



**UNIVERSITÉ
DE LYON**



**LABORATOIRE
HUBERT CURIEN**

UMR • CNRS • 5516 • SAINT-ETIENNE

Extension d'un corpus d'articles scientifiques
par recherche de similarités sémantiques :
application à une problématique des sciences du sport

Atelier Fouille de Textes / Text Mine 2017

Grenoble, le 24 janvier 2017

Fabrice MUHLENBACH

Univ Lyon, UJM-Saint-Etienne, CNRS

Laboratoire Hubert Curien UMR 5516

courriel : fabrice.muhlenbach@univ-st-etienne.fr

Plan de la présentation

- Présentation de l'équipe et des soutiens financiers
- Problématique :
de la difficulté d'utiliser les ressources des bibliothèques numériques dans un contexte de recherche pluridisciplinaire
- Fouille de textes et similarité sémantique
- La bibliothèque numérique ISTEX
- Cas d'application : la rotation mentale
- Approche, expérimentations et premiers résultats

• Équipe



Fabrice
MUHLENBACH



UNIVERSITÉ
DE LYON



LABORATOIRE
HUBERT CURIE

UMR • CNRS • 5516 • SAINT-ETIENNE



Hussein
AL-NATSHEH



Lucie
MARTINET



UNIVERSITÉ
LUMIÈRE
LYON 2
UNIVERSITÉ DE LYON



Djamel A.
ZIGHED



Fabien
RICO



Patrick
FARGIER



Raphaël
MASSARELLI



• Soutiens financiers

Ce travail a été réalisé grâce au soutien financier du *Programme Avenir Lyon Saint-Etienne* de l'Université de Lyon dans le cadre du programme « Investissements d'Avenir » (ANR-11-007)



La thèse d'Hussein AL-NATSHEH est soutenue par une allocation doctorale de recherche de la Région Auvergne-Rhône-Alpes

La Région
Auvergne-Rhône-Alpes



COMMUNAUTÉS
DE RECHERCHE
ACADÉMIQUE
Auvergne-Rhône-Alpes



T.I.C. ET USAGES
INFORMATIQUES
INNOVANTS

Le travail de post-doctorante de Lucie MARTINET a été financé dans le cadre des chantiers d'usage d'ISTEX (programme « Investissements d'Avenir » initié par le MESR) : **3ST**

ISTEX
L'excellence documentaire pour tous

Extension d'un corpus d'articles scientifiques par recherche de similarités sémantiques

F. Muhlenbach

Text Mine 2017
Grenoble, 24/01/2017

Plan de la présentation

- Présentation de l'équipe et des soutiens financiers
- **Problématique :**
de la difficulté d'utiliser les ressources des bibliothèques numériques dans un contexte de recherche pluridisciplinaire
- Fouille de textes et similarité sémantique
- La bibliothèque numérique ISTEX
- Cas d'application : la rotation mentale
- Approche, expérimentations et premiers résultats

Problématique

Constat

- accès des chercheurs à des masses d'informations (bibliothèques numériques d'articles scientifiques en ligne)
- exploration des documents scientifiques limitée à la communauté d'appartenance de chaque chercheur

Proposition

- extension de l'exploration bibliographique au-delà de la communauté d'appartenance
- → contexte pluri- et trans-disciplinaire

Problématique

Défis

- grande taille des bibliothèques numériques
- hétérogénéité des données
- complexité du langage naturel
- pour le chercheur : limitations cognitives + manque de temps
 → incapacité à pouvoir embrasser des concepts issus :
 - d'articles anciens pourtant pertinents
 (focalisation sur l'axe diachronique limitée aux articles scientifiques les plus récents)
 - d'articles venant de disciplines complémentaires
 (focalisation sur l'axe synchronique limitée à la communauté scientifique d'appartenance)

Problématique

Conséquence

→ le saut **quantitatif** en masse d'information apportée par les bibliothèques numériques ne se traduit pas vraiment en saut **qualitatif** pour le chercheur qui souhaite exploiter ces documents

Plan de la présentation

- Présentation de l'équipe et des soutiens financiers
- Problématique :
de la difficulté d'utiliser les ressources des bibliothèques numériques dans un contexte de recherche pluridisciplinaire
- Fouille de textes et similarité sémantique
- La bibliothèque numérique ISTEX
- Cas d'application : la rotation mentale
- Approche, expérimentations et premiers résultats

Fouille de textes et similarité sémantique

Approches classiques en fouille de textes

- transformation de textes en format analysable par des techniques de fouille de données
- méthodes quantitatives appliquées à la linguistique
- préparation des textes (lemmatisation, minuscules...)
- construction de la matrice termes-documents
- fréquence du terme, fréquence inverse du document (tf-idf)
- recherche des termes les plus fréquents
- recherche des associations les plus fréquentes
- visualisation par nuage de mots
- classification de textes
- construction de modèles thématiques (*topic modeling*)

Fouille de textes et similarité sémantique

Comment estimer la correspondance sémantique ?

- études de mesures de similarité entre textes
→ base de tests : *Exercices de Style* de Raymond Queneau



- même histoire racontée de 99 façons différentes (« Litotes », « Rétrograde », « Surprises », « Rêve », « Hésitations »)
- textes de l'essai de Queneau mélangés avec un corpus d'histoires courtes non pertinentes (ex. négatifs)
- comparaison de différentes méthodes (mesures de similarité / indices de RI)

Plan de la présentation

- Présentation de l'équipe et des soutiens financiers
- Problématique :
de la difficulté d'utiliser les ressources des bibliothèques numériques dans un contexte de recherche pluridisciplinaire
- Fouille de textes et similarité sémantique
- **La bibliothèque numérique ISTEX**
- Cas d'application : la rotation mentale
- Approche, expérimentations et premiers résultats

La bibliothèque numérique ISTEX



ISTEX

L'excellence documentaire pour tous

Objectifs

- ISTEX = initiative d'excellence en information scientifique et technique
- offrir à l'ensemble de la communauté de l'enseignement supérieur et de la recherche, un accès en ligne aux collections rétrospectives de la littérature scientifique dans toutes les disciplines en engageant une politique nationale d'acquisition massive de documentation

Avantages

- **égalité** territoriale et institutionnelle d'accès à l'information
- **pluridisciplinarité** : ensemble des champs scientifiques
- **complémentarité** : accès unifié à toute la documentation
- **économie** : licences nationales → économies d'échelle

Extension d'un corpus d'articles scientifiques par recherche

de similarités sémantiques

F. Muhlenbach

Text Mine 2017

Grenoble, 24/01/2017

13

Plan de la présentation

- Présentation de l'équipe et des soutiens financiers
- Problématique :
de la difficulté d'utiliser les ressources des bibliothèques numériques dans un contexte de recherche pluridisciplinaire
- Fouille de textes et similarité sémantique
- La bibliothèque numérique ISTEX
- **Cas d'application : la rotation mentale**
- Approche, expérimentations et premiers résultats

Cas d'application : la rotation mentale

Pourquoi les sciences du sport ?

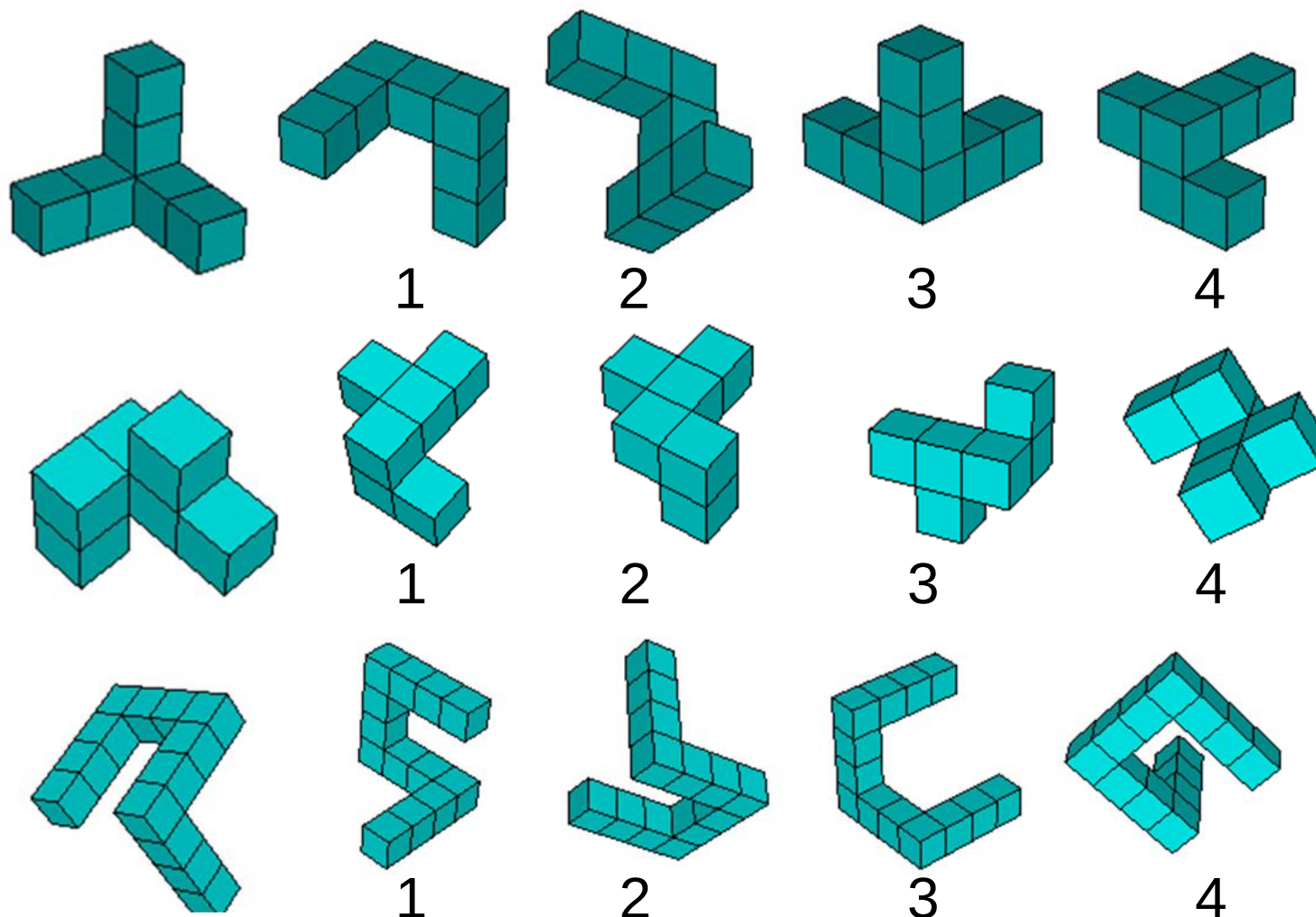
- par définition, les sciences du sport sont l'ensemble des sciences qui ont pour but la connaissance des différents aspects des pratiques sportives
- elles ont pour objet de déterminer des théories concernant la pratique physique (reconnaissance de lois et constantes, et repérer des principes généralisables à un ensemble de phénomènes)
- domaine par nature **pluridisciplinaire** : les pratiques sportives constituent un objet d'étude qui peut être abordé par différentes disciplines telles que la physiologie, la psychologie ou la psycho-sociologie

Cas d'application : la rotation mentale

Qu'est-ce que la rotation mentale ?

- la rotation mentale consiste en la capacité à faire tourner mentalement l'image d'un objet en 2 ou en 3 dimensions
- forme particulière d'imagerie mentale ou d'imagerie motrice nécessitant une structuration de l'espace
- implication des processus moteurs : une bonne représentation mentale requiert la capacité à travailler l'image mentale visuelle et à la faire tourner mentalement
- tâche de rotation mentale classique : indiquer le plus rapidement possible si deux images en 2 ou 3 dimensions, présentées sous différents angles, sont identiques ou différentes

Cas d'application : la rotation mentale



Cas d'application : la rotation mentale

Caractéristiques de la rotation mentale

- opération mentale 1 : rotation d'au moins une des figures dans un des plans de l'espace pour la superposer avec l'autre afin de pouvoir juger de leur similitude ou différence
- opération mentale 2 : s'imaginer se déplacer soi-même en tournant autour de l'objet afin de le visualiser sous un angle différent pour pouvoir effectuer le jugement de similitude
- études : recherche des mécanismes sous-jacents de la rotation mentale et ses liens avec la performance et l'expérience motrice ; évolution des capacités individuelles à la suite d'un entraînement et transfert vers d'autres domaines d'expertise (capacités motrices et intellectuelles)

Cas d'application : la rotation mentale

Recherches sur la rotation mentale

- étude des bases neuro-fonctionnelles de la rotation mentale à l'aide de l'imagerie cérébrale (IRMf, TEP, MEG, EEG)
- liens entre les capacités de traitement d'une image mentale (construction, transformation et manipulation d'une image visuelle) et les processus moteurs permettant de mettre en mouvement et de faire tourner cette image mentale
- la rotation mentale est un phénomène complexe :
 - pas de spécialisations exclusives des aires cérébrales
 - lien avec la réussite à l'école
 - existence de différences garçons/filles sur les performances
 - la pratique du sport d'équipe augmente les performances

Plan de la présentation

- Présentation de l'équipe et des soutiens financiers
- Problématique :
de la difficulté d'utiliser les ressources des bibliothèques numériques dans un contexte de recherche pluridisciplinaire
- Fouille de textes et similarité sémantique
- La bibliothèque numérique ISTEX
- Cas d'application : la rotation mentale
- **Approche, expérimentations et premiers résultats**

Approche, expérimentations, résultats

Approche

- système d'extension du corpus :
 - en entrée : des articles scientifiques portant sur un sujet
 - en sortie : des articles scientifiques associés au sujet

- contraintes :
 - articles issus de disciplines scientifiques différentes
 - articles présentant des « pépites » (la fouille de données, telle que définie par D. J. Hand en 2000, est la découverte de structures **intéressantes**, **inattendues** ou **précieuses** dans les grands ensembles de données)
 - meilleurs résultats que ceux obtenus par l'approche de la RI

Approche, expérimentations, résultats

Processus général

ISTEX bibliothèque
numérique
L'excellence documentaire pour tous

corpus de base

articles sources



portant sur
un sujet donné
(exemples fournis
par l'utilisateur)

Systeme
d'extension
du corpus



articles associés
au sujet



thème 1

...



thème i

Approche, expérimentations, résultats

Fonctionnement du système d'extension du corpus

- représentation sémantique des documents par des vecteurs denses : décomposition en valeurs singulières, analyse sémantique latente
- apprentissage supervisé avec 2 classes (exemples positifs et exemples négatifs)
- prédiction effectuée sur tous les autres articles de la bibliothèque numérique
- tri par « top k » des articles prédits
- regroupement des articles par thème suivant l'allocation de Dirichlet latente (LDA)

Approche, expérimentations, résultats

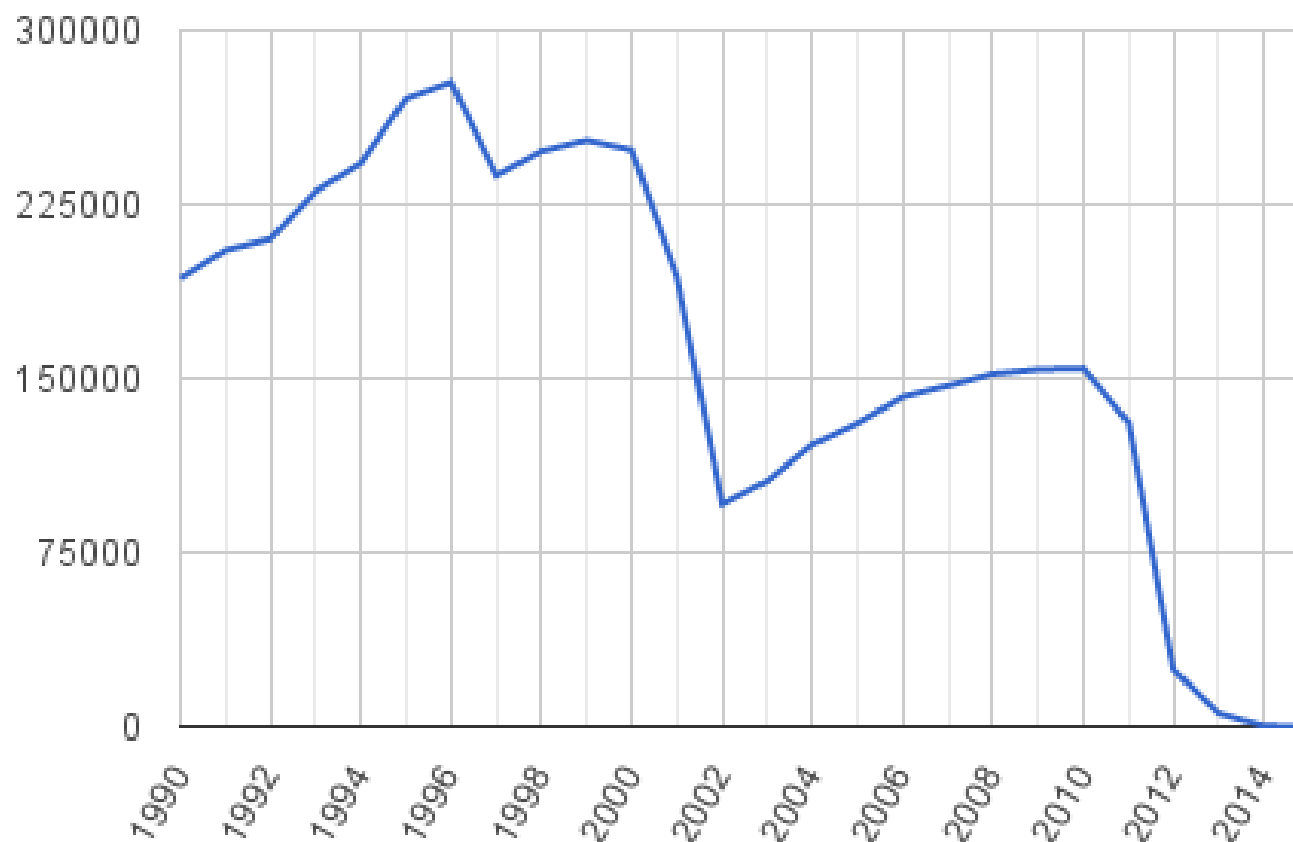
Préparation : sélection des articles

- articles avec méta-données complètes (titre, résumé de 35 à 500 mots, auteurs...), assez récents et en anglais
- articles avec un nombre de pages compris entre 3 et 60
- articles de recherche, de journaux ou d'actes de conférence (pas de table des matières ou d'index, pas de posters...)
- utilisation de l'interface de programmation (API) d'ISTEX :

```
$istex-api-harvester -q "publicationDate:[1990 2016]
AND language:("eng" OR "unknown")
AND pdfPageCount:[3 60] AND abstractWordCount:[35 500]
AND genre:("research_article" OR "conference[eBooks]" OR "article")"
```


Approche, expérimentations, résultats

Articles issus de la bibliothèque numérique (en nb/an)



Approche, expérimentations, résultats

Détail des opérations du processus

- transformation des articles sélectionnés en représentation par sac de mots (matrice creuse)
- extraction des caractéristiques sémantiques descriptives des documents par modèles de vectorisation sémantique (Doc2Vec, décomposition en valeurs singulières, analyse sémantique latente)
- construction d'un modèle d'apprentissage supervisé (ici, classement par forêts aléatoires)
- utilisation du modèle pour retrouver des documents pertinents sans les mots clés du sujet (ici, "mental rotation")
- tri des documents par pertinence, tests par des experts

Approche, expérimentations, résultats

Expérimentations

- sujet : « rotation mentale », un centre d'intérêt de la recherche des sciences du sport, combinant les disciplines :
 - neurosciences (aires cérébrales impliquées)
 - psychologie (compétences cognitives, développement...)
 - physiologie (habiletés motrices)
- comparaison de la méthode proposée avec la méthode "more_like_this" d'*Elasticsearch* (indexation et RI)
- affichage des résultats combinant les retours des deux approches (notre méthode et *Elasticsearch*)
- présentation des résultats aux experts du domaine (vérité terrain permettant de quantifier les résultats)

Approche, expérimentations, résultats

Articles utilisés comme exemples positifs

- articles fournis par les utilisateurs experts du domaine (sciences du sport) sur le thème "mental rotation"
- ajout d'articles issus de la bibliothèque numérique avec les mots clés "mental rotation" :

```
wget https://api.istex.fr/document/?q="mental rotation"
AND publicationDate:[1990 2016]
AND language:("eng" OR "unknown")
AND pdfPageCount:[3 60]
AND abstractWordCount:[35 500]
AND genre:("research-article" OR "conference [eBooks]"
OR "article" )&size=2000
```

Approche, expérimentations, résultats

Résultats préliminaires

- sur les articles proposés par notre méthode, identification de thèmes considérés comme pertinents par les experts lorsque les articles sont regroupés en 2 thèmes :
 - un thème où sont présents les mots "motor", "task", "orientation", "stimuli", donc ayant plutôt trait à la partie expérimentale (psychologie cognitive) de la rotation mentale
 - un thème où apparaissent les mots "spatial ability", "visual", "mental rotation", "performance", "sex/age/profession differences", donc plutôt des éléments ayant trait aux aspects comportementaux ou sociaux de la rotation mentale

Approche, expérimentations, résultats

Résultats préliminaires

- des analyses plus poussées font apparaître des phrases clés dans ces thèmes, phrases qui sont liées :
 - soit aux approches expérimentales (p. ex. les potentiels évoqués ou la stimulation magnétique transcrânienne) ;
 - soit aux phénomènes enregistrés (p. ex. la négativité de discordance) ;
 - soit aux conséquences comportementales (p. ex. le trouble du déficit de l'attention avec hyperactivité) ;
 - soit les aires cérébrales impliquées (p. ex. le lobule lingual ou le cortex périrhinal)...

Approche, expérimentations, résultats

Résultats préliminaires

- résultats de notre approche (1000 articles issus d'ISTEX)
→ sélection des articles scientifiques ne présentant pas l'expression "mental rotation" en titre ou résumé, échantillon de 30 articles (sur 959) présentés à 2 experts
- 5 articles issus du bas du classement (entre la 909 et la 959^{ème} place) ont été considérés comme non pertinents (3/5) ou à la pertinence non évaluable par les experts (2/5)
- sur les 25 articles faisant partie du haut du classement :
 - 9 ont été considérés comme pertinents par les 2 experts
 - 13 ont été considérés comme non pertinents
 - 3 articles n'ont pas pu être évalués (ou désaccord)

Conclusions et perspectives

- travail encore en cours (l'évaluation manuelle de la pertinence des résultats par des experts prend du temps)
- résultats préliminaires encourageants
- découvertes d'articles « surprenants » : articles associés au sujet mais ne comportant pas l'expression "mental rotation" dans le texte, par exemple lien entre des tâches de poursuites oculaires (qui sont un ensemble de tâches d'imagerie motrice), l'attention et la schizophrénie
- perspectives : améliorations possibles de la méthode
 - test d'autres représentations sémantiques des documents par vecteurs denses (par exemple, par plongement lexical)
 - autres méthodes d'apprentissage supervisé (p. ex., SVM)
 - application à d'autres sujets de recherche pluridisciplinaire



**UNIVERSITÉ
DE LYON**



**LABORATOIRE
HUBERT CURIEN**

UMR • CNRS • 5516 • SAINT-ETIENNE

Extension d'un corpus d'articles scientifiques
par recherche de similarités sémantiques :
application à une problématique des sciences du sport

Atelier Fouille de Textes / Text Mine 2017

Grenoble, le 24 janvier 2017

Fabrice MUHLENBACH

Univ Lyon, UJM-Saint-Etienne, CNRS

Laboratoire Hubert Curien UMR 5516

courriel : fabrice.muhlenbach@univ-st-etienne.fr