Kodex : comment organiser les résultats d'une recherche web par analyse de graphe

Benoit Gaillard, Emmanuel Navarro

5 décembre 2011

Intervention en deux parties

- Détection de communautés sur un graphe biparti documents-mots (Emmanuel Navarro)
- Labellisation endogène de clusters de documents (Benoît Gaillard)

Google

japan

Everything

Images

Maps

Videos

News

Shopping

More

All results

Sites with images

More search tools

Japan - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/**Japan** +1

Japan Listen/dʒə'pæn/ (**Japanese**: 日本 Nihon or Nippon; formally 日本国 About this sound Nippon-koku or Nihon-koku, literally, the State of **Japan**) is an **...**

Names of Japan - Geography of Japan - History of Japan - Culture

Japan National Tourism Organization Web Site

www.jnto.go.jp/ +1

JNTO is involved in a broad range of activities promoting travel to **Japan** through various activities overseas as well as tourism-promoting activities in **Japan**.

japan-guide.com - **Japan** Travel and Living Guide

www.**japan**-guide.com/ +1

Everything about modern and traditional **Japan** with emphasis on travel and living related information.

Japan maps.google.com

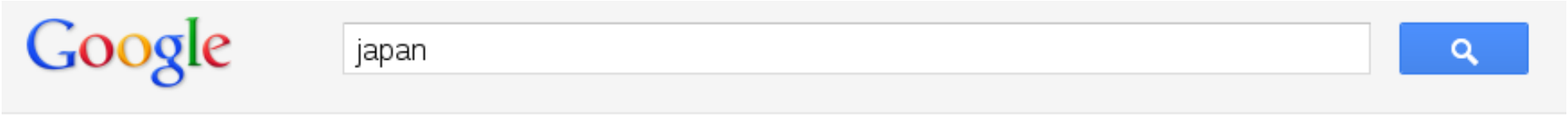
Japan Travel Information and Travel Guide - Lonely Planet

www.lonelyplanet.com/**japan** +1

Japan tourism and travel information including facts, maps, history, culture, transport and weather in **Japan**. Find popular places to visit in **Japan** - Lonely Planet.

Japan News - Earthquake, Tsunami and Nuclear Crisis (2011)

topics.nytimes.com › World › Countries and Territories +1



Everything
Images
Maps
Videos
News
Shopping
More

Japan - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Japan** +1
Japan Listen/dʒə'pæn/ (**Japanese**: 日本 Nihon or Nippon; formally 日本国 About this sound Nippon-koku or Nihon-koku, literally, the State of **Japan**) is an **...**
Names of Japan - Geography of Japan - History of Japan - Culture

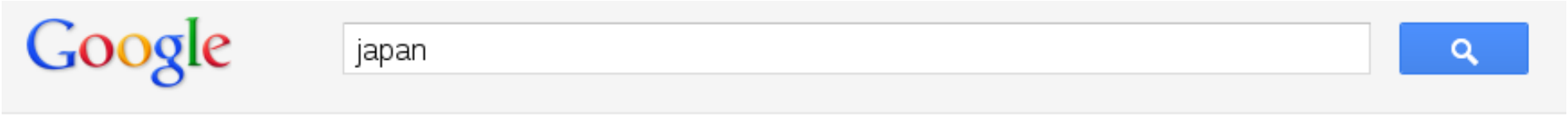
Japan National Tourism Organization Web Site
www.jnto.go.jp/ +1
JNTO is involved in a broad range of activities promoting travel to **Japan** through various activities overseas as well as tourism-promoting activities in **Japan**.

japan-guide.com - **Japan** Travel and Living Guide
www.**japan**-guide.com/ +1

- Liste ordonnée !
- Pourtant :
  - Polysémie des requêtes,
  - Structure dans la listes des documents.
- Rendre visible cette structure à l'utilisateur,
- Lui permettre de raffiner sa recherche.

Japan Travel Information and Travel Guide - Lonely Planet
www.lonelyplanet.com/**japan** +1
Japan tourism and travel information including facts, maps, history, culture, transport and weather in **Japan**. Find popular places to visit in **Japan** - Lonely Planet.

Japan News - Earthquake, Tsunami and Nuclear Crisis (2011)
topics.nytimes.com › World › Countries and Territories +1



Everything
Images
Maps
Videos
News
Shopping
More

Japan - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Japan** +1
Japan Listen/dʒə'pæn/ (**Japanese**: 日本 Nihon or Nippon; formally 日本国 About this sound Nippon-koku or Nihon-koku, literally, the State of **Japan**) is an **...**
Names of Japan - Geography of Japan - History of Japan - Culture

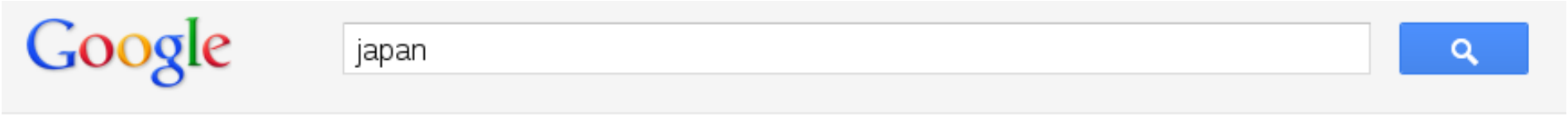
Japan National Tourism Organization Web Site
www.jnto.go.jp/ +1
JNTO is involved in a broad range of activities promoting travel to **Japan** through various activities overseas as well as tourism-promoting activities in **Japan**.

japan-guide.com - **Japan** Travel and Living Guide
www.**japan**-guide.com/ +1

- Liste ordonnée !
- Pourtant :
  - Polysémie des requêtes,
  - Structure dans la listes des documents.
- Rendre visible cette structure à l'utilisateur,
- Lui permettre de raffiner sa recherche.

Japan Travel Information and Travel Guide - Lonely Planet
www.lonelyplanet.com/**japan** +1
Japan tourism and travel information including facts, maps, history, culture, transport and weather in **Japan**. Find popular places to visit in **Japan** - Lonely Planet.

Japan News - Earthquake, Tsunami and Nuclear Crisis (2011)
topics.nytimes.com › World › Countries and Territories +1



Everything
Images
Maps
Videos
News
Shopping
More

Japan - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Japan** +1
Japan Listen/dʒə'pæn/ (**Japanese**: 日本 Nihon or Nippon; formally 日本国 About this sound Nippon-koku or Nihon-koku, literally, the State of **Japan**) is an **...**
Names of Japan - Geography of Japan - History of Japan - Culture

Japan National Tourism Organization Web Site
www.jnto.go.jp/ +1
JNTO is involved in a broad range of activities promoting travel to **Japan** through various activities overseas as well as tourism-promoting activities in **Japan**.

japan-guide.com - **Japan** Travel and Living Guide
www.**japan**-guide.com/ +1

- Liste ordonnée !
- Pourtant :
  - Polysémie des requêtes,
  - Structure dans la listes des documents.
- Rendre visible cette structure à l'utilisateur,
- Lui permettre de raffiner sa recherche.

Japan Travel Information and Travel Guide - Lonely Planet
www.lonelyplanet.com/**japan** +1
Japan tourism and travel information including facts, maps, history, culture, transport and weather in **Japan**. Find popular places to visit in **Japan** - Lonely Planet.

Japan News - Earthquake, Tsunami and Nuclear Crisis (2011)
topics.nytimes.com › World › Countries and Territories +1

- Liste ordonnée !
- Pourtant :
  - Polysémie des requêtes,
  - Structure dans la listes des documents.
- Rendre visible cette structure à l'utilisateur,
- Lui permettre de raffiner sa recherche.

japan search options

30 results for "japan"

japanese, china, asia, history, government,

travel, guide, sightseeing, tourism, information,

news, business, network, photo, sports,

airlines, flight, purchase, japan domestic flight, jal group,

nuclear, tsunami, crisis, nuclear power plant, earthquake,

Japan - Wikipedia
Find information about the history, politics, geography, economy, and culture of **Japan**.
http://en.wikipedia.org/wiki/Japan

Japan National Tourism Organization |
Japan is situated in northeastern Asia between the North Pacific and the Sea of **Japan**. **... Japan** consists of four major islands, surrounded by more than 4,000 **...**
http://www.jnto.go.jp/eng/

Japan Today
Japan Today is an international news network covering news, politics, business, sports, technology, and more.
http://www.japantoday.com/

VISIT **JAPAN** 2011
It presents **Japan's** history and culture, covers events, festivities, tourism, food and shopping, and provides other kinds of helpful information. Come...
http://www.visitjapan.jp/

Japan-guide.com
Everything about modern and traditional **Japan** with emphasis on travel and living related information.
http://www.japan-guide.com/

Official Tourism Guide for **Japan** Travel
Japan National Tourist Organization offers information on transportation, lodging, restaurants, tourist attractions, culture, history, festivals, and **...**
http://www.japantravelinfo.com/

japan search options

30 results for "japan"

-->

japanese, china, asia, history, government,

travel, guide, sightseeing, tourism, information,

news, business, network, photo, sports,

airlines, flight, purchase, japan domestic flight, jal group,

nuclear, tsunami, crisis, nuclear power plant, earthquake,

Japan.org
SENDAI, **Japan** — A strong new earthquake rattled **Japan's** northeast Monday as the government urged more people living near a tsunami-crippled nuclear **...**
http://www.japan.org/

Japan News - Earthquake, Tsunami and Nuclear Crisis (2011)
World news about **Japan**, including breaking news about the March 11, 2011 earthquake and tsunami and the nuclear crisis from The New York Times.
http://topics.nytimes.com/top/news/international/countriesandterritories/japan/index.html

Embassy of **Japan**, Addis Ababa, ETHIOPIA
Japan has the experience of a relatively fast recovery from the devastation of the war and has overcome various natural disasters in the past. **...**
http://www.et.emb-japan.go.jp/

Japan: News & Videos about **Japan** - CNN.com
The nuclear crisis wreaks havoc on one of **Japan's** prized exports: green tea. **...** Matador's destination expert on **Japan** lays out the country's avoidable attractions ...
http://topics.cnn.com/topics/Japan

Special report: **Japan's** throwaway nuclear workers | Reuters
FUKUSHIMA, **Japan** (Reuters) - A decade and a half before it blew apart in a hydrogen blast that punctuated the worst nuclear accident since Chernobyl, **...**
http://www.reuters.com/article/2011/06/24/us-japan-nuclear-idUSTRE75N18A20110624

Récapitulatif

- Organiser les résultats d'une recherche web
- Méthode : clustering de graphe
- Projet dans le cadre du programme Quaero [1]

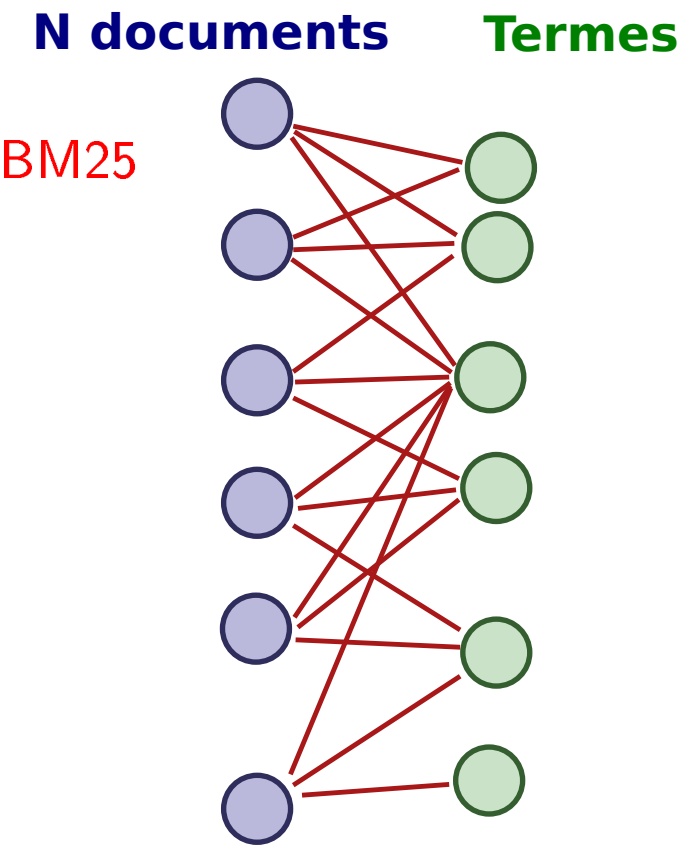
.. et donc :

- Comment construire un graphe ? un "bigraphe" ?
- Comment fonctionne le clustering ?
- (comment donner des étiquettes aux clusters ?)
- Comment évaluer la méthode ?

1. 1er version du projet présentée à Coria2011 [?]
2. Nouvelle version en cours de développement...

1. http://www.quaero.org/

Construction du graphe Documents/Mots



Pour chaque document : les 50 termes de plus fort BM25

Filtres :

- $df => 100$
- $df_{local} <= 90\%$
- $df_{local} > 1$

Pondération par : $tf.idf_{local} = tf.\log(\frac{1}{df_{local}})$

Note : on a un graphe **biparti** (projection possible).

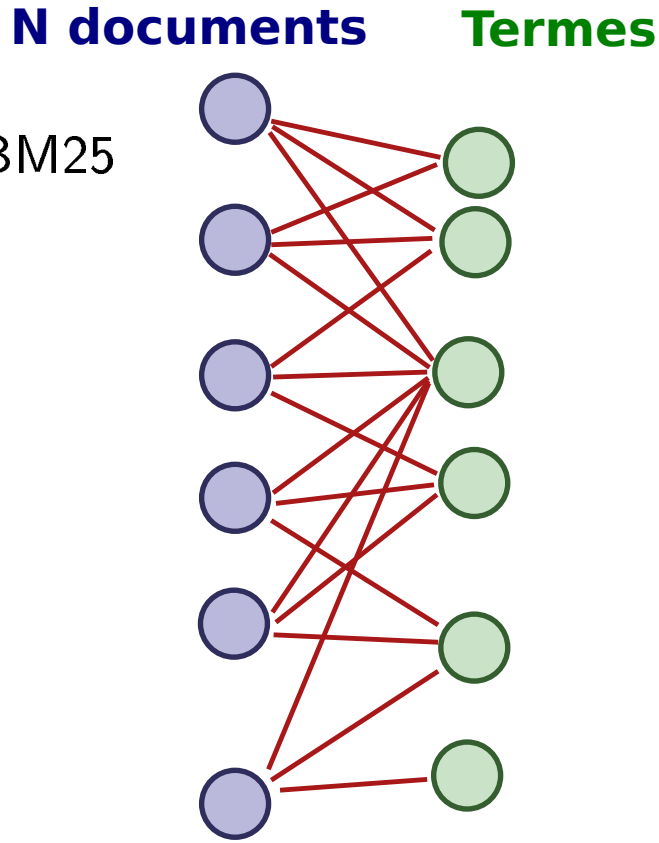
Construction du graphe Documents/Mots

Pour chaque document : les 50 termes de plus fort BM25

Filtres :

- $df => 100$
- $df_{local} <= 90\%$
- $df_{local} > 1$

Pondération par : $tf.idf_{local} = tf.\log(\frac{1}{df_{local}})$



Note : on a un graphe **biparti** (projection possible).

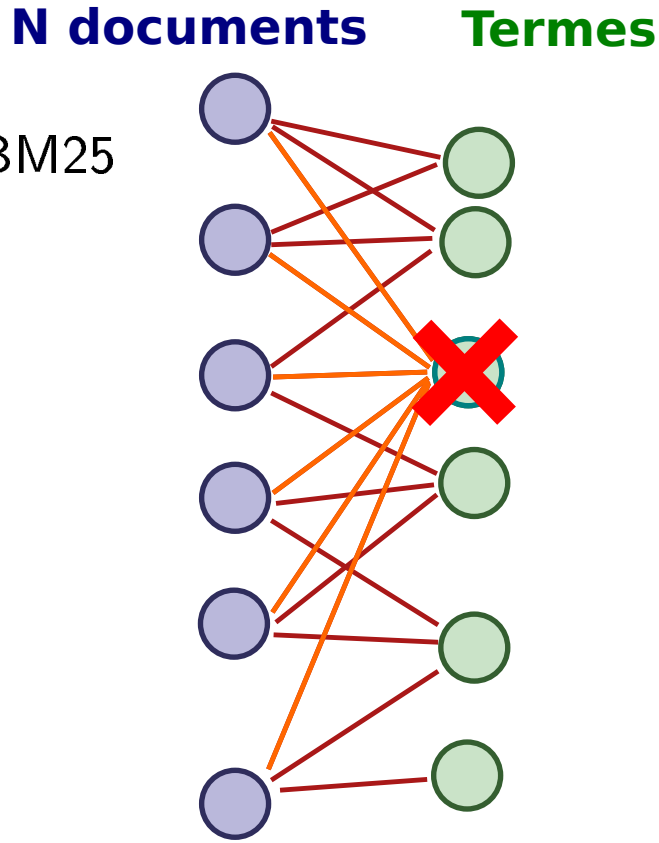
Construction du graphe Documents/Mots

Pour chaque document : les 50 termes de plus fort BM25

Filtres :

- $df => 100$
- $df_{local} <= 90\%$
- $df_{local} > 1$

Pondération par : $tf.idf_{local} = tf.\log(\frac{1}{df_{local}})$



Note : on a un graphe **biparti** (projection possible).

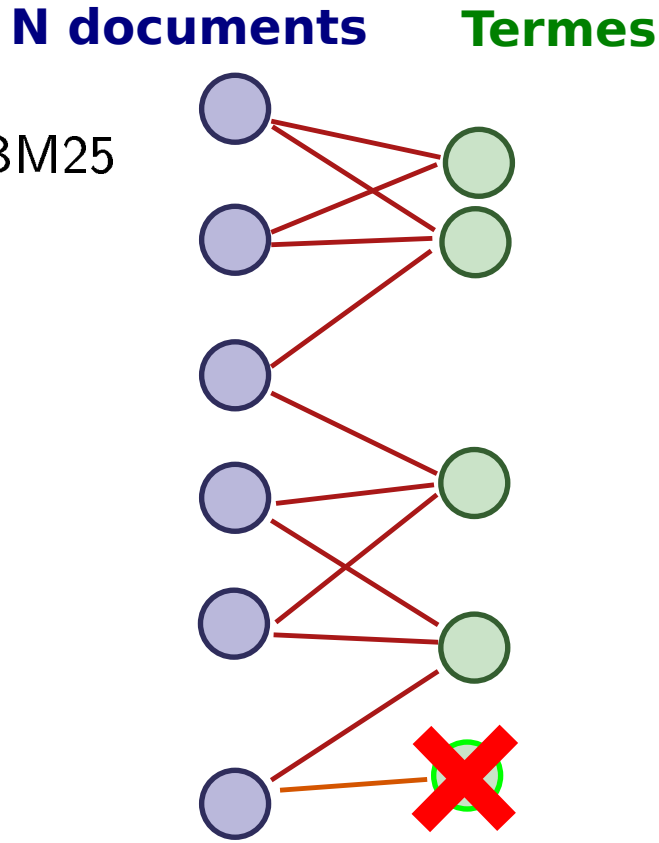
Construction du graphe Documents/Mots

Pour chaque document : les 50 termes de plus fort BM25

Filtres :

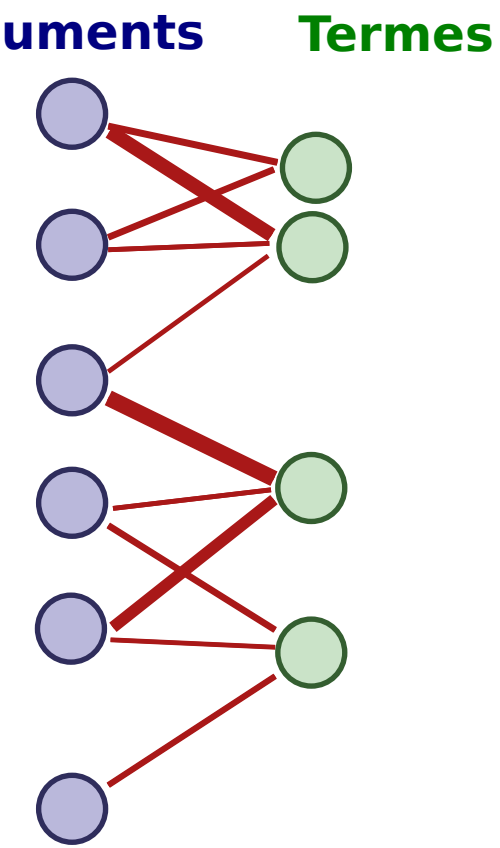
- $df => 100$
- $df_{local} <= 90\%$
- $df_{local} > 1$

Pondération par : $tf.idf_{local} = tf.\log(\frac{1}{df_{local}})$



Note : on a un graphe **biparti** (projection possible).

Construction du graphe Documents/Mots



Pour chaque document : les 50 termes de plus fort BM25

Filtres :

- $df => 100$
- $df_{local} <= 90\%$
- $df_{local} > 1$

Pondération par : $tf.idf_{local} = tf.\log(\frac{1}{df_{local}})$

Note : on a un graphe **biparti** (projection possible).

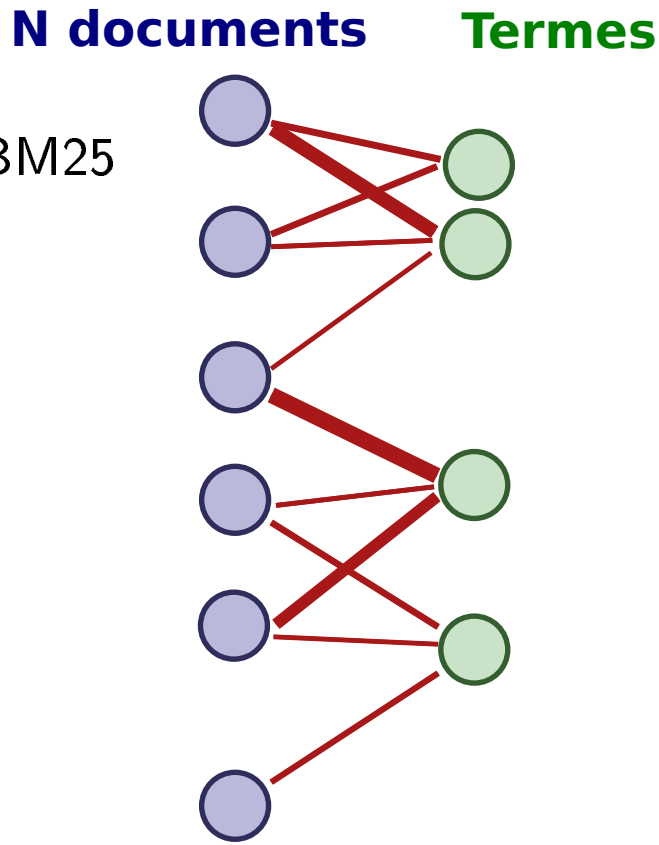
Construction du graphe Documents/Mots

Pour chaque document : les 50 termes de plus fort BM25

Filtres :

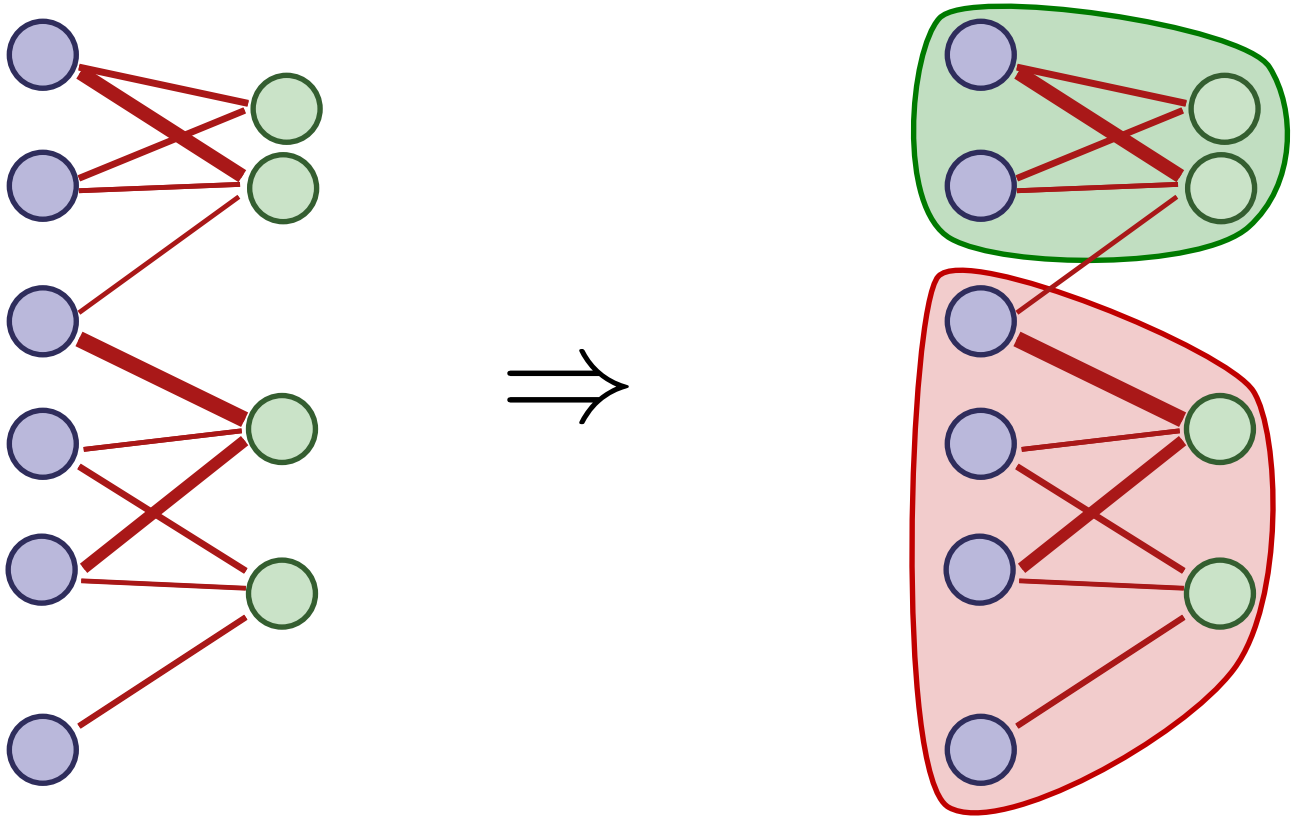
- $df => 100$
- $df_{local} <= 90\%$
- $df_{local} > 1$

Pondération par : $tf.idf_{local} = tf.\log(\frac{1}{df_{local}})$



Note : on a un graphe **biparti** (projection possible).

clustering du graphe ?



Clustering de graphe *ou* Détection de communautés

Le problème : repérer les structures mésoscopiques porteuses de sens.

graphes lexicaux : *concepts*,
graphes de documents : *thèmes*,
graphes sociaux : *communautés*, *familles*, *clubs*, *etc...*

communauté $\simeq$ *cluster* :

Sous-ensemble de sommets
"plus fortement" connectés entre eux qu'au reste du graphe.

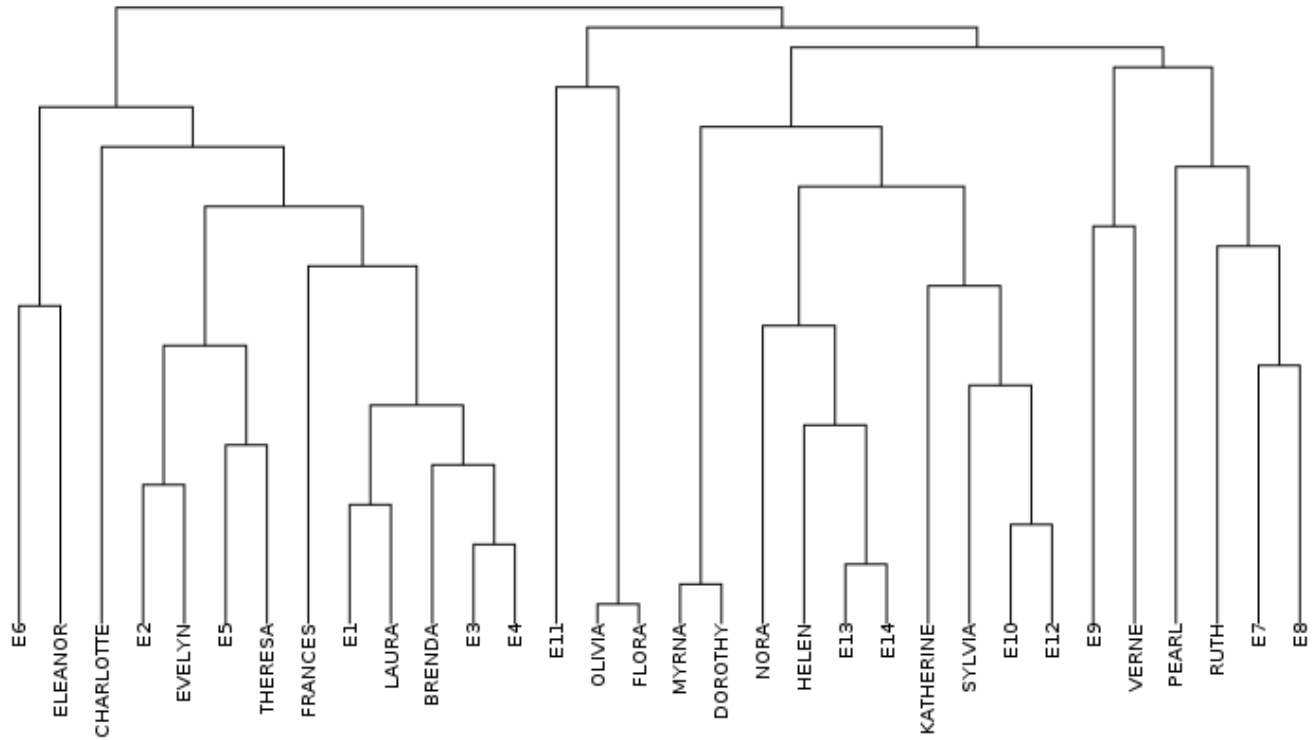
mais pas de définition formelle générale.

- énormément de méthodes proposées depuis 2003 (après découverte des propriétés communes des grands graphes de terrain)
- héritier des problèmes de *clustering* (data mining) et de *partitionnement de graphes* (informatique),
- apprentissage non supervisé.

Article fondateur : [?]

Une méthode parmi (beaucoup) d'autres : Walktrap

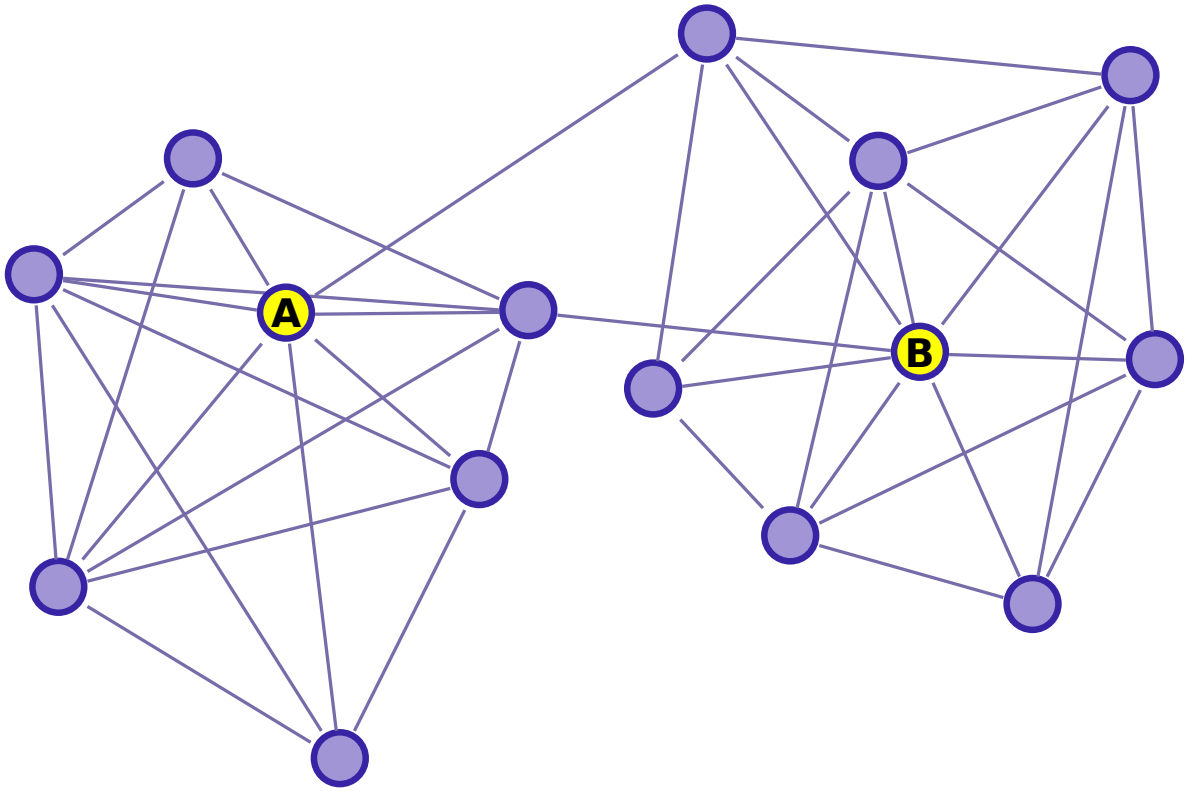
1. Calcul d'une similarité entre sommets,
2. Clustering hiérarchique,
3. Coupe du dendrogramme (optimisation de la modularité).



Comment calculer une distance/similarité pertinente entre les sommets d'un graphe ?

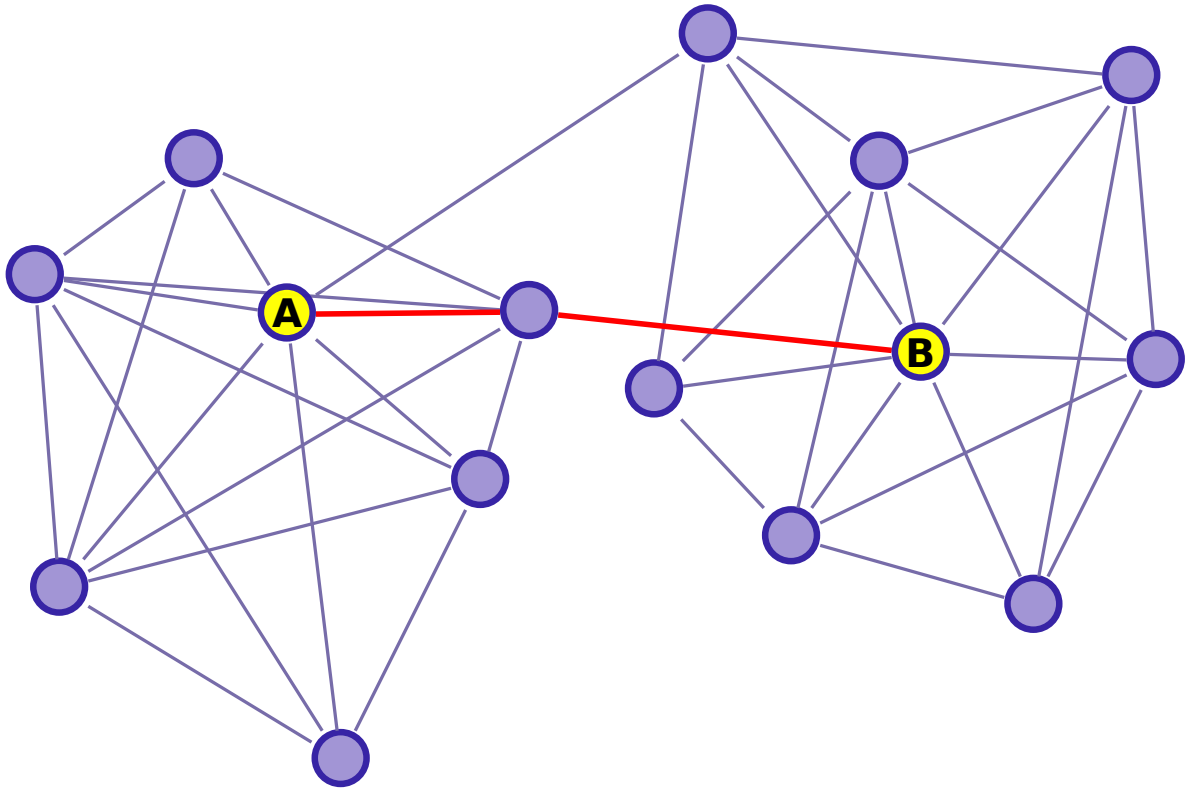
- dans graphe uni-parti,
- puis dans un graphe biparti.

Comment mesurer une similarité entre sommets ?



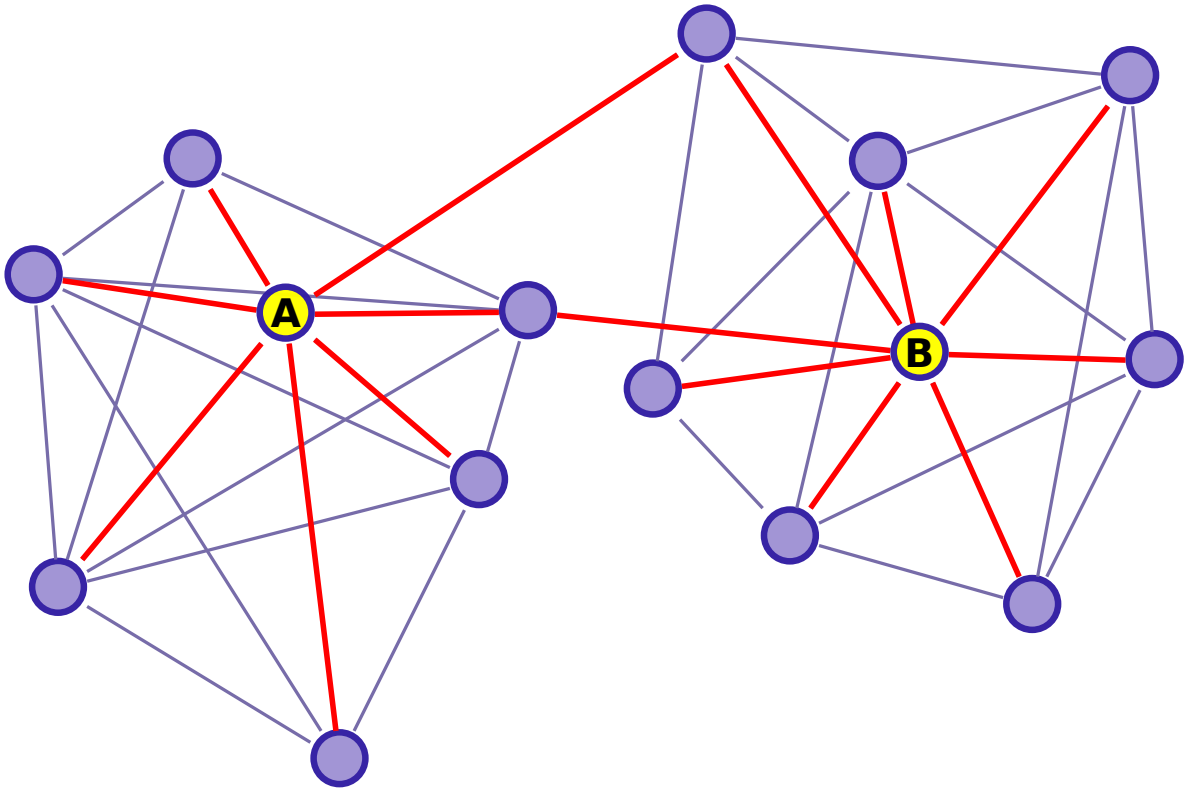
Quel est la similarité entre A et B (par exemple) ?

Comment mesurer une similarité entre sommets ?



Distance géodésique : nombre d'arêtes du plus court chemin.
MAIS les graphes sont des *Petits Mondes* : $d_T(A, B) \precsim 4$
La mesure n'est pas discriminatoire.

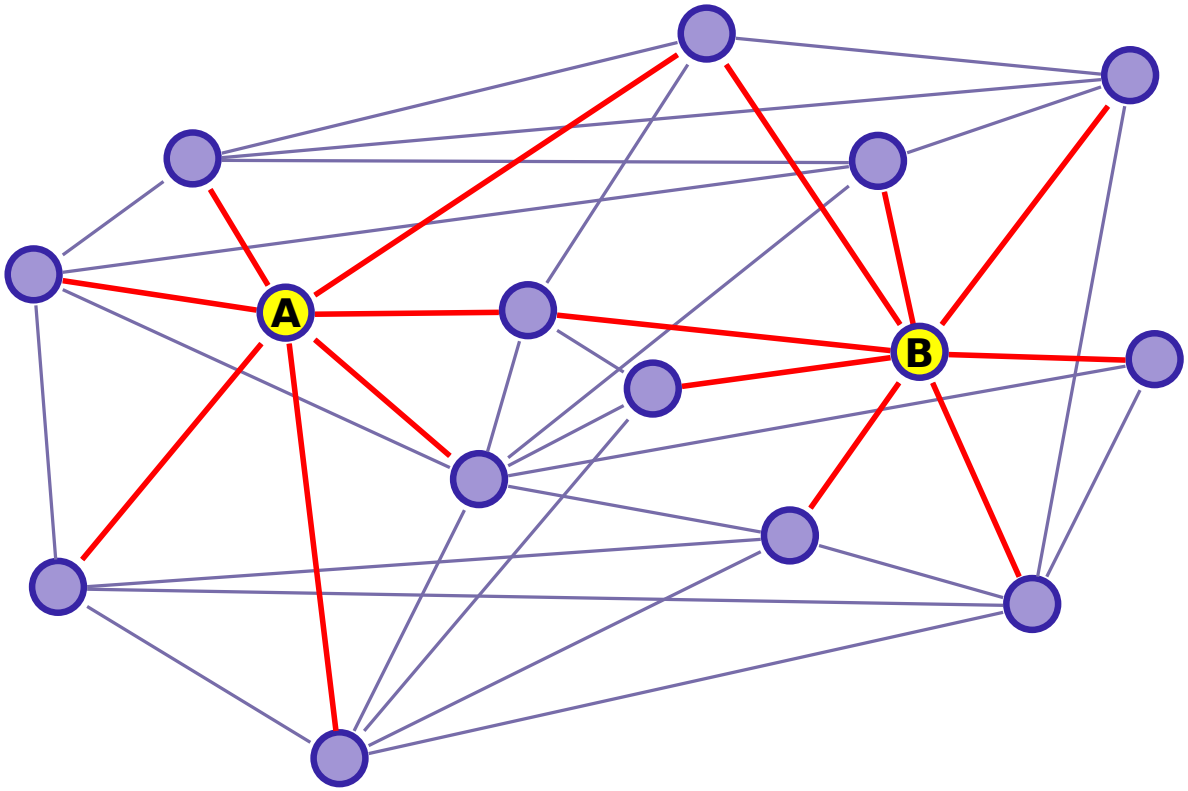
Comment mesurer une similarité entre sommets ?



On peut utiliser les vecteurs d'adjacence ? (modèle "vectoriel")

$$A = [1,1,1,1,1,1,1,1,0,0,0,0,0,0,0]$$
$$B = [0,0,0,0,0,0,1,1,1,1,1,1,1,1,1]$$

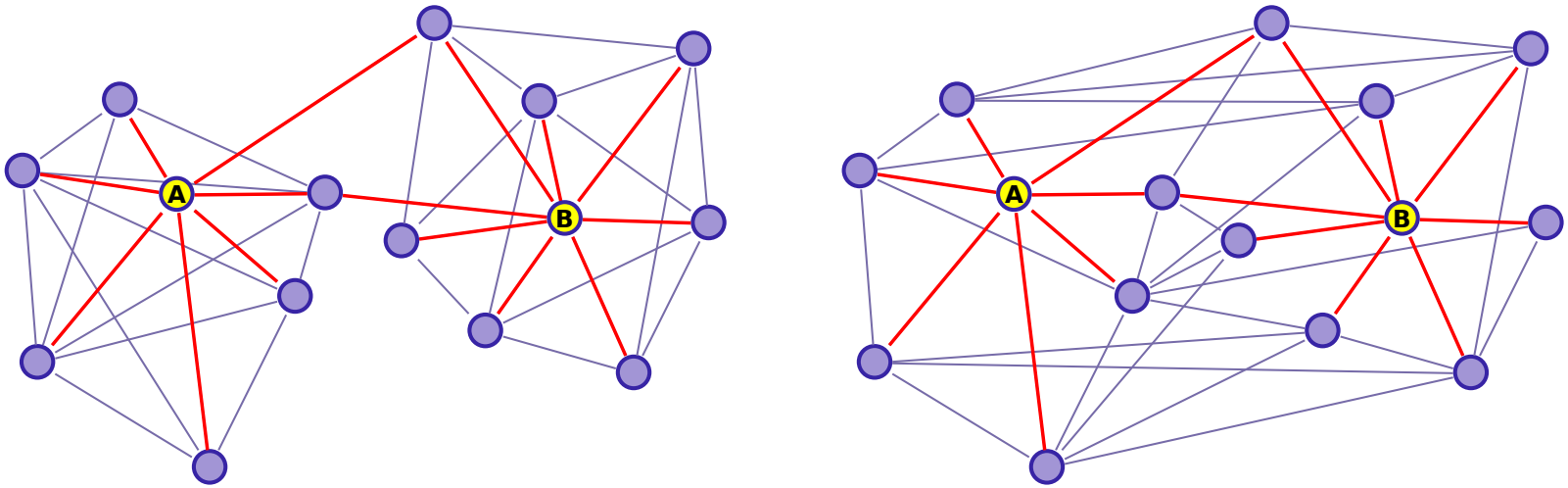
Comment mesurer une similarité entre sommets ?



On peut utiliser les vecteurs d'adjacence ? (modèle "vectoriel")

$$A = [1,1,1,1,1,1,1,1,0,0,0,0,0,0,0]$$
$$B = [0,0,0,0,0,0,1,1,1,1,1,1,1,1,1]$$

Comment mesurer une similarité entre sommets ?

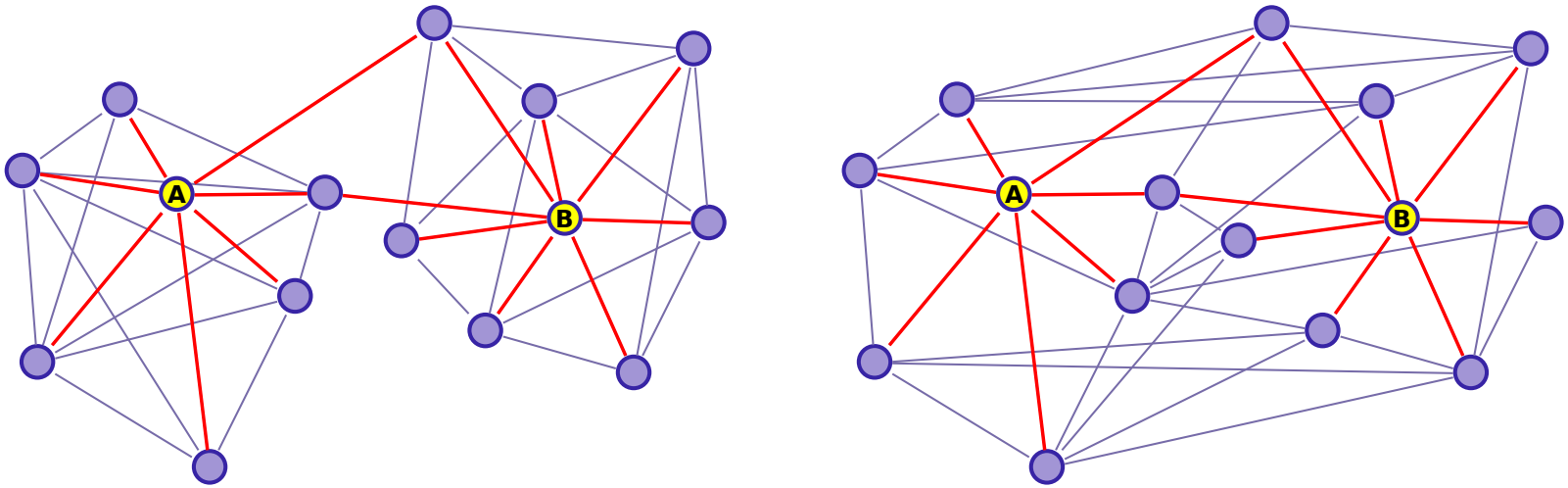


Tout le graphe a changé... sauf $d(A, B)$! !

Vecteurs d'adjacence = **information locale**

ignore la **topologie** du graphe !

Comment mesurer une similarité entre sommets ?

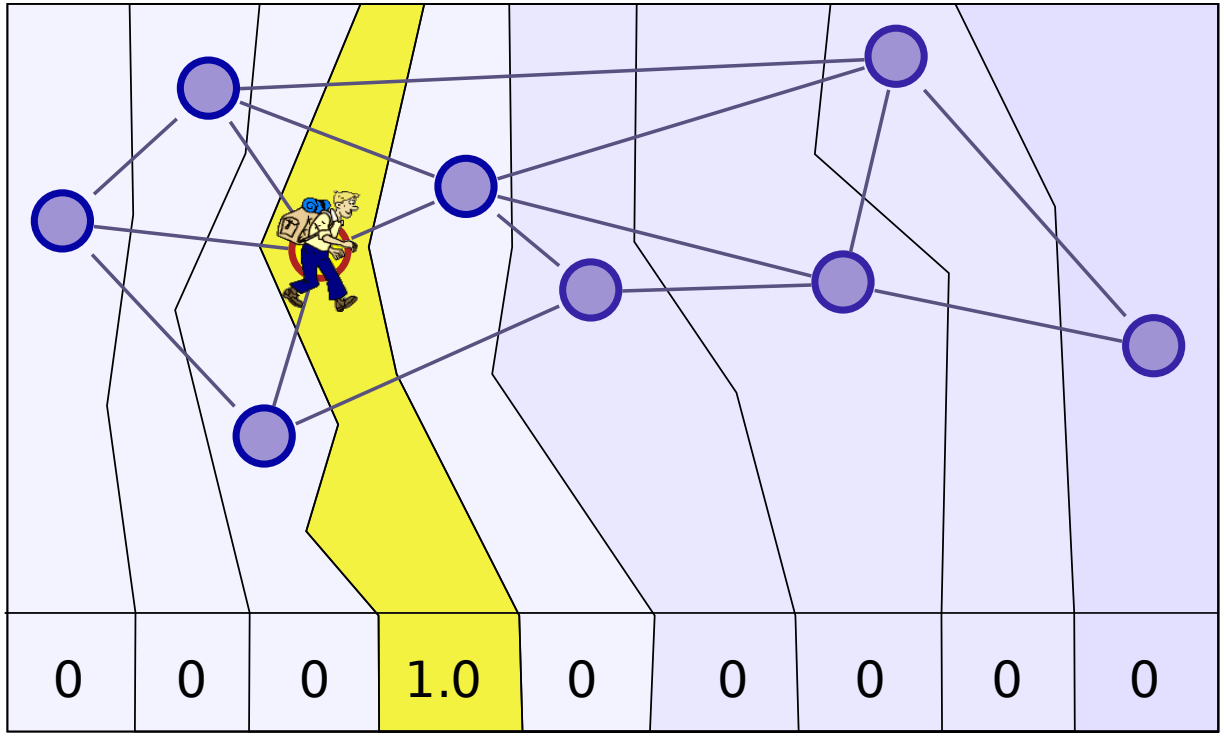


Tout le graphe a changé... sauf $d(A, B)$! !
Vecteurs d'adjacence = **information locale**
ignore la **topologie** du graphe !

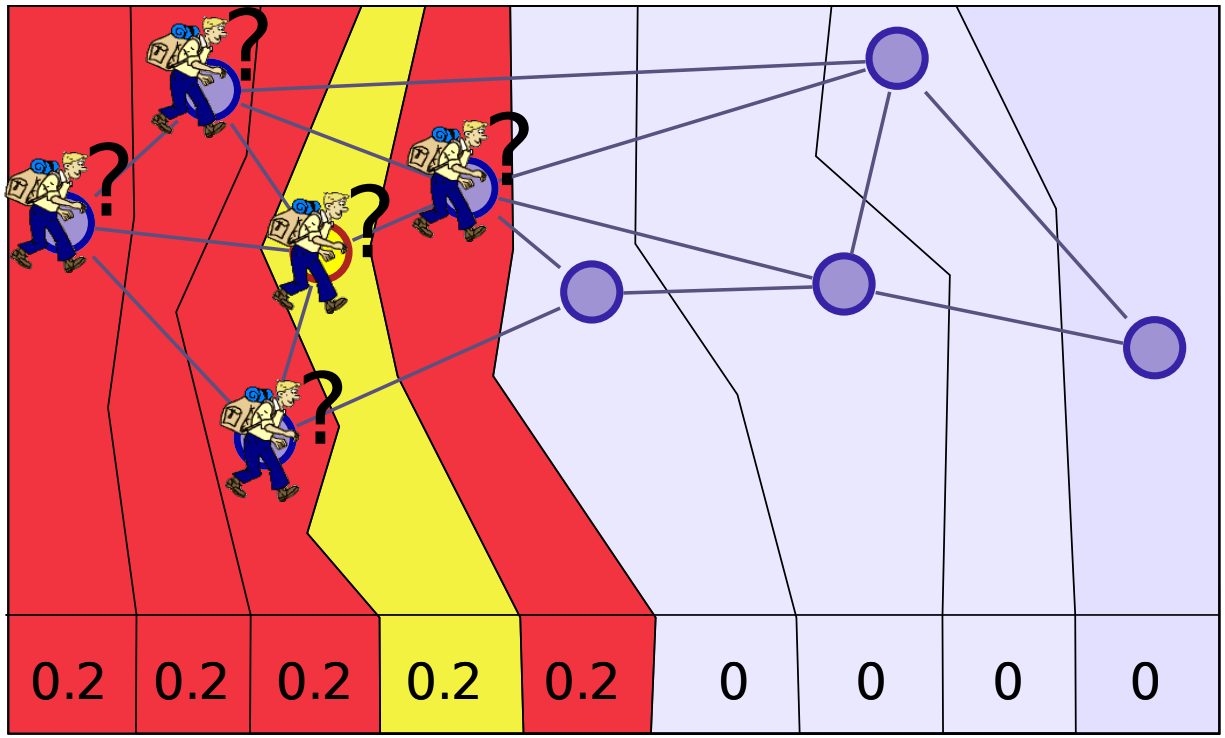
Les **marcheurs aléatoires** à la rescousse !

Chaque sommet est représenté par la **distribution de probabilité** résultant d'une **marche aléatoire courte** partant de ce sommet.

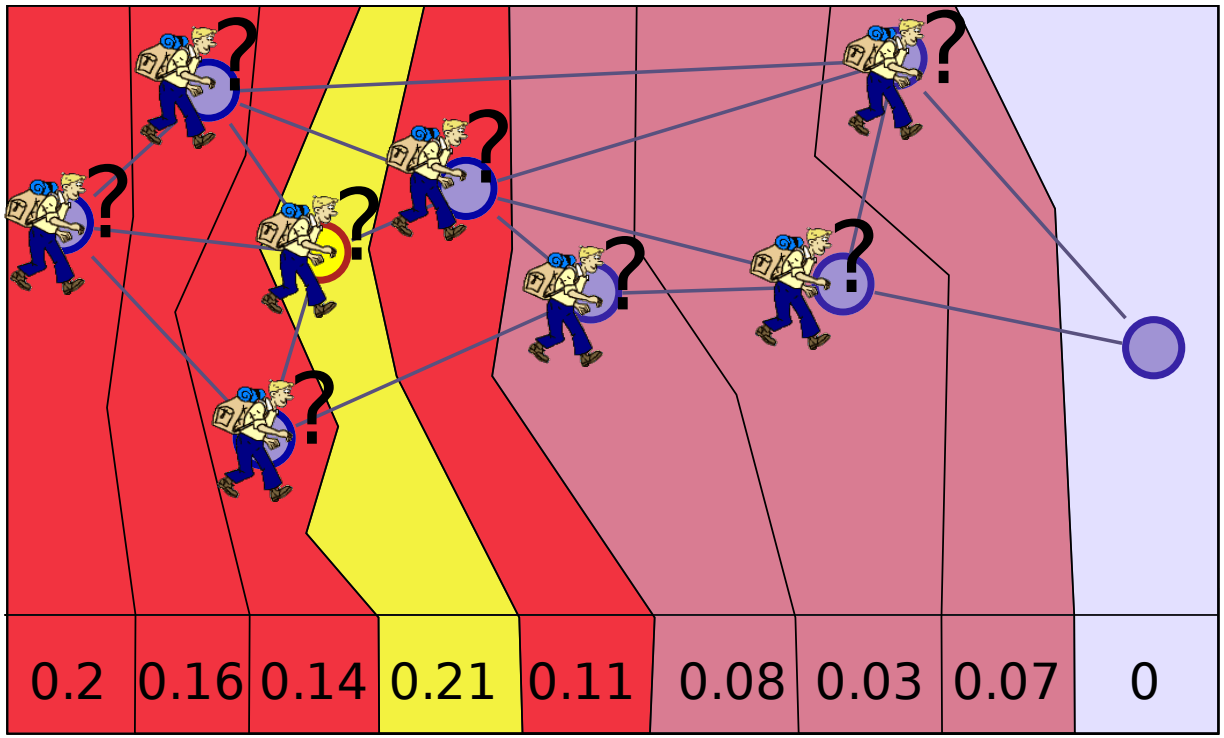
[?]



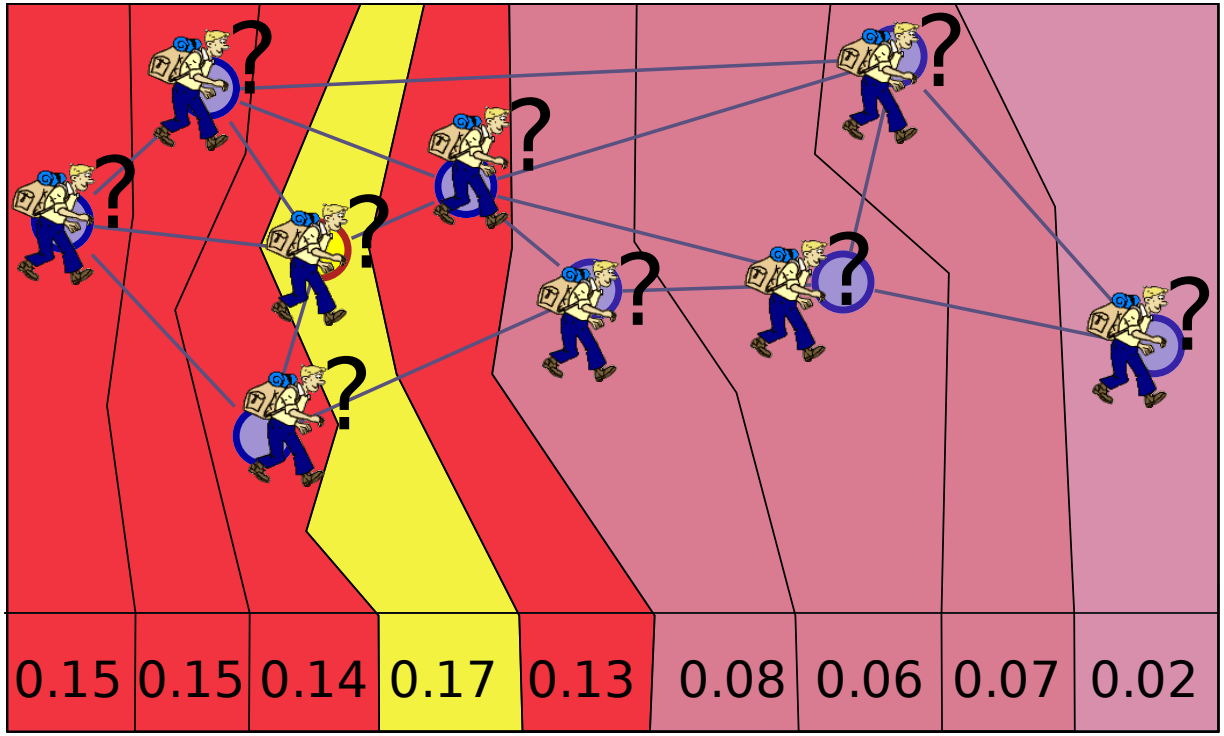
$$t = 0\ , \quad C(u,t) = [0,0,0,1.0,0,0,0,0,0]$$



$$t = 1 \ , \quad C(u,t) = [0.2, 0.2, 0.2, 0.2, 0.2, 0, 0, 0, 0, 0]$$

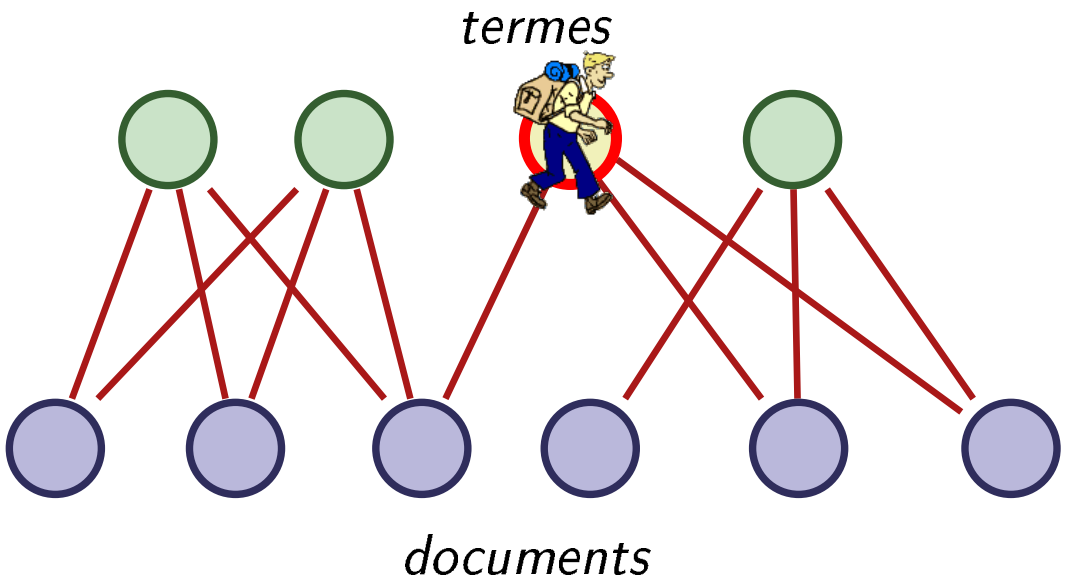


$t = 2$, $C(u, t) = [0.2, 0.16, 0.14, 0.21, 0.11, 0.08, 0.03, 0.07, 0]$



$$t = 3 \ , \quad C(u,t) = [0.15, 0.15, 0.14, 0.17, 0.13, 0.08, 0.06, 0.07, 0.02]$$

Et sur un graphe biparti ?



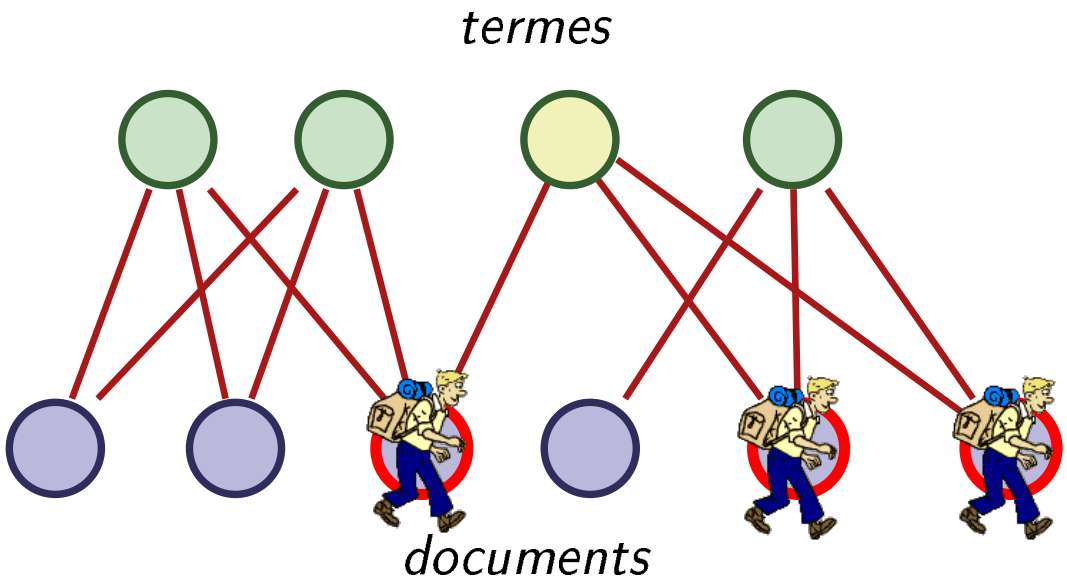
$$\begin{cases} C(u,t) &= [x,x,x,x,0,0,0,0,0,0] \\ C(u,t+1) &= [0,0,0,0,x,x,x,x,x,x] \end{cases}$$

$$\begin{cases} C(u \in termes,t) &= [x,x,x,x,0,0,0,0,0,0] \\ C(v \in documents,t) &= [0,0,0,0,x,x,x,x,x,x] \end{cases}$$

Problème de périodicité :

- pas de convergence !
- Comment comparer *termes* et *documents* ?

Et sur un graphe biparti ?



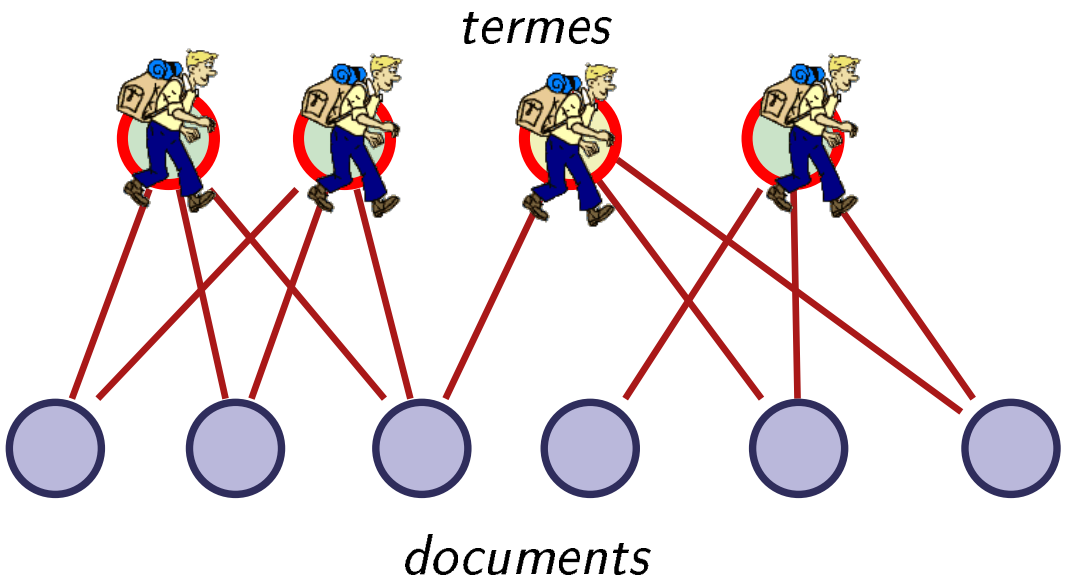
$$\begin{cases} C(u,t) &= [x,x,x,x,0,0,0,0,0,0] \\ C(u,t+1) &= [0,0,0,0,x,x,x,x,x,x] \end{cases}$$

$$\begin{cases} C(u \in \mathit{termes},t) &= [x,x,x,x,0,0,0,0,0,0] \\ C(v \in \mathit{documents},t) &= [0,0,0,0,x,x,x,x,x,x] \end{cases}$$

Problème de périodicité :

- pas de convergence !
- Comment comparer *termes* et *documents* ?

Et sur un graphe biparti ?



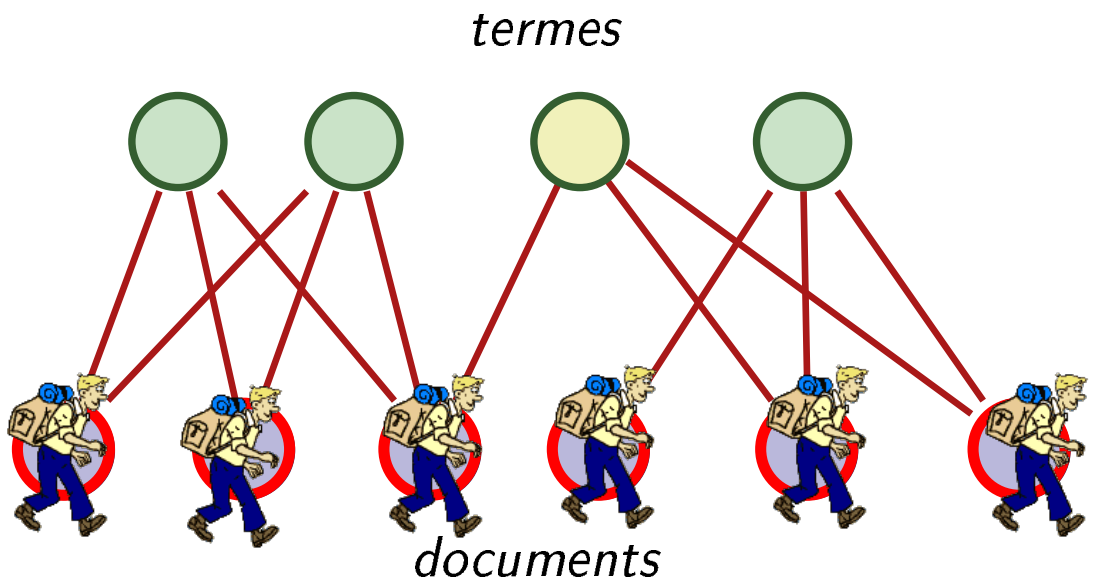
$$\begin{cases} C(u,t) &= [x,x,x,x,0,0,0,0,0,0] \\ C(u,t+1) &= [0,0,0,0,x,x,x,x,x,x] \end{cases}$$

$$\begin{cases} C(u \in \mathit{termes},t) &= [x,x,x,x,0,0,0,0,0,0] \\ C(v \in \mathit{documents},t) &= [0,0,0,0,x,x,x,x,x,x] \end{cases}$$

Problème de périodicité :

- pas de convergence !
- Comment comparer *termes* et *documents* ?

Et sur un graphe biparti ?



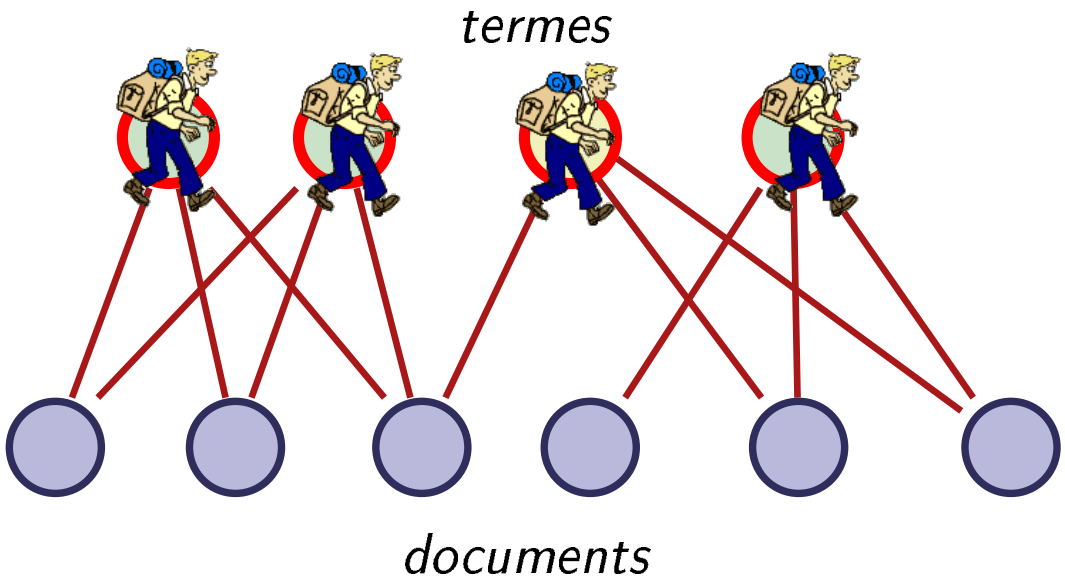
$$\begin{cases} C(u,t) &= [x,x,x,x,0,0,0,0,0,0] \\ C(u,t+1) &= [0,0,0,0,x,x,x,x,x,x] \end{cases}$$

$$\begin{cases} C(u \in \mathit{termes},t) &= [x,x,x,x,0,0,0,0,0,0] \\ C(v \in \mathit{documents},t) &= [0,0,0,0,x,x,x,x,x,x] \end{cases}$$

Problème de périodicité :

- pas de convergence !
- Comment comparer *termes* et *documents* ?

Et sur un graphe biparti ?



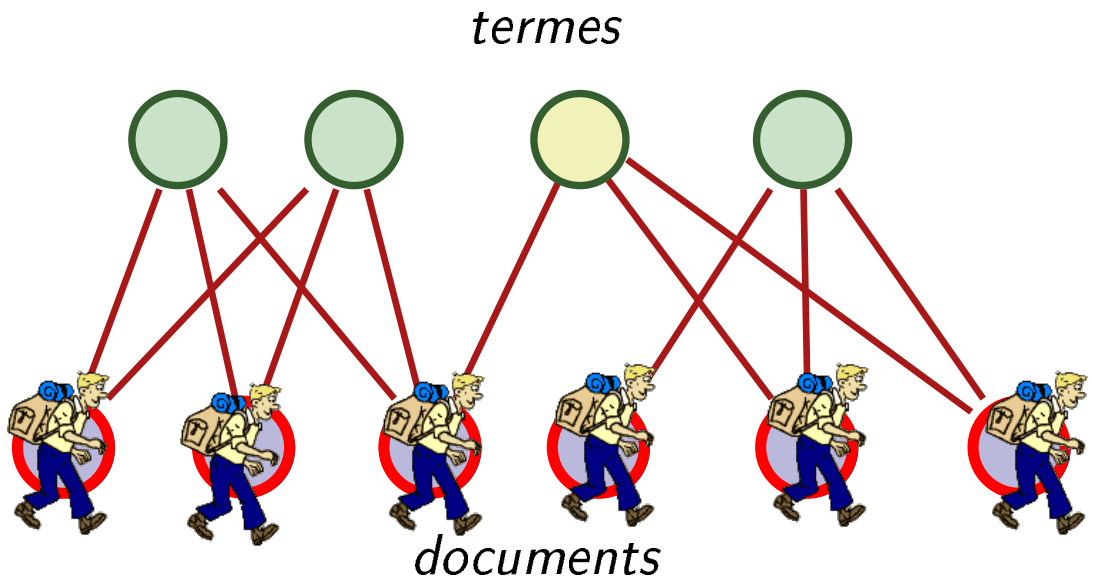
$$\begin{cases} C(u,t) &= [x,x,x,x,0,0,0,0,0,0] \\ C(u,t+1) &= [0,0,0,0,x,x,x,x,x,x] \end{cases}$$

$$\begin{cases} C(u \in termes,t) &= [x,x,x,x,0,0,0,0,0,0] \\ C(v \in documents,t) &= [0,0,0,0,x,x,x,x,x,x] \end{cases}$$

Problème de périodicité :

- pas de convergence !
- Comment comparer *termes* et *documents* ?

Et sur un graphe biparti ?



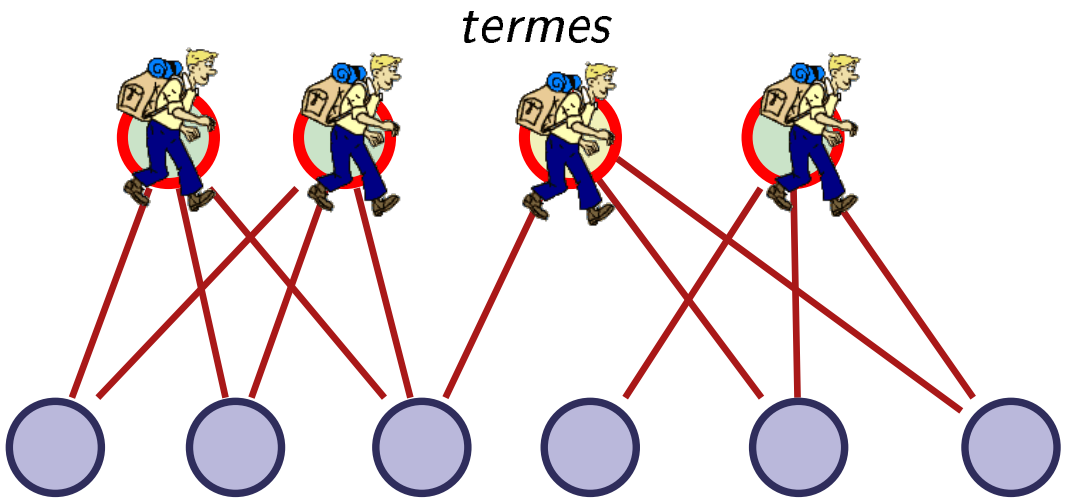
$$\begin{cases} C(u,t) &= [x,x,x,x,0,0,0,0,0,0] \\ C(u,t+1) &= [0,0,0,0,x,x,x,x,x,x] \end{cases}$$

$$\begin{cases} C(u \in termes, t) &= [x,x,x,x,0,0,0,0,0,0] \\ C(v \in documents, t) &= [0,0,0,0,x,x,x,x,x,x] \end{cases}$$

Problème de périodicité :

- pas de convergence !
- Comment comparer *termes* et *documents* ?

Et sur un graphe biparti ?



$$\begin{cases} C(u,t) &= [x,x,x,x,0,0,0,0,0,0] \\ C(u,t+1) &= [0,0,0,0,x,x,x,x,x,x] \end{cases}$$

$$\begin{cases} C(u \in \mathit{termes},t) &= [x,x,x,x,0,0,0,0,0,0] \\ C(v \in \mathit{documents},t) &= [0,0,0,0,x,x,x,x,x,x] \end{cases}$$

Problème de périodicité :

- pas de convergence !
- Comment comparer *termes* et *documents* ?

Une solution simple...

- Balades de longueur $2t$ pour les *termes*
  $\Leftrightarrow$ **balades de longueur** t sur un graphe projeté pondéré

- Balades de longueur $2t+1$ pour les *documents*
  $\Leftrightarrow$ **barycentre** des coordonnées **de leurs voisins** ($\in$ *termes*)

Evaluation (1/2)

hypothèse : A partir d'un ensemble de documents restitués par un moteur de recherche pour une requête, Kodex est capable de construire un cluster contenant un maximum de documents pertinents.

Evaluation :

Rappel/Précision du "meilleur" des clusters,
par rapport au Rappel/Précision de la liste ordonnée.

Quelle est la qualité du meilleur cluster par rapport au même nombre de documents dans la liste ordonnée ?

Evaluation (2/2)

Collection : Quaero-P2, **2.6 millions de pages Web** (.fr), aspirées par Exalead,

25 topics : requête textuelle et un texte explicitant le besoin en information. Besoins réels (logs de requêtes chez Exalead), ex :

- Hypertyroidie
- *On cherche les conséquences et traitements (médicaments, opération, ...) de l'hypertyroidie. Les documents ne traitant que des causes de la maladie ne sont pas pertinents.*

SRI utilisé : Terrier [2]

Run	R	P	F_1
Kodex	**0,4461**	**0,4728**	**0,3210**
Terrier	0,3295	0,3709	0,2628
Gain de Kodex *versus* Terrier	35 %*	27 %	22 %*

Table: Moyenne de R, P et F_1 pour Terrier et Kodex, calculée sur 25 topics. Une astérisque indique que l'amélioration est statistiquement significative selon le test t de Student pairé et bilatéral avec $p < 0,05$

2. `http://terrier.org/`

Perspective, travail en cours...

travail présenté = 1er version du Kodex,
Plusieurs limites :

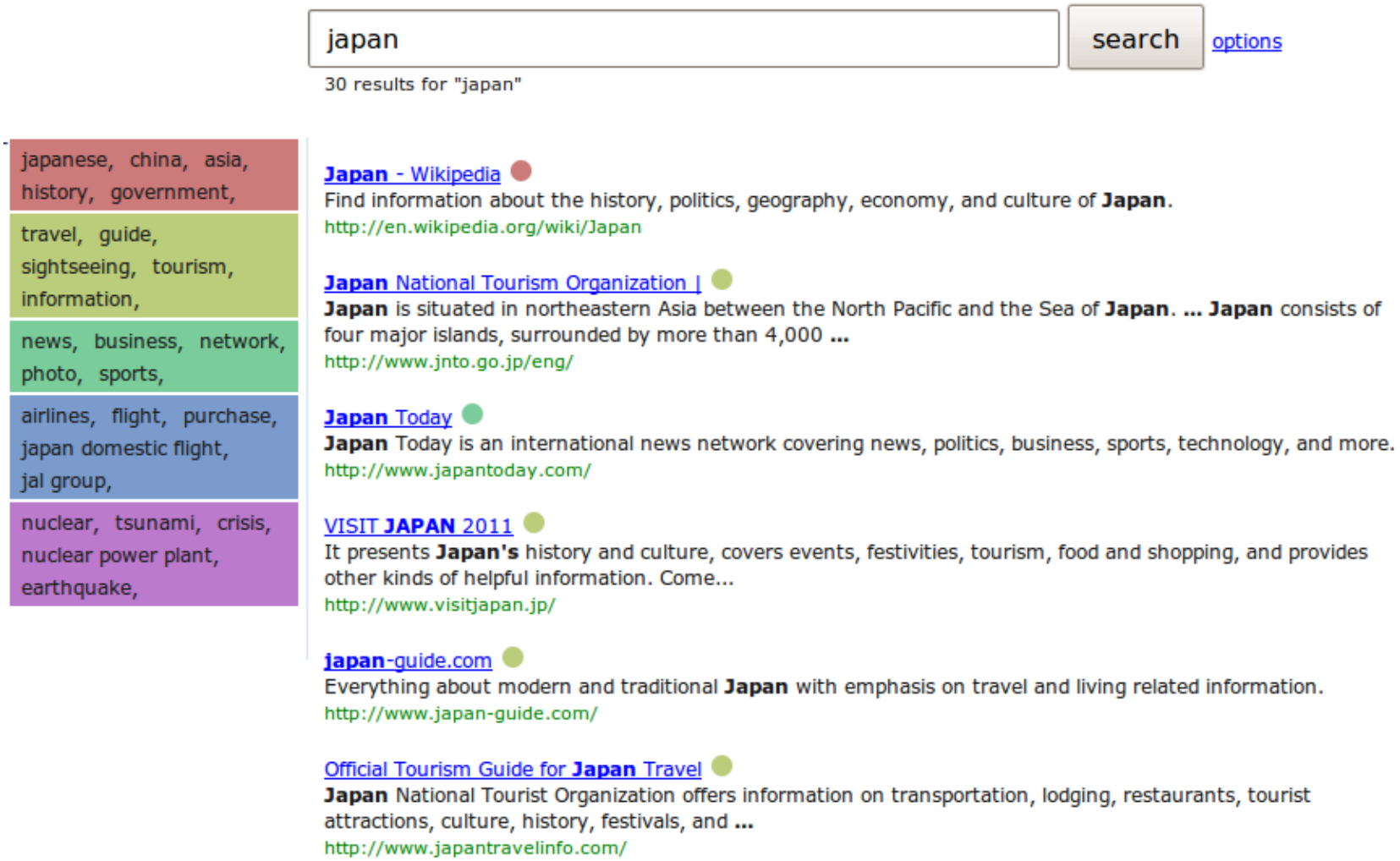
- labélisation (termes des clusters) non suffisante et non évalué,
- pas de **recouvrement** entre les clusters,
- évaluation : ne nous dit vraiment pas si le découpage est pertinent.

Etiquetage endogène de grappes de documents

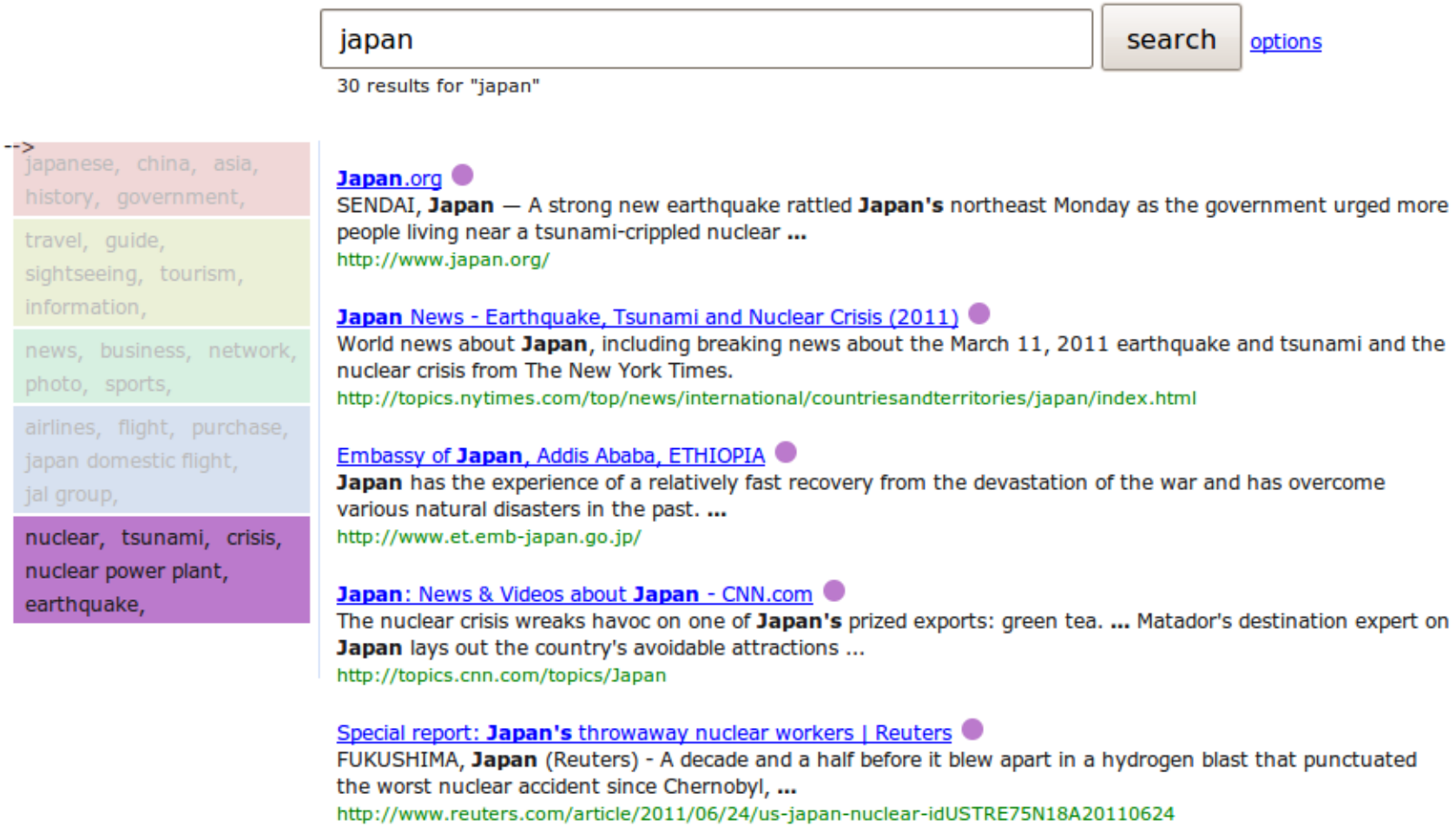
Sommaire

- Problématique : identifier le meilleur cluster
- Travaux connexes : questions en suspens
- Approche proposée :
  - Facteur Descriptif, Facteur Discriminant et Eloquence
  - Ressource endogène pour labels externes
- Expériences
- Conclusion

Des résultats regroupés par thématiques : *clusters*



Identifier ces thématiques sans explorer les clusters



Labelliser des clusters :

Quelques mots pour décrire un ensemble de documents :

- Décrire de quoi parlent les documents du cluster
- Distinguer les documents du cluster des autres résultats
- Etre compréhensibles par l'utilisateur

Ni trop général, ni trop spécifique :

- Mots **les plus fréquents** : peu informatifs car trop communs. [?],[?]

Exemple de mots fréquents

"forum","2008"

- Les plus **informatifs** (important au sens RI) : termes spécialisés trop spécifiques ([?]).

Exemple de termes trop spécialisés

- *"iode 131"* pour *"nucléaire"*
- *"Chelem"* apparaît 25 fois dans un article sur Agassi, vs tennis 13.
- *"tag, text, linguist, lexicon, corpus, tagger, word, syntax, grammar."* pour *"natural language processing"* [?]

Description vs. discrimination :

Informer du thème général ou d'un aspect de ce thème [?] :

Descriptive or Discriminative power [?]

- *Descriptive power* : plus fréquent dans cluster que dans collection totale
- *Discriminative power* : plus fréquent dans cluster que dans les résultats de la requête (extended log-likelihood, [?])

Ressources externes

Informatifs généraux peuvent être externes aux documents

Labels provenant de ressources externes :

- [?] : Wikipédia
- [?] : Wordnet
- [?] : Dictionnaire

Limites des approches classiques

- Ressources *exogènes* (Wordnet, Wikipedia, dictionnaire) : questions de qualité et de couverture.
- Labels externes **moins spécialisés** : **hyperonymes** ou catégories générales dans des ressources externes, mais sélection de candidats sans critère explicite d'***Eloquence*** :
  Exemple :
  [**?**] : *"judges"* = Information Mutuelle, scores de RI.

Approche pour Kodex - description générale

- Critère explicite d'**Eloquence** : **E**
- Candidats informatifs : *Pertinence* à 2 dimensions :
  - **Facteur *Descriptif*** : **I**
  - **Facteur *Discriminant*** : **D**
- **Ressource endogène** pour étiquettes (labels) externes

Liste de candidats

Valeur de chaque candidat :

$$V = \alpha.I + \beta.D + \gamma.E$$

Sélection des N candidats de plus haute valeur

Mesures de pertinence issues de la RI

Principe simple :

- $fratio(t,d) = \frac{Freq(t,d)}{Freq(t,Col)} = \frac{tf(t,d)}{dlen} / \frac{CTF(t)}{ColLen}$
- tf :term frequency, CTF : Collection term frequency, $ColLen$: Longueur de la Collection, DF Document Frequency

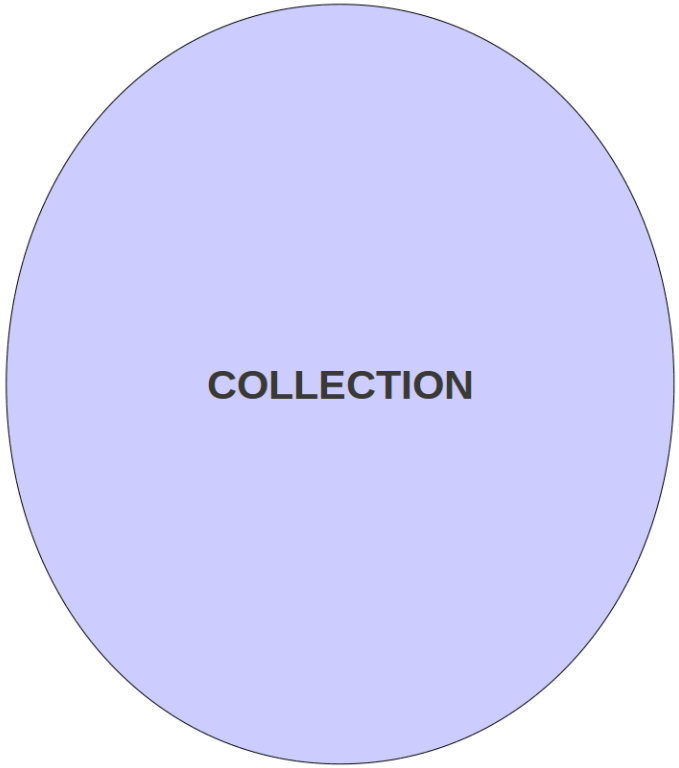
Application en Recherche d'Information :

- $tfidf(t,d) = tf(t,d).log(\frac{N_{docs}}{DF(t)})$
- $tfidfplus(t,d) = tf(t,d).log(\frac{ColLen}{CTF(t)})$
- $\times \frac{1}{len(d)}$, longueur du document d
- bm25 (okapi)...

Pertinence : facteur descriptif, facteur discriminant

Facteur Descriptif d'un terme pour un cluster :

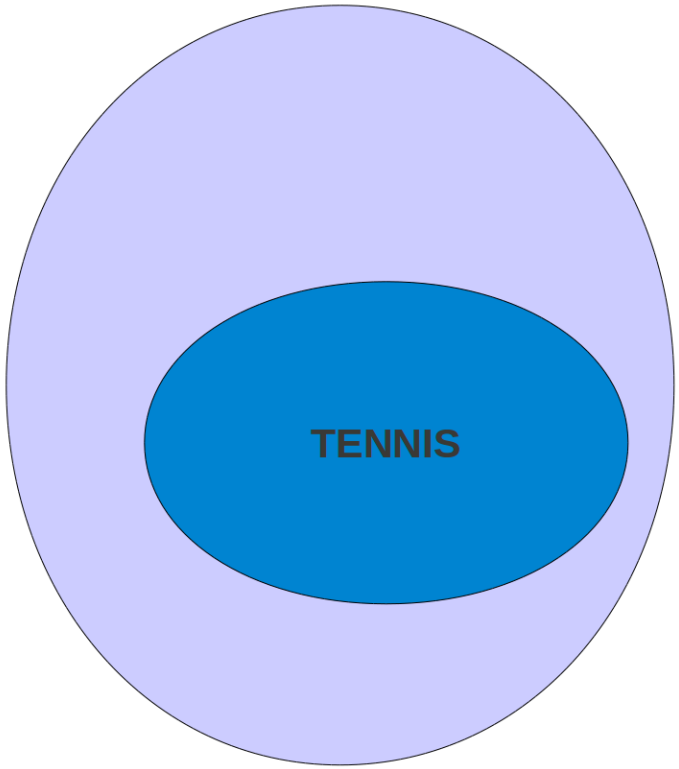
- $I(w) = \frac{Freq(w, cluster)}{Freq(w, collection)}$



Pertinence : facteur descriptif, facteur discriminant

Facteur Descriptif d'un terme pour un cluster :

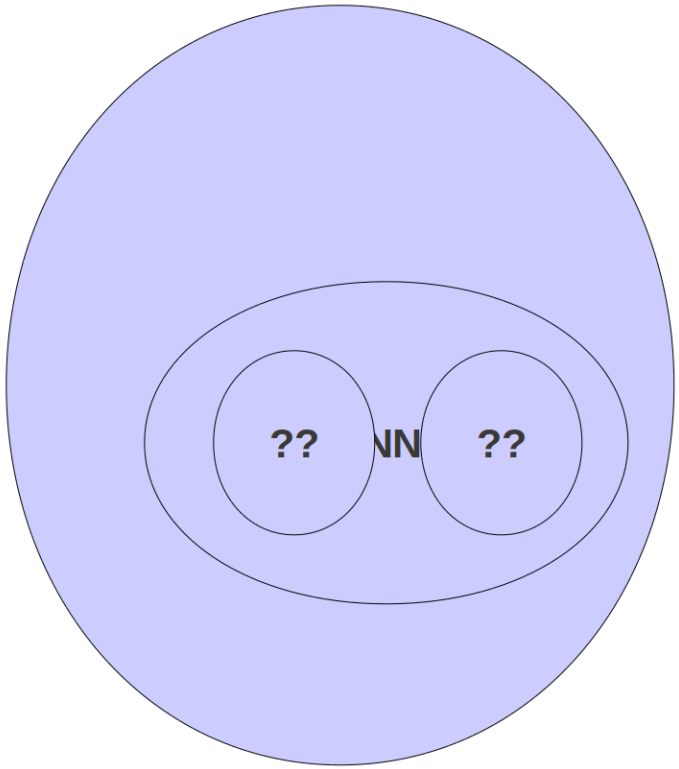
- $I(w) = \frac{Freq(w, cluster)}{Freq(w, collection)}$



Pertinence : facteur descriptif, facteur discriminant

Facteur Descriptif d'un terme pour un cluster :

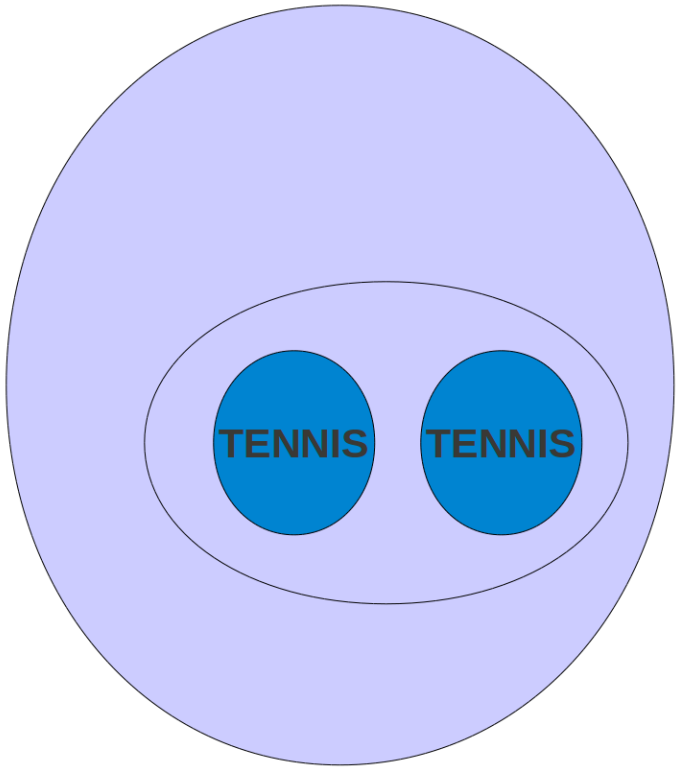
- $I(w) = \frac{Freq(w, cluster)}{Freq(w, collection)}$



Pertinence : facteur descriptif, facteur discriminant

Facteur Descriptif d'un terme pour un cluster :

- $I(w) = \frac{Freq(w, cluster)}{Freq(w, collection)}$



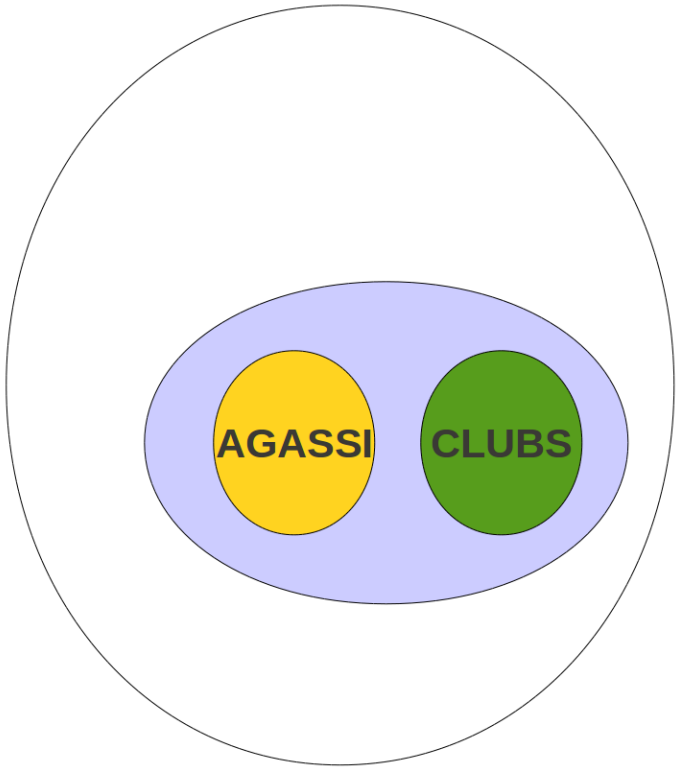
Pertinence : facteur descriptif, facteur discriminant

Facteur Descriptif d'un terme pour un cluster :

- $I(w) = \frac{Freq(w, cluster)}{Freq(w, collection)}$

Facteur Discriminant d'un terme pour un cluster :

- $D(w) = \frac{Freq(w, cluster)}{Freq(w, results)}$



Eloquence

- Approches classiques : hypernonymes
- Hyperonymes : voisins de plus fort degré dans un graphe de synonymie [**?**]
- Fréquence dans le corpus (lien avec l'apprentissage du vocabulaire) :
  - $E(w) = Freq(w, collection)$

Descripteurs généraux externes au cluster

Occurrences un texte sur une manifestation d'agriculteurs :

- ours=5 ; candidat=1
- politique=0 ; écologie=0

Trouver des Unités Lexicales :

- Proches sémantiquement
- Ressources exogènes : dictionnaires, Wordnet, Wikipedia. couverture lexicale, imprécisions.

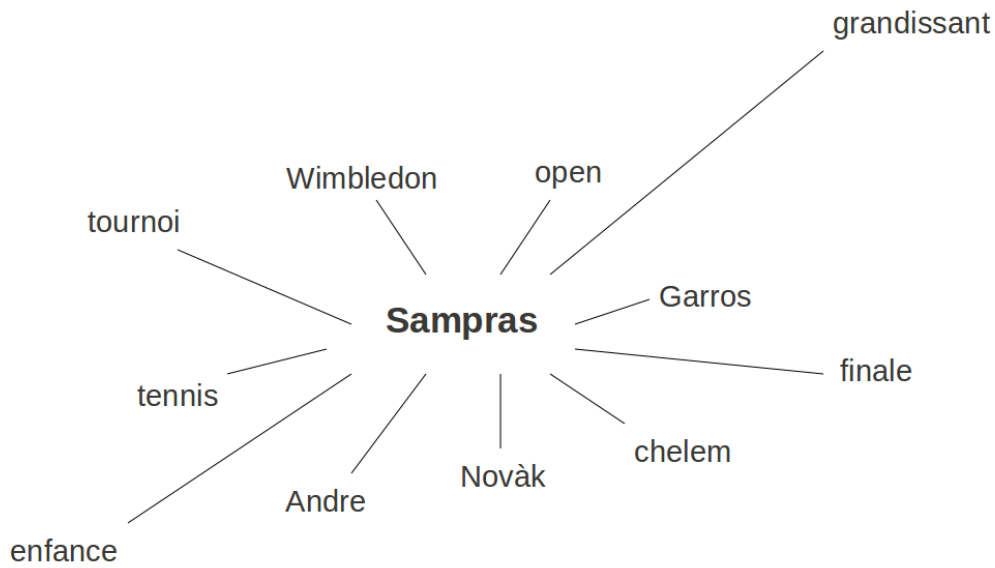
Objectif : décrire la thématique du cluster

- Moins strict que "le même sens" : Même contexte thématique
- Collocations "dans un même texte"

Ressource endogène : Graphe de co-textualité

G = (V,E), graphe pondéré :

- V : les UL de l'index : couverture lexicale parfaite
- E : arêtes pondérées par la cotextualité de leurs extrémités
- $E(A,B) = log(\frac{P(A|B)}{P(B)})$



Interprétation du poids des arêtes :

Tests : Collection de 10 textes dont l'article Wikipédia Agassi.
Trouver B dans un document est un bon indice de la présence de A.

- $P(A) = \frac{N_{doc}(A)}{N_{doc}}$.....$P(tennis) = \frac{1}{10}$
- $P(A|B) = \frac{N_{doc}(A.B)}{N_{doc}(B)}$.....$P(tennis|Sampras) = \frac{1}{1}$
- $E(A, B) = E(B, A)$.....$E(tennis, Sampras) = 10$
- $E(A, B) =$ information mutuelle ponctuelle (Pointwise Mutual Information, PMI)

Marches aléatoires thématiques

On trouve les candidats par marches aléatoires sur le graphe de co-textualité :
Les descripteurs externes sont des proxèmes des mots des clusters :

- Proxèmes : $w \rightarrow \{w_i\}$ (prox)
- Pertinence : $P(w_i) = P(w).Prox(w \rightarrow w_i)$
- Pertinence : un facteur Descriptif :$I(w)$, un facteur Discriminant : $D(w)$

Récapitulation :

Valeur d'un candidat w :

$$V(w) = \alpha.I + \beta.D + \gamma.E$$

Récapitulation :

Valeur d'un candidat w :

$V(w) = \alpha.I + \beta.D + \gamma.E$

Facteur descriptif :

$I(w) = \sum_{o \in cluster} tfidf(o, cluster, collection).Prox(o \rightarrow w)$

Récapitulation :

Valeur d'un candidat w :

$V(w) = \alpha.I + \beta.D + \gamma.E$

Facteur descriptif :
$I(w) = \sum_{o \in cluster} tfidf(o, cluster, collection).Prox(o \rightarrow w)$

Facteur discriminant :
$D(w) = \sum_{o \in cluster} tfidf(o, cluster, results).Prox(o \rightarrow w)$

Récapitulation :

Valeur d'un candidat w :

$V(w) = \alpha.I + \beta.D + \gamma.E$

Facteur descriptif :
$I(w) = \sum_{o \in cluster} tfidf(o, cluster, collection).Prox(o \rightarrow w)$

Facteur discriminant :
$D(w) = \sum_{o \in cluster} tfidf(o, cluster, results).Prox(o \rightarrow w)$

Eloquence :
$E(w) = log(CTF(w))$

Indexation de la collection

- Pertinence mesurée sur la base d'un index document $\leftrightarrow$ "mot"
- Unité Lexicale pour l'indexation :
  - Racine (Porter Stemmer)
  - Troncature
  - Lemmatiser (Treetagger)
  - Syntagmes
- Optimisations : Prétraitements vs. traitements en ligne

Expériences : Evaluation

- But : ergonomie
- Evaluation de la clusterisation = R,P,F du **meilleur cluster** (cf. première partie)
- -> Evaluer le **cluster choisi** par l'utilisateur sur la base des labels
- Littérature : Comparer les résultats du système à des thématiques groupées et étiquetées manuellement

Expériences : paramètres à étudier

- $V = \alpha.I + \beta.D + \gamma.E$:importance relative de chacune des 3 dimensions
  - α pour le facteur descriptif
  - β pour le facteur discriminant
  - γ pour l'éloquence)
- $log(\frac{P(A|B)}{P(B)})$: valeurs négatives. Seuil ou proxémie avec des liens inhibiteurs ?
- Pertinence : tfidf, bm25, information...
- Eloquence : meilleurs critères, corrélation degré/fréquence

Conclusion

Travail en cours : pas encore de résultats.
Nouveautés :

- Paramètre Eloquence
- Graphe de co-textualité

Bientôt des expériences

Carmel, D., Roitman, H., and Zwerdling, N. (2009).

Enhancing cluster labeling using wikipedia.

In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '09, pages 139–146, New York, NY, USA. ACM.

de Winter, W. and de Rijke, M. (2007).

Identifying facets in query-biased sets of blog posts.

In Proceedings Int. Conf. on Weblogs and Social Media (ICWSM-2007), pages 251–254.

Fukumoto, F. and Suzuki, Y. (2011).

Cluster labeling based on concepts in machine-readable dictionaries.

In Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1371–1375, Chiang Mai, Thailand.

Gaume, B. (2004).

Balades aléatoires dans les petits mondes lexicaux.

I3 Information Interaction Intelligence, 4(2).

Geraci, F., Pellegrini, M., Maggini, M., and Sebastiani, F. (2007).

Cluster generation and labeling for web snippets : A fast, accurate hierarchical solution.

Internet Mathematics, 3(4) :413–443.

Girvan, M. and Newman, M. E. J. (2002).

Community structure in social and biological networks.

Proceedings of the National Academy of Sciences of the United States of America, 99(12) :7821–7826.

Navarro, E., Chudy, Y., Gaume, B., Cabanac, G., and Pinel-Sauvagnat, K. (2011).

Kodex ou comment organiser les résultats d'une recherche d'information par détection de communautés sur un graphe biparti ?

In CORIA'11, Avignon, pages 25–40. ARIA.

Popescul, A. and Ungar, L. H. (2000).

Automatic labeling of document clusters.

Treeratpituk, P. and Callan, J. (2006).

Automatically labeling hierarchical clusters.

In Fortes, J. A. B. and Macintosh, A., editors, DG.O, volume 151 of ACM International Conference Proceeding Series, pages 167–176. Digital Government Research Center.

Tseng, Y.-H., Lin, C.-J., Chen, H.-H., and Lin, Y.-I. (2006).

Toward generic title generation for clustered documents.

In Ng, H., Leong, M.-K., Kan, M.-Y., and Ji, D., editors, Information Retrieval Technology, volume 4182 of Lecture Notes in Computer Science, pages 145–157. Springer Berlin / Heidelberg.