H 15M 54€

**CHAMPIONNAT DU MONDE DE**  
TV 11 OCTOBRE - Premier  
Quintessence des années 80

**RAE GREENBURD - Présent**  
Le 28/09/17 - Cabaret Boulevard, Paris  
Quintessence des années 80  
002 20H 15M 54€

**VIVIZ L'AUTOMOBILE**

**JON BELLION**  
Samedi 14/10  
AU TRÉBOND  
PREVENTE

**CONCERT**

**TOP DES VENTES**

1 - JULIEN DORÉ  
EN TOURNEE  
23 Mar au 10 Mai 2017

2 - MONDIAL DE L'AUTOMOBILE 2017  
EN TOURNEE  
2 au 24 oct. 2016

3 - KIDS UNITED  
EN TOURNEE  
2 oct. 2016 au 12 juil. 2017

4 - SCORPANO  
EN TOURNEE  
8 mai au 22 oct. 2017

5 - FC NANTES / AS SAINT ETIENNE  
STADE DE LA BEAUJOURNE - NANTES  
21 sep. 2016

6 - M. POKORIA  
EN TOURNEE  
3 mars au 23 mai 2017

7 - LES 3 MOUSquetaires

**ticketmaster®**

Recherchez un spectacle, un événement, une salle ...

MUSIQUE ARTS & SPECTACLES SPORT FAMILLE & LOISIRS RÉGIONS BONS PLANS

0,00 €

1 - Réserver 2 - Payer 3 - Coordonnées 4 - Paiement 5 - Confirmation

**KIDS UNITED** ★★★★★  
CONCERT - VARIÉTÉ ET CHANSON FRANÇAISE  
Artiste: KIDS UNITED

ZENITH SUD MONTEPELLIER - Alerte Email Nouveautés  
Domaine de Grammont Av Albert Einstein 34000 MONTEPELLIER - FRANCE

billetcollecteur®

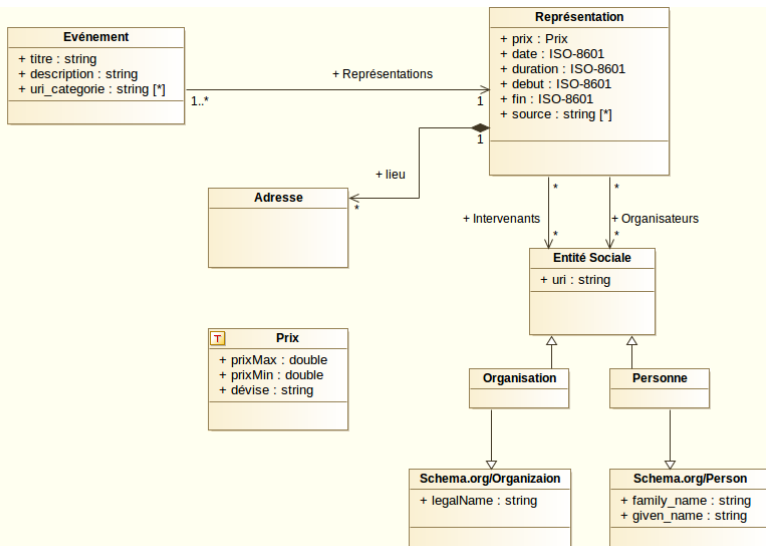
BILLETTERIE  
100 % Garantie  
100 % Officielle

e-ticket ✓  
placement ✓

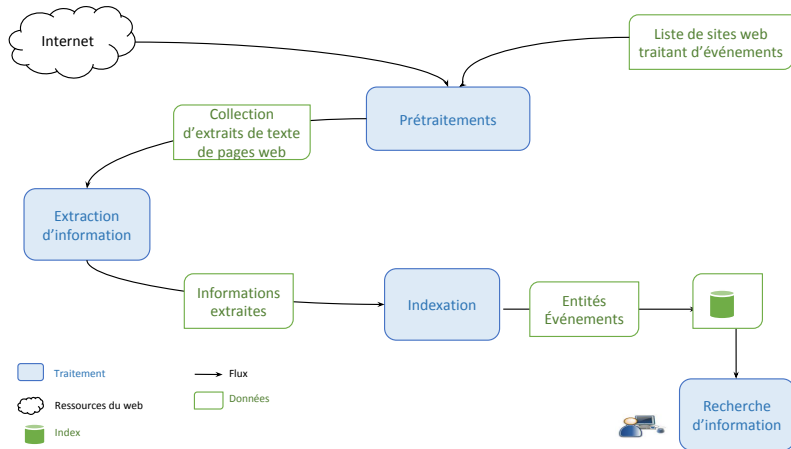
Présentation	Placement et tarifs	Avis des internautes	PMR	Pages Offert!
<p>Le 21 mai, les enfants de KIDS UNITED donneront leur tout premier concert déjà complet sur la scène mythique de L'Olympia. Une consécration pour ces artistes de 8 à 15 ans dont le premier album cassette est tête des ventes depuis 2 mois. Ils seront en tournée* dans toute la France à la rentrée.</p> <p>*Une partie des bénéfices de la vente des billets est reversée à l'UNICEF afin de permettre aux équipes sur le terrain de poursuivre et de renforcer leurs actions pour protéger les enfants vulnérables.</p> <p>Réservez vos places de concert pour : KIDS UNITED - ZENITH SUD MONTEPELLIER Le prix des places est compris entre : 39.00€ et 99.00€ Date : samedi 29 octobre 2016 Vous disposez par ailleurs du service e-ticket pour imprimer vos billets à domicile dès la fin de commande pour KIDS UNITED ainsi que du plan de salle interactif pour choisir vos places dans le lieu : ZENITH SUD MONTEPELLIER.</p> <p><a href="#">Moins d'infos</a></p>				

# Architecture du service Cogniseach Event

# Modèle de représentation d'événement



## Enchaînement des modules de l'architecture du service






## Problématique : illustration

**INFOCONCERT.COM** Artistes, villes, salles, festivals  MON INFOCONCERT   

### MAIN SQUARE

EDITION 2017  
DU 30 JUIN 2017 AU 02 JUILLET 2017 À ARRAS (62)  
13ème édition  
[LIRE LA PRÉSENTATION →](#)

**ALERTE FESTIVAL**

 **J'aime** Soyez le premier de vos amis à indiquer que vous aimez ça.

Dernière News : FESTIVAL / Le Main Square annonce une nouvelle vague de noms : Jain, La Femme, Kungs et une quinzaine d'autres artistes rejoint Radiohead, Major Lazer, System Of A Down

À l'affiche : RADIOHEAD / JAIN / SYSTEM OF A DOWN (SOAD) / LA FEMME / DIE ANTWOORD / MAJOR LAZER / BIFFY CLYRO

**ticketmaster®**  

MUSIQUE ARTS & SPECTACLES SPORT FAMILLE & LOISIRS RÉGIONS BONS PLANS 

 0.00 €

1 - Réservation 2 - Panier 3 - Coordonnées 4 - Paiement 5 - Confirmation

### MAIN SQUARE FESTIVAL 2017 - PASS 1 JOUR

★★★★★

A partir du **30 juin 2017**

FESTIVAL - FESTIVAL MUSIQUE  
Artiste : MAJOR LAZER , RADIOHEAD , M... +

**LA CITADELLE - QUARTIER DE TURENNE**   
BVD DU GENERAL DE GAULLE  
62000 ARRAS - FRANCE

**billetcollector™**  
Le billet souvenir pour tous les fans  
**DISPONIBLE POUR CET ÉVÈNEMENT**  
\*Prépayé au moment de votre réservation



## 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 :

$$o_1 = \langle p_{11}, p_{12}, \dots, p_{1n} \rangle$$

$$o_2 = \langle p_{21}, p_{22}, \dots, p_{2n} \rangle$$

- ▶ 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

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

Plusieurs approches de pondération :

- ▶ Pondération empirique
- ▶ *Prioritized Agregation*
- ▶ Utilisation des méta-heuristiques (Algorithmes génétiques ...)
- ▶ Exploitation des fonctions logiques

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

## Calcul des poids par étalonnage

- ▶ Elle ré-utilise la combinaison linéaire

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

- ▶ 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  $s_i(p_{1i}, p_{2i})$

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

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

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

- ▶ Les  $w_i$  sont déterminés par descente de gradient

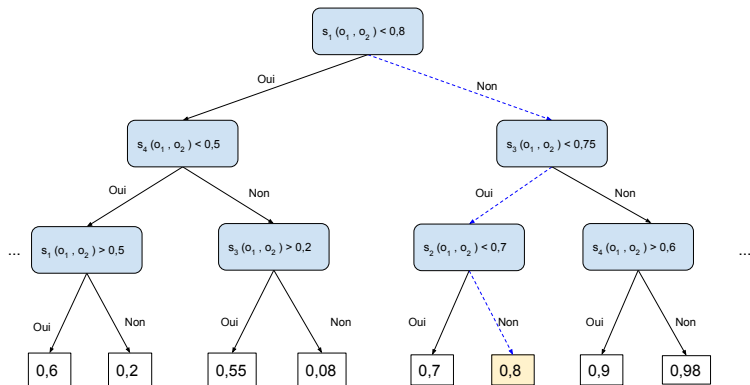
## Régression basée sur les arbres de décision : principe

- ▶ 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

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



## Illustration de la mise en œuvre des arbres



# Application au calcul de similarité entre événements sociaux

## Formalisation

- ▶ Un événement est défini par :

$$e = \langle t, Cat_e, AP_e \rangle$$

- ▶ Une représentation est définie par :

$$r = \langle l, d, h \rangle, r \in R_e$$

- ▶ 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

$$s(o_1, o_2) = \text{somme des } s_i \text{ non nuls} * \text{nombre de } s_i \text{ non nuls}$$

- ▶ 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	<b>89,09</b>	84,48
Rappel	84,48	<b>86,21</b>	84,48	84,48
F <sub>1</sub> -mesure	81,67	82,64	<b>86,73</b>	84,48
Exactitude	78,00	79,00	<b>85,00</b>	82,00

## Résultats & Analyses : Scénario 1

**Scénario 1** : toutes les propriétés renseignées

<b>Approches</b>	<b>CombMNZ</b>	<b>Etalonnage</b>	<b>Logistique</b>	<b>Arbres</b>
<b>Précision</b>	<b>92,86</b>	86,67	100	<b>92,86</b>
<b>Rappel</b>	<b>92,86</b>	<b>92,86</b>	85,71	<b>92,86</b>
<b>F<sub>1</sub>-mesure</b>	<b>92,86</b>	89,66	92,31	<b>92,86</b>
<b>Exactitude</b>	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	<b>86,05</b>	81,82
Rappel	81,82	<b>84,09</b>	<b>84,09</b>	81,82
F <sub>1</sub> -mesure	78,26	80,43	<b>85,06</b>	81,82
Exactitude	74,03	76,62	<b>83,12</b>	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