LABORATOIRE HUBERT CURIEN

UMR • CNRS • 5516 • SAINT-ETIENNE

NEOTEX

Exploration de documents Textuels d'un domaine par un Néophyte

Séminaire technique « Chantiers d'usage » d'ISTEX

26 Avril 2016

Contexte

L'équipe du projet NeoTex :

- Permanents :
  - Christine Largeron
  - Michel Beigbeder

- Post-doc 10 mois (septembre 2016 – juin 2017)
  - Bissan Audeh

- Stages : 5 mois
  - Ayman Alazizi (Avril 2016 – Août 2016)
  - ? (Février 2017 – Juin 2017)

Contexte

La problématique :

Pour explorer un nouveau domaine, un néophyte

- ne connaît pas
  - les mots clés du domaine
  - les « spécialistes » de ce domaine
  - les articles de référence

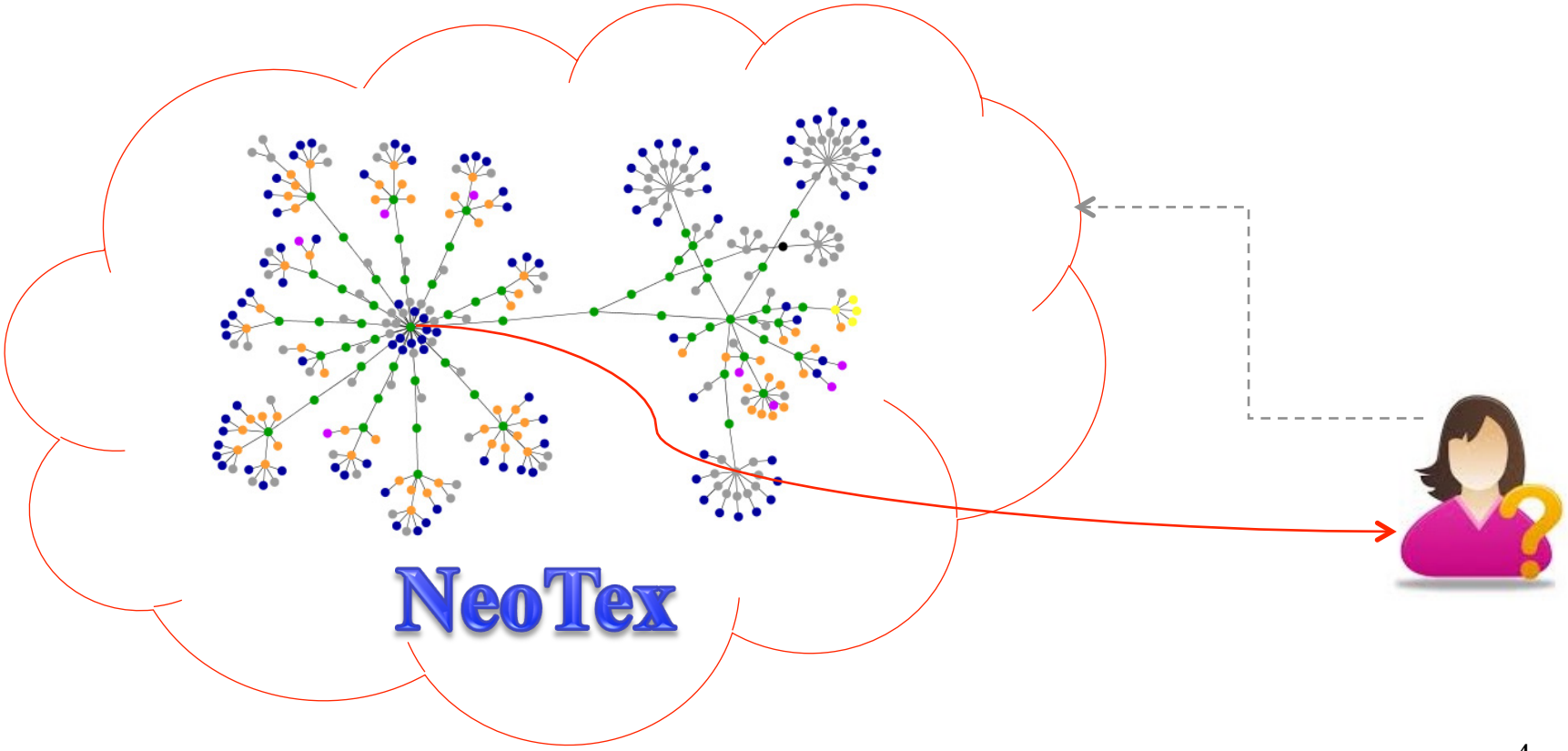
- utilise une requête courte et peu précise

- risque de passer beaucoup de temps sur des articles très spécifiques, peu utiles pour démarrer son étude

Contexte

Objectifs du projet NeoTex :

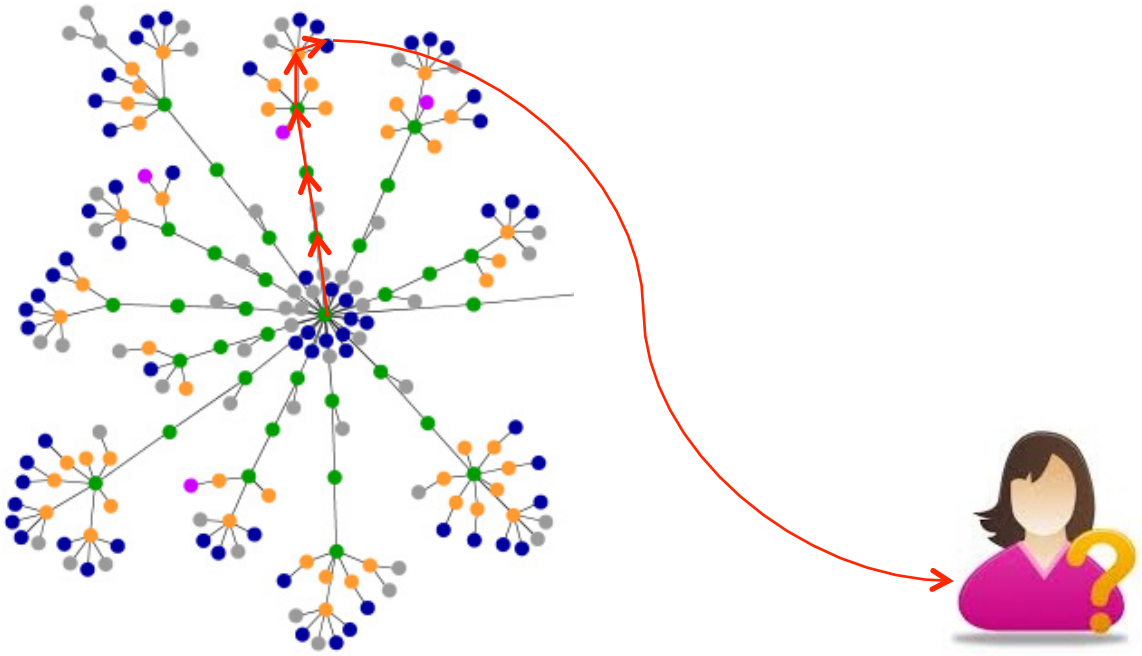
1. En utilisant la requête d'un néophyte, pouvoir proposer « les bons documents » à lire dans le cadre d'une recherche exploratoire



Contexte

Objectifs du projet NeoTex :

2. Soutenir l'utilisateur qui souhaite approfondir ses connaissances dans une direction spécifique en proposant une liste de documents adaptés à ses besoins



NeoTex

Contexte

Les verrous du projet NeoTex :

- Qu'est ce qu'un « bon document » ?
  - Le plus récent ?
  - Le plus cité ?
  - Celui qui cite de nombreuses références ?
  - Celui qui cite des références variées ?

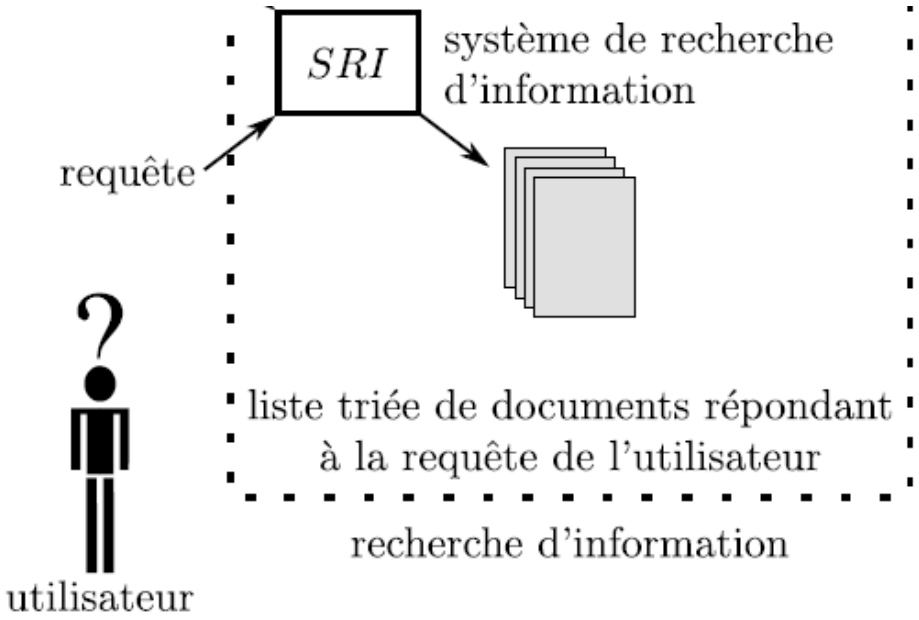
- Comment trouver un « bon document » ?

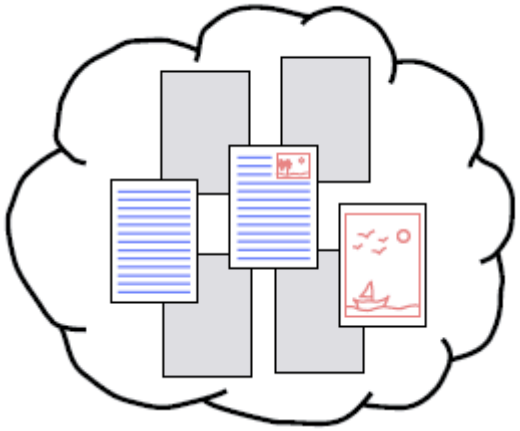
- Comment valider la stratégie de recherche d'un tel document ?

Cadre méthodologique

Idée de base du projet NeoTex :

Combiner l'information textuelle contenue dans le document et l'information contextuelle associée à ce document





Collection ISTEX de documents reliés entre eux :

- cite
- est écrit par le même auteur
- est publié dans le même journal
- etc.

Cadre méthodologique

Domaines de recherche du projet NeoTex :

Recherche d'information (Salton et Mc Gill 1986, Manning 2008)

- Identifier les documents dans la collection ISTEX répondant à la requête de l'utilisateur

Fouille de réseaux sociaux ou de texte (Aggarwal 2011, Aggarwal 2012)

- Représenter les relations entre les documents, auteurs, titres sous forme de graphes
  - R1 : document – document (cite/est cité par)
  - R2 : auteur – auteur (co-écrit)
  - R3 : auteur – document (est l'auteur de)
  - R4 : document – titre (est apparu dans)
- Mesurer l'importance des documents à l'aide d'indicateurs de centralité (Freeman 1979, Wasserman 1994, Newman 2001, Borgatti 2006)

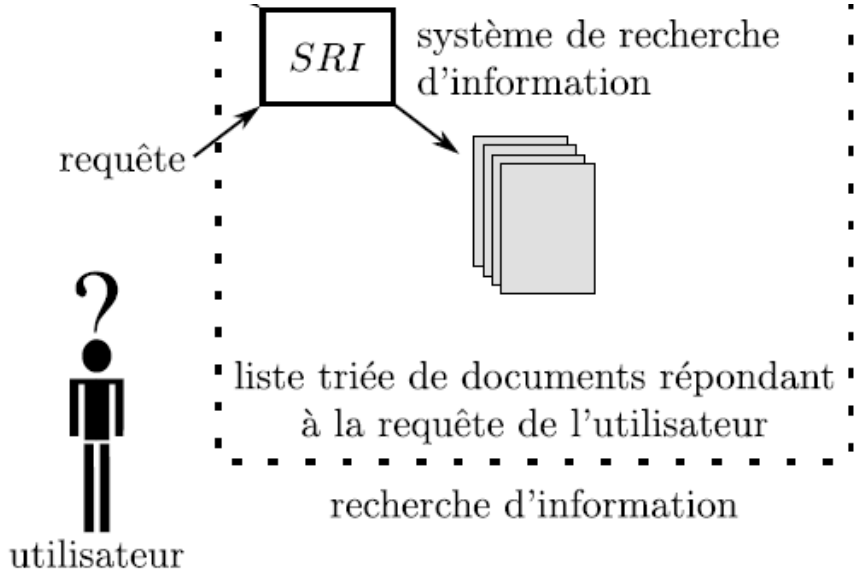
Cadre méthodologique

Stratégie du projet NeoTex

1. Identifier les documents répondant le mieux aux besoins informationnels de l'utilisateur exprimés à travers sa requête

➡ Tâche classique de RI

Qui fournit un score basé sur le contenu S_C(d,q)



Besoin :

- accès au score du moteur d'indexation Lucene
- à défaut, classement basé uniquement sur le contenu du document

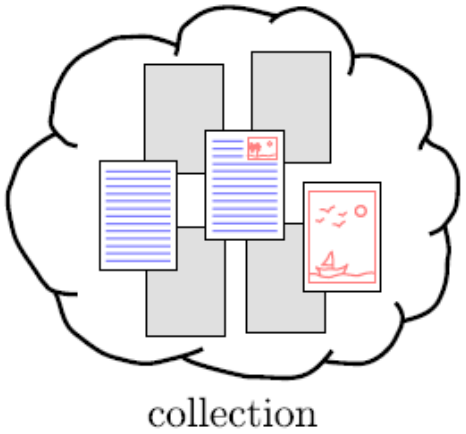
Contexte

Stratégie du projet NeoTex :

2. Identifier les documents les plus pertinents de la liste compte tenu de leurs relations aux autres documents de la collection

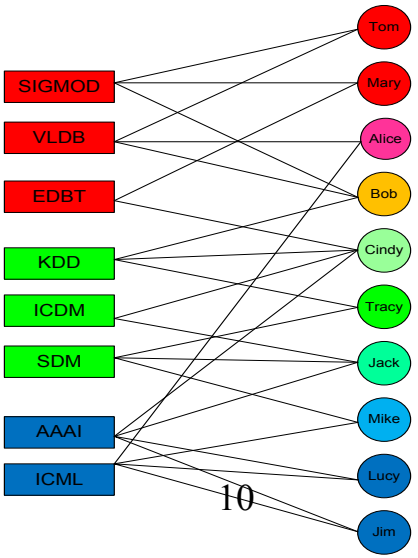
➡ Tâche classique de *social/graph mining*

- Construction de graphes en fonction de différentes relations : citations, auteurs, etc.



collection

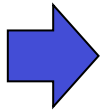
Caractérisation des documents par rapport à leur position dans la collection complète



Contexte

Stratégie du projet NeoTex :

2. Identifier les documents les plus pertinents de la liste
   - Calcul de mesures de centralité caractérisant « l'importance » d'un document dans le graphe (Brin 1989, Kleinberg 1999)



Fourni un score pour chaque document en fonction de la relation R : S_R(d)

Contexte

2. Stratégie du projet NeoTex :

- Centralité de degré : nombre de voisins directs d'un nœud

  ►Nombre de fois où un document cite les autres (article d'état de l'art)

  ►Nombre de fois où un document est cité par les autres

- Centralité de valeur propre (*pagerank*)

  ►Document cité par un grand nombre de documents eux-mêmes très cités

- Centralité d'intermédiarité (*betweenness*) : nombre de plus courts chemins entre toutes les paires de sommets qui passent par ce nœud

  ►Document inter-domaine

Contexte

Stratégie du projet NeoTex :

- réordonner la liste des documents compte tenu de leur importance dans la collection.
  - combinaison linéaire des scores basés sur le contenu et la position

$$S\_G(d) = \lambda\, S\_C(q,d) + (1-\lambda)\, S\_R(d)$$

  - par apprentissage automatique
  - etc.

Documents triés en fonction du contenu S_C	Documents triés en fonction de la position S_R	Documents triés en fonction du score global S_G
D1	D2	D2
D2	D4	D3
D3	D3	D4
D4	D1	D1

Cadre méthodologique

Approche en trois étapes :

Phase 1

- Attribution d'un score de qualité s*(d) à chaque document d de la collection D de documents ISTEX publiés pendant la période [1950–2005]

s*(d) = nombre de citations de d jusqu'en 2016 (vérité terrain)

Phase 2

- Recherche des documents répondant à la requête q parmi les documents ISTEX de D publiés [1950–2005]
- Obtention d'un score S_G(d), $d \in D$

Phase 3

- Comparaison des scores s*(d) et S_G(d) ou comparaison des classements
- Évaluation en termes de précision, rappel, F-mesure, tests Kendall et Spearman

Avancement du travail :

1. Construction d'une collection d'évaluation *i.e.* un corpus avec vérité terrain

D ={d; date_publi(d) ∈ [1950, 2005]}

∀d ∈ D, s*(d) : nombre de citations de d jusqu'en 2016

Mais est ce une « bonne » vérité terrain et avec quelle base l'obtenir ?

2. Construction du graphe de documents basé sur les citations

Mais comment gérer la mise en correspondance des titres des documents ?

➡ Désambiguïsation d'entités nommées (Largeron 2009)

Avancement du travail : Évaluation des résultats

Idée sous jacente :

Si un article publié en t est un article « important » pour la communauté, il aura un nombre élevé de citations à t+n

Construction de vérité(s) terrain :

- pour chaque document d appartenant à un échantillon E de 250 documents extraits d'ISTEX
- recherche du nombre de citations de d dans plusieurs bases :
- ISTEX 2016 : c_I2016(d) ISTEX 2005 : c_I2005(d)
  - *Web of science* (API – droit d'accès)
  - *Microsoft Academic* (droit d'accès)
  - *Citeseer*
  - *Google scholar* : C_GS(d)

Avancement du travail :

Comparaison des vérité(s) terrain :

- Coefficients de corrélation entre les nombres de citation sur ISTEX 2005, ISTEX 2016, Google Scholar

	ISTEX2005	ISTEX2016	X2016.2005	Googlescholar2016
ISTEX2005	1.00000000	0.9920707	0.6278366	0.03751336
ISTEX2016	0.99207073	1.0000000	0.7206814	0.13475571
X2016.2005	0.62783659	0.7206814	1.0000000	0.62761929
Googlescholar2016	0.03751336	0.1347557	0.6276193	1.00000000

Avancement du travail :

Comparaison des vérité(s) terrain :

- tests de Spearman et Kendall pour comparer les classements basés sur c_I2005, c_I2016, c_GS

➢ Concordance des classements avec un risque de 5 ou 1%

	R_2005_3	R_2016_3	R_Diff_3	R_Googlescholar_3
R_2005_3	1.00	0.80	0.21	0.36
R_2016_3	0.80	1.00	0.65	0.64
R_Diff_3	0.21	0.65	1.00	0.73
R_Googlescholar_3	0.36	0.64	0.73	1.00

n= 250

	R_2005_3	R_2016_3	R_Diff_3	R_Googlescholar_3
R_2005_3	1.0000000	0.6594378	0.1531566	0.2494137
R_2016_3	0.6594378	1.0000000	0.4937189	0.4776546
R_Diff_3	0.1531566	0.4937189	1.0000000	0.5591968
R_Googlescholar_3	0.2494137	0.4776546	0.5591968	1.0000000

Avancement du travail :

Construction du graphe à partir des titres (titre d'article cité ou citant)

- Récolte des métadonnées Jason des documents de 1950 à 2005 avec au moins une référence avec un auteur (API ISTEX)
  - 110 giga-octets
  - 6,8 millions d'unité documentaires

```
 "corpusName": "elsevier",
      "author": [
        {
          "name": "C.-P. Adler",
          "affiliations": [
            "Pathologisches Institut der Albert-Ludwigs-Universität
(Ludwig-Aschoff-Haus), Freiburg i. Br. (Direktor: Prof. Dr. W.
Sandritter)"
        ...
      "title": "DNS in Kinderherzen. Biochemische und
zytophotometrische Untersuchungen",
        ...
      "refBibs": [
        {
          "host": {
            "author": [],
            "title": "Polyploidisierung und Zellzahl im menschlichen
Herzen"
          }
        },
        ...
```


"DNS in Kinderherzen. Biochemische und zytophotometrische Untersuchungen"

Cite

"Polyploidisierung und Zellzahl im menschlichen Herzen"

Le graphe de titres

- 118 millions de lignes
  - 58 millions de valeurs différentes
- 6,8 millions de titres citants
- 111 millions arcs (ou de titres cités)
- 11 giga-octets

```
DNS in Kinderherzen. Biochemische und zytophotometrische Untersuchungen
        Polyploidisierung und Zellzahl im menschlichen Herzen
        Postmortale DNS-Veränderungen im Herzmuskel
        Cell number in human heart in atrophy, hypertrophy, and under [...]
        Der DNS-Gehalt in wachsenden Menschenherzen
        Hormonal influences in the regulation of cardiac performance
        L'acide désoxy-ribonucléique du noyau cellulaire, dépositaire [...]
```

Les titres

- Certains titres sont étranges

```
>?

Abstract

000 <,
```

- Certains titres sont très longs (le plus long : plus de 28 000 caractères, environ six pages A4)

```
000,000 100.0% Source: World Christian Trends Demographic Future s for
Christianity and the World Religions . Todd M. Johnson 1

000 -000 clinical perspectives 593 Adults with cystic fibrosis: meeting the
challenge BYE reviews 598 Diagnosis of primary hyperparathyroidism:
controversies, practical issues and the need for Australian guidelines
```

Titres similaires

- Erreur d’extraction
- Caractères « équivalents » : par exemple - – − —
- Erreur d’orthographe
- Abréviation
- Mots manquants
- etc.

Similarité de titres

- Similarité de titres
  - distance d'édition
  - distance de Jaccard entre les ensembles de n-grammes
  - distance *soft-Jaccard* (Largeron 2009)
  - etc.

- Comparer les titres deux à deux : impossible
  - $58\,000\,000^2/2$ comparaisons : un demi-siècle pour une comparaison par microseconde

Hachage des titres

- Méthode de hachage : *Locality sensitive hashing* (Gionis, Indyk et Motwani 1999) décrit dans *Mining Massive Data Sets* (Ullman, Rajaraman, Leskovec 2011)

- Résultat : des regroupements de titres (parce qu'il y a plusieurs fonctions de hachage)

- Temps de calcul (et d'entrées-sorties) : environ une demi-journée

Un regroupement de titres

```
Optical control of GaAs MESFET
Optical control of GaAs MESFETs
Optical control of GaAs MESFET's

TSH-induced hyperthyroidism and acromegaly due to pituitary tumour (abstr)
TSH-induced hyperthyroidism and acromegaly due to pituitary tumour (Abstr) Baylis
PH: Case of hyperthyroidism due to a chromophobe adenoma

Int. Conf. on Occupational Radiation Safety in Mining
Proc. Int. Conf. on Occupational Radiation Safety in Mining

Morphological responses of macrobenthic polychaetes to low oxygen on the Oman
continental slope NW Arabian Sea
Morphological responses of macrobenthic polychaetes to low oxygen on the Oman
continental slope, NW Arabian Sea

Prognostic value of histological and clinical factors in 56 patients with
gastrointestinal lymphoma
Prognostic value of histological and clinical factors in 56 patients with
gastrointestinal lymphomas

Environmental organic chemistry of oceans, fjords and anoxic basins
Evnironmental organic chemistry of oceans, fjords and anoxic basins
```

Suite du travail (1/2)

Sur le graphe de titre

- Utiliser tous les regroupements
- Évaluer le résultat
- Intégration dans la plateforme

Sur la construction des graphes

- Choix des relations
- Construction effective

Suite du travail (2/2)

Sur les mesures de centralité

- Choix et/ou définition de mesures
- Implémentation

Sur le calcul du score global

- Choix des mesures de centralité les plus prédictives par apprentissage automatique

Sur l'évaluation du système sur notre collection de test

- Construction de requêtes test à partir de thèses publiés en 2005

Nos besoins

Définition d'un « bon document » pour un néophyte

- Avis d'un documentaliste

Avis sur notre choix de vérité terrain

Le score BM25 calculé par Lucene

Données d'utilisation réelles

Sur l'évaluation du système sur notre collection de test

Merci pour votre attention

et pour votre aide.

Exemple de 6-grammes sur un titre

```
000,000 100.0% Source: World Christian Trends Demographic Future s for
Christianity and the World Religions . Todd M. Johnson 1

000,00
00,000
0,000
,000 1
000 10
00 100
0 100.
 100.0
100.0%
00.0%
0.0% S
.0% So
0% Sou
% Sour
 Sourc
Source
ource:
...
```

Les étapes

- Construction et hachage des k-grammes des titres
  - taille du résultat : ~70 giga-octets
- Regroupement par valeur de hachage
  - tri (de ~70 giga-octets)
- Hachage par minimum (*minhashing*)
  - 100 permutations
  - mémoire : 100 × 58 entiers : 22 giga-octets
- Temps de calcul : environ une demi-journée