Webscale Semantic Indexing Project

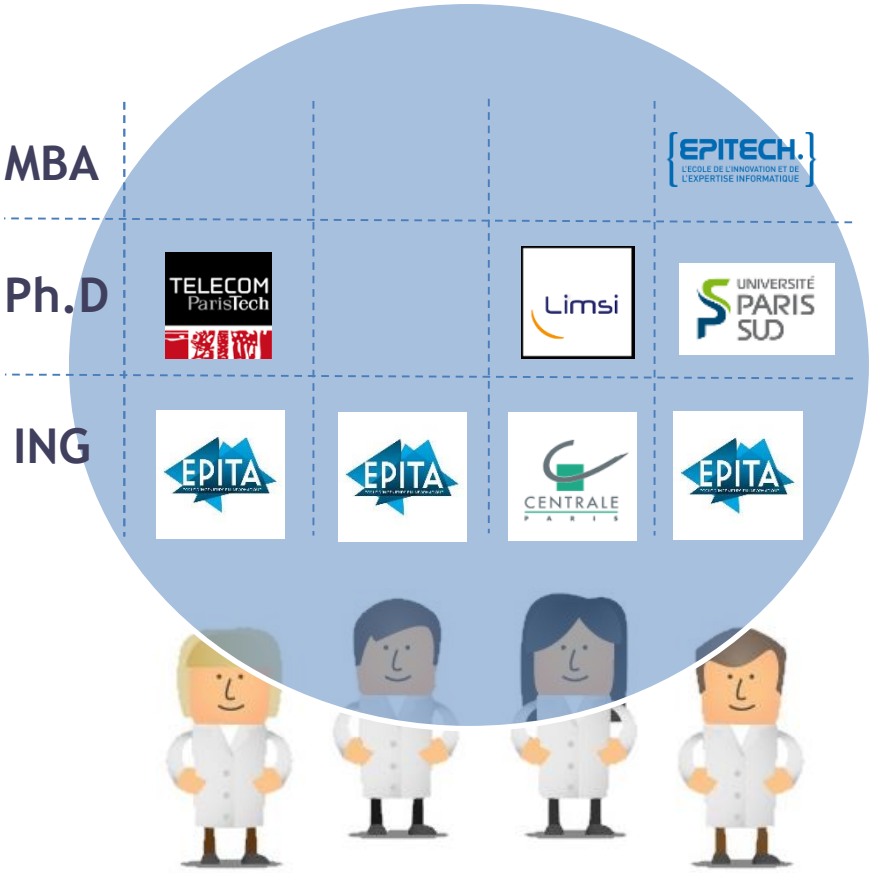
can we index the web with embeddings ?

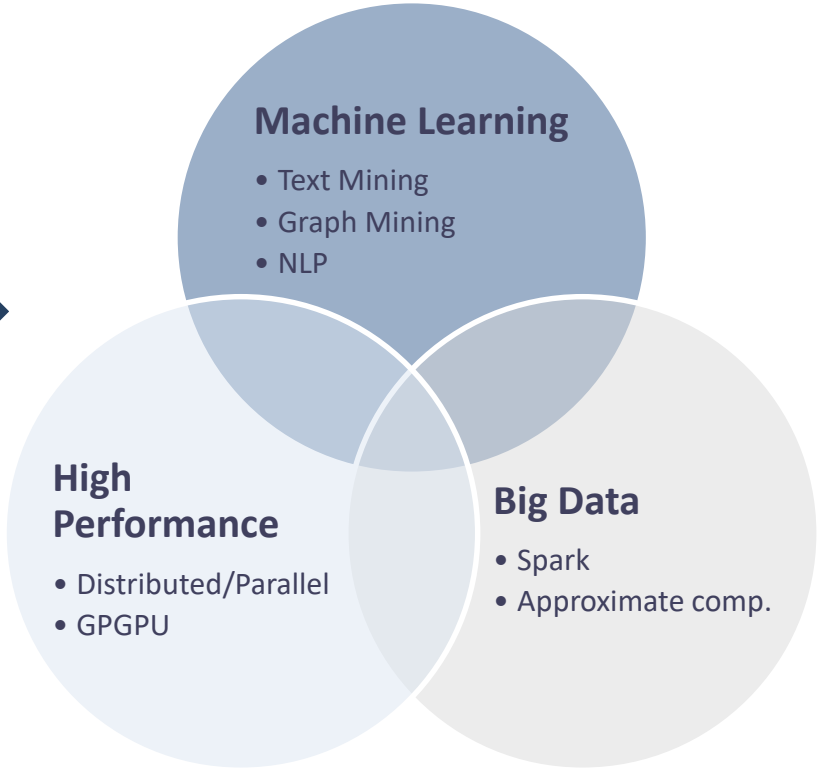
Présentation Atelier RISE

15 mai 2018

Guillaume PITEL

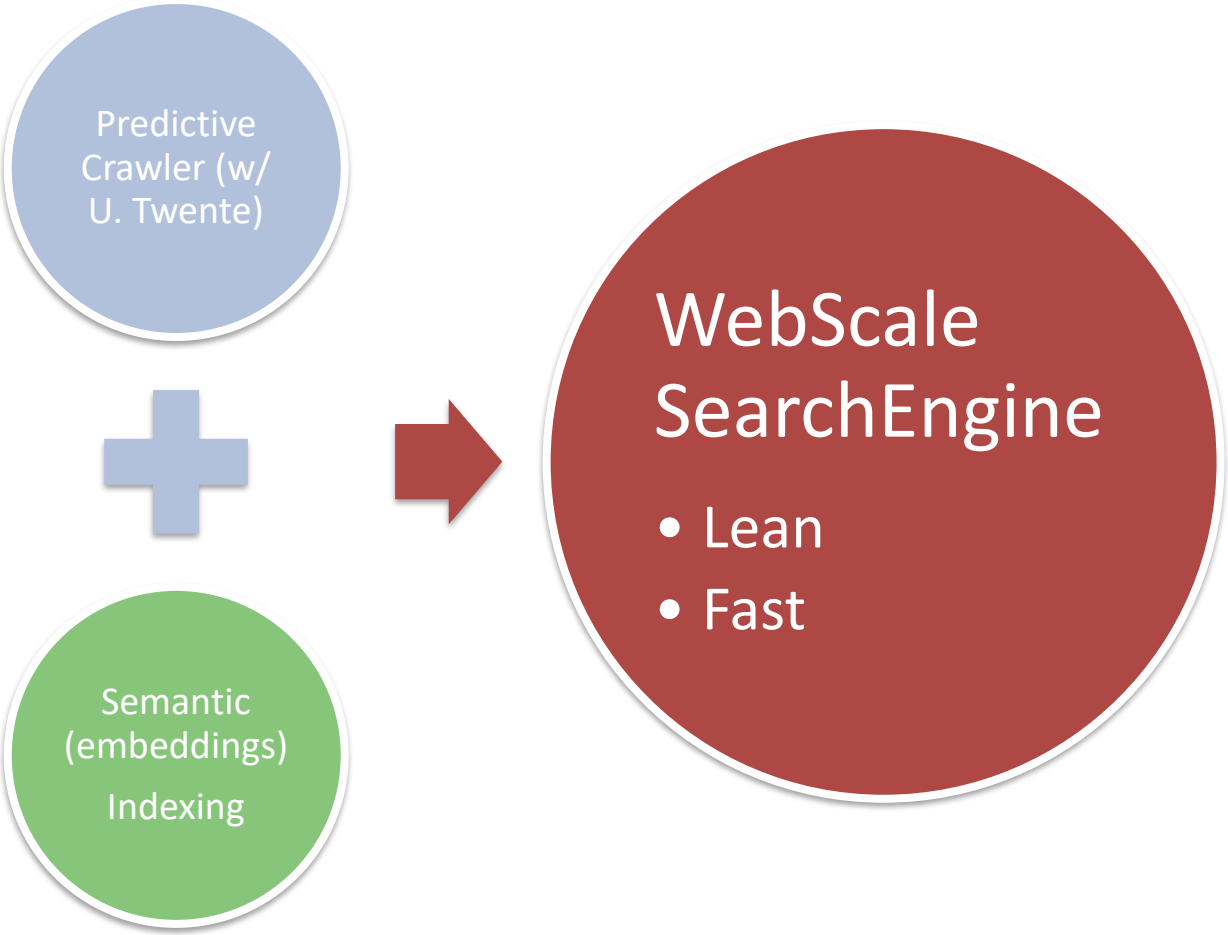
eXenSa : the team





… and the skills

eXenSa : le projet



eXenSa : les technos existantes

A l'origine, créé suite à une découverte en Machine Learning, eXenSa a développé des algorithmes aux performances souvent 10 à 100 fois supérieures à l'état de l'art avec une pertinence équivalente ou supérieure.

NCISC-Batch & AP	Cout-Min Tree Sketch	NCISC-Online	Fast Index+Query	Crawling
•Factorisation rapide •**Text/Graph/Link Behaviour** •**Side Information** •Amélioration qualitative par normalisation	•Comptage Approximatif pour données Zipf-like •10 fois plus compact •**Plus rapide que des solutions « natives »**	•**Apprentissage en continu** •Hybride Neural/Algébrique	•Indexation Approximative ultra-rapide •Pour la recherche de plus proches voisins avec des recall > 95%	•Large-scale •Intégré avec mesures de graphe (TrustRank/Centrality, etc.)

eXenGine Import → Prepare → Analyze → Index → Serve (API)

…implémentés dans une solution complète

Pour répondre à des besoins diversifiés…

- Classification
- Segmentation
- Recommandation
- Prédiction
- …

Tâches

- Sémantique
- Comportemental
- Relationnel

Data

- Moteurs de recherche
- Réseaux sociaux
- E-Marketing
- E-Reputation
- E-Commerce

Applications

Wolters Kluwer TARGET2SELL LeGuide.com AUGURE criteo

…validés par nos clients

NCISC-Batch (factorisation de matrice)

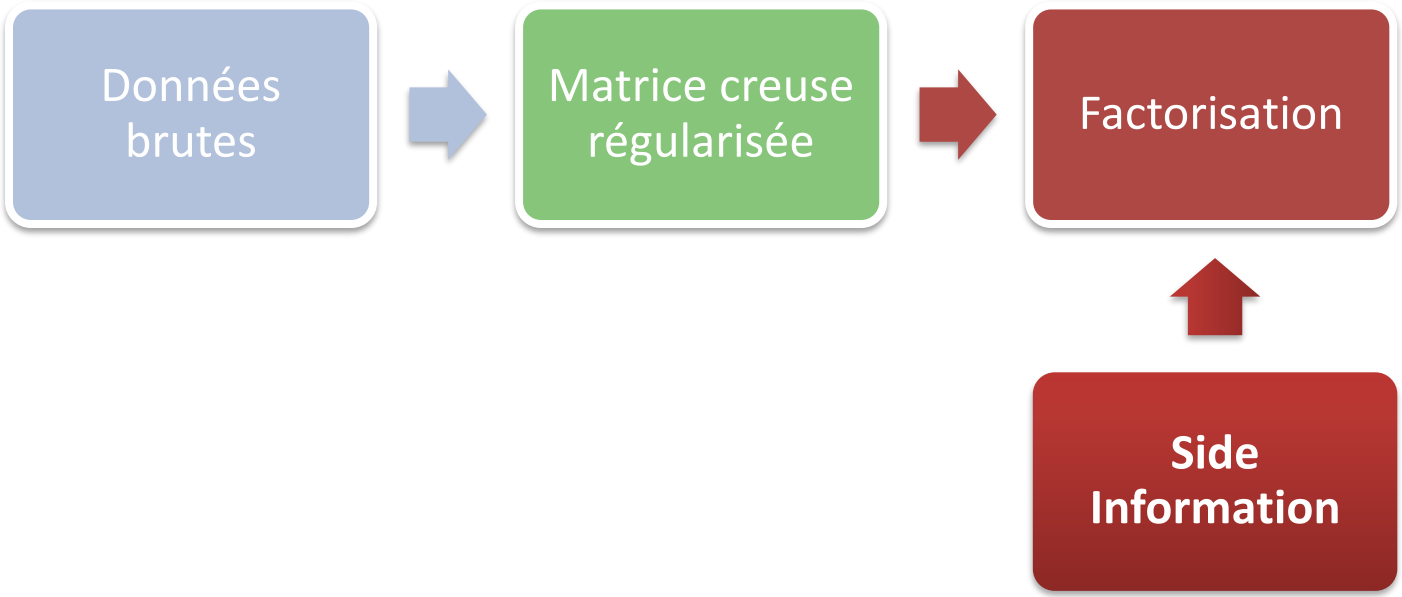
Alternatives

- **Batch**: Alternating Least Squares, Singular vector decomposition
- **Online**: Gradient Descent, Autoencoders, Word2vec / Paragraph2Vec

Points forts

- Très rapide (Wikipedia : 2-5min vs 30-60min with Word2vec)
- Hyperparamètres de régularisation & normalisation internes

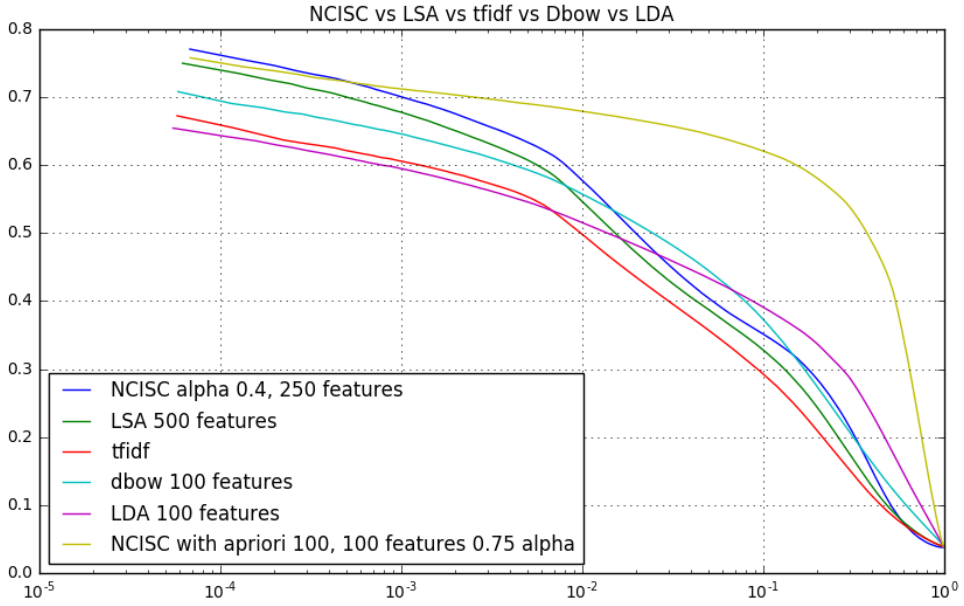
NCISC-AP (apriori)

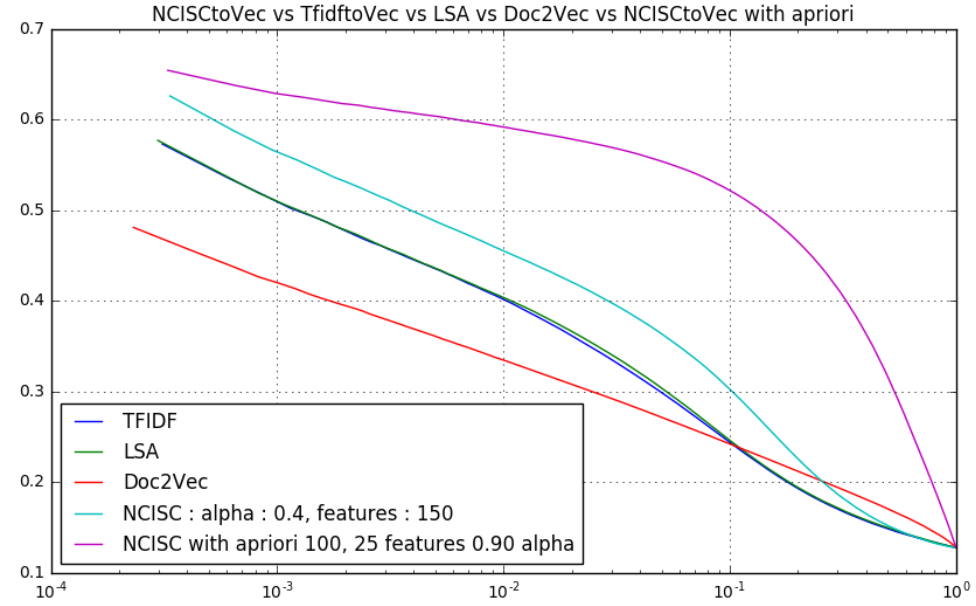


Points forts

- Permet d'injecter de la « side information » arbitraire offrant des améliorations de performance ciblées
- Coût négligeable dans la factorisation

Performance (Doc2Vec)





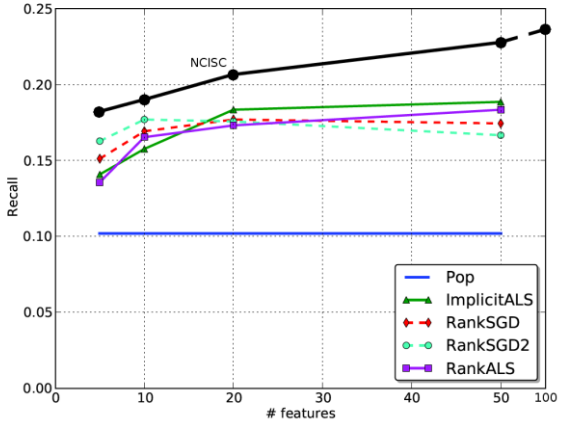


Figure 6: Recall on the Netflix data set

Count-Min Tree Sketch

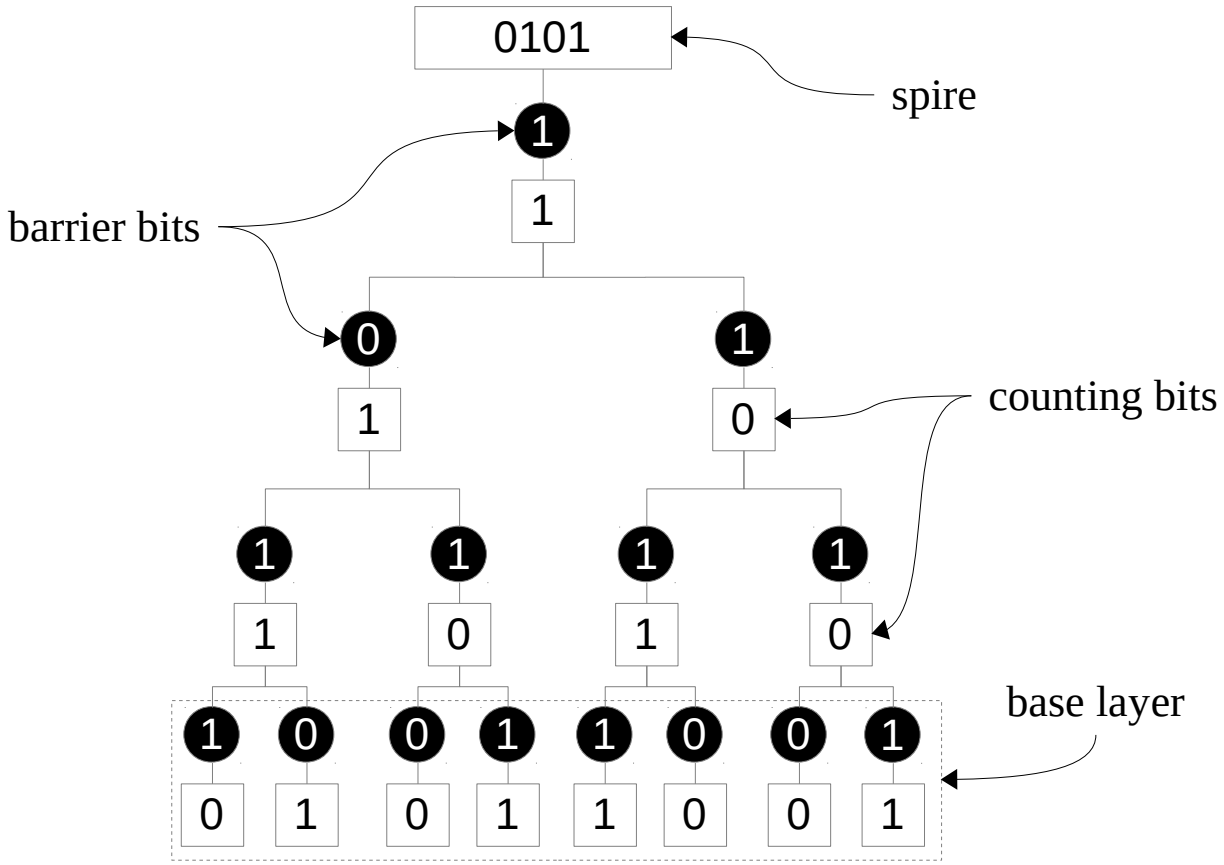
Problème du comptage d'événements	Solution classique	Solution approximative

• Mots + n-grammes
• Liens entre pages
• Likes/Views
	• Stockage Clef-Valeur
	• Distribution / réduction / join
	• Environ 8 fois la taille du vocabulaire
		• Count-Min Sketch (5*voc)
		• Et nos améliorations :
		  • Count-Min Log (2.5*voc)
		  • Count-Min Tree (données power ou Zipf) (0.8*voc)

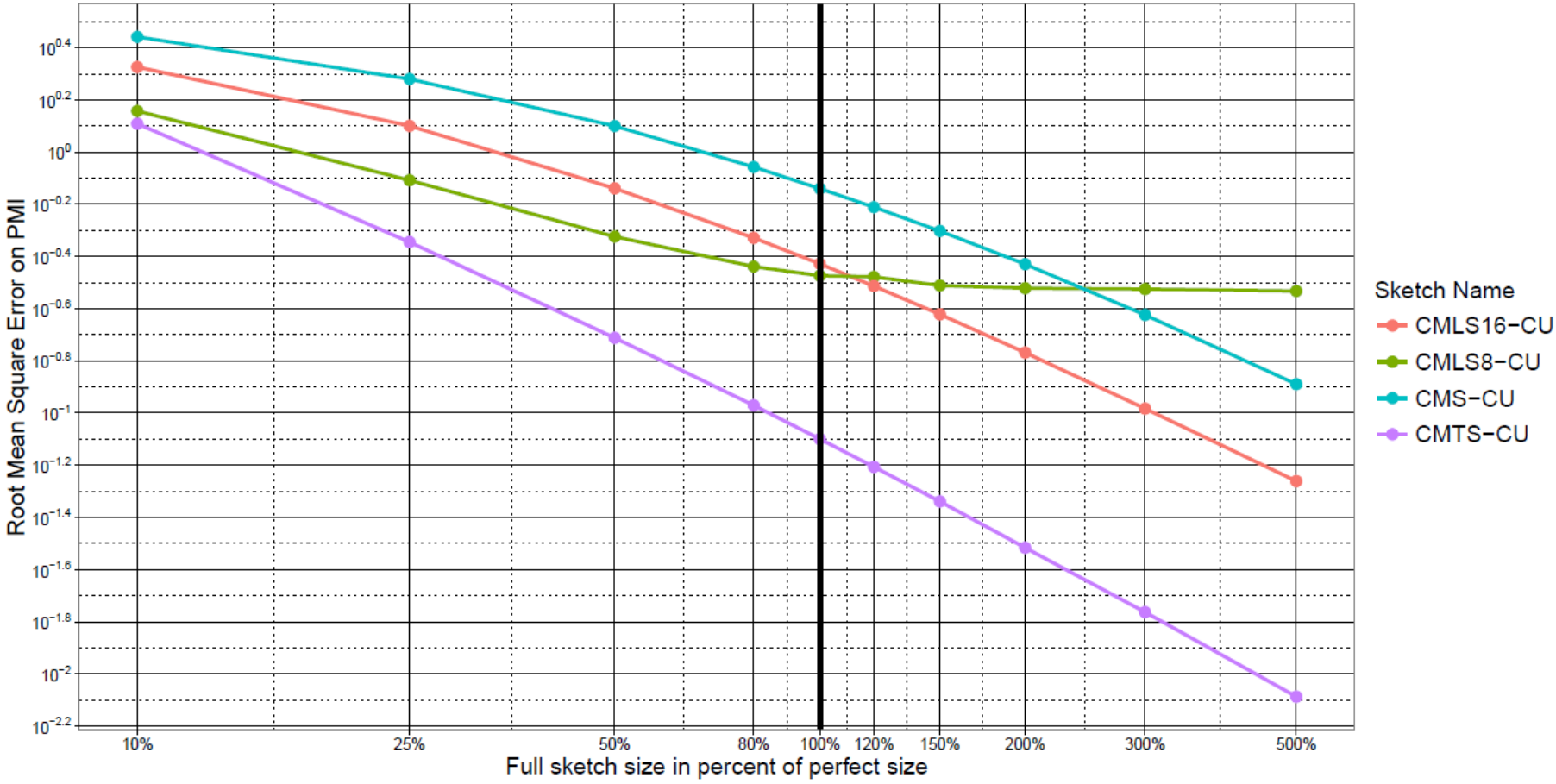
Points forts

- Permet de distribuer localement les counts sur les nœuds compute et service
- Vitesse >= hashmap native

Count-Min Tree Sketch



Count-Min Tree Sketch



Comptage approximatif

Projet de recherche

- Réduire le temps d'accès (et l'occupation mémoire)
- Objectif : combiner notre approche avec des fingerprints
- Idée : plus le nombre stocké est grand, plus le fingerprint est grand, réduisant les risques de collisions

Indexation

Problème du plus proche voisin

- Vecteurs représentant les entités (documents, mots, etc.)

Solution classique

- Recherche exhaustive en O(n)

Solution approximative

- Locality sensitive Hashing
- Notre solution :
  - Multi-probe KLSH avec clustering hiérarchique

Points forts

- Maximise le ratio qualité / temps d'indexation * temps de query
- Excellent passage à l'échelle

Indexation

Projet de recherche

- Remplacer la mesure de distance (euclidenne ou cosinus) par un score arbitraire pouvant notamment intégrer du ranking
  - Objectif : ne pas avoir à faire 2 passes (similarité + rang) lors de la query
  - Idée : se baser sur du soft clustering pour que certains éléments (typiquement ceux avec les scores élevés) soient présents dans plusieurs clusters
- Autoriser l'indexation continue avec la mise à jour des clusters (aujourd'hui les clusters sont fixes à l'issue d'une première passe)

NCISC-Online

Points forts

- Apprentissage Hybride algébrique / gradient descent
- Pour les cooccurrences, équivalent à SGNS sans le coût du Negative Sampling
- Pour les docs, 1 seule passe par document, apprentissage du modèle de mot en simultané

Vitesse

- Crawl ~30TB compressed, 1.2B pages, 450B tokens
  Cluster : 30x16cores // Storage AWS::S3
- Count : 1,3h
  Learn : 1h
- Generate embeddings : 0,75h

NCISC-Online

Projet de recherche

- Compression des embeddings (binarisation)
- Compression hiérarchique (représentation d'un élément par plusieurs vecteurs combinés)
- Résolution des problématiques temporelles : comment faire évoluer les représentants des éléments tout en étant capables de conserver les index passés (avec U. Twente)

Crawling Prédictif (avec U. Twente)

Performances actuelles

- Environ 1000 pages stockées/secondes pour un serveur 4 cœurs + 32GB ram

Objectif

- Prédire les fréquences de crawl des pages (projet avec U. Twente)
- Pour réduire les fetch inutiles
- Et avoir un délai de rafraîchissement minimal

Moyens

- Indicateurs de dynamique d'évolution du graphe
- Indicateurs endogènes
- Prédiction par transfert via des features sémantiques

Cas d'usage et limites

Moteurs de recherche par similarité

- Wikinsights.org
- Sites.exensa.net
- D'autres démos sur les brevets, articles scientifiques

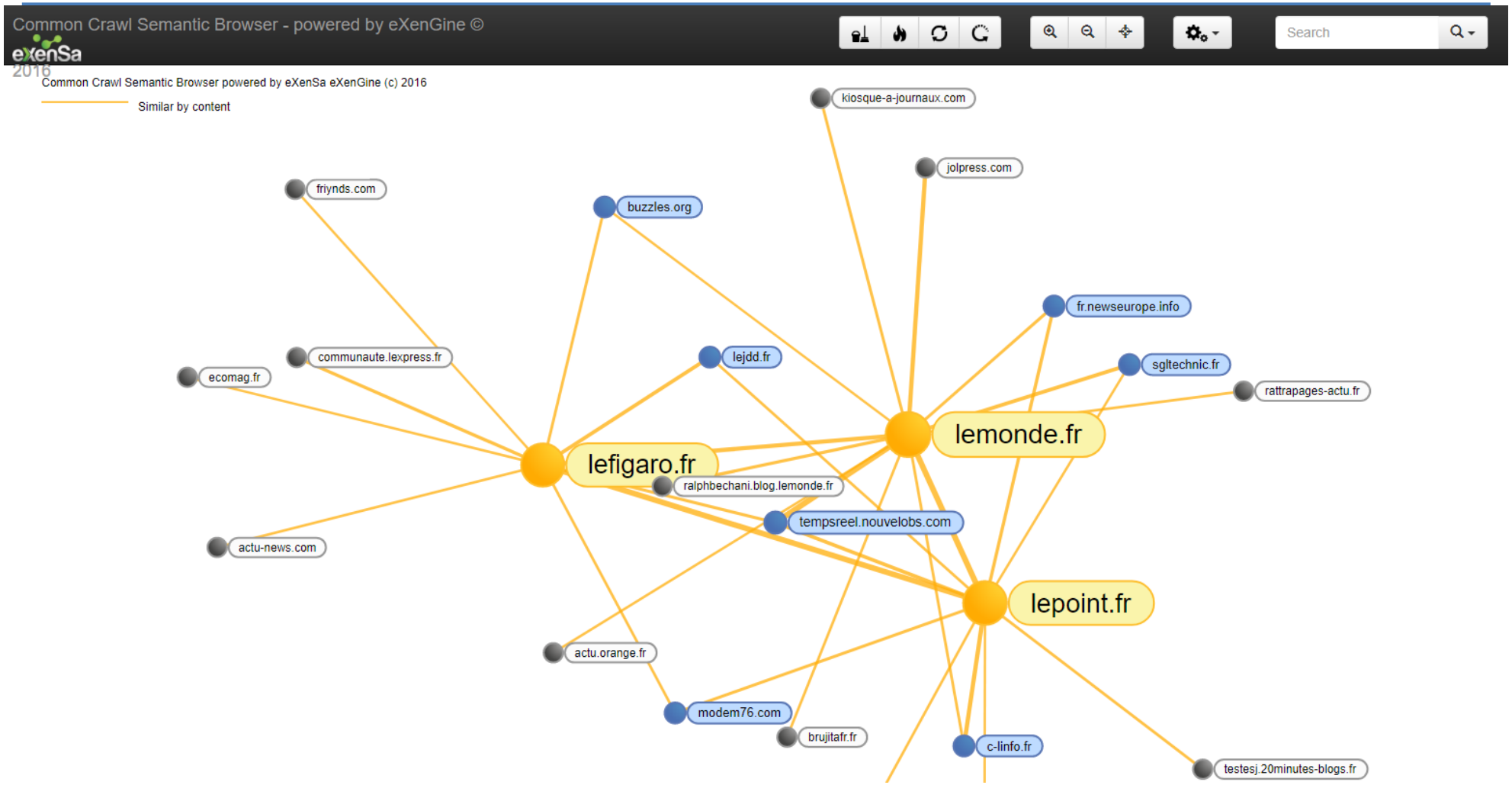
Limites

- On remarque de vrais problèmes selon la taille des documents, y compris sur la similarité doc/doc
- Les requêtes « mots vers document » ne fonctionnent bien que sur des petits documents

Projet

- Découper les documents en « snippets » hiérarchiques, incluant les éléments clefs jusqu'au paragraphe
- Exemple : « site, url-path, title, h1, h2, list item, text chunk »

Crawl / site-level similarity



Conclusion

Crawling

- Add page update prediction for improved and fresher index

Continuous Indexation

- Allow cluster updates to adapt to new documents and index continuous modificaiton

Structures des documents

- Adapt page representation to work with embeddings
- Manage smart weighting of documents with internal information (for instance taking into account the position of a snippet inside the page)

Conclusion

Des remarques ?

Des questions ?

On recherche des partenaires !

Merci !