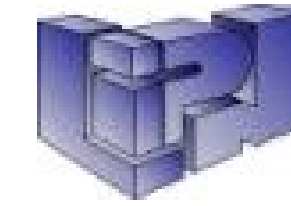


GraphCorpus : un outil d'aide à l'analyse de corpus



Laboratoire d'Informatique de Paris Nord
Université Paris 13 et CNRS UMR 7030
93430 Villetaneuse

Problématique

Analyse automatique de corpus (dépêches)

- Comment **regrouper les documents qui traitent d'un même événement** sans effectuer de traitements lourds ?

Ce type d'analyse présente un intérêt pour :

- La **veille d'informations** (possibilité d'avoir une vue synthétique du traitement d'un événement par un média)
- La **synthèse automatique** (regrouper les documents qui traitent d'un même événement peut constituer le point de départ d'une tâche de synthèse)

Aspects applicatifs

Donner aux analystes des outils pour les aider à prendre connaissance de différents aspects d'un corpus :

Sa structure

- Combien d'événements traités
- Combien de documents traitent d'un même événement les différents événements ont-ils des liens entre eux

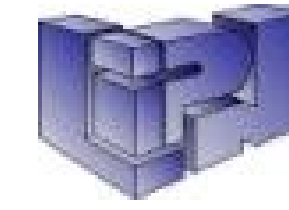
Son contenu (description des événements)

- Cette aide doit être amenée de manière automatique, sans intervention de l'utilisateur.

Techniques utilisées

- Les documents sont organisés dans un **graphe dont les arêtes sont pondérées**.
- Le **poids des arêtes est fonction du nombre d'entités nommées** que les documents partagent
- Les entités nommées sont typées : **un poids est affecté à chaque type d'entité nommée** (pour l'instant manuellement) selon l'aspect discriminant du type (une date est-elle plus importante qu'un nom de personne ou qu'un nom de lieu pour décrire un événement ?)
- L'organisation du graphe est calculée grâce à l'**algorithme SOM (Self-Organizing Map)** dont l'entrée est la matrice de similarité des documents qui composent le corpus.

GraphCorpus : un outil d'aide à l'analyse de corpus



Laboratoire d'Informatique de Paris Nord
Université Paris 13 et CNRS UMR 7030
93430 Villetaneuse

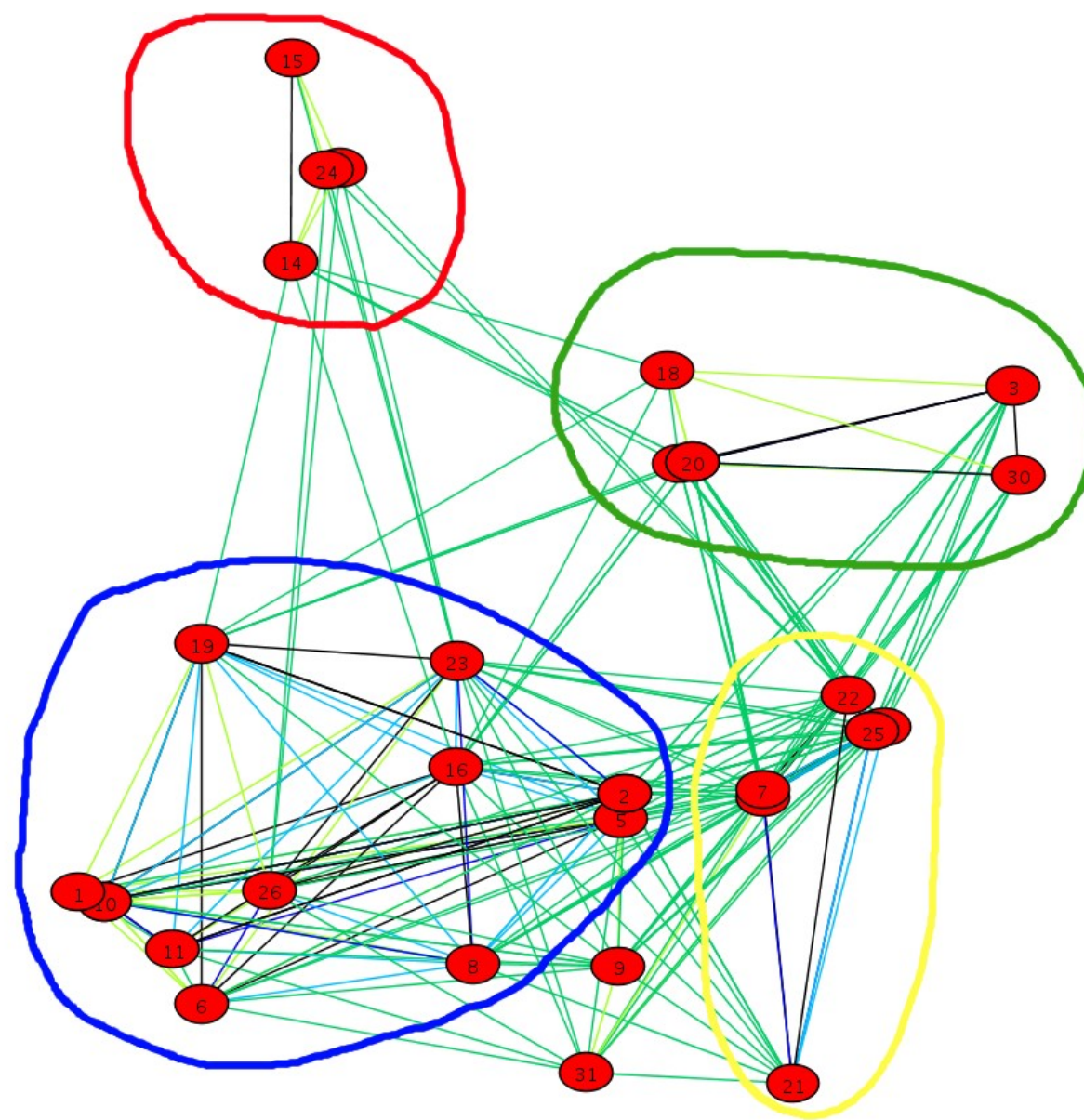
Les réalisations

Les **entités nommées** sont analysées grâce au logiciel **TagEN** développé au LIPN

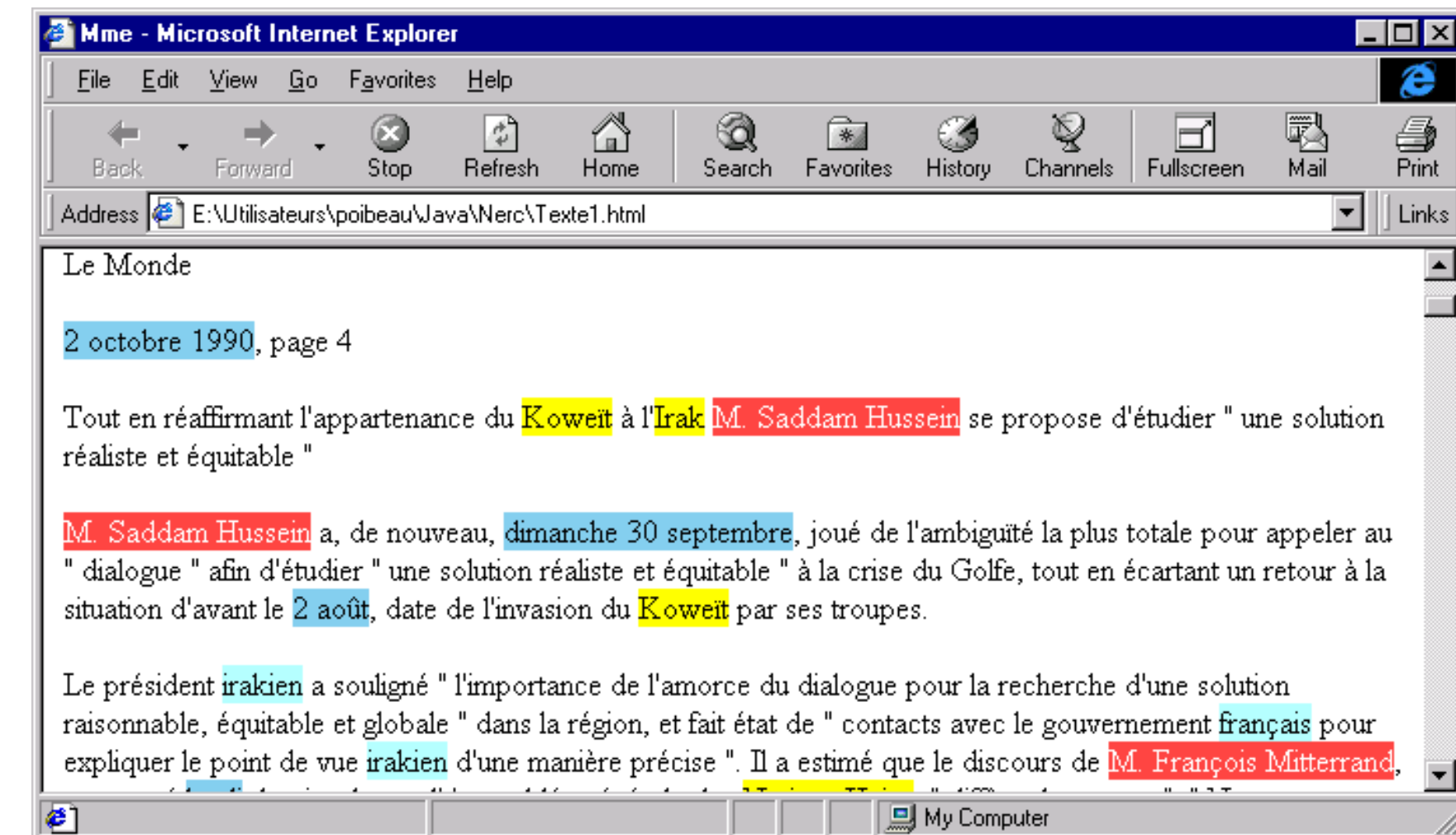
(<http://www-lipn.univ-paris13.fr/~poibeau/tagen.html>)

A partir de cette analyse, **GraphCorpus** fournit une **cartographie** sous forme de **carte SOM**, grâce à GraphExplore

<http://graphexplore.cagp.duke.edu>)



GraphCorpus : Cartographie des liens entre documents au sein d'un corpus (chaque point en rouge correspond à un document ; la proximité des points, le nombre et la couleur des liens sont autant de facteurs pertinents pour l'analyste).



TagEN : Outil d'analyse et de visualisation des entités au sein d'un document

Perspectives

- Déterminer automatiquement le poids des types d'entités nommées
- Proposer un *clustering* automatique des documents
- Proposer une description d'un cluster grâce à des nuages de mots (termes et entités nommées récupérés dans les différents documents qui composent un cluster)