Recherche d'information dynamique

Robin Joganah, étudiant à la maîtrise
Directeurs de recherche :
Pr. Luc Lamontagne
Pr. Richard Khoury

UNIVERSITÉ LAVAL

Sommaire

- Définition de la recherche d'information (IR)
- Évaluation d'un système
- Introduction de la compétition TREC
- Prétraitement du corpus
- Processus de recherche d'information
- Entités nommées et segmentation en phrase
- Traitement du retour de l'utilisateur
- Résultats et travaux en cours
- Conclusion

Qu'est-ce que la recherche d'information ?

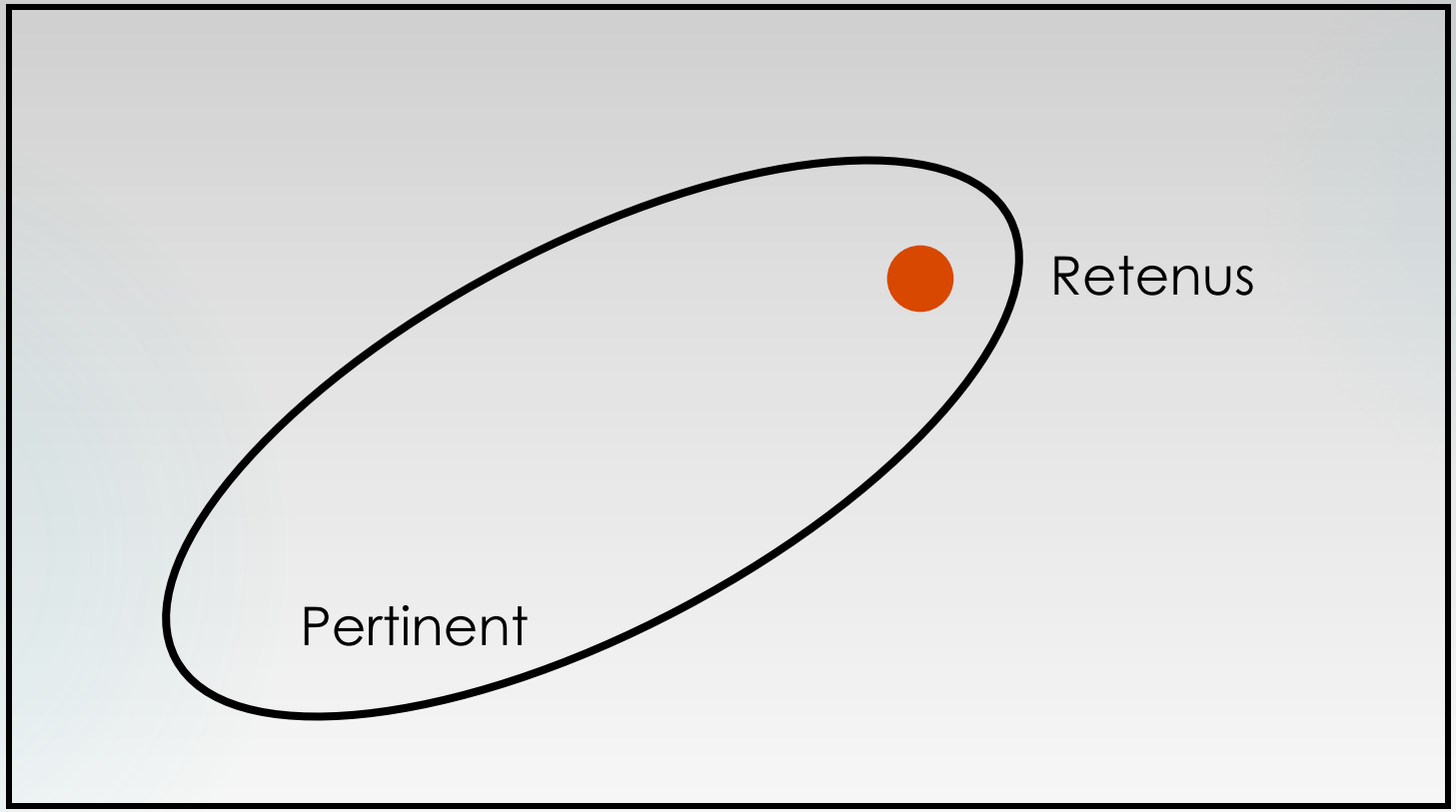
- C'est l'action de trouver un document de nature non structurée qui satisfasse un besoin d'information dans une large collection.

- Visible tous les jours avec les moteurs de recherche en ligne (Google, Yahoo)

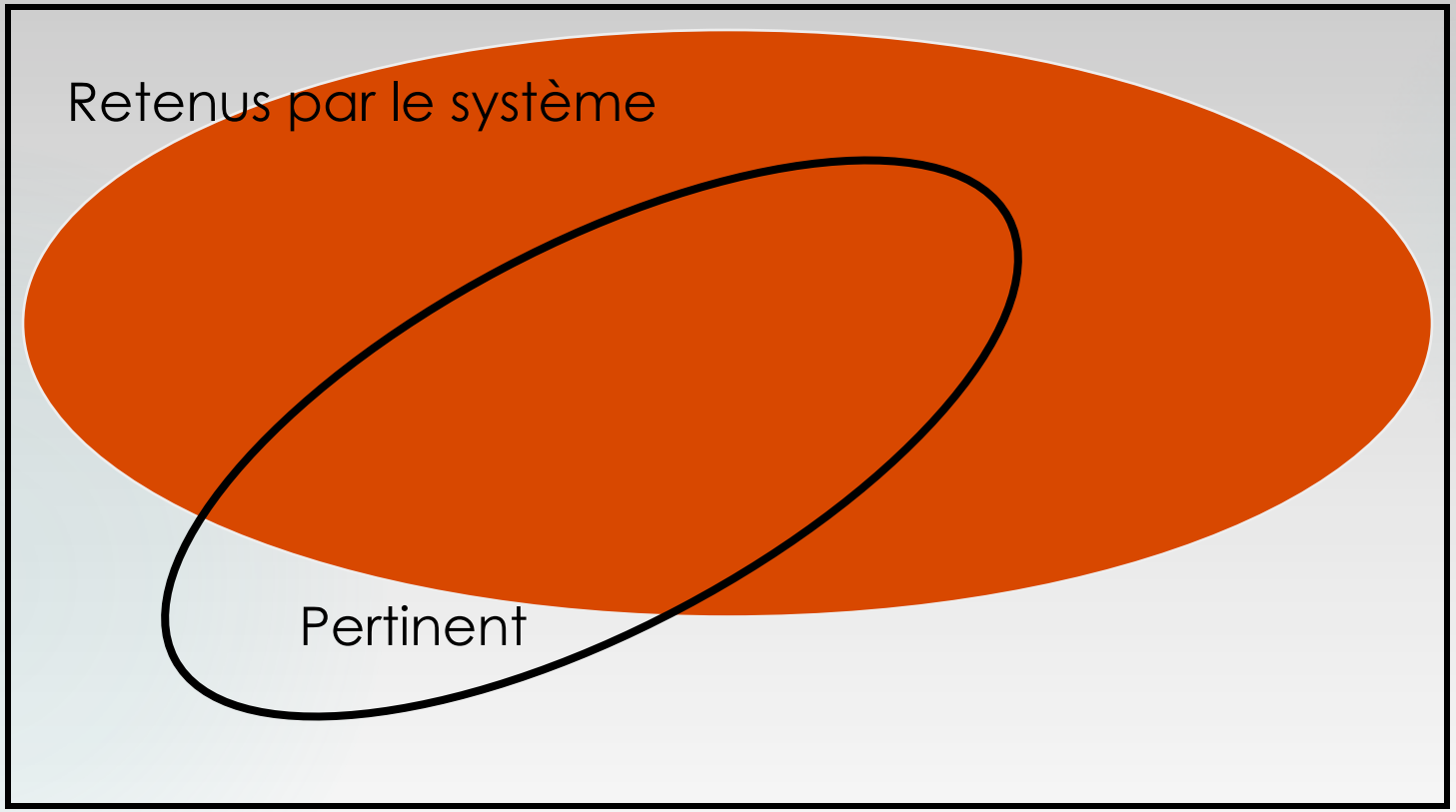
Évaluation d'un système de recherche d'information

- Rappel
  - Proportion de documents pertinents retenus
- Précision
  - Proportion de documents retenus qui sont pertinents

Évaluation d'un système de recherche d'information



Évaluation d'un système de recherche d'information

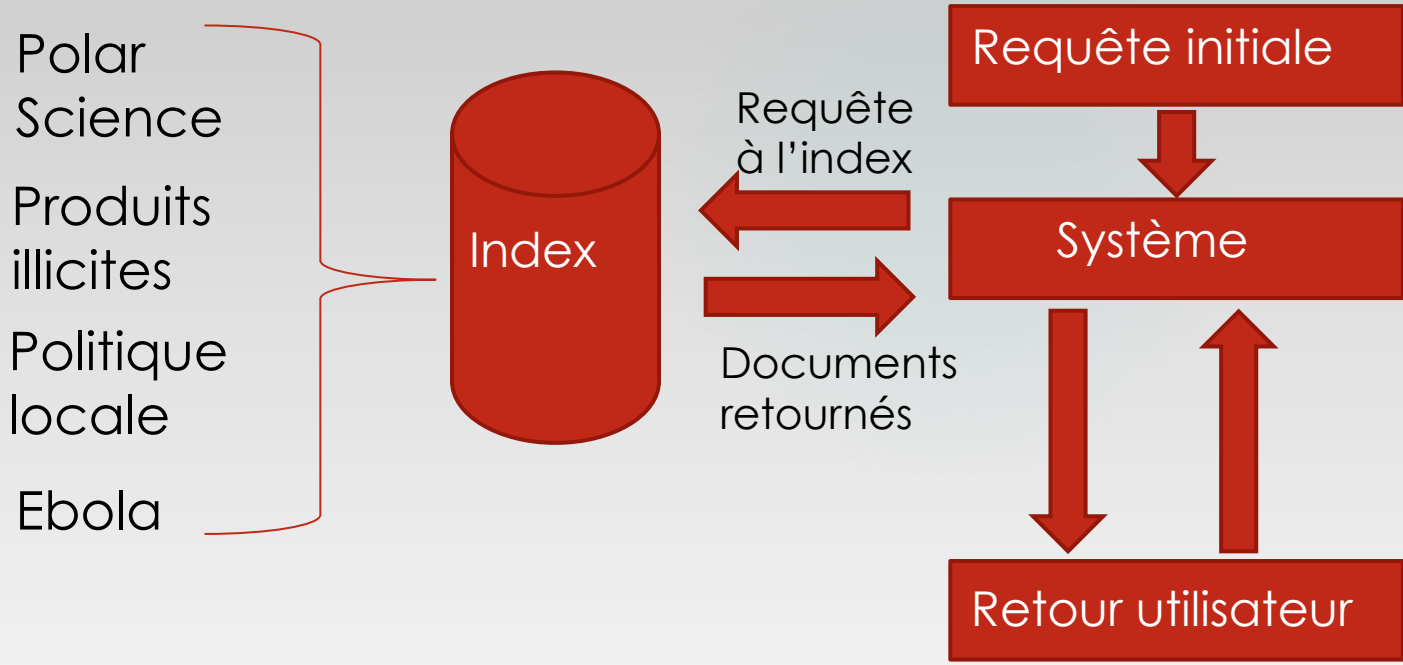


Introduction de la compétition TREC

- Compétition / Conférence de recherche d'information depuis 1992
- Avancées du domaine grâce à la création de corpus permettant d'évaluer et de comparer les systèmes

Objectifs de la compétition

- Trouver les intérêts sous-jacents de l'utilisateur (sous-sujets)
- Modélisation d'un domaine
- Exploitation du retour de l'utilisateur
- Naviguer au travers de nombreux documents non jugés



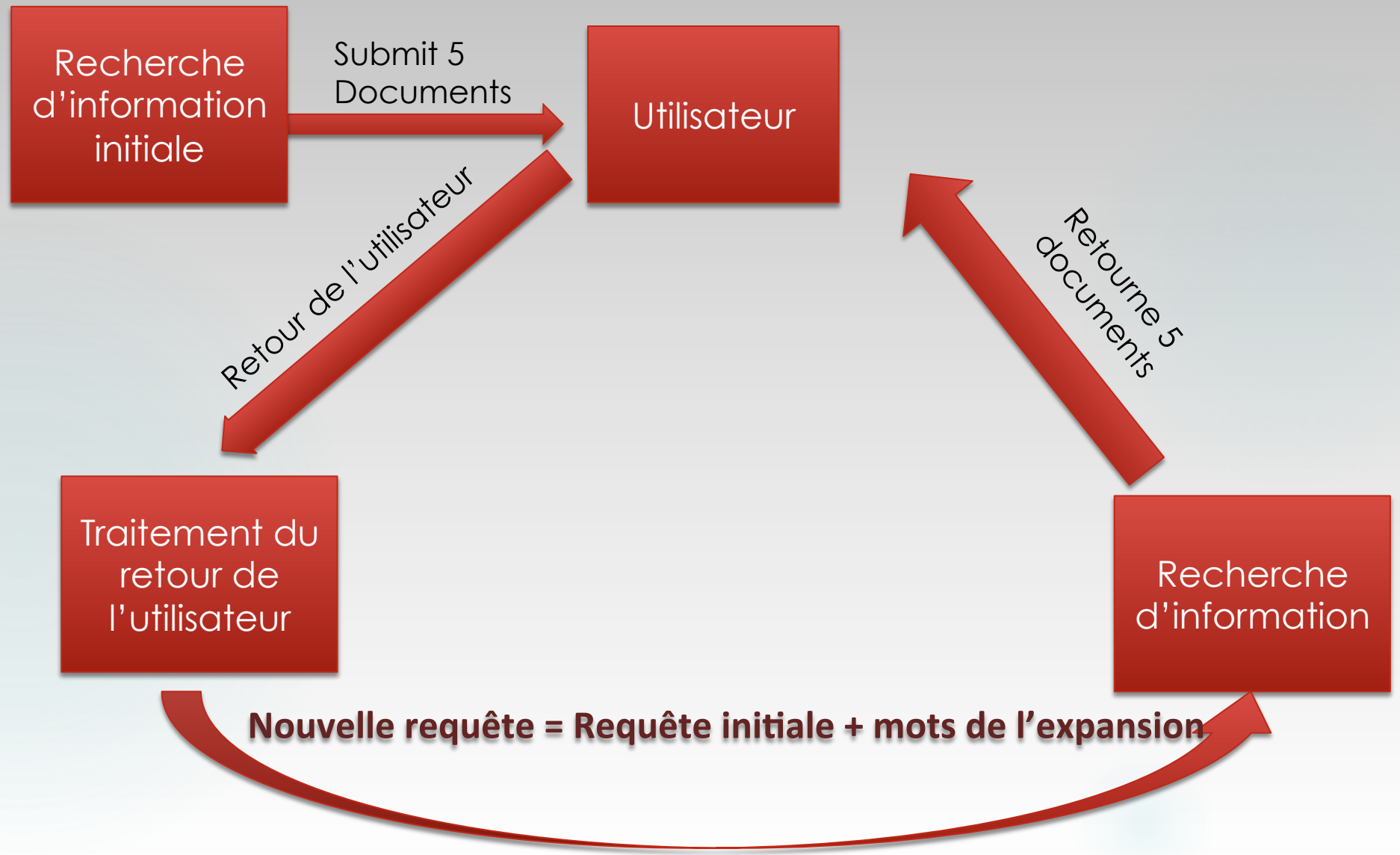
Exemple de Requête

- healthcare impacts of ebola
  - Les sous-sujets doivent être découverts :
    - healthcare insurance
    - healthcare costs of ebola
    - provider impacts

Exemple d'interaction avec l'utilisateur simulé


```
[
 {
     "topic_id": "DD15-1"
     "ranking_score": "123",
     "on_topic": "1",
     "doc_id": "1335424206-b5476b1b8bf25b179bcf92cfda23d975",
     "subtopics": [
         {
             "passage_text": "this is a passage of relevant text from the document 'stream_id', relevant to the
             "rating": 3,
             "subtopic_id": "DD15-1.4",
         }
     ],
 },
 { ... }
]
```

Vue globale du système



Méthode d'évaluation d'un système dynamique

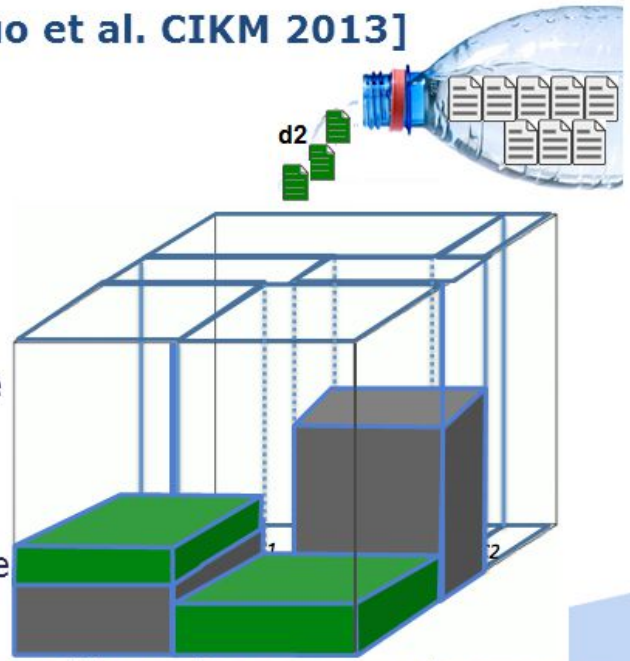
- Les mesures classiques ne capturent pas la couverture des intérêts de l'utilisateur
- Introduction du CubeTest
  - Mesure du gain d'information pour chaque sous-sujet
  - Les passages pertinents remplissent le cube
  - Un score de CubeTest est meilleur pour un système qui remplit moyennement plusieurs cubes (sous-sujets) qu'un système remplissant pleinement une minorité de sous-sujets.

Méthode d'évaluation d'un système dynamique

Evaluation - Cube Test

[Luo et al. CIKM 2013]

- An empty task cube for a search task with multiple subtopics
- A stream of "document water" fills into the task cube
- A new coming relevant document will increase waters in all its relevant subtopics
- The total height of the water in one cuboid represents the accumulated relevance gain for a subtopic
- There is a cap for Gains
- Total volume in the task Cube is the total Gain



- **Cube Test (CT) calculates the rates of how fast a search system can fill up the task cube as much as possible**

$$CT(Q,D) = \frac{1}{|D|}\sum_{t}\frac{Gain(Q,D^t)}{Time(D^t)}$$

Prétraitement des données

- Le jeu de données concernant le virus Ebola contient près de 200 000 pages web:
    - Bruit important pour le K-Means / LDA
    - Introduction d'erreurs de sujet à cause des information extérieures à l'article

- BoilerPipe permet d'extraire uniquement un article d'une page

Prétraitement

Igloolik survivor calls rescuers 'heroes'

Adds he is 'so sorry' for sergeant who sacrificed his life to save them

CBC News Posted: Oct 31, 2011 4:00 PM CT | Last Updated: Oct 31, 2011 5:06 PM CT

Related Stories

- Quebec airman's body returns to base
- Airman dies after Igloolik rescue effort

One of the survivors of last week's fatal rescue mission near Igloolik, Nunavut, describes his rescuers as heroes.

"The one who came to us truly saved our lives, bailing water from the raft. We were helpless and he saved us," said David Aqqiaruq.

The tragic day began when Aqqiaruq and his 17-year-old son Leslie went walrus hunting last Wednesday, about 90 minutes from Igloolik.

David Aqqiaruq told the CBC that those who rescued him and his son near Igloolik last week are heroes. (CBC)

Both are experienced on the land, and both survived a rescue about two years ago from the same spot – the Fury and Hecla Strait between Baffin Island and the Melville Peninsula.

"Since when I was really young, I've been out with my dad," said Leslie.

The two set out in good weather Wednesday morning and were soon successful, bringing down a walrus to take back to Igloolik to share with their family and the community.

But the weather deteriorated. The winds rose, temperatures fell and sea ice began to form.

"We tried to go home, but the ice was too thick and we couldn't move. It was really a big surprise."

Weather

Severe weather warnings or watches in effect for:

Sachs Harbour Ulukhaktok

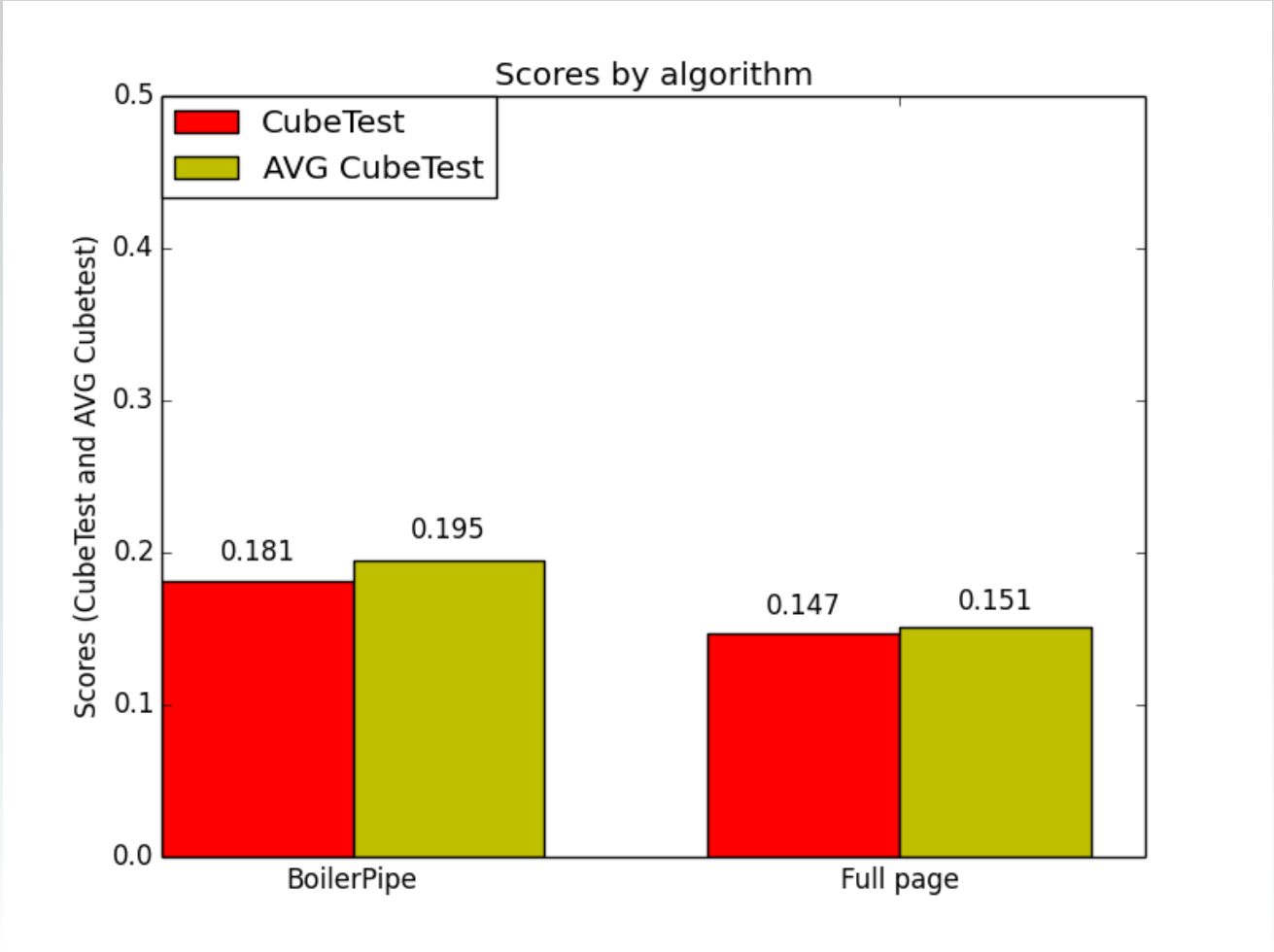
Whitehorse	Yellowknife	Inuvik	Iqaluit	Kuujjuaq
-18°C	-15°C	-29°C	-25°C	-10°C

More Weather

Contaminated mine 'an embarrassment to Canada', says Yukon judge

Ground patrol to investigate mysterious Arctic 'ping' sound

Comparaison du système avec ou sans prétraitement sur les requêtes concernant Ebola



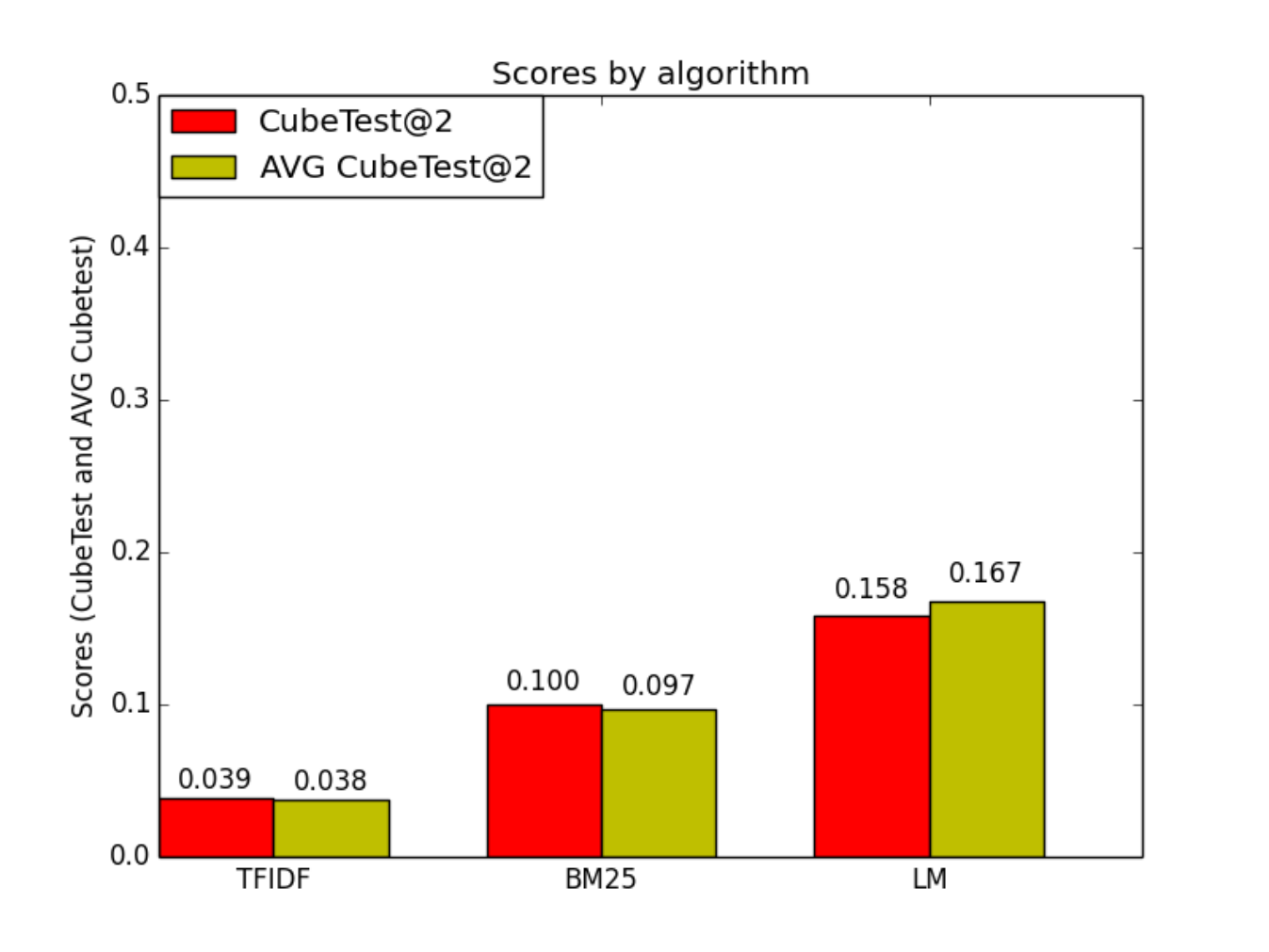
Processus de recherche d'information

- Moteur de recherche : Solr avec différentes mesures de similarités
- Algorithmes de diversification:
  - Latent Dirichlet Allocation (LDA)
  - K-Means

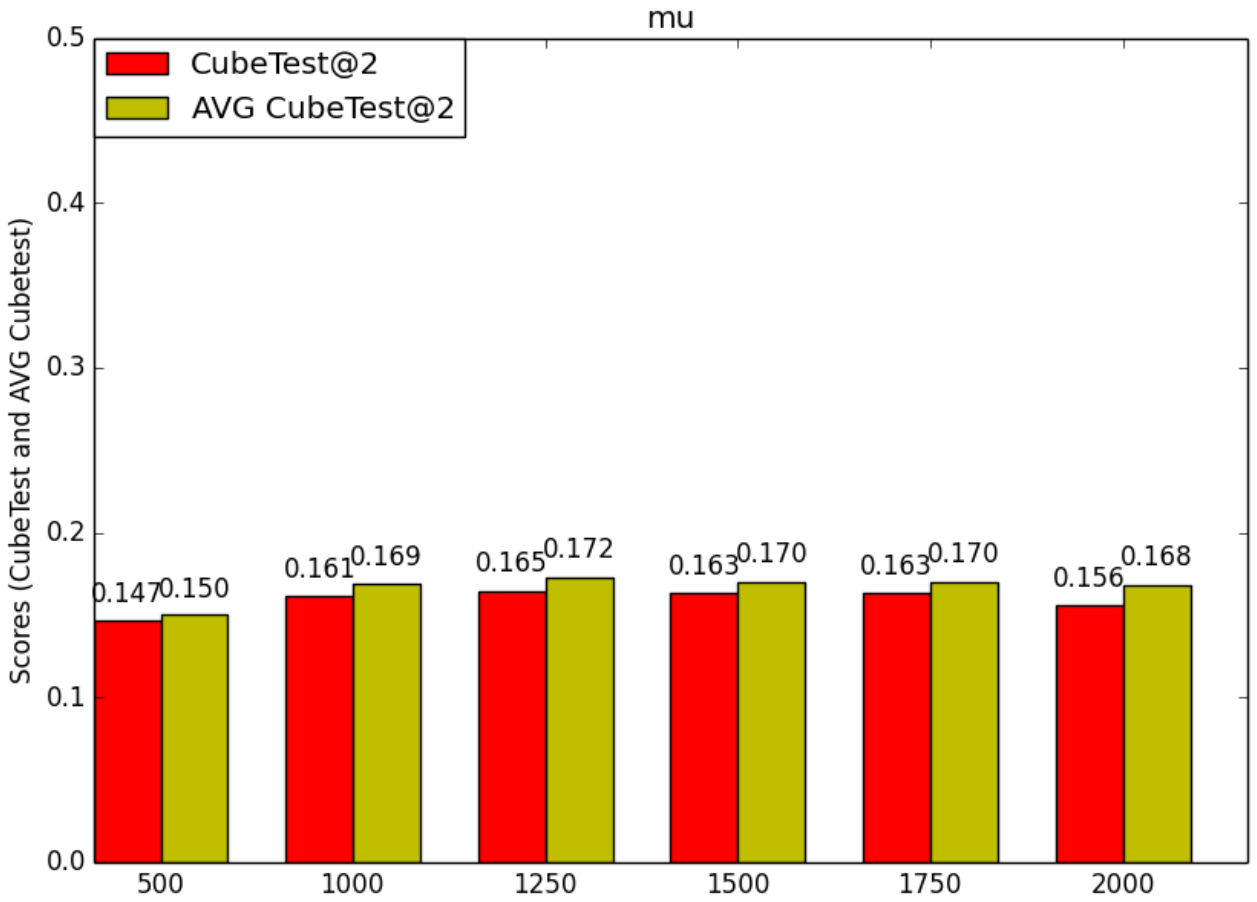
Mesures de similarités

- Modèle vectoriel : TF * IDF
- Modèle Probabiliste : Okapi BM25
- Language Model (Bayes) : LM
  - Unigramme

Comparaison des mesures de similarités

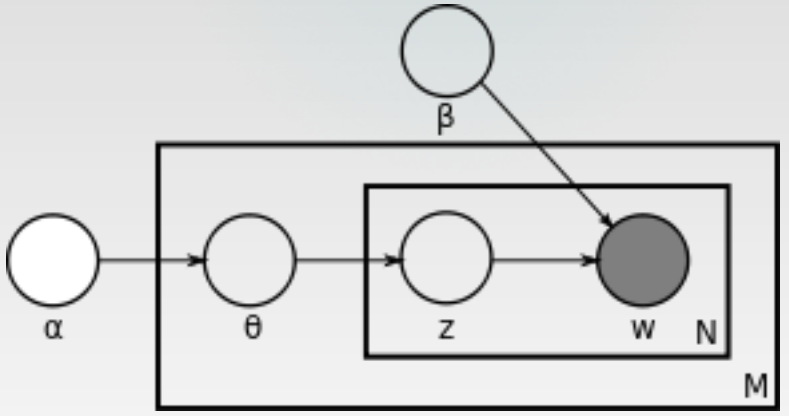


Prior LM dirichlet



Latent Dirichlet Allocation (LDA)

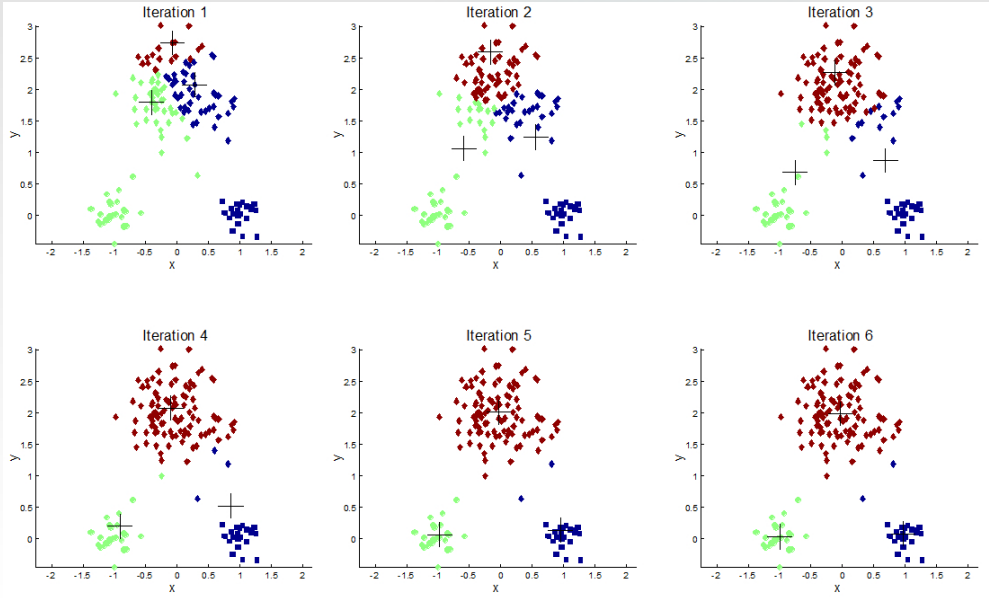
- Modélisation de topics
- Choix du nombre de topics fixé
- Extraction des mots les plus probables de chaque Topic
- Expansion de requête avec chaque topic



Pr. David Blei (Columbia University)

K-means Clustering

- À partir d'une recherche le moteur nous retourne les documents les plus pertinents, maisi ils peuvent être très similaires
- Besoin de divserification pour couvrir des topics différents
- Choix du nombre de groupes déterminé (Nombre de sous-topics visé)



Segmentation et sélection de phrases

- Les articles peuvent n'avoir qu'une partie de leur contenu pertinente par rapport à l'intérêt de notre utilisateur

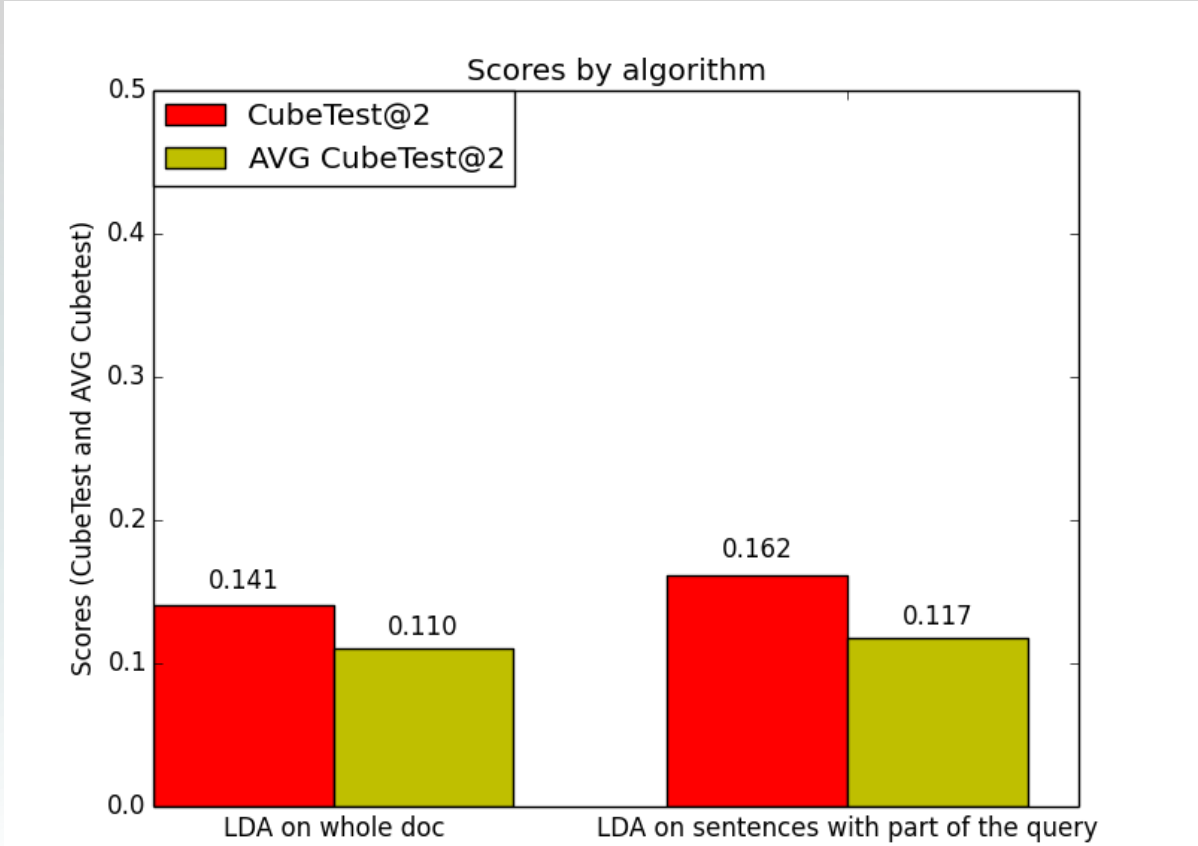
- La segmentation en phrase peut nous permettre de nous concentrer sur le contenu autour de notre sujet

- Le fait de se concentrer sur les éléments peut nous permettre de diminuer le bruit global du jeu de données, et d'avoir une meilleure modélisation. (Mais aussi perte d'information)

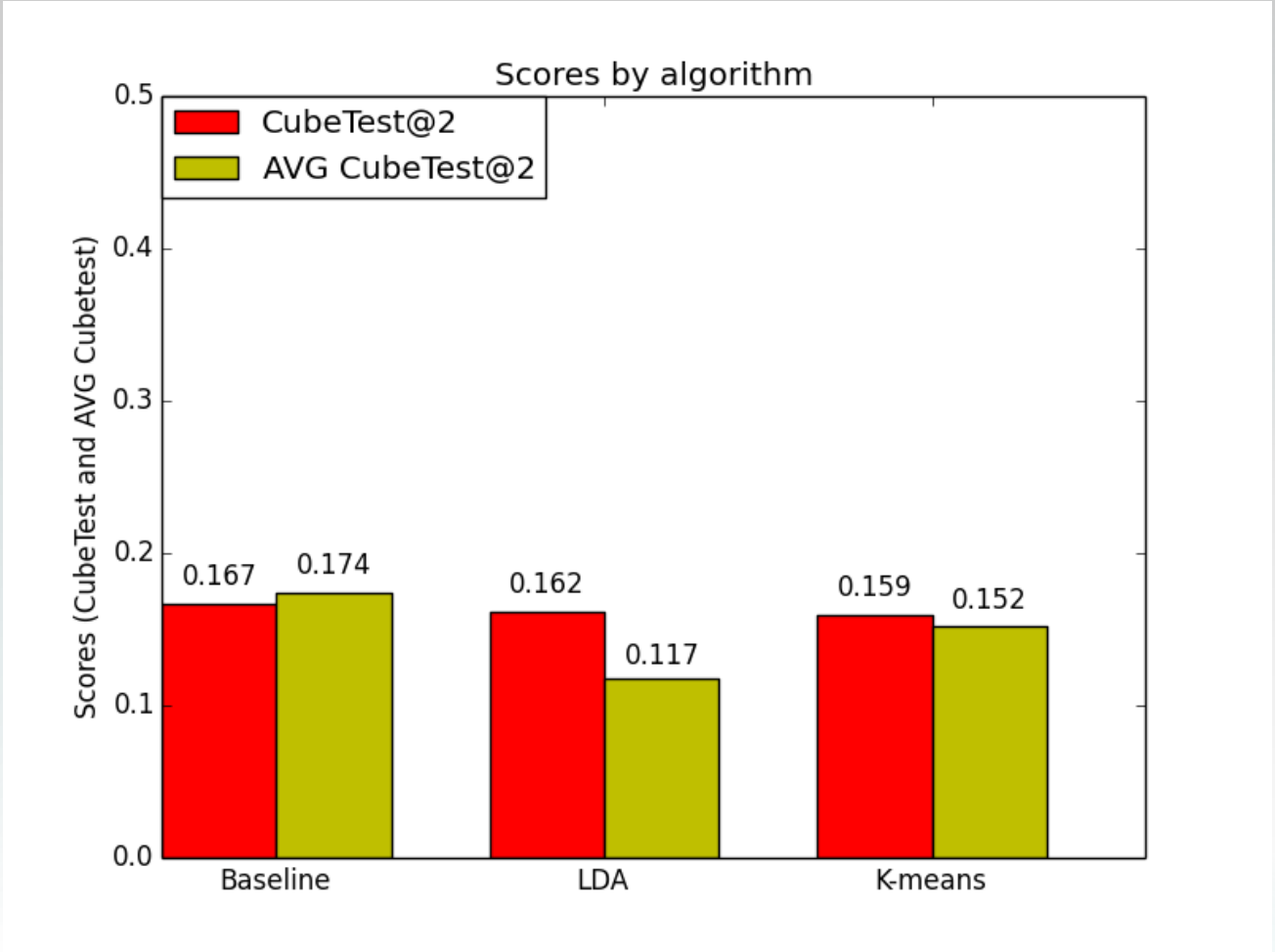
Segmentation et sélection de phrases

- Example with « DANIEL BEREHULAK » query :
  - Daniel Berehulak is a freelance photographer represented by Reportage by Getty Images
  - For the past six weeks photographer Daniel Beherulak has been covering the virus' deadly spread for the New York Times
  - When Berehulak first arrived in Monrovia on Aug. 22 – armed with 300 pairs of gloves 35 Personal Protective Equipment suits goggles surgical face masks hand sanitizers and countless rolls of tape – he met with Getty Images photographer John Moore who had been covering the Ebola crisis for a week.

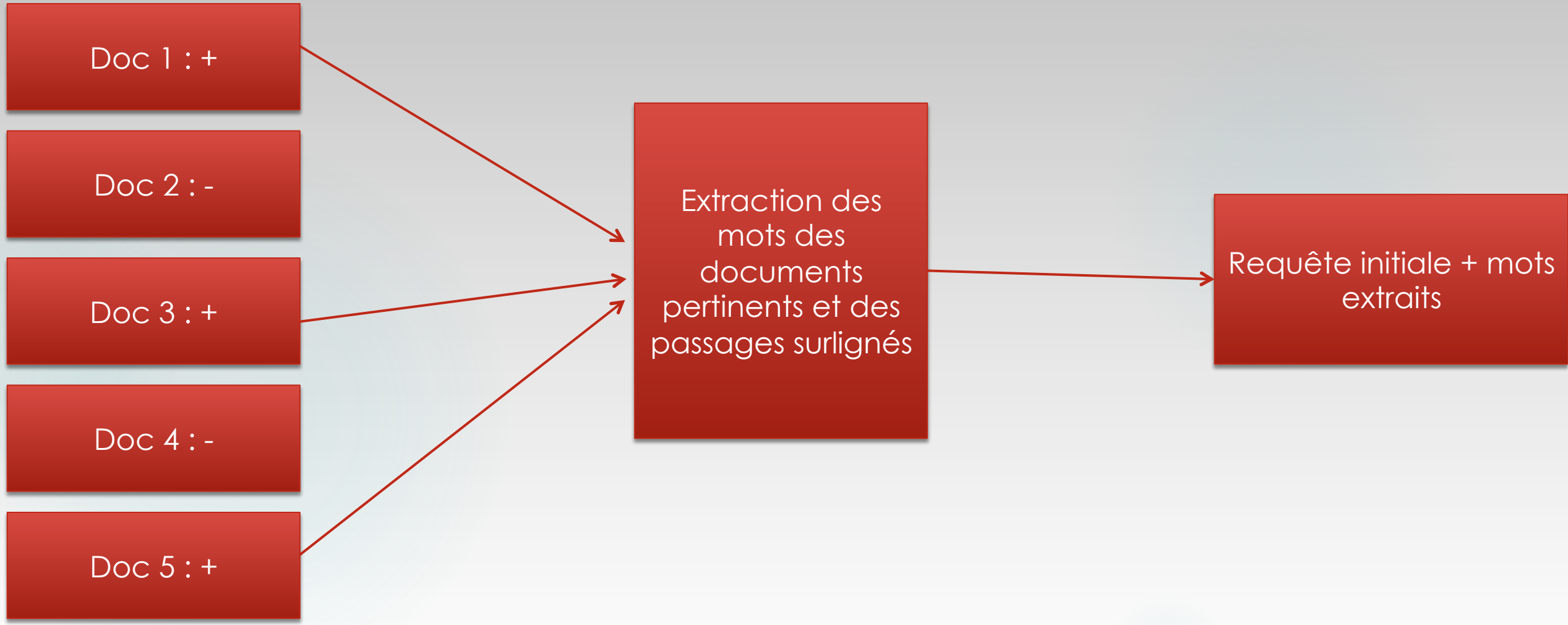
Comparaison des résultats de l'algorithme LDA appliqué à tout le document vs sur les phrases pertinentes



Comparaison des algorithmes de diversification avec notre système de base



Feedback processing



Reconnaissance d'entités nommées

- Les passages de textes peuvent être longs
- Il faut extraire des informations qui ont un certain sens
- Les entités nommées sont multi-domaine (Entités géopolitiques, organisations , personnes)

Reconnaissance d'entités nommées

Retour sur l'exemple précédent

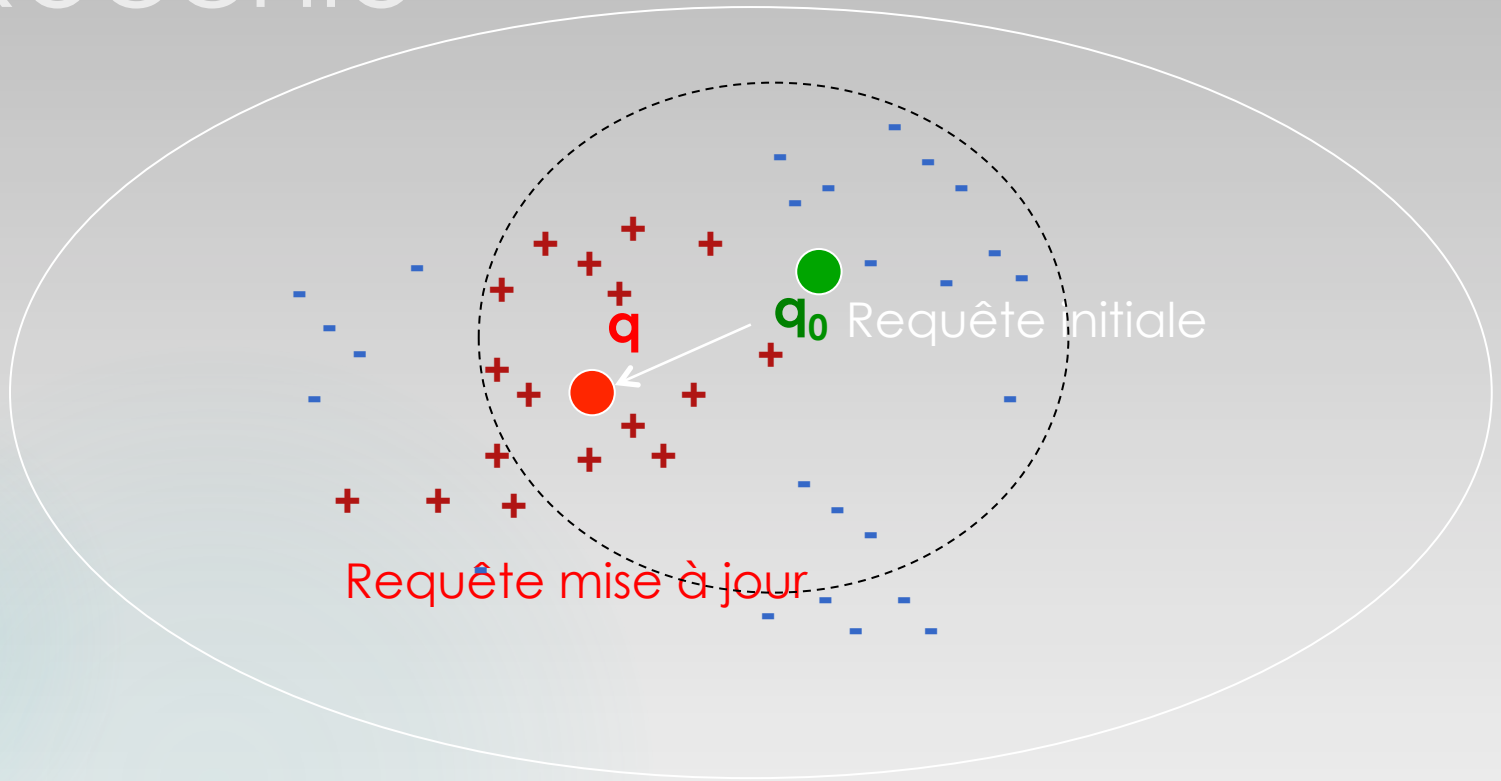
- When Berehulak first arrived in Monrovia on Aug. 22 – armed with 300 pairs of gloves 35 Personal Protective Equipment suits goggles surgical face masks hand sanitizers and countless rolls of tape – he met with Getty Images photographer John Moore who had been covering the Ebola crisis for a week.

Rocchio

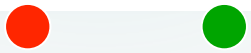
- Rocchio Relevance Feedback
- Trouver les mots commun des documents pertinents tout en soustrayant les mots apparaissant dans les documents non pertinents
- (a=1, b=0.75, c=0.25)

$$\overrightarrow{Q_m} = (a \cdot \overrightarrow{Q_o}) + \left(b \cdot \frac{1}{|D_r|} \cdot \sum_{\overrightarrow{D_j} \in D_r} \overrightarrow{D_j}\right) - \left(c \cdot \frac{1}{|D_{nr}|} \cdot \sum_{\overrightarrow{D_k} \in D_{nr}} \overrightarrow{D_k}\right)$$

Rocchio

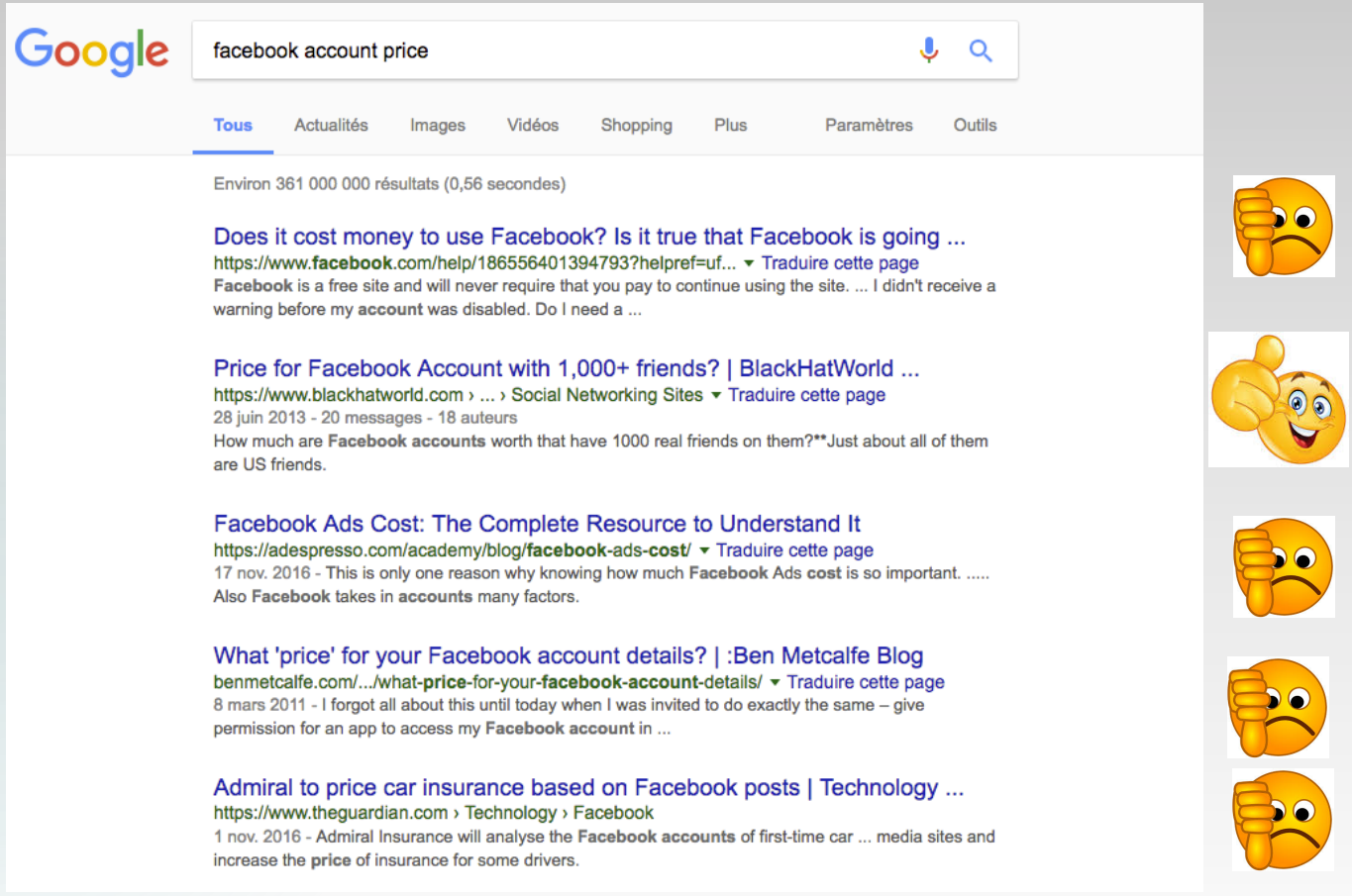


$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$



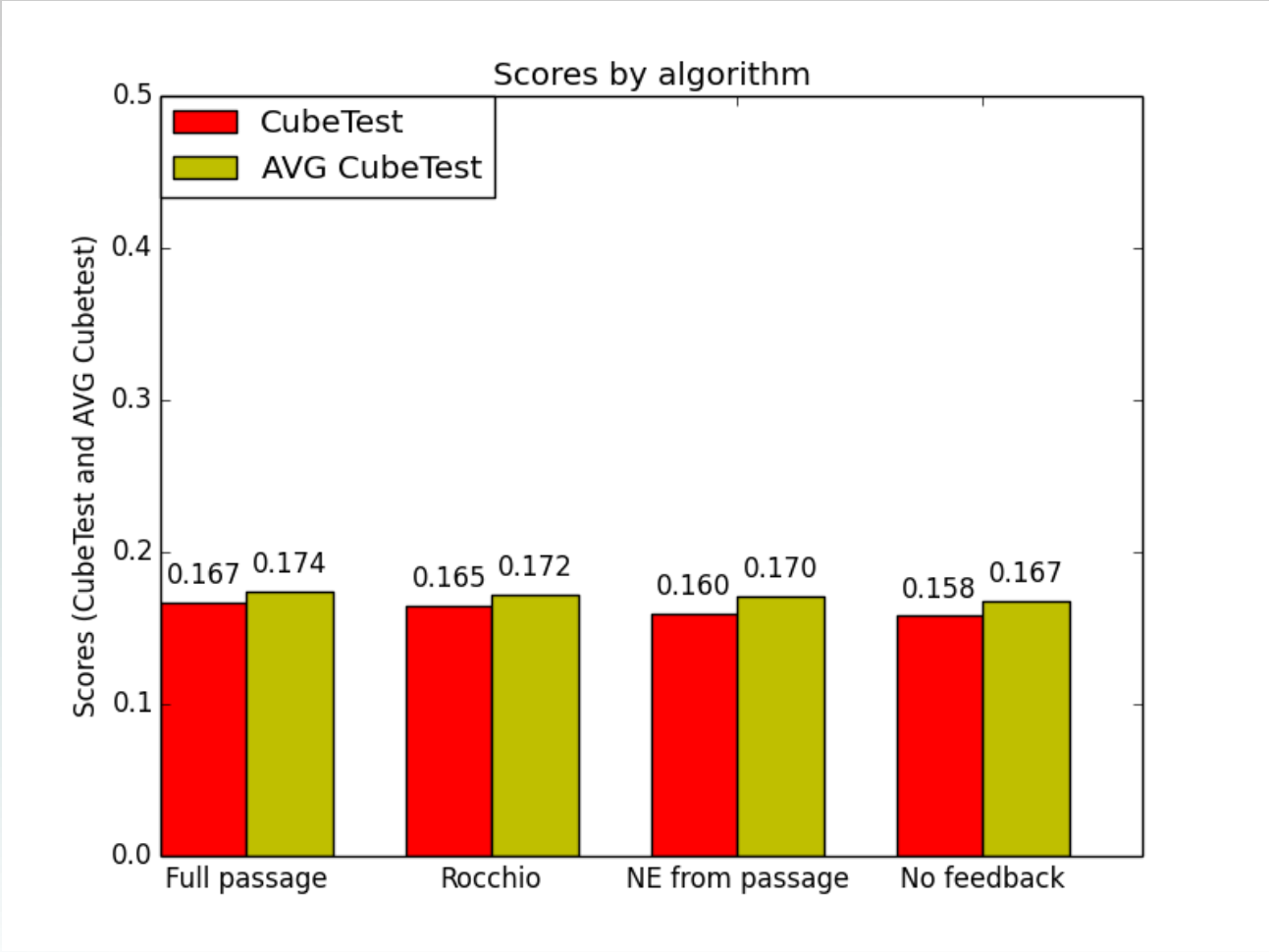
Exemple de recherche interactive

Facebook account price



Relancer la recherche…

Comparaison du retour de l'utilisateur



Comparaison avec l'état de l'art

Conclusion

- Le prétraitement avec BoilerPipe montre une amélioration intéressante du système
- La segmentation en phrase est capable d'améliorer le système en utilisant un algorithme de diversification
- Les algorithmes de diversifications n'ont pas l'impact espéré sur le système
- Travaux futurs:
  - Modélisation des domaines avec Word2Vec
  - Travailler à un critère d'arrêt intelligent (mesure du gain par topic)

Questions ?