AVANT-PROJET : CITATION FOCUSER (CITEFOX)

- But : mettre en place des méthodes de ciblage de citation entre documents citant et documents cités,
- Mettre en place des techniques à la frontière de l'état de l'art dans le domaine du résumé de communauté:
  - Résumé du contenu des documents basé sur la compétition de blocs

  ⇒ Théorie de la maximisation des traits,
  - Mesures de mise en correspondance textuelles

  ⇒ Etat de l'art,
  - Expansion de requêtes

  ⇒ Méthodes de propagation d'activation,
  - Extraction des citations

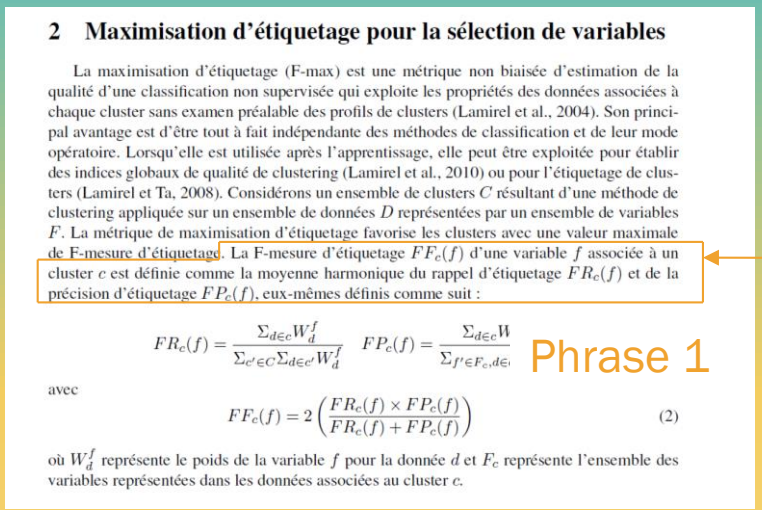
  ⇒ Méthodes d'extraction d'entités nommées ? (à voir).

- Travailler à titre d'exemple sur des données-test issues d'un challenge international (CL-SCISumm 2016).

Intervenants : Jean-Charles LAMIREL (Synalp-LORIA), Hazem AL ZIED (ATILF), Nicolas DUGUE (Université du Mans).

EXTRACTION ET CIBLAGE DES CITATIONS

EGC 2014, Lamirel et al.



2 Maximisation d'étiquetage pour la sélection de variables

La maximisation d'étiquetage (F-max) est une métrique non biaisée d'estimation de la qualité d'une classification non supervisée qui exploite les propriétés des données associées à chaque cluster sans examen préalable des profils de clusters (Lamirel et al., 2004). Son principal avantage est d'être tout à fait indépendante des méthodes de classification et de leur mode opératoire. Lorsqu'elle est utilisée après l'apprentissage, elle peut être exploitée pour établir des indices globaux de qualité de clustering (Lamirel et al., 2010) ou pour l'étiquetage de clusters (Lamirel et Ta, 2008). Considérons un ensemble de clusters C résultant d'une méthode de clustering appliquée sur un ensemble de données D représentées par un ensemble de variables F. La métrique de maximisation d'étiquetage favorise les clusters avec une valeur maximale de F-mesure d'étiquetage. La F-mesure d'étiquetage $FF_c(f)$ d'une variable f associée à un cluster c est définie comme la moyenne harmonique du rappel d'étiquetage $FR_c(f)$ et de la précision d'étiquetage $FP_c(f)$, eux-mêmes définis comme suit :

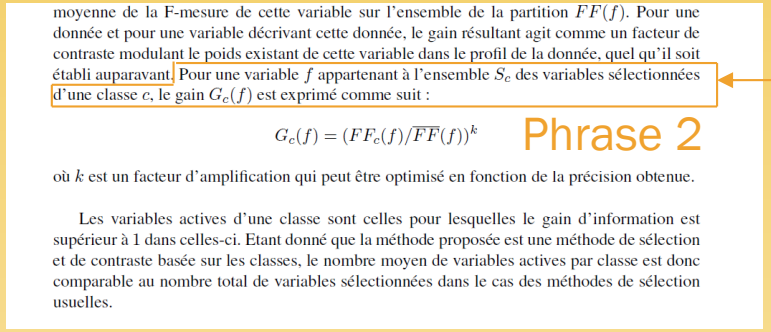
$$FR_c(f) = \frac{\Sigma_{d \in c} W_d^f}{\Sigma_{c' \in C} \Sigma_{d \in c'} W_d^f} \qquad FP_c(f) = \frac{\Sigma_{d \in c} W}{\Sigma_{f' \in F_c, d \in}}$$

Phrase 1

avec

$$FF_c(f) = 2\left(\frac{FR_c(f) \times FP_c(f)}{FR_c(f) + FP_c(f)}\right) \quad (2)$$

où W_d^f représente le poids de la variable f pour la donnée d et F_c représente l'ensemble des variables représentées dans les données associées au cluster c.



moyenne de la F-mesure de cette variable sur l'ensemble de la partition $FF(f)$. Pour une donnée et pour une variable décrivant cette donnée, le gain résultant agit comme un facteur de contraste modulant le poids existant de cette variable dans le profil de la donnée, quel qu'il soit établi auparavant. Pour une variable f appartenant à l'ensemble S_c des variables sélectionnées d'une classe c, le gain $G_c(f)$ est exprimé comme suit :

$$G_c(f) = (FF_c(f)/\overline{FF}(f))^k$$

Phrase 2

où k est un facteur d'amplification qui peut être optimisé en fonction de la précision obtenue.

Les variables actives d'une classe sont celles pour lesquelles le gain d'information est supérieur à 1 dans celles-ci. Etant donné que la méthode proposée est une méthode de sélection et de contraste basée sur les classes, le nombre moyen de variables actives par classe est donc comparable au nombre total de variables sélectionnées dans le cas des méthodes de sélection usuelles.

Les cahiers du numérique 2016, Dugué et al.

Ciblage

Rappel, Précision, F-mesure

Extraction terminologique

…. La F-mesure de traits combine les indices de Rappel et de Précision (Lamirel et al., 2014) ….

Ciblage

Contraste, Gain, Information, Classe

Extraction terminologique

…. Le gain d'information, ou contraste, exprime la capacite du traits a caractériser une classe (Lamirel et al., 2014) ….

Résumé de communauté

Phrase 1. Phrase 2.

CAS DU CHALLENGE CL-SCISUMM

- Corpus de 10 articles de référence dans le domaine biomédical,
- Un ensemble de plus de 10 articles citant accompagnés du contexte des citations est associé avec chaque article cité,
- Un Gold (citation-meilleur phrase/paragraphe du document cité) est construit par expertise.
- Extraction terminologique et résumé du document citant par maximisation d'étiquetage,
- Extraction terminologique des termes dans les phrases support des citations,
- (Propagation d'activation pour enrichir les termes des citations),
- Analyse de la densité de contraste générée par les termes des citations dans les blocs du document cité,
- Ciblage des meilleures paragraphes/phrases par mesure de similarité,
- Adjonction des phrases sélectionnées au résumé de communauté,
- Mesure de validité des résultats en exploitant des méthodes de la classe ROUGE.

Une expérimentation préalable a été menée sur du texte non structuré.

ETAT D'AVANCEMENT

- Nous avons présenté une méthodologie de trouver les points d'impacts des citations de documents citant dans les documents cités,
- Cette méthodologie a été testé avec succès dans le cadre du challenge CL-SCISumm 2016,
- Le système obtenu s'est avéré être le plus efficace parmi les 17 systèmes proposés dans le challenge (avec une très forte différence en terme de rappel),
- Le système ne nécessite pas de source de connaissance externe pour apprendre (contrairement aux systèmes concurrents) et possède des capacités naturelles d'élimination de la redondance,
- La méthodologie et les test ont été publiés dans un journal international (IJDL 2017),
- Un problème important dans le cadre ISTEX est celui du repérage cohérent des citations et d'extraction du contexte,
- Cette proposition de projet reste à financer (le financement de l'étude préalable a été opéré dans le cadre du CPER LCHN).

AVANT-PROJET : NOUVEAUX PARADIGMES SCIENTIFIQUES

- But : explorer un corpus de données scientifiques en mesurant les changements de sujets incluant la récurrence de sujets (alternance de citations et d'oublis,
- Travailler sur un corpus de données issues de la base mutti-éditeurs ISTEX gérée par l'INIST,
- Mettre en place des techniques à la frontière de l'état de l'art :
  - Distances de compromis entre la généralité et la discrimination

  ⇒ Théorie de la maximisation des traits,
  - Travailler avec des vues multiples et des mécanismes de généralisation en ligne

  ⇒ Paradigme MVDA,
  - Intégrer la visualisation

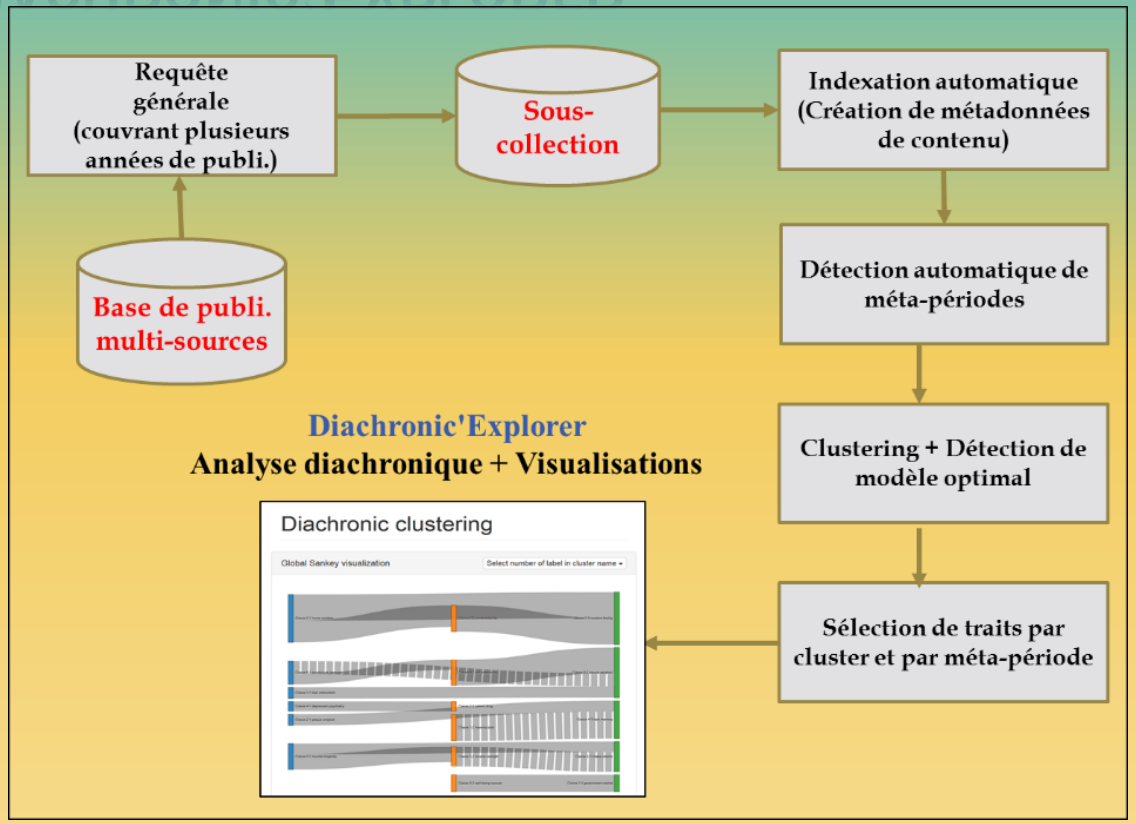
  => Approche Diachronic'Explorer,
  - Intégrer les informations produites par les entités nommées dans le processus (en cours),
- Travailler à titre d'exemple sur des données du domaine de l'astronomie (en cours).

Intervenants : Jean-Charles LAMIREL (Synalp-LORIA), Denis MAUREL (Université de Tours), Anubhav GUPTA (DIST-CNRS & Université de Tours).

Analyse diachronique et navigation dans les données multisources

ISTEX-R - WP1 & Diachronic'Explorer

La figure présente le déroulement de l'approche Diachronic'Explorer complète jusqu'à la visualisation



La méthode ne présente pas les inconvénients des méthodes d'extraction de sujets usuelles, comme LDA (Blei et al. 2003) : sujets imprécis et dépendants du processus d'optimisation utilisé, non applicabilité à l'échelle des documents, ...

L'indexation automatique peut être remplacée par le processus d'extraction de métadonnées basé sur la maximisation des traits.

Cas des données astromoniques

- Corpus d'env. 500000 articles sur le thème général de l'astronomie de issues de la base ISTEX,
- Période couvrant 189 ans,
- Les données sont étiquetées par les entités nommées, noms de lieu, nom de personnes, dates
  => Unitex/CacSys [Maurel et al. 2016].
- Identification des articles les plus cités,
- Analyse du contenu par extraction automatique des métadonnées et isolement de sujet centraux (ex: big bang, théorie des cordes ….),
- Extraction du contexte des citations (phrases dans lesquelles les citations apparaissent) [Al Zied et al. 2017],
- Mesure du cumul de contraste/période généré sur les sujets centraux par le contexte des citations,
- Visualisation des variabilités temporelles,
- Mise en parallèle avec une analyse directe basée sur le clustering et sur MVDA.

Cas des données astromoniques

Un exemple d'annotations dans un article :

18CBC6B00F42A58322ECB8DA287F002C5D828ACD - Notepad

File Edit Format View Help

```
                <persName change="#Unitex-3.2.0-alpha" resp="istex-rd" scheme="http://persName-entity.lod.istex.fr
                        <term>Martin Heidegger</term>
                        <fs type="statistics">
                                <f name="frequency">
                                        <numeric>1</numeric>
                                </f>
                        </fs>
                </persName>
        </annotationBlock>
        <annotationBlock corresp="text" xmls="https://wwww.tei-c.org/ns/1.0">
                <persName change="#Unitex-3.2.0-alpha" resp="istex-rd" scheme="http://persName-entity.lod.istex.fr
                        <term>Marcus Hellyer</term>
                        <fs type="statistics">
                                <f name="frequency">
                                        <numeric>2</numeric>
                                </f>
                        </fs>
                </persName>
        </annotationBlock>
        <annotationBlock corresp="text" xmls="https://wwww.tei-c.org/ns/1.0">
                <persName change="#Unitex-3.2.0-alpha" resp="istex-rd" scheme="http://persName-entity.lod.istex.fr
                        <term>Francisco Suárez</term>
                        <fs type="statistics">
                                <f name="frequency">
                                        <numeric>9</numeric>
                                </f>
                        </fs>
                </persName>
        </annotationBlock>
        <annotationBlock corresp="text" xmls="https://wwww.tei-c.org/ns/1.0">
                <persName change="#Unitex-3.2.0-alpha" resp="istex-rd" scheme="http://persName-entity.lod.istex.fr
                        <term>Raymond Bullman</term>
```

- Nous avons présenté une méthodologie permettant d'analyser les données de manière diachronique à partir des données étiquetées par des entités nommées,
- Le principe de cette méthodologie a été présenté à la conférence ACFAS 2017 (Montréal),
- Le but est d'analyser les effets de récurrence liés aux nouveaux paradigmes scientifiques,
- Le principe général de l'approche repose sur des techniques récemment expérimentées avec succès,
- Les données étiquetées restent cependant en cours de traitement,
- Un problème important est celui du repérage cohérent des citations dont la syntaxe varie en fonction des périodes de temps,
- Le repérage du contexte des citations reste également à traiter,
- Cette proposition de projet reste à financer.