

Fouille de textes et quelques applications

Solen Quiniou

`solen.quiniou@univ-nantes.fr`

Meetup Machine Learning

Lundi 2 octobre 2017



Plan de la présentation

- 1 Introduction
- 2 Traitements de données textuelles
- 3 Fouille de motifs séquentiels
 - Extraction de motifs séquentiels sous contraintes
 - Extraction de motifs séquentiels émergents
 - Application : analyse stylistique de textes
 - Application : analyse de publications biomédicales
- 4 Fouille de graphes pour l'exploration de grands textes
 - Contexte
 - Modèle linguistique de Hoey
 - Fouille de graphes enrichis
- 5 Conclusion
- 6 Références

Plan du cours

- 1 Introduction
- 2 Traitements de données textuelles
- 3 Fouille de motifs séquentiels
- 4 Fouille de graphes pour l'exploration de grands textes
- 5 Conclusion
- 6 Références

Introduction (1)

- Fouille de données (*Data Mining*)
 - ▶ Extraction d'informations, découverte de connaissances à partir de gros volumes de données, mise en évidence de règles potentiellement invisibles pour un analyste humain
- Différents types de données
 - ▶ Données nominales
 - ★ Nombre de cas dénombrable, pas de relation d'ordre entre les valeurs
 - ▶ Données ordinales
 - ★ Nombre de cas dénombrable, relation d'ordre entre les valeurs
 - ▶ Données numériques ou continues
 - ★ Nombre de cas théoriquement infini, relation d'ordre entre les valeurs
- Différentes techniques pour différents objectifs
 - ▶ Description : trouver un résumé des données plus intelligible
→ *Statistique descriptive, analyse factorielle*
 - ▶ Structuration : identifier des groupes représentant des entités particulières
→ *Classification dont clustering, apprentissage non supervisé*
 - ▶ Explication : prédire les valeurs d'un attribut à partir d'autres attributs
→ *Régression, apprentissage supervisé*
 - ▶ Association : trouver les ensembles de descripteurs les plus corrélés
→ *Règles d'association, motifs*

Introduction (1)

- Fouille de données (*Data Mining*)
 - ▶ Extraction d'informations, découverte de connaissances à partir de gros volumes de données, mise en évidence de règles potentiellement invisibles pour un analyste humain
- Différents types de données
 - ▶ Données nominales
 - ★ Nombre de cas dénombrable, pas de relation d'ordre entre les valeurs
 - ▶ Données ordinales
 - ★ Nombre de cas dénombrable, relation d'ordre entre les valeurs
 - ▶ Données numériques ou continues
 - ★ Nombre de cas théoriquement infini, relation d'ordre entre les valeurs
- Différentes techniques pour différents objectifs
 - ▶ Description : trouver un résumé des données plus intelligible
→ *Statistique descriptive, analyse factorielle*
 - ▶ Structuration : identifier des groupes représentant des entités particulières
→ *Classification dont clustering, apprentissage non supervisé*
 - ▶ Explication : prédire les valeurs d'un attribut à partir d'autres attributs
→ *Régression, apprentissage supervisé*
 - ▶ Association : trouver les ensembles de descripteurs les plus corrélés
→ *Règles d'association, motifs*

Introduction (1)

- Fouille de données (*Data Mining*)
 - ▶ Extraction d'informations, découverte de connaissances à partir de gros volumes de données, mise en évidence de règles potentiellement invisibles pour un analyste humain
- Différents types de données
 - ▶ Données nominales
 - ★ Nombre de cas dénombrable, pas de relation d'ordre entre les valeurs
 - ▶ Données ordinales
 - ★ Nombre de cas dénombrable, relation d'ordre entre les valeurs
 - ▶ Données numériques ou continues
 - ★ Nombre de cas théoriquement infini, relation d'ordre entre les valeurs
- Différentes techniques pour différents objectifs
 - ▶ Description : trouver un résumé des données plus intelligible
→ *Statistique descriptive, analyse factorielle*
 - ▶ Structuration : identifier des groupes représentant des entités particulières
→ *Classification dont clustering, apprentissage non supervisé*
 - ▶ Explication : prédire les valeurs d'un attribut à partir d'autres attributs
→ *Régression, apprentissage supervisé*
 - ▶ Association : trouver les ensembles de descripteurs les plus corrélés
→ *Règles d'association, motifs*

Introduction (1)

- Fouille de données (*Data Mining*)
 - ▶ Extraction d'informations, découverte de connaissances à partir de gros volumes de données, mise en évidence de règles potentiellement invisibles pour un analyste humain
- Différents types de données
 - ▶ Données nominales
 - ★ Nombre de cas dénombrable, pas de relation d'ordre entre les valeurs
 - ▶ Données ordinales
 - ★ Nombre de cas dénombrable, relation d'ordre entre les valeurs
 - ▶ Données numériques ou continues
 - ★ Nombre de cas théoriquement infini, relation d'ordre entre les valeurs
- Différentes techniques pour différents objectifs
 - ▶ Description : trouver un résumé des données plus intelligible
→ *Statistique descriptive, analyse factorielle*
 - ▶ Structuration : identifier des groupes représentant des entités particulières
→ *Classification dont clustering, apprentissage non supervisé*
 - ▶ Explication : prédire les valeurs d'un attribut à partir d'autres attributs
→ *Régression, apprentissage supervisé*
 - ▶ Association : trouver les ensembles de descripteurs les plus corrélés
→ *Règles d'association, motifs*

Introduction (1)

- Fouille de données (*Data Mining*)
 - ▶ Extraction d'informations, découverte de connaissances à partir de gros volumes de données, mise en évidence de règles potentiellement invisibles pour un analyste humain
- Différents types de données
 - ▶ Données nominales
 - ★ Nombre de cas dénombrable, pas de relation d'ordre entre les valeurs
 - ▶ Données ordinales
 - ★ Nombre de cas dénombrable, relation d'ordre entre les valeurs
 - ▶ Données numériques ou continues
 - ★ Nombre de cas théoriquement infini, relation d'ordre entre les valeurs
- Différentes techniques pour différents objectifs
 - ▶ Description : trouver un résumé des données plus intelligible
→ *Statistique descriptive, analyse factorielle*
 - ▶ Structuration : identifier des groupes représentant des entités particulières
→ *Classification dont clustering, apprentissage non supervisé*
 - ▶ Explication : prédire les valeurs d'un attribut à partir d'autres attributs
→ *Régression, apprentissage supervisé*
 - ▶ Association : trouver les ensembles de descripteurs les plus corrélés
→ *Règles d'association, motifs*

Introduction (1)

- Fouille de données (*Data Mining*)
 - ▶ Extraction d'informations, découverte de connaissances à partir de gros volumes de données, mise en évidence de règles potentiellement invisibles pour un analyste humain
- Différents types de données
 - ▶ Données nominales
 - ★ Nombre de cas dénombrable, pas de relation d'ordre entre les valeurs
 - ▶ Données ordinales
 - ★ Nombre de cas dénombrable, relation d'ordre entre les valeurs
 - ▶ Données numériques ou continues
 - ★ Nombre de cas théoriquement infini, relation d'ordre entre les valeurs
- Différentes techniques pour différents objectifs
 - ▶ Description : trouver un résumé des données plus intelligible
→ *Statistique descriptive, analyse factorielle*
 - ▶ Structuration : identifier des groupes représentant des entités particulières
→ *Classification dont clustering, apprentissage non supervisé*
 - ▶ Explication : prédire les valeurs d'un attribut à partir d'autres attributs
→ *Régression, apprentissage supervisé*
 - ▶ Association : trouver les ensembles de descripteurs les plus corrélés
→ *Règles d'association, motifs*

Introduction (1)

- Fouille de données (*Data Mining*)
 - ▶ Extraction d'informations, découverte de connaissances à partir de gros volumes de données, mise en évidence de règles potentiellement invisibles pour un analyste humain
- Différents types de données
 - ▶ Données nominales
 - ★ Nombre de cas dénombrable, pas de relation d'ordre entre les valeurs
 - ▶ Données ordinales
 - ★ Nombre de cas dénombrable, relation d'ordre entre les valeurs
 - ▶ Données numériques ou continues
 - ★ Nombre de cas théoriquement infini, relation d'ordre entre les valeurs
- Différentes techniques pour différents objectifs
 - ▶ Description : trouver un résumé des données plus intelligible
 - *Statistique descriptive, analyse factorielle*
 - ▶ Structuration : identifier des groupes représentant des entités particulières
 - *Classification dont clustering, apprentissage non supervisé*
 - ▶ Explication : prédire les valeurs d'un attribut à partir d'autres attributs
 - *Régression, apprentissage supervisé*
 - ▶ Association : trouver les ensembles de descripteurs les plus corrélés
 - *Règles d'association, motifs*

Introduction (1)

- Fouille de données (*Data Mining*)
 - ▶ Extraction d'informations, découverte de connaissances à partir de gros volumes de données, mise en évidence de règles potentiellement invisibles pour un analyste humain
- Différents types de données
 - ▶ Données nominales
 - ★ Nombre de cas dénombrable, pas de relation d'ordre entre les valeurs
 - ▶ Données ordinales
 - ★ Nombre de cas dénombrable, relation d'ordre entre les valeurs
 - ▶ Données numériques ou continues
 - ★ Nombre de cas théoriquement infini, relation d'ordre entre les valeurs
- Différentes techniques pour différents objectifs
 - ▶ Description : trouver un résumé des données plus intelligible
 - *Statistique descriptive, analyse factorielle*
 - ▶ Structuration : identifier des groupes représentant des entités particulières
 - *Classification dont clustering, apprentissage non supervisé*
 - ▶ Explication : prédire les valeurs d'un attribut à partir d'autres attributs
 - *Régression, apprentissage supervisé*
 - ▶ Association : trouver les ensembles de descripteurs les plus corrélés
 - *Règles d'association, motifs*

Introduction (2)

- Fouille de textes (*Text Mining*)

- ▶ Ensemble des techniques et méthodes destinées au traitement automatique de données textuelles en langage naturel
- ▶ Acquisition de connaissances à partir de corpus textuels (grandes collections de textes ayant des caractéristiques communes), après pré-traitement des textes

- Exemples de corpus

- ▶ Revues de presse, dépêches AFP, articles scientifiques
- ▶ Transcriptions d'entretiens téléphoniques
- ▶ CV, lettres de motivation, lettres de réclamation
- ▶ Courriers électroniques, tweets, articles de blogs
- ▶ ...

- Exemples d'applications

- ▶ Classification de courriers électroniques : détecter automatiquement les *spam*
- ▶ Résumé automatique : sélectionner les phrases représentatives d'un texte voire les reformuler
- ▶ Fouille d'opinion : évaluer si un texte contient un avis positif ou négatif
- ▶ Détection de thème : trouver le(s) thème(s) abordé(s) dans un texte
- ▶ ...

Introduction (2)

- Fouille de textes (*Text Mining*)

- ▶ Ensemble des techniques et méthodes destinées au traitement automatique de données textuelles en langage naturel
- ▶ Acquisition de connaissances à partir de corpus textuels (grandes collections de textes ayant des caractéristiques communes), après pré-traitement des textes

- Exemples de corpus

- ▶ Revues de presse, dépêches AFP, articles scientifiques
- ▶ Transcriptions d'entretiens téléphoniques
- ▶ CV, lettres de motivation, lettres de réclamation
- ▶ Courriers électroniques, tweets, articles de blogs
- ▶ ...

- Exemples d'applications

- ▶ Classification de courriers électroniques : détecter automatiquement les *spam*
- ▶ Résumé automatique : sélectionner les phrases représentatives d'un texte voire les reformuler
- ▶ Fouille d'opinion : évaluer si un texte contient un avis positif ou négatif
- ▶ Détection de thème : trouver le(s) thème(s) abordé(s) dans un texte
- ▶ ...

Introduction (2)

- Fouille de textes (*Text Mining*)

- ▶ Ensemble des techniques et méthodes destinées au traitement automatique de données textuelles en langage naturel
- ▶ Acquisition de connaissances à partir de corpus textuels (grandes collections de textes ayant des caractéristiques communes), après pré-traitement des textes

- Exemples de corpus

- ▶ Revues de presse, dépêches AFP, articles scientifiques
- ▶ Transcriptions d'entretiens téléphoniques
- ▶ CV, lettres de motivation, lettres de réclamation
- ▶ Courriers électroniques, tweets, articles de blogs
- ▶ ...

- Exemples d'applications

- ▶ Classification de courriers électroniques : détecter automatiquement les *spam*
- ▶ Résumé automatique : sélectionner les phrases représentatives d'un texte voire les reformuler
- ▶ Fouille d'opinion : évaluer si un texte contient un avis positif ou négatif
- ▶ Détection de thème : trouver le(s) thème(s) abordé(s) dans un texte
- ▶ ...

Introduction (2)

- Fouille de textes (*Text Mining*)

- ▶ Ensemble des techniques et méthodes destinées au traitement automatique de données textuelles en langage naturel
- ▶ Acquisition de connaissances à partir de corpus textuels (grandes collections de textes ayant des caractéristiques communes), après pré-traitement des textes

- Exemples de corpus

- ▶ Revues de presse, dépêches AFP, articles scientifiques
- ▶ Transcriptions d'entretiens téléphoniques
- ▶ CV, lettres de motivation, lettres de réclamation
- ▶ Courriers électroniques, tweets, articles de blogs
- ▶ ...

- Exemples d'applications

- ▶ Classification de courriers électroniques : détecter automatiquement les *spam*
- ▶ Résumé automatique : sélectionner les phrases représentatives d'un texte voire les reformuler
- ▶ Fouille d'opinion : évaluer si un texte contient un avis positif ou négatif
- ▶ Détection de thème : trouver le(s) thème(s) abordé(s) dans un texte
- ▶ ...

Introduction (2)

- Fouille de textes (*Text Mining*)

- ▶ Ensemble des techniques et méthodes destinées au traitement automatique de données textuelles en langage naturel
- ▶ Acquisition de connaissances à partir de corpus textuels (grandes collections de textes ayant des caractéristiques communes), après pré-traitement des textes

- Exemples de corpus

- ▶ Revues de presse, dépêches AFP, articles scientifiques
- ▶ Transcriptions d'entretiens téléphoniques
- ▶ CV, lettres de motivation, lettres de réclamation
- ▶ Courriers électroniques, tweets, articles de blogs
- ▶ ...

- Exemples d'applications

- ▶ Classification de courriers électroniques : détecter automatiquement les *spam*
- ▶ Résumé automatique : sélectionner les phrases représentatives d'un texte voire les reformuler
- ▶ Fouille d'opinion : évaluer si un texte contient un avis positif ou négatif
- ▶ Détection de thème : trouver le(s) thème(s) abordé(s) dans un texte
- ▶ ...

Introduction (2)

- Fouille de textes (*Text Mining*)

- ▶ Ensemble des techniques et méthodes destinées au traitement automatique de données textuelles en langage naturel
- ▶ Acquisition de connaissances à partir de corpus textuels (grandes collections de textes ayant des caractéristiques communes), après pré-traitement des textes

- Exemples de corpus

- ▶ Revues de presse, dépêches AFP, articles scientifiques
- ▶ Transcriptions d'entretiens téléphoniques
- ▶ CV, lettres de motivation, lettres de réclamation
- ▶ Courriers électroniques, tweets, articles de blogs
- ▶ ...

- Exemples d'applications

- ▶ Classification de courriers électroniques : détecter automatiquement les *spam*
- ▶ Résumé automatique : sélectionner les phrases représentatives d'un texte voire les reformuler
- ▶ Fouille d'opinion : évaluer si un texte contient un avis positif ou négatif
- ▶ Détection de thème : trouver le(s) thème(s) abordé(s) dans un texte
- ▶ ...

Introduction (2)

- Fouille de textes (*Text Mining*)

- ▶ Ensemble des techniques et méthodes destinées au traitement automatique de **données textuelles en langage naturel**
- ▶ Acquisition de connaissances à partir de **corpus textuels** (grandes collections de textes ayant des caractéristiques communes), **après pré-traitement des textes**

- Exemples de corpus

- ▶ Revues de presse, dépêches AFP, articles scientifiques
- ▶ Transcriptions d'entretiens téléphoniques
- ▶ CV, lettres de motivation, lettres de réclamation
- ▶ Courriers électroniques, tweets, articles de blogs
- ▶ ...

- Exemples d'applications

- ▶ **Classification de courriers électroniques** : détecter automatiquement les *spam*
- ▶ **Résumé automatique** : sélectionner les phrases représentatives d'un texte voire les reformuler
- ▶ **Fouille d'opinion** : évaluer si un texte contient un avis positif ou négatif
- ▶ **Détection de thème** : trouver le(s) thème(s) abordé(s) dans un texte
- ▶ ...

Introduction (2)

- Fouille de textes (*Text Mining*)

- ▶ Ensemble des techniques et méthodes destinées au traitement automatique de données textuelles en langage naturel
- ▶ Acquisition de connaissances à partir de corpus textuels (grandes collections de textes ayant des caractéristiques communes), après pré-traitement des textes

- Exemples de corpus

- ▶ Revues de presse, dépêches AFP, articles scientifiques
- ▶ Transcriptions d'entretiens téléphoniques
- ▶ CV, lettres de motivation, lettres de réclamation
- ▶ Courriers électroniques, tweets, articles de blogs
- ▶ ...

- Exemples d'applications

- ▶ Classification de courriers électroniques : détecter automatiquement les *spam*
- ▶ Résumé automatique : sélectionner les phrases représentatives d'un texte voire les reformuler
- ▶ Fouille d'opinion : évaluer si un texte contient un avis positif ou négatif
- ▶ Détection de thème : trouver le(s) thème(s) abordé(s) dans un texte
- ▶ ...

Introduction (2)

- Fouille de textes (*Text Mining*)

- ▶ Ensemble des techniques et méthodes destinées au traitement automatique de **données textuelles en langage naturel**
- ▶ Acquisition de connaissances à partir de **corpus textuels** (grandes collections de textes ayant des caractéristiques communes), **après pré-traitement des textes**

- Exemples de corpus

- ▶ Revues de presse, dépêches AFP, articles scientifiques
- ▶ Transcriptions d'entretiens téléphoniques
- ▶ CV, lettres de motivation, lettres de réclamation
- ▶ Courriers électroniques, tweets, articles de blogs
- ▶ ...

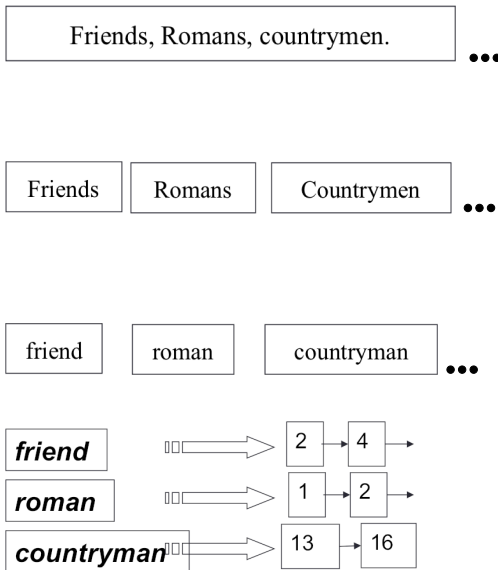
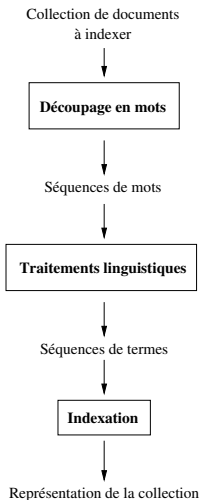
- Exemples d'applications

- ▶ **Classification de courriers électroniques** : détecter automatiquement les *spam*
- ▶ **Résumé automatique** : sélectionner les phrases représentatives d'un texte voire les reformuler
- ▶ **Fouille d'opinion** : évaluer si un texte contient un avis positif ou négatif
- ▶ **Détection de thème** : trouver le(s) thème(s) abordé(s) dans un texte
- ▶ ...

Plan du cours

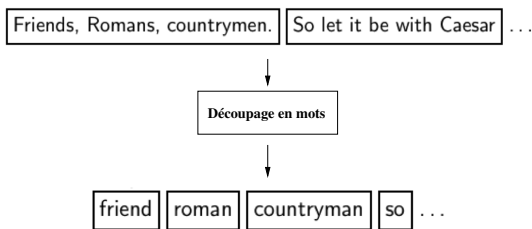
- 1 Introduction
- 2 Traitements de données textuelles**
- 3 Fouille de motifs séquentiels
- 4 Fouille de graphes pour l'exploration de grands textes
- 5 Conclusion
- 6 Références

Schéma du traitement de données textuelles



Découpage en mots (*tokenization*)

- Le **découpage en mots** (*tokenization*) consiste à découper une séquence de caractères en des unités appelées mots (*tokens*).



- On s'appuie généralement sur les espaces et la ponctuation (la ponctuation est également supprimée).
- Chaque mot peut ensuite subir d'autres traitements linguistiques.

Découpage du texte en mots – problèmes

- Il n'est pas toujours facile de savoir où couper les mots.
 - ▶ Problème des apostrophes
 - ★ *l'ensemble, aujourd'hui* : un ou deux mots ?
 - ▶ Problème des tirets
 - ★ *state-of-the-art, compte-rendu* : combien de mots ?
 - ▶ Problème des espaces
 - ★ *San Francisco* : un ou deux mots ?
 - ▶ Problème des nombres
 - ★ *127.0.0.1, 15/09/2013*
- Selon les langages, le découpage en mots ainsi que les autres pré-traitements peuvent être encore plus difficiles.
 - ▶ Problème des langues agglutinantes (par exemple, l'allemand), des alphabets non latin, des langues avec différents sens de lecture. . .

Normalisation

- Après le découpage des mots, il faut normaliser les mots des documents. Il y a en effet des cas où les mots ne sont pas exactement les mêmes mais on veut quand même les mettre en correspondance.
 - ▶ Par exemple, *U.S.A.*, *USA* et *U.S.*
- Pour cela, nous définissons des **classes d'équivalence**.
 - ▶ *U.S.A.*, *USA* et *U.S.* appartiendront à la même classe d'équivalence.
- Il existe différentes techniques pour définir ces classes d'équivalence.

Accents, signes diacritiques et casse

- Accents et signes diacritiques

- ▶ On peut les supprimer dans la plupart des cas.

- ★ Peu d'impact en anglais : *cliché* et *cliche*.
- ★ Plus d'impact en espagnol, par exemple : *peña* et *pena*.
- ★ Substitution de lettres en allemand (traitement du umlaut) : *Universität* → *Universitaet*.

- Traitement de la casse

- ▶ On met généralement tous les mots en minuscule.
- Mais perte d'information quand on cherche les entités nommées d'un texte (c'est-à-dire les noms de personnes, lieux ou organisations)

Lemmatisation

- La **lemmatisation** réduit les formes fléchies des mots à leur forme de base.
 - ▶ *suis, es, est, sommes, êtes, sont* → *être*.
 - « *nous sommes arrivés* » devient ainsi « *nous être arriver* ».
- Pour effectuer la lemmatisation, il faut faire une réduction appropriée à la forme de base du dictionnaire. Cela nécessite un **étiqueteur morpho-syntaxique** qui dépend de la langue.
 - ▶ *je souris* → *je sourire*.
 - ▶ *la souris* → *le souris*.

Racinisation (*stemming*)

- La **racinisation** (*stemming*) réduit les mots à leur « racine » en coupant généralement la fin des mots. Elle peut dépendre du langage.
 - ▶ *automate, automatisme, automatique* → *automat*.
- **Algorithme de Porter** [Por80]
 - ▶ C'est l'algorithme le plus utilisé pour l'anglais (mais il existe pour d'autres langues).
 - ▶ Il utilise des conventions et cinq phases de réductions (ensembles de commandes).
 - ★ Exemple de commande : *supprimer le « ement » final s'il reste plus d'un caractère après réduction.*
 - ★ Exemple de convention : *parmi les commandes, utiliser celle qui s'applique au plus long suffixe.*
- D'autres algorithmes existent, comme l'**algorithme de Krovetz** [Kro93] : il utilise un dictionnaire et ne coupe que les mots n'y apparaissant pas.

Suppression des mots vides (*stop list*)

- Un **mot vide** (*stop word*) est un mot qui n'apporte pas d'information utile.
- Une **liste de mots vides** (*stop list*) est une liste de mots vides à supprimer dans les documents.
 - ▶ **Prépositions** : *de, sur...*
 - ▶ **Déterminants** : *le, une...*
 - ▶ **Pronoms** : *je, vous...*
 - ▶ **Quelques adverbes et adjectifs** : *déjà, plusieurs...*
 - ▶ **Quelques noms et verbes** : *faire, mois...*
- Il existe plusieurs listes de mots vides standards.
 - ▶ La liste SMART contient 571 mots anglais.
- Généralement, la suppression des mots vides améliore l'efficacité des applications de fouille de textes mais il peut également être utile de les conserver dans certains cas.

Plan du cours

1 Introduction

2 Traitements de données textuelles

3 Fouille de motifs séquentiels

- Extraction de motifs séquentiels sous contraintes
- Extraction de motifs séquentiels émergents
- Application : analyse stylistique de textes
- Application : analyse de publications biomédicales

4 Fouille de graphes pour l'exploration de grands textes

5 Conclusion

6 Références

Plan du cours

1 Introduction

2 Traitements de données textuelles

3 Fouille de motifs séquentiels

- **Extraction de motifs séquentiels sous contraintes**
- Extraction de motifs séquentiels émergents
- Application : analyse stylistique de textes
- Application : analyse de publications biomédicales

4 Fouille de graphes pour l'exploration de grands textes

- Contexte
- Modèle linguistique de Hoey
- Fouille de graphes enrichis

5 Conclusion

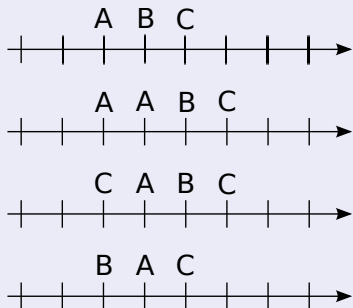
6 Références

Fouille de motifs séquentiels

- **Ordre temporel entre les données**

- ▶ **Données** : séquences d'items ou d'itemsets (ensemble d'items)
- ▶ **Motif** : sous-séquence d'items ou d'itemsets

Base de 4 séquences d'items

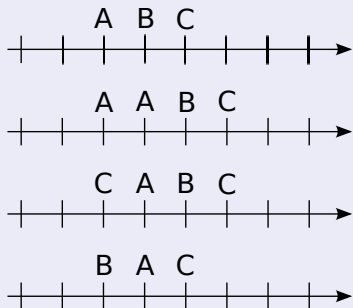


Base de séquences d'itemsets

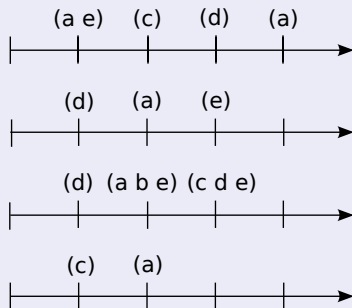
Fouille de motifs séquentiels

- **Ordre temporel entre les données**
 - ▶ **Données** : séquences d'items ou d'itemsets (ensemble d'items)
 - ▶ **Motif** : sous-séquence d'items ou d'itemsets

Base de 4 séquences d'items



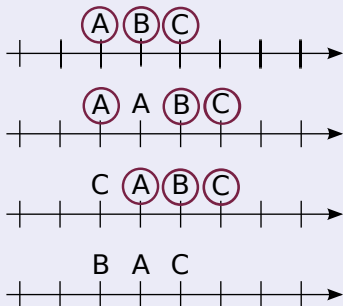
Base de séquences d'itemsets



Fouille de motifs séquentiels

- **Ordre temporel entre les données**
 - ▶ **Données** : séquences d'items ou d'itemsets (ensemble d'items)
 - ▶ **Motif** : sous-séquence d'items ou d'itemsets

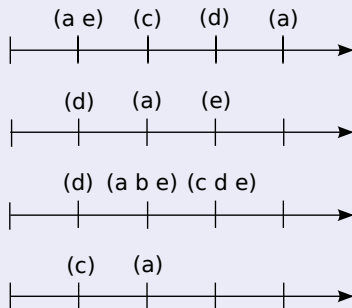
Base de 4 séquences d'items



- Exemple de motif : $\langle ABC \rangle$

→ présent dans 3 séquences

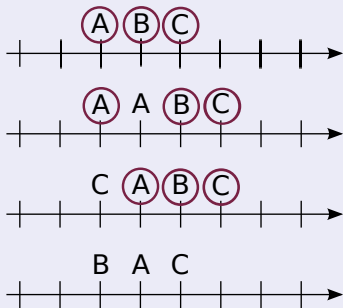
Base de séquences d'itemsets



Fouille de motifs séquentiels

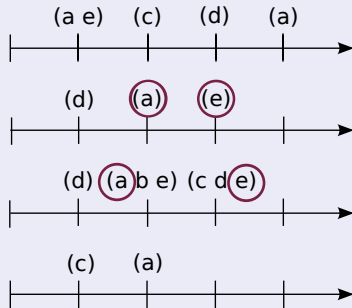
- **Ordre temporel entre les données**
 - ▶ **Données** : séquences d'items ou d'itemsets (ensemble d'items)
 - ▶ **Motif** : sous-séquence d'items ou d'itemsets

Base de 4 séquences d'items



- Exemple de motif : $\langle ABC \rangle$
→ présent dans 3 séquences

Base de séquences d'itemsets



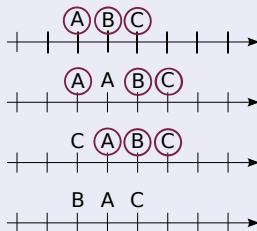
- Exemple de motif : $\langle (a)(e) \rangle$
→ présent dans 2 séquences

Extraction de motifs séquentiels fréquents

- **Support** d'un motif : nombre de séquences dans lequel le motif apparaît



Base de séquences d'items

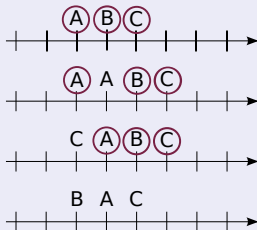


- $sup(\langle ABC \rangle) = 3$; $supRel(\langle ABC \rangle) = 0,75$

Extraction de motifs séquentiels fréquents

- **Support relatif** d'un motif : nombre de séquences dans lequel le motif apparaît, normalisé par le nombre total de séquences

Base de séquences d'items

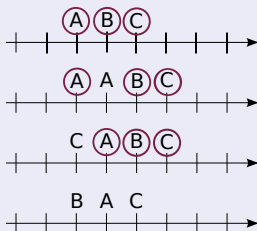


- $sup(\langle ABC \rangle) = 3$; $supRel(\langle ABC \rangle) = 0,75$

Extraction de motifs séquentiels fréquents

- **Support relatif** d'un motif : nombre de séquences dans lequel le motif apparaît, normalisé par le nombre total de séquences
- Extraction de motifs **fréquents** : extraction de tous les motifs M tels que $sup(M) \geq minsup$ ($minsup$, un seuil prédéfini)

Base de séquences d'items

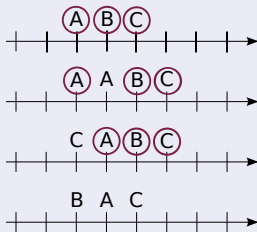


- $minsup = 3$
 - ▶ $MF = \{\langle A \rangle, \langle B \rangle, \langle C \rangle, \langle AB \rangle, \langle BC \rangle, \langle AC \rangle, \langle ABC \rangle\}$

Extraction de motifs séquentiels sous contraintes

- Extraction de motifs fréquents **sous contraintes** : extraction des motifs M tels que $\text{sup}(M) \geq \text{minsup}$ et M vérifie toutes les contraintes
 - ▶ **Contrainte de longueur** : longueur minimale du motif
 - ▶ **Contrainte de clôture** : un motif M est **clos** si aucun sur-motif n'a exactement le même support
 - ▶ **Contrainte de $\text{gap}[X, Y]$** : nombre d'items entre chaque item du motif (au moins X items et au plus Y items)
 - ▶ ...

Base de séquences d'items

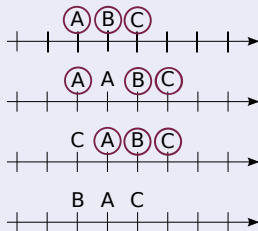


- $\text{minsup} = 3$
 - ▶ $MF = \{\langle A \rangle, \langle B \rangle, \langle C \rangle, \langle AB \rangle, \langle BC \rangle, \langle AC \rangle, \langle ABC \rangle\}$

Extraction de motifs séquentiels sous contraintes

- Extraction de motifs fréquents **sous contraintes** : extraction des motifs M tels que $sup(M) \geq minsup$ et M vérifie toutes les contraintes
 - ▶ **Contrainte de longueur** : longueur minimale du motif
 - ▶ **Contrainte de clôture** : un motif M est **clos** si aucun sur-motif n'a exactement le même support
 - ▶ **Contrainte de $gap[X, Y]$** : nombre d'items entre chaque item du motif (au moins X items et au plus Y items)
 - ▶ ...

Base de séquences d'items

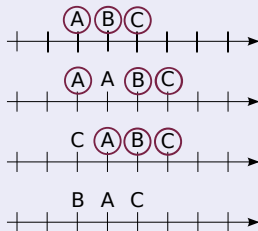


- $minsup = 3 \wedge minlong = 2$
 - ▶ $MF = \{\langle A \rangle, \langle B \rangle, \langle C \rangle, \langle AB \rangle, \langle BC \rangle, \langle AC \rangle, \langle ABC \rangle\}$

Extraction de motifs séquentiels sous contraintes

- Extraction de motifs fréquents **sous contraintes** : extraction des motifs M tels que $sup(M) \geq minsup$ et M vérifie toutes les contraintes
 - ▶ **Contrainte de longueur** : longueur minimale du motif
 - ▶ **Contrainte de clôture** : un motif M est **clos** si aucun sur-motif n'a exactement le même support
 - ▶ **Contrainte de $gap[X, Y]$** : nombre d'items entre chaque item du motif (au moins X items et au plus Y items)
 - ▶ ...

Base de séquences d'items

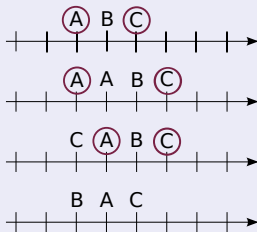


- $minsup = 3 \wedge minlong = 2 \wedge clos$
 - ▶ $MF = \{\langle A \rangle, \langle B \rangle, \langle C \rangle, \langle AB \rangle, \langle BC \rangle, \langle AC \rangle, \langle ABC \rangle\}$

Extraction de motifs séquentiels sous contraintes

- Extraction de motifs fréquents **sous contraintes** : extraction des motifs M tels que $sup(M) \geq minsup$ et M vérifie toutes les contraintes
 - ▶ **Contrainte de longueur** : longueur minimale du motif
 - ▶ **Contrainte de clôture** : un motif M est **clos** si aucun sur-motif n'a exactement le même support
 - ▶ **Contrainte de $gap[X, Y]$** : nombre d'items entre chaque item du motif (au moins X items et au plus Y items)
 - ▶ ...

Base de séquences d'items



- $minsup = 3 \wedge minlong = 2 \wedge clos \wedge gap[1, 2]$
 - ▶ $MF = \{\langle A \rangle, \langle B \rangle, \langle C \rangle, \langle AB \rangle, \langle BC \rangle, \langle AC \rangle, \langle ABC \rangle\}$

Fouille de motifs séquentiels – exemple

Base de séquences d'items

- 1 *the wind of the night...*
- 2 *the big noises of the ocean...*
- 3 *the very dark light of the night...*
- 4 *the wind is blowing...*

Base de séquences d'itemsets

- 1 *(the DET) (wind NC) (of PREP) (the DET) (night NC)...*
- 2 *(the DET) (big ADJ) (noises noise NC) (of PREP) (the DET) (ocean NC)...*
- 3 *(the DET) (very ADV) (dark ADJ) (light NC) (of PREP) (the DET) (night NC)...*
- 4 *(the DET) (wind NC) (is be V PRES 3S) (blowing blow V PARPRES)...*

- « *the of the* » est un motif d'items.
- « *(the) (of PREP) (the) (NC)* » est un motif d'itemsets.

Fouille de motifs séquentiels – exemple

Base de séquences d'items

- 1 *the wind **of the** night...*
- 2 *the big noises **of the** ocean...*
- 3 *the very dark light **of the** night...*
- 4 *the wind is blowing...*

Base de séquences d'itemsets

- 1 *(the DET) (wind NC) (of PREP) (the DET) (night NC)...*
- 2 *(the DET) (big ADJ) (noises noise NC) (of PREP) (the DET) (ocean NC)...*
- 3 *(the DET) (very ADV) (dark ADJ) (light NC) (of PREP) (the DET) (night NC)...*
- 4 *(the DET) (wind NC) (is be V PRES 3S) (blowing blow V PARPRES)...*

- « *the of the* » est un motif d'items.
- « *(the) (of PREP) (the) (NC)* » est un motif d'itemsets.

Fouille de motifs séquentiels – exemple

Base de séquences d'items

- 1 *the wind **of the** night...*
- 2 *the big noises **of the** ocean...*
- 3 *the very dark light **of the** night...*
- 4 *the wind is blowing...*

Base de séquences d'itemsets

- 1 (***the*** DET) (*wind* NC) (***of PREP***) (***the*** DET) (*night NC*)...
- 2 (***the*** DET) (*big* ADJ) (*noises noise* NC) (***of PREP***) (***the*** DET) (*ocean NC*)...
- 3 (***the*** DET) (*very* ADV) (*dark* ADJ) (*light* NC) (***of PREP***) (***the*** DET) (*night NC*)...
- 4 (*the* DET) (*wind* NC) (*is be* V PRES 3S) (*blowing blow* V PARPRES)...

- « *the of the* » est un motif d'items.
- « (*the*) (*of PREP*) (*the*) (*NC*) » est un motif d'itemsets.

Plan du cours

1 Introduction

2 Traitements de données textuelles

3 Fouille de motifs séquentiels

- Extraction de motifs séquentiels sous contraintes
- **Extraction de motifs séquentiels émergents**
- Application : analyse stylistique de textes
- Application : analyse de publications biomédicales

4 Fouille de graphes pour l'exploration de grands textes

- Contexte
- Modèle linguistique de Hoey
- Fouille de graphes enrichis

5 Conclusion

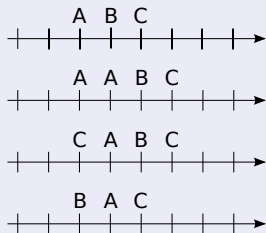
6 Références

Extraction de motifs séquentiels émergents

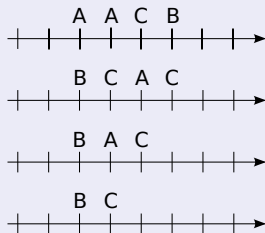
- Taux de croissance d'un motif M dans une base B_1 par rapport à B_2 :

$$\text{TauxCroiss}_{B_1/B_2}(M) = \begin{cases} \infty & \text{si } \text{supRel}_{B_2}(M) = 0 \\ \frac{\text{supRel}_{B_1}(M)}{\text{supRel}_{B_2}(M)} & \text{sinon} \end{cases}$$

Base de séquences d'items :
 B_1



Base de séquences d'items :
 B_2

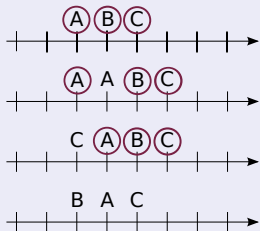


Extraction de motifs séquentiels émergents

- Taux de croissance d'un motif M dans une base B_1 par rapport à B_2 :

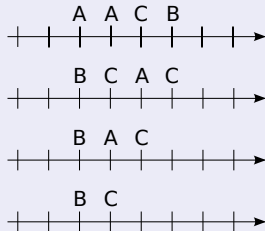
$$\text{TauxCroiss}_{B_1/B_2}(M) = \begin{cases} \infty & \text{si } \text{supRel}_{B_2}(M) = 0 \\ \frac{\text{supRel}_{B_1}(M)}{\text{supRel}_{B_2}(M)} & \text{sinon} \end{cases}$$

Base de séquences d'items :
 B_1



$$\text{TauxCroiss}_{B_1/B_2}(\langle ABC \rangle) = \infty$$

Base de séquences d'items :
 B_2

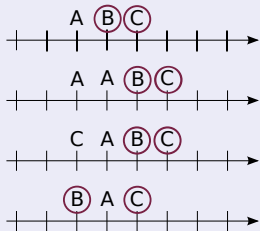


Extraction de motifs séquentiels émergents

- Taux de croissance d'un motif M dans une base B_1 par rapport à B_2 :

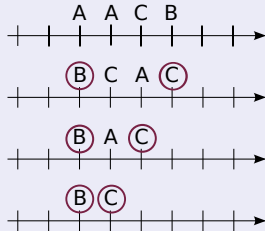
$$\text{TauxCroiss}_{B_1/B_2}(M) = \begin{cases} \infty & \text{si } \text{supRel}_{B_2}(M) = 0 \\ \frac{\text{supRel}_{B_1}(M)}{\text{supRel}_{B_2}(M)} & \text{sinon} \end{cases}$$

Base de séquences d'items :
 B_1



$$\text{TauxCroiss}_{B_1/B_2}(\langle BC \rangle) = \frac{1}{0,75} = 1,33$$

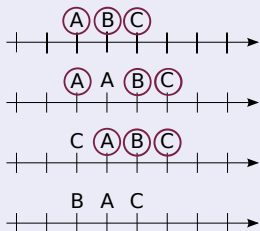
Base de séquences d'items :
 B_2



Extraction de motifs séquentiels émergents

- Taux de croissance d'un motif M dans une base B_1 par rapport à B_2
- Extraction des motifs émergents de B_1 par rapport à B_2 : extraction des motifs M tels que $TauxCroiss_{B_1/B_2}(M) \geq \rho$ (ρ , un seuil prédéfini)

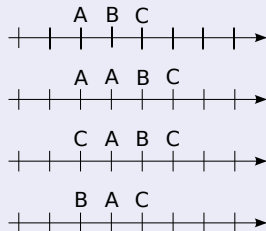
Base de séquences d'items : B_1



• $\rho = 2 \wedge minsup = 3$

► $MEF = \{\langle AB \rangle, \langle ABC \rangle\}$

Base de séquences d'items : B_2



• $\rho = 2 \wedge minsup = 3$

► $MEF = \emptyset$

Motifs séquentiels émergents – exemple

Base B_1

- 1 *the wind of the night...*
- 2 *the big noises of the ocean...*
- 3 *the very dark light of the night...*
- 4 *the wind is blowing...*

Base B_2

- 1 *the wind is blowing...*
- 2 *the cold wind blows...*
- 3 *the nights are over...*
- 4 *the wind is warm...*

- Le motif « *the wind* » n'est émergent ni dans B_1 , ni dans B_2 (si le seuil est $\sigma = 2$).
- Le motif « *the wind of night* » est émergent infini dans la base B_1 .

Motifs séquentiels émergents – exemple

Base B_1

- 1 *the wind* of the night...
- 2 *the big noises of the ocean...*
- 3 *the very dark light of the night...*
- 4 *the wind* is blowing...

Base B_2

- 1 *the wind* is blowing...
- 2 *the cold wind* blows...
- 3 *the nights are over...*
- 4 *the wind* is warm...

- Le motif « *the wind* » n'est émergent ni dans B_1 , ni dans B_2 (si le seuil est $\sigma = 2$).
- Le motif « *the wind of night* » est émergent infini dans la base B_1 .

Motifs séquentiels émergents – exemple

Base B_1

- 1 *the wind of the night...*
- 2 *the big noises of the ocean...*
- 3 *the very dark light of the night...*
- 4 *the wind is blowing...*

Base B_2

- 1 *the wind is blowing...*
- 2 *the cold wind blows...*
- 3 *the nights are over...*
- 4 *the wind is warm...*

- Le motif « *the wind* » n'est émergent ni dans B_1 , ni dans B_2 (si le seuil est $\sigma = 2$).
- Le motif « *the wind of night* » est **émergent infini** dans la base B_1 .

Plan du cours

1 Introduction

2 Traitements de données textuelles

3 Fouille de motifs séquentiels

- Extraction de motifs séquentiels sous contraintes
- Extraction de motifs séquentiels émergents
- **Application : analyse stylistique de textes**
- Application : analyse de publications biomédicales

4 Fouille de graphes pour l'exploration de grands textes

- Contexte
- Modèle linguistique de Hoey
- Fouille de graphes enrichis

5 Conclusion

6 Références

Contexte du travail

Objectifs

Extraction automatique de **patrons linguistiques caractéristiques** de genres de texte (analyse stylistique)

→ Pas de choix *a priori* des patrons étudiés

Contributions [QCCL12a, QCCL12b]

- Utilisation de la **fouille de motifs séquentiels** [AS95]
 - Motifs séquentiels considérés : motifs d'items et motifs d'itemsets
- Extraction de **motifs émergents** [DL99]
 - Focalisation sur les motifs caractéristiques d'un genre de texte
- Validation linguistique de motifs extraits (pour la poésie)

Contexte du travail

Objectifs

Extraction automatique de **patrons linguistiques caractéristiques** de genres de texte (analyse stylistique)

→ Pas de choix *a priori* des patrons étudiés

Contributions [QCCL12a, QCCL12b]

- Utilisation de la **fouille de motifs séquentiels** [AS95]
 - ▶ Motifs séquentiels considérés : motifs d'items et motifs d'itemsets
- Extraction de **motifs émergents** [DL99]
 - Focalisation sur les motifs caractéristiques d'un genre de texte
- Validation linguistique de motifs extraits (pour la poésie)

Extraction de patrons linguistiques caractéristiques



Méthodologie

- ➊ **Pré-traitement des corpus** : un corpus par genre de texte étudié
 - ▶ Mise en minuscule des mots et découpage en séquences aux ponctuations de l'ensemble :
{ « . », « ? », « ! », « ... », « ; », « : », « , », » }
 - ▶ Étiquetage des mots avec leur forme, leur lemme et leur catégorie morpho-syntaxique (3 traits)
- ➋ Extraction des motifs fréquents de chaque corpus
- ➌ Sélection des motifs caractéristiques de chaque corpus
- ➍ Interprétation linguistique des motifs émergents

Extraction de patrons linguistiques caractéristiques



Méthodologie

- ➊ Pré-traitement des corpus : un corpus par genre de texte étudié
- ➋ Extraction des motifs fréquents de chaque corpus
 - ▶ Motifs d'items : chaque mot est décrit par son lemme
 - ★ Exemple de motif : $\langle le \ du \ qui \rangle$
 - ▶ Motifs d'itemsets : chaque mot est décrit par ses 3 traits
 - ★ Exemple de motif : $\langle (le)(NC)(du)(NC)(qui)(V) \rangle$
- ➌ Sélection des motifs caractéristiques de chaque corpus
- ➍ Interprétation linguistique des motifs émergents

Extraction de patrons linguistiques caractéristiques



Méthodologie

- 1 Pré-traitement des corpus : un corpus par genre de texte étudié
- 2 Extraction des motifs fréquents de chaque corpus
- 3 **Sélection des motifs caractéristiques** de chaque corpus
 - Calcul des motifs émergents de chaque corpus par rapport à chacun des autres corpus
 - ★ Un motif M est **émergent** pour un corpus C_i s'il est émergent par rapport à chaque corpus C_j (avec $j \neq i$)
- 4 Interprétation linguistique des motifs émergents

Extraction de patrons linguistiques caractéristiques



Méthodologie

- ➊ Pré-traitement des corpus : un corpus par genre de texte étudié
- ➋ Extraction des motifs fréquents de chaque corpus
- ➌ Sélection des motifs caractéristiques de chaque corpus
- ➍ **Interprétation linguistique des motifs émergents**
 - ▶ Sélection par le linguiste des motifs intéressants découverts automatiquement
 - ▶ Interprétation par le linguiste des motifs sélectionnés

Corpus et paramètres de l'extraction de motifs

Données

- **Corpus français** : textes de 1800 à 1900 du CNRTL
 - ▶ Étiquetage avec Cordial (35 catégories morpho-syntaxiques)

Corpus	Nb auteurs	Nb œuvres	Nb séquences	Nb mots
<i>Poésie</i>	27	48	151 116	1 167 422
<i>Correspondance</i>	5	9	234 997	1 562 543
<i>Roman</i>	37	52	663 860	5 105 240

Paramètres

- Extraction des motifs
 - ▶ Motifs d'items : utilisation de dmt4 [NR07]
 - ▶ $minsup = 0,001\%$; $minlong = 2$, $maxlong = 20$; différents *gaps*
 - ▶ Motifs d'itemsets : utilisation de Clospan [YHA03]
 - ▶ $minsup = 0,15\%$; pas d'autres contraintes disponibles
- Sélection des motifs émergents : $\rho = 1,001$

Corpus et paramètres de l'extraction de motifs

Données

- Corpus français : textes de 1800 à 1900 du CNRTL
 - ▶ Étiquetage avec Cordial (35 catégories morpho-syntaxiques)

Corpus	Nb auteurs	Nb œuvres	Nb séquences	Nb mots
<i>Poésie</i>	27	48	151 116	1 167 422
<i>Correspondance</i>	5	9	234 997	1 562 543
<i>Roman</i>	37	52	663 860	5 105 240

Paramètres

- Extraction des motifs
 - ▶ **Motifs d'items** : utilisation de dmt4 [NR07]
 - ★ $minsup = 0,001 \%$; $minlong = 2$, $maxlong = 20$; différents *gaps*
 - ▶ **Motifs d'itemsets** : utilisation de Clospan [YHA03]
 - ★ $minsup = 0,15 \%$; pas d'autres contraintes disponibles
- Sélection des motifs émergents : $\rho = 1,001$

Corpus et paramètres de l'extraction de motifs

Données

- Corpus français : textes de 1800 à 1900 du CNRTL
 - ▶ Étiquetage avec Cordial (35 catégories morpho-syntaxiques)

Corpus	Nb auteurs	Nb œuvres	Nb séquences	Nb mots
<i>Poésie</i>	27	48	151 116	1 167 422
<i>Correspondance</i>	5	9	234 997	1 562 543
<i>Roman</i>	37	52	663 860	5 105 240

Paramètres

- Extraction des motifs
 - ▶ Motifs d'items : utilisation de dmt4 [NR07]
 - ★ $minsup = 0,001\%$; $minlong = 2$, $maxlong = 20$; différents *gaps*
 - ▶ Motifs d'itemsets : utilisation de Clospan [YHA03]
 - ★ $minsup = 0,15\%$; pas d'autres contraintes disponibles
- Sélection des motifs émergents : $\rho = 1,001$

Analyse quantitative des motifs

Corpus	Motifs d'items avec contrainte <i>gap</i>				Motifs d'itemsets
	[1, 1]	[1, 2]	[1, 3]	[1, 5]	
<i>Poésie</i>	18 816 (30,7 %)	37 933 (27,0 %)	55 762 (24,3 %)	86 901 (22,6 %)	2 245 326 (11,4 %)
<i>Correspondance</i>	16 936 (50,2 %)	36 849 (50,7 %)	56 755 (50,4 %)	96 549 (50,0 %)	10 128 288 (57,4 %)
<i>Roman</i>	78 210 (6,1 %)	175 645 (5,3 %)	282 967 (4,9 %)	512 647 (4,6 %)	11 681 913 (71,2 %)
Total	113 962 (16,7 %)	250 427 (15,3 %)	395 484 (14,2 %)	696 097 (13,2 %)	24 055 527 (59,8 %)

Résultats : nb de motifs (et ratio de motifs émergents)

- Intérêt des motifs émergents
 - ▶ Réduction du nombre de motifs à analyser
- Intérêt de la contrainte *gap*
 - ▶ Compromis entre nombre de motifs et pertinence des motifs émergents
- Motifs d'items vs. itemsets
 - ▶ Beaucoup plus de motifs d'itemsets mais pas de contraintes fixées

Analyse quantitative des motifs

Corpus	Motifs d'items avec contrainte <i>gap</i>				Motifs d'itemsets
	[1, 1]	[1, 2]	[1, 3]	[1, 5]	
<i>Poésie</i>	18 816 (30,7 %)	37 933 (27,0 %)	55 762 (24,3 %)	86 901 (22,6 %)	2 245 326 (11,4 %)
<i>Correspondance</i>	16 936 (50,2 %)	36 849 (50,7 %)	56 755 (50,4 %)	96 549 (50,0 %)	10 128 288 (57,4 %)
<i>Roman</i>	78 210 (6,1 %)	175 645 (5,3 %)	282 967 (4,9 %)	512 647 (4,6 %)	11 681 913 (71,2 %)
Total	113 962 (16,7 %)	250 427 (15,3 %)	395 484 (14,2 %)	696 097 (13,2 %)	24 055 527 (59,8 %)

Résultats : nb de motifs (et ratio de motifs émergents)

- Intérêt des motifs émergents
 - ▶ Réduction du nombre de motifs à analyser
- Intérêt de la contrainte *gap*
 - ▶ Compromis entre nombre de motifs et pertinence des motifs émergents
- Motifs d'items vs. itemsets
 - ▶ Beaucoup plus de motifs d'itemsets mais pas de contraintes fixées

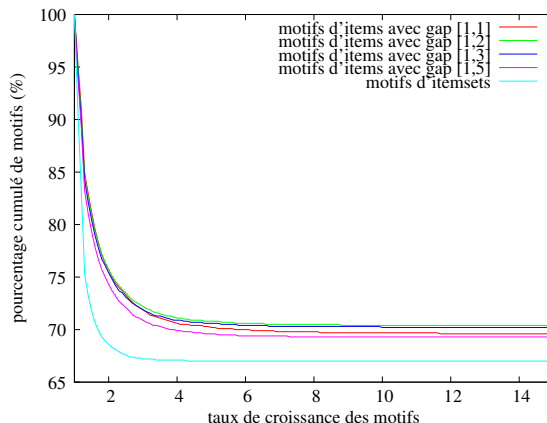
Analyse quantitative des motifs

Corpus	Motifs d'items avec contrainte <i>gap</i>				Motifs d'itemsets
	[1, 1]	[1, 2]	[1, 3]	[1, 5]	
<i>Poésie</i>	18 816 (30,7 %)	37 933 (27,0 %)	55 762 (24,3 %)	86 901 (22,6 %)	2 245 326 (11,4 %)
<i>Correspondance</i>	16 936 (50,2 %)	36 849 (50,7 %)	56 755 (50,4 %)	96 549 (50,0 %)	10 128 288 (57,4 %)
<i>Roman</i>	78 210 (6,1 %)	175 645 (5,3 %)	282 967 (4,9 %)	512 647 (4,6 %)	11 681 913 (71,2 %)
Total	113 962 (16,7 %)	250 427 (15,3 %)	395 484 (14,2 %)	696 097 (13,2 %)	24 055 527 (59,8 %)

Résultats : nb de motifs (et ratio de motifs émergents)

- Intérêt des motifs émergents
 - ▶ Réduction du nombre de motifs à analyser
- Intérêt de la contrainte *gap*
 - ▶ Compromis entre nombre de motifs et pertinence des motifs émergents
- Motifs d'items vs. itemsets
 - ▶ Beaucoup plus de motifs d'itemsets mais pas de contraintes fixées

Analyse quantitative des motifs émergents



Résultats sur les motifs émergents

- Majorité des motifs avec un **taux de croissance infini**
→ Ces motifs n'apparaissent que dans un seul corpus

Analyse quantitative des motifs : conclusions

Motifs analysés par la suite, d'un point de vue linguistique

- Motifs émergents avec taux de croissance infini
 - ▶ Motifs d'items
 - ★ Contrainte de longueur : au moins 3 items
 - ★ Contrainte de *gap* : [1,3]
 - ▶ Motifs d'itemsets
 - ★ Contrainte de longueur : au moins 3 itemsets
 - ★ Motifs avec à la fois des formes de mots/lemmes et des catégories morpho-syntaxiques

Analyse stylistique de motifs d'items émergents

Motifs d'items	Exemples de séquences correspondantes
est * un * qui	est -ce un goëland qui bat de l'aile ? ta grâce est comme un luth qui vibre au fond du bois
le * de * qui * dans	le vent du soir qui meurt dans le feuillage la brise du soir qui pleure dans des branches de coudrier le grand bruit du rêveur océan qui parle dans la nuit la feuille des forêts qui tourne dans la bise

Résultats sur *Poésie*

- Intérêt de la contrainte *gap* : généralisation
 - ▶ Remplissage du *gap* (représenté par *) avec un nombre de mots qui peut être différent pour 2 instances du même motif
- Intérêt des motifs d'items
 - ▶ Structures grammaticales relativement indépendantes du lexique
 - Éléments fixes : mots grammaticaux
 - Éléments variables : mots lexicaux

Analyse stylistique de motifs d'items émergents

Motifs d'items	Exemples de séquences correspondantes
est * un * qui	est -ce un goëland qui bat de l'aile ? ta grâce est comme un luth qui vibre au fond du bois
le * de * qui * dans	le vent du soir qui meurt dans le feuillage la brise du soir qui pleure dans des branches de coudrier le grand bruit du rêveur océan qui parle dans la nuit la feuille des forêts qui tourne dans la bise

Résultats sur *Poésie*

- Intérêt de la contrainte *gap* : généralisation
 - ▶ Remplissage du *gap* (représenté par *) avec un nombre de mots qui peut être différent pour 2 instances du même motif
- Intérêt des motifs d'items
 - ▶ Structures grammaticales relativement indépendantes du lexique
 - ★ Éléments fixes : mots grammaticaux
 - ★ Éléments variables : mots lexicaux

Analyse stylistique de motifs d'itemsets émergents

Motifs d'items	Motifs d'itemsets
est * un * qui	est-ce un N qui V est comme un N qui V
le * de * qui * dans	le N de N qui V dans le N

Résultats sur *Poésie*

- Un ou plusieurs motifs d'itemsets pour un même motif d'items
- Intérêt des motifs d'itemsets
 - Catégories morpho-syntaxiques des éléments variables connues
→ Obtention directe de *patrons grammaticaux*
 - Motifs d'itemsets à rapprocher des *cadres collocationnels*
(collocations sur des unités grammaticales) [RS91]

Analyse stylistique de motifs d'itemsets émergents

Motifs d'items	Motifs d'itemsets
est * un * qui	est-ce un N qui V est comme un N qui V
le * de * qui * dans	le N de N qui V dans le N

Résultats sur *Poésie*

- Un ou plusieurs motifs d'itemsets pour un même motif d'items
- Intérêt des motifs d'itemsets
 - ▶ Catégories morpho-syntaxiques des éléments variables connues
→ Obtention directe de **patrons grammaticaux**
 - Motifs d'itemsets à rapprocher des *cadres collocationnels* (collocations sur des unités grammaticales) [RS91]

Plan du cours

1 Introduction

2 Traitements de données textuelles

3 Fouille de motifs séquentiels

- Extraction de motifs séquentiels sous contraintes
- Extraction de motifs séquentiels émergents
- Application : analyse stylistique de textes
- **Application : analyse de publications biomédicales**

4 Fouille de graphes pour l'exploration de grands textes

- Contexte
- Modèle linguistique de Hoey
- Fouille de graphes enrichis

5 Conclusion

6 Références

Schéma général des applications

- La fouille de textes sur des **publications scientifiques** suit l'architecture globale suivante (tiré de [TCM12]).

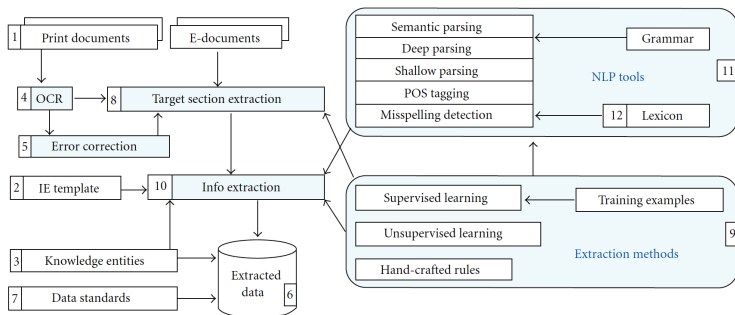


FIGURE 2: A reference system architecture for an example IE system. Numbers correspond to the text.

→ Des états de l'art dans ce domaine sont également donnés dans [TCM12, SLC⁺16] et on s'intéressera plutôt à l'utilisation de la fouille de motifs séquentiels dans la suite.

Extractions de relation biomédicales

Objectifs

Extraction automatique de **relations entre gènes et maladies rares** dans les résumés d'articles de PubMed.

→ Pas de choix *a priori* des patrons étudiés

Contributions [BCC⁺12]

- Utilisation de la **fouille de motifs séquentiels d'itemsets**.
- Approche similaire à l'approche pour l'extraction de motifs stylistiques
- Définition et utilisation de **contraintes supplémentaires** pendant le processus de fouille de motifs.

Extractions de relation biomédicales

Objectifs

Extraction automatique de **relations entre gènes et maladies rares** dans les résumés d'articles de PubMed.

→ Pas de choix *a priori* des patrons étudiés

Contributions [BCC⁺12]

- Utilisation de la **fouille de motifs séquentiels d'itemsets**.
- Approche similaire à l'approche pour l'extraction de motifs stylistiques
- Définition et utilisation de **contraintes supplémentaires** pendant le processus de fouille de motifs.

Méthodologie et résultats

Contraintes supplémentaires pendant la fouille de motifs

- **Contrainte d'appartenance** : le motif doit contenir les items GENE et DISEASE.
- **Contrainte de portée** : le nombre maximal d'itemsets est fixé entre le premier et le dernier d'une séquence
- **Contrainte d'association** : un item est associé à un autre (par exemple, pour chaque verbe d'étiquette VB, on lui associe son lemme).

Résultats

- Corpus de 17 527 phrases extraites des résumés de PubMed et contenant au moins un nom de gène et un nom de maladie rare.
- Meilleurs résultats obtenus avec un support minimal relatif de 0,05% et un gap [1, 10] :
 - Rappel : 65% (moins bon rappel si pas de contrainte de gap)
 - Précision : 66%

→ Exemple de patron extrait : (DISEASE) (BE VBP) (JJ) (IN) (FACTOR NN) (GENE)

Méthodologie et résultats

Contraintes supplémentaires pendant la fouille de motifs

- **Contrainte d'appartenance** : le motif doit contenir les items GENE et DISEASE.
- **Contrainte de portée** : le nombre maximal d'itemsets est fixé entre le premier et le dernier d'une séquence
- **Contrainte d'association** : un item est associé à un autre (par exemple, pour chaque verbe d'étiquette VB, on lui associe son lemme).

Résultats

- Corpus de 17 527 phrases extraites des résumés de PubMed et contenant au moins un nom de gène et un nom de maladie rare.
- Meilleurs résultats obtenus avec un support minimal relatif de 0,05% et un gap [1, 10] :
 - ▶ Rappel : 65% (moins bon rappel si pas de contrainte de gap)
 - ▶ Précision : 66%

→ **Exemple de patron extrait** : (DISEASE) (BE VBP) (JJ) (IN) (FACTOR NN) (GENE)

Plan du cours

- 1 Introduction
- 2 Traitements de données textuelles
- 3 Fouille de motifs séquentiels
- 4 **Fouille de graphes pour l'exploration de grands textes**
 - Contexte
 - Modèle linguistique de Hoey
 - Fouille de graphes enrichis
- 5 Conclusion
- 6 Références

Plan du cours

1 Introduction

2 Traitements de données textuelles

3 Fouille de motifs séquentiels

- Extraction de motifs séquentiels sous contraintes
- Extraction de motifs séquentiels émergents
- Application : analyse stylistique de textes
- Application : analyse de publications biomédicales

4 Fouille de graphes pour l'exploration de grands textes

- **Contexte**
- Modèle linguistique de Hoey
- Fouille de graphes enrichis

5 Conclusion

6 Références

Contexte du travail

Objectifs

Extraction automatique de sous-parties cohérentes de textes représentés sous forme de graphes (exploration de textes)

→ Conserver la structure en graphe du texte dans les sous-parties

Contributions [QCCL12a, QCCL13]

- Utilisation de la fouille de graphes sous contraintes linguistiques
 - ▶ Représentation du texte sous forme de graphe par application du modèle linguistique de Hoey [Hoe91]
 - ▶ Extraction de sous-graphes de texte cohérents par la fouille de collections de k -PC homogènes (CoHoP) [MRG12]
- Validation linguistique de la cohérence des sous-parties extraites

Contexte du travail

Objectifs

Extraction automatique de sous-parties cohérentes de textes représentés sous forme de graphes (exploration de textes)

→ Conserver la structure en graphe du texte dans les sous-parties

Contributions [QCCL12a, QCCL13]

- Utilisation de la **fouille de graphes sous contraintes linguistiques**
 - ▶ Représentation du texte sous forme de graphe par application du **modèle linguistique de Hoey** [Hoe91]
 - ▶ Extraction de sous-graphes de texte cohérents par la fouille de **collections de k -PC homogènes** (CoHoP) [MRG12]
- Validation linguistique de la cohérence des sous-parties extraites

Plan du cours

1 Introduction

2 Traitements de données textuelles

3 Fouille de motifs séquentiels

- Extraction de motifs séquentiels sous contraintes
- Extraction de motifs séquentiels émergents
- Application : analyse stylistique de textes
- Application : analyse de publications biomédicales

4 Fouille de graphes pour l'exploration de grands textes

- Contexte
- **Modèle linguistique de Hoey**
- Fouille de graphes enrichis

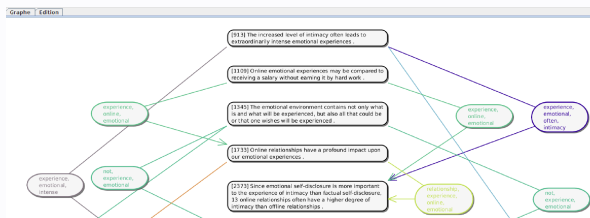
5 Conclusion

6 Références

Modèle linguistique de Hoey

- Modèle fondé sur les **répétitions lexicales** : répétition stricte, répétition du lemme, relation de synonymie, relation d'hyper/hyponymie, reprise anaphorique...
- Réseau phrastique : ensemble d'au moins 3 phrases tel que chaque phrase est appariée avec au moins une phrase de l'ensemble
 - ▶ Appariement de 2 phrases si au moins 3 unités lexicales communes
- Hypotexte : ensemble des réseaux phrastiques d'un texte
 - Sorte de résumé du texte (suppression des phrases non appariées)

Extrait d'un réseau phrastique

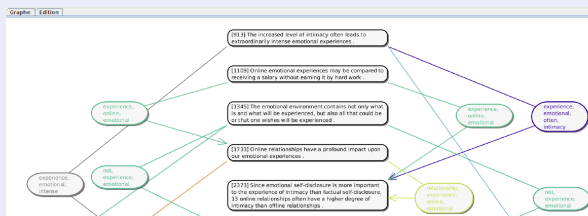


Répétition lexicale : répétition des lemmes des lexèmes des phrases →
Lexèmes : noms, adjectifs, verbes ou adverbes

Modèle linguistique de Hoey

- Modèle fondé sur les **répétitions lexicales**
- **Réseau phrastique** : ensemble d'au moins 3 phrases tel que chaque phrase est appariée avec au moins une phrase de l'ensemble
 - ▶ **Appariement** de 2 phrases si au moins 3 unités lexicales communes
- **Hypotexte** : ensemble des réseaux phrastiques d'un texte
→ Sorte de résumé du texte (suppression des phrases non appariées)

Extrait d'un réseau phrastique

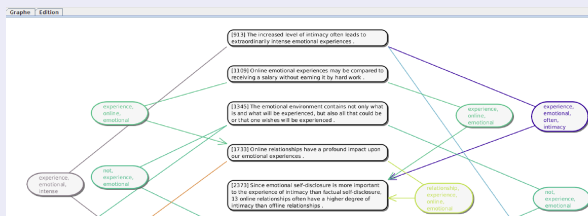


Répétition lexicale : répétition des lemmes des lexèmes des phrases →
Lexèmes : noms, adjectifs, verbes ou adverbes

Modèle linguistique de Hoey

- Modèle fondé sur les **répétitions lexicales**
- **Réseau phrastique** : ensemble d'au moins 3 phrases tel que chaque phrase est appariée avec au moins une phrase de l'ensemble
 - ▶ **Appariement** de 2 phrases si au moins 3 unités lexicales communes
- **Hypotexte** : ensemble des réseaux phrastiques d'un texte
→ Sorte de résumé du texte (suppression des phrases non appariées)

Extrait d'un réseau phrastique



Répétition lexicale : répétition des lemmes des lexèmes des phrases →
Lexèmes : noms, adjectifs, verbes ou adverbes

Résultats expérimentaux sur les réseaux phrastiques

Textes étudiés

- Deux grands textes en anglais (textes expositifs)
 - ▶ “*The Origin of Speech*” [Mac08] : 416 pages
 - ▶ “*Love Online: Emotions on the Internet*” [BZ04] : 302 pages

Texte	Nb phrases	Nb appariements	Nb réseaux phrastiques	% phrases dans hypotexte
<i>Speech</i>	5 308	50 277	2	75,6%
<i>Love</i>	5 571	131 497	2	79,0%

Résultats sur les réseaux phrastiques

- Forte cohésion lexicale des phrases des textes
 - ▶ Beaucoup d'appariements entre phrases (entre 15 et 30, en moyenne)
 - ▶ Peu de réseaux phrastiques
 - ▶ Hypotexte contenant beaucoup de phrases des textes

Résultats expérimentaux sur les réseaux phrastiques

Textes étudiés

- Deux grands textes en anglais (textes expositifs)
 - ▶ “*The Origin of Speech*” [Mac08] : 416 pages
 - ▶ “*Love Online: Emotions on the Internet*” [BZ04] : 302 pages

Texte	Nb phrases	Nb appariements	Nb réseaux phrastiques	% phrases dans hypotexte
<i>Speech</i>	5 308	50 277	2	75,6%
<i>Love</i>	5 571	131 497	2	79,0%

Résultats sur les réseaux phrastiques

- Forte cohésion lexicale des phrases des textes
 - ▶ Beaucoup d'appariements entre phrases (entre 15 et 30, en moyenne)
 - ▶ Peu de réseaux phrastiques
 - ▶ Hypotexte contenant beaucoup de phrases des textes

Résultats expérimentaux sur les réseaux phrastiques

Textes étudiés

- Deux grands textes en anglais (textes expositifs)
 - ▶ “*The Origin of Speech*” [Mac08] : 416 pages
 - ▶ “*Love Online: Emotions on the Internet*” [BZ04] : 302 pages

Texte	Nb phrases	Nb appariements	Nb réseaux phrastiques	% phrases dans hypotexte
<i>Speech</i>	5 308	50 277	2	75,6%
<i>Love</i>	5 571	131 497	2	79,0%

Résultats sur les réseaux phrastiques

- Forte cohésion lexicale des phrases des textes
 - ▶ Beaucoup d'appariements entre phrases (entre 15 et 30, en moyenne)
 - ▶ Peu de réseaux phrastiques
 - ▶ Hypotexte contenant beaucoup de phrases des textes

Résultats expérimentaux sur les réseaux phrastiques

Textes étudiés

- Deux grands textes en anglais (textes expositifs)
 - ▶ “*The Origin of Speech*” [Mac08] : 416 pages
 - ▶ “*Love Online: Emotions on the Internet*” [BZ04] : 302 pages

Texte	Nb phrases	Nb appariements	Nb réseaux phrastiques	% phrases dans hypotexte
<i>Speech</i>	5 308	50 277	2	75,6%
<i>Love</i>	5 571	131 497	2	79,0%

Résultats sur les réseaux phrastiques

- Forte cohésion lexicale des phrases des textes
 - ▶ Beaucoup d'appariements entre phrases (entre 15 et 30, en moyenne)
 - ▶ Peu de réseaux phrastiques
 - ▶ Hypotexte contenant beaucoup de phrases des textes
→ Difficulté à afficher l'hypotexte entier pour l'analyser d'où la nécessité d'en extraire des sous-parties cohérentes

Plan du cours

1 Introduction

2 Traitements de données textuelles

3 Fouille de motifs séquentiels

- Extraction de motifs séquentiels sous contraintes
- Extraction de motifs séquentiels émergents
- Application : analyse stylistique de textes
- Application : analyse de publications biomédicales

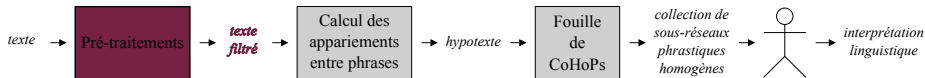
4 Fouille de graphes pour l'exploration de grands textes

- Contexte
- Modèle linguistique de Hoey
- **Fouille de graphes enrichis**

5 Conclusion

6 Références

Proposition : fouille de réseaux phrastiques



Méthodologie

1 Pré-traitement du texte

- ▶ Mise en minuscule des mots du texte et découpage en séquences aux ponctuations de l'ensemble : { « . », « ? », « ! », « : » }
- ▶ Représentation de chaque phrase par les lemmes de ses lexèmes

2 Construction de la représentation du texte sous forme de graphe

3 Extraction de sous-parties de texte cohérentes

4 Interprétation linguistique des sous-réseaux phrastiques

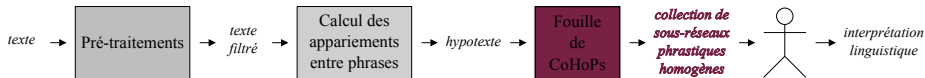
Proposition : fouille de réseaux phrastiques



Méthodologie

- ① Pré-traitement du texte
- ② Construction de la représentation du texte sous forme de graphe
 - ▶ Application du **modèle linguistique de Hoey**
→ Construction de l'hypotexte
- ③ Extraction de sous-parties de texte cohérentes
- ④ Interprétation linguistique des sous-réseaux phrastiques

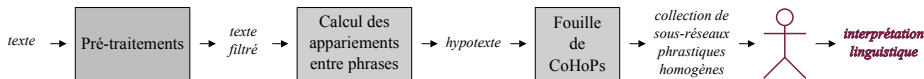
Proposition : fouille de réseaux phrastiques



Méthodologie

- ➊ Pré-traitement du texte
- ➋ Construction de la représentation du texte sous forme de graphe
- ➌ Extraction de sous-parties de texte cohérentes
 - ▶ Hypotexte considéré comme un graphe enrichi
 - ▶ Utilisation de l'algorithme d'extraction des CoHoP pour extraire des collections de sous-réseaux phrastiques homogènes (CoHoSS)
- ➍ Interprétation linguistique des sous-réseaux phrastiques

Proposition : fouille de réseaux phrastiques

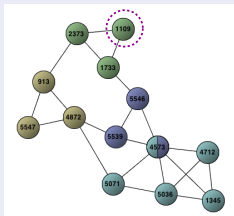


Méthodologie

- ➊ Pré-traitement du texte
- ➋ Construction de la représentation du texte sous forme de graphe
- ➌ Extraction de sous-parties de texte cohérentes
- ➍ **Fouille de réseaux phrastiques**
 - ▶ Sélection par le linguiste des sous-réseaux intéressants découverts automatiquement
 - ▶ Interprétation par le linguiste des sous-réseaux sélectionnés

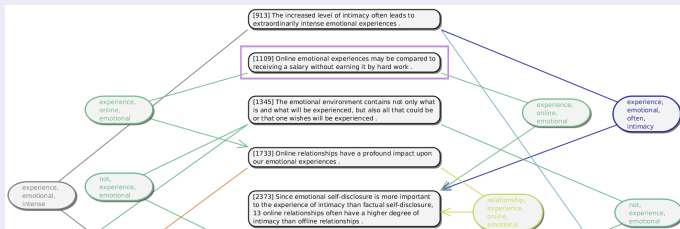
Extraction de sous-réseaux : fouille de CoHoP

Exemple de CoHoP



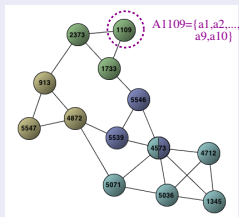
- Sommet du graphe \leftrightarrow phrase du texte
- Attributs du sommet \leftrightarrow lexèmes de la phrase
- CoHoP (collection de k -PC homogènes) \leftrightarrow CoHoSS (collection de sous-réseaux phrastiques homogènes)
 - k -PC (k -clique percolée) \leftrightarrow sous-réseau phrastique

Extrait de la CoHoSS correspondante



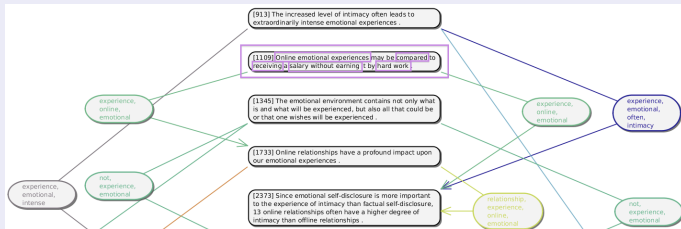
Extraction de sous-réseaux : fouille de CoHoP

Exemple de CoHoP



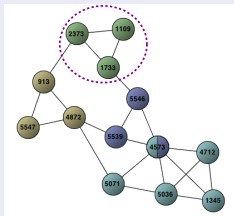
- Sommet du graphe \leftrightarrow phrase du texte
- Attributs du sommet \leftrightarrow lexèmes de la phrase
- CoHoP (collection de k -PC homogènes) \leftrightarrow CoHoSS (collection de sous-réseaux phrastiques homogènes)
 - k -PC (k -clique percolée) \leftrightarrow sous-réseau phrastique

Extrait de la CoHoSS correspondante



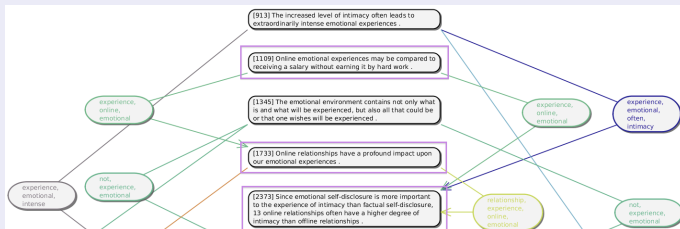
Extraction de sous-réseaux : fouille de CoHoP

Exemple de CoHoP



- Sommet du graphe \leftrightarrow phrase du texte
- Attributs du sommet \leftrightarrow lexèmes de la phrase
- CoHoP (collection de k -PC homogènes) \leftrightarrow CoHoSS (collection de sous-réseaux phrastiques homogènes)
 - k -PC (k -clique percolée) \leftrightarrow sous-réseau phrastique

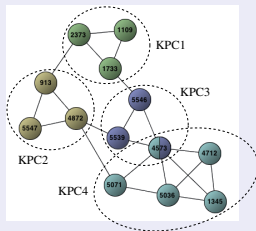
Extrait de la CoHoSS correspondante



Fouille de graphes enrichis : motifs de type CoHoP

- **k -cliques percolées** : version relâchée des k -cliques [DPV05]
 - ▶ Union des k -cliques connectées par chevauchement de $k - 1$ sommets
- **Collection de k -PC homogènes** [MRG12] : ensemble de sommets tel que (k , α et γ sont des entiers prédéfinis)
 - ▶ tous les sommets partagent au moins α attributs
 - ▶ la collection contient au moins γ k -PC
 - ▶ toutes les k -PC avec les mêmes attributs sont dans la collection

Exemple de CoHoP($k = 3$)

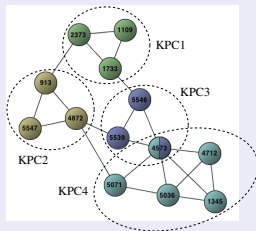


- **4 k -PC ($k = 3$)**
 - ▶ KPC_1 , KPC_2 et KPC_3 avec une k -clique
 - ▶ KPC_4 avec 5 k -cliques

Fouille de graphes enrichis : motifs de type CoHoP

- **k -cliques percolées** : version relâchée des k -cliques [DPV05]
 - ▶ Union des k -cliques connectées par chevauchement de $k - 1$ sommets
- **Collection de k -PC homogènes** [MRG12] : ensemble de sommets tel que (k , α et γ sont des entiers prédéfinis)
 - ▶ tous les sommets partagent au moins α attributs
 - ▶ la collection contient au moins γ k -PC
 - ▶ toutes les k -PC avec les mêmes attributs sont dans la collection

Exemple de CoHoP($k = 3$)

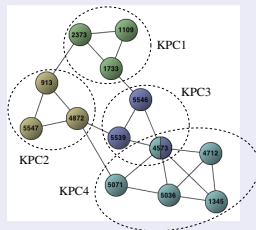


- **4 k -PC ($k = 3$)**
 - ▶ KPC_1 , KPC_2 et KPC_3 avec une k -clique
 - ▶ KPC_4 avec 5 k -cliques

Fouille de graphes enrichis : motifs de type CoHoP

- **k -cliques percolées** : version relâchée des k -cliques [DPV05]
 - ▶ Union des k -cliques connectées par chevauchement de $k - 1$ sommets
- **Collection de k -PC homogènes** [MRG12] : ensemble de sommets tel que (k , α et γ sont des entiers prédéfinis)
 - ▶ tous les sommets partagent au moins α attributs
 - ▶ la collection contient au moins γ k -PC
 - ▶ toutes les k -PC avec les mêmes attributs sont dans la collection

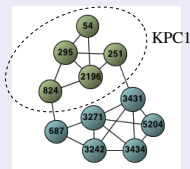
Exemple de CoHoP($k = 3$)



- **4 k -PC ($k = 3$)**
 - ▶ KPC_1 , KPC_2 et KPC_3 avec une k -clique
 - ▶ KPC_4 avec 5 k -cliques

Exemple de CoHoSS : interprétation linguistique

CoHoP extraite ($k = 3$)



CoHoSS extraite à partir de {*adaptation*}

- 2 sous-réseaux phrastiques
 - ▶ KPC_1 : phénomène d'adaptation
 - ▶ KPC_2 : spécialisation de l'hémisphère gauche
- Empan important : 5 150 phrases

Sous-réseau de la CoHoSS correspondante

[54] I take the standpoint of an **evolutionary** biologist who, according to Mayr (1982), "studies the forces that bring about changes in faunas and floras ... [and] studies the steps by which have **evolved** the miraculous **adaptations** so characteristic of every aspect of the organic world" (pp.69 – 70).

[251] An important connotation of the tinkering metaphor, for Jacob, is that **adaptations** exploit whatever is available in order to respond successfully to selection pressures, whether or **not** they originally **evolved** for the use they're now put to.

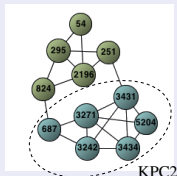
[295] "**language** cannot be as novel as it seems, for **evolutionary adaptation** does **not evolve** out of the blue" (p.7).

[824] Indeed, the same claim about the genes could be made for organisms without **language** and culture, because the **evolutionary** process **involves adaptation** to a particular niche.

[2196] "**language** cannot be as novel as it seems, for **evolutionary adaptations** do **not evolve** out of the blue" (Bickerton, 1990, p.7).

Exemple de CoHoSS : interprétation linguistique

CoHoP extraite ($k = 3$)



CoHoSS extraite à partir de {adaptation}

- 2 sous-réseaux phrastiques
 - KPC_1 : phénomène d'adaptation
 - KPC_2 : spécialisation de l'hémisphère gauche
- Empan important : 5 150 phrases

Sous-réseau de la CoHoSS correspondante

[687] In my **view**, **speech** is an **adaptation** that made the rich message-sending **capacity** of spoken **language** possible.

[3242] The most prevalent **view** of the **origin** of the **hand** – mouth relationship in the latter part of the last century was that the **adaptation** in tool use which occurred in **Homo habilis** about 2 million years ago led to a **left-hemispheric** specialization for manual "praxis" (basically motor skill) and that the first **language** was a gestural **language** built on this basis.

[3271] This led to the **conclusion** that the **origin** of the human **left-hemispheric** praxic specialization, commonly thought to be a basis for the **left-hemisphere speech capacity**, cannot be attributed to the tool-use **adaptation**...

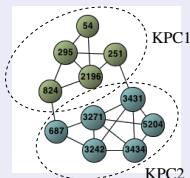
[3431] One implication of the **origin** of a **left-hemisphere** routine-action-control **specialization** in early vertebrates is that this already-existing **left-hemisphere** action **specialization** may have been put to use in the form of the right-side dominance associated with the clinging and leaping motor **adaptation** characteristic of everyday...

[3434] If so, then the **left-hemisphere** action-control **capacity** favoring right-sided **postural** support may have triggered the asymmetric reaching **adaptation** favoring the **hand** on the side less dominant for postural support – the **left hand** – before the manual-predation **specialization** in vertical clingers and leapers,...

[5204] As evidence for the highly specialized nature of this emergent **adaptation**, he cites the **conclusion** of the **postural origins** theory that left-hand preferences for prehension **evolved** in **prosimians** (see Chapter 10).

Exemple de CoHoSS : interprétation linguistique

CoHoP extraite ($k = 3$)



CoHoSS extraite à partir de {adaptation}

- 2 sous-réseaux phrastiques
 - ▶ KPC_1 : phénomène d'adaptation
 - ▶ KPC_2 : spécialisation de l'hémisphère gauche
- Empan important : 5 150 phrases

Sous-réseau de la CoHoSS correspondante

[687] In my **view**, **speech** is an **adaptation** that made the rich message-sending **capacity** of spoken **language** possible.

[3242] The most prevalent **view** of the **origin** of the **hand** – mouth relationship in the latter part of the last century was that the **adaptation** in tool use which occurred in **Homo habilis** about 2 million years ago led to a **left-hemispheric** specialization for manual "praxis" (basically motor skill) and that the first **language** was a gestural **language** built on this basis.

[3271] This led to the **conclusion** that the **origin** of the human **left-hemispheric** praxic specialization, commonly thought to be a basis for the **left-hemisphere speech capacity**, cannot be attributed to the tool-use **adaptation**...

[3431] One implication of the **origin** of a **left-hemisphere** routine-action-control **specialization** in early vertebrates is that this already-existing **left-hemisphere** action **specialization** may have been put to use in the form of the right-side dominance associated with the clinging and leaping motor **adaptation** characteristic of everyday...

[3434] If so, then the **left-hemisphere** action-control **capacity** favoring right-sided **postural** support may have triggered the asymmetric reaching **adaptation** favoring the **hand** on the side less dominant for postural support – the **left hand** – before the manual-predation **specialization** in vertical clingers and leapers,...

[5204] As evidence for the highly specialized nature of this emergent **adaptation**, he cites the **conclusion** of the **postural origins** theory that left-hand preferences for prehension **evolved** in **prosimians** (see Chapter 10).

Conclusion sur les CoHoSS extraites

Conclusion sur le choix des valeurs de k , α et γ

- Paramètre k

- Ajustement du **degré de cohésion lexicale** des sous-réseaux phrastiques
- ⇒ Plus k est grand, plus la cohésion lexicale est grande

- Paramètre α

- Ajustement du nombre minimal d'unités lexicales partagées par les phrases des CoHoSS. En pratique :
- ⇒ α assez petit pour une analyse sur une thématique
- ⇒ α grand, émergence de thématiques plus spécifiques

- Paramètre γ

- Ajustement du nombre minimal de sous-réseaux phrastiques dans les CoHoSS
- ⇒ Plus γ est grand, plus les CoHoSS sont grosses mais peu nombreuses

Conclusion sur les CoHoSS extraites

Conclusion sur le choix des valeurs de k , α et γ

- Paramètre k

- Ajustement du **degré de cohésion lexicale** des sous-réseaux phrastiques
- ⇒ Plus k est grand, plus la cohésion lexicale est grande

- Paramètre α

- Ajustement du **nombre minimal d'unités lexicales partagées** par les phrases des CoHoSS. En pratique :
- ⇒ α assez petit pour une analyse sur une thématique
- ⇒ α grand, émergence de thématiques plus spécifiques

- Paramètre γ

- Ajustement du **nombre minimal de sous-réseaux phrastiques** dans les CoHoSS
- ⇒ Plus γ est grand, plus les CoHoSS sont grosses mais peu nombreuses

Conclusion sur les CoHoSS extraites

Conclusion sur le choix des valeurs de k , α et γ

- Paramètre k

- Ajustement du **degré de cohésion lexicale** des sous-réseaux phrastiques
- ⇒ Plus k est grand, plus la cohésion lexicale est grande

- Paramètre α

- Ajustement du **nombre minimal d'unités lexicales partagées** par les phrases des CoHoSS. En pratique :
- ⇒ α assez petit pour une analyse sur une thématique
- ⇒ α grand, émergence de thématiques plus spécifiques

- Paramètre γ

- Ajustement du **nombre minimal de sous-réseaux phrastiques** dans les CoHoSS
- ⇒ Plus γ est grand, plus les CoHoSS sont grosses mais peu nombreuses

Plan du cours

- 1 Introduction
- 2 Traitements de données textuelles
- 3 Fouille de motifs séquentiels
- 4 Fouille de graphes pour l'exploration de grands textes
- 5 Conclusion**
- 6 Références

Conclusion

- Fouille de motifs séquentiels
 - ▶ Extraction de motifs d'items ou d'itemsets à partir de bases de séquences
 - ▶ Différentes contraintes possibles pour limiter le nombre de motifs extraits
- Application à l'analyse stylistique de textes littéraires
 - ▶ Motifs d'items avec *gap* : capacité de généralisation et extraction automatique de patrons avec éléments fixes et variables
 - ▶ Motifs d'itemsets : catégorie morpho-syntaxique des éléments variables automatiquement connue → obtention directe de patrons grammaticaux
- Application à l'analyse de publications biomédicales
 - ▶ Motifs d'itemsets avec contraintes supplémentaires d'appartenance, de portée et d'association → obtention directe de patrons linguistiques
- Fouille de graphes
 - ▶ Représentation du texte sous forme de graphe par application du modèle linguistique de Hoey
 - ▶ Extraction de collections de sous-réseaux phrastiques homogènes (CoHoSS) : phrases appariées et partageant un ensemble de lexèmes

Conclusion

- Fouille de motifs séquentiels

- ▶ Extraction de motifs d'items ou d'itemsets à partir de bases de séquences
- ▶ Différentes contraintes possibles pour limiter le nombre de motifs extraits

- Application à l'analyse stylistique de textes littéraires

- ▶ Motifs d'items avec *gap* : capacité de généralisation et extraction automatique de patrons avec éléments fixes et variables
- ▶ Motifs d'itemsets : catégorie morpho-syntaxique des éléments variables automatiquement connue → obtention directe de patrons grammaticaux

- Application à l'analyse de publications biomédicales

- ▶ Motifs d'itemsets avec contraintes supplémentaires d'appartenance, de portée et d'association → obtention directe de patrons linguistiques

- Fouille de graphes

- ▶ Représentation du texte sous forme de graphe par application du modèle linguistique de Hoey
- ▶ Extraction de collections de sous-réseaux phrastiques homogènes (CoHoSS) : phrases appariées et partageant un ensemble de lexèmes

Conclusion

- Fouille de motifs séquentiels

- ▶ Extraction de motifs d'items ou d'itemsets à partir de bases de séquences
- ▶ Différentes contraintes possibles pour limiter le nombre de motifs extraits

- Application à l'analyse stylistique de textes littéraires

- ▶ Motifs d'items avec *gap* : capacité de généralisation et extraction automatique de patrons avec éléments fixes et variables
- ▶ Motifs d'itemsets : catégorie morpho-syntaxique des éléments variables automatiquement connue → obtention directe de patrons grammaticaux

- Application à l'analyse de publications biomédicales

- ▶ Motifs d'itemsets avec contraintes supplémentaires d'appartenance, de portée et d'association → obtention directe de patrons linguistiques

- Fouille de graphes

- ▶ Représentation du texte sous forme de graphe par application du modèle linguistique de Hoey
- ▶ Extraction de collections de sous-réseaux phrastiques homogènes (CoHoSS) : phrases appariées et partageant un ensemble de lexèmes




Conclusion

- Fouille de motifs séquentiels
 - ▶ Extraction de motifs d'items ou d'itemsets à partir de bases de séquences
 - ▶ Différentes contraintes possibles pour limiter le nombre de motifs extraits
- Application à l'analyse stylistique de textes littéraires
 - ▶ Motifs d'items avec *gap* : capacité de généralisation et extraction automatique de patrons avec éléments fixes et variables
 - ▶ Motifs d'itemsets : catégorie morpho-syntaxique des éléments variables automatiquement connue → obtention directe de patrons grammaticaux
- Application à l'analyse de publications biomédicales
 - ▶ Motifs d'itemsets avec contraintes supplémentaires d'appartenance, de portée et d'association → obtention directe de patrons linguistiques
- Fouille de graphes
 - ▶ Représentation du texte sous forme de graphe par application du *modèle linguistique de Hoey*
 - ▶ Extraction de *collections de sous-réseaux phrastiques homogènes* (CoHoSS) : phrases appariées et partageant un ensemble de lexèmes








Plan du cours

- 1 Introduction
- 2 Traitements de données textuelles
- 3 Fouille de motifs séquentiels
- 4 Fouille de graphes pour l'exploration de grands textes
- 5 Conclusion
- 6 Références**

Références I

-  R. Agrawal and R. Srikant, *Mining sequential patterns*, Proc. of ICDE, 1995, pp. 3–14.
-  Nicolas Béchet, Peggy Cellier, Thierry Charnois, Bruno Crémilleux, and Marie-Christine Jaulent, *Sequential pattern mining to discover relations between genes and rare diseases.*, CBMS, IEEE Computer Society, 2012, pp. 1–6.
-  A. Ben-Ze'ev, *Love Online : Emotions on the Internet*, Cambridge Univ. Pr., 2004.
-  G. Dong and J. Li, *Efficient minig of emerging patterns : Discovering trends and differences*, Proc. of SIGKDD, 1999, pp. 43–52.
-  I. Derenyi, G. Palla, and T. Vicsek, *Clique percolation in random networks*, Physical Review Letters **94** (2005).
-  M. Hoey, *Patterns of Lexis in Text*, Describing English Language, Oxford Univ. Pr., 1991.

Références II

-  R. Krovetz, *Viewing Morphology as an Inference Process*, Proc. of ACM SIGIR Conference, 1993, pp. 191–202.
-  P. MacNeilage, *The Origin of Speech*, UOP Oxford, 2008.
-  P.-N. Mougél, C. Rigotti, and O. Gandrillon, *Finding collections of k -clique percolated components in attributed graphs*, Proc. of PAKDD, 2012.
-  M. Nanni and C. Rigotti, *Extracting trees of quantitative serial episodes*, Proc. of KDID, 2007, pp. 170–188.
-  M.F. Porter, *An Algorithm for Suffix Stripping*, Program **14** (1980), no. 3, 130–137.
-  S. Quiniou, P. Cellier, T. Charnois, and D. Legallois, *Fouille de données pour la stylistique : cas des motifs séquentiels émergents*, Actes de JADT (Liège, Belgique), Juin 2012, pp. 821–833.
-  ———, *What about sequential data mining techniques to identify linguistic patterns for stylistics ?*, Proc. of CICLing (New Delhi, India), March 2012, pp. 166–177.

Références III



_____, *Graph mining under linguistic constraints to explore large texts*, Proc. of CICLing, 2013.



A.J. Renouf and J.M. Sinclair, *English corpus linguistics : Studies in honour of jan svartvik*, ch. Collocational Frameworks in English, pp. 128–143, Longman, 1991.



Ayush Singhal, Robert Leaman, Natalie L. Catlett, Thomas Lemberger, Johanna R. McEntyre, Shawn W. Polson, Ioannis Xenarios, Cecilia N. Arighi, and Zhiyong Lu, *Pressing needs of biomedical text mining in biocuration and beyond : opportunities and challenges.*, Database (2016).



Anne E. Thessen, Hong Cui, and Dmitry Mozzherin, *Applications of natural language processing in biodiversity science.*, Adv. Bioinformatics **391574** (2012), 1–17.



X. Yan, J. Han, and R. Afshar, *Clospan : Mining closed sequential patterns in large databases*, Proc. of SDM, 2003.