Chantier d'usage: NeoTEX

B. Audeh, M. Beigbeder, C. Largeron

Laboratoire Hubert Curien

7 juin 2017

L'équipe

- Bissan Audeh, post-doctorante 10 mois
  - septembre 2016–juin 2017
- Michel Beigbeder
  - École des mines de Saint-Étienne, Laboratoire Hubert Curien
- Christine Largeron
  - Université de Saint-Étienne, Laboratoire Hubert Curien
- Ayman Alazizi, stagiaire Master 2
  - avril–août 2016
- Diana Ramirez, stagiaire Master 2
  - février–juin 2017

Contexte: un néophyte

Exploration de documents **T**extuels d'un domaine par un **Néo**phyte

Recommandation/**recherche** de documents pour l'initiation d'une recherche. Le néophyte

- ne connaît pas les mots-clés du domaine
- ne connaît pas les experts du domaine
- **ne connaît pas les articles de référence**

Buts et défis du projet

- Concevoir, implémenter, tester un cadre de travail et des modèles de recherche orientés vers les besoins d'un néophyte
- Défis
  1. définir les besoins d'un néophyte
     - qu'est-ce qui est pertinent pour un néophyte?
  2. concevoir des modèles qui retrouvent les documents pertinents au sens de 1.
  3. tester
     - construire des vérités terrains
     - obtenir les données

Plan

- panorama des bibliothèques numériques (BN)
- état de l'art
- le système NeoTEX
- les vérités terrains
- la construction du graphe de citation
- les modèles
- les expériences

Contexte: bibliothèques numériques (BN)

- collecter les documents
- mettre à disposition des documents
- gérer les droits
- gérer les usages
- numériser
- archiver, préserver
- *créer des méta-informations*
- **rechercher**
- *recommander*
- naviguer

Contexte: une bibliothèque numérique

- Recherche sur les méta-informations
  - auteur, titre, journal, date, ISSN, ISBN, etc.
  - recherche orientée vers ce que l'on connaît
- plus récemment aussi sur le résumé ou sur le texte intégral

Comparaison de quelques BN académiques (1/2): données

		données de rech.	références	citations
Google Scholar	universelle	texte intégral	références	citations
Microsoft Academic	universelle	texte intégral	références	citations
CiteseerX	informatique	texte intégral	références	citations
ACM DL	informatique	texte intégral	références	citations
dblp	informatique	méta-informations	–	–
Web of Science	universelle	méta-informations	références	citations

Vocabulaire:

- **références**: liens **sortants**
- **citations**: liens **entrants**

Comparaison de quelques BN académiques (2/2): services

		tri | présentation
Google Scholar	par champ	pert. | date
Microsoft Academic	$\pm$	pert. | date | cit.
CiteseerX	par champ	pert. | date | cit.
ACM DL	par champ	pert. | date | cit. | usage
dblp	par champ	date
Web of Science	par champ	date | cit. | usage | auteur | venue

liens cliquables

raffinements

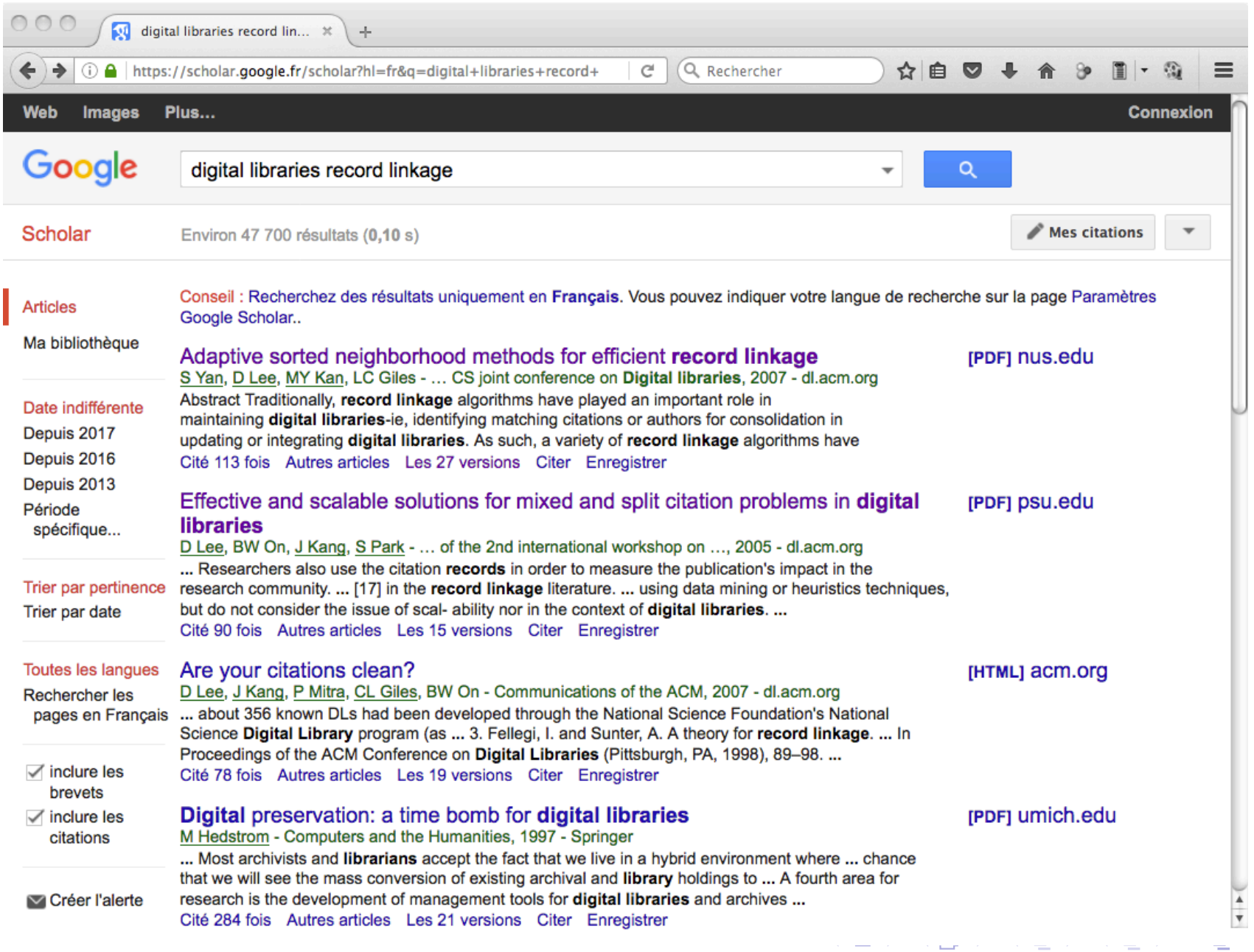
accès aux contenus, versions

informations sur les auteurs

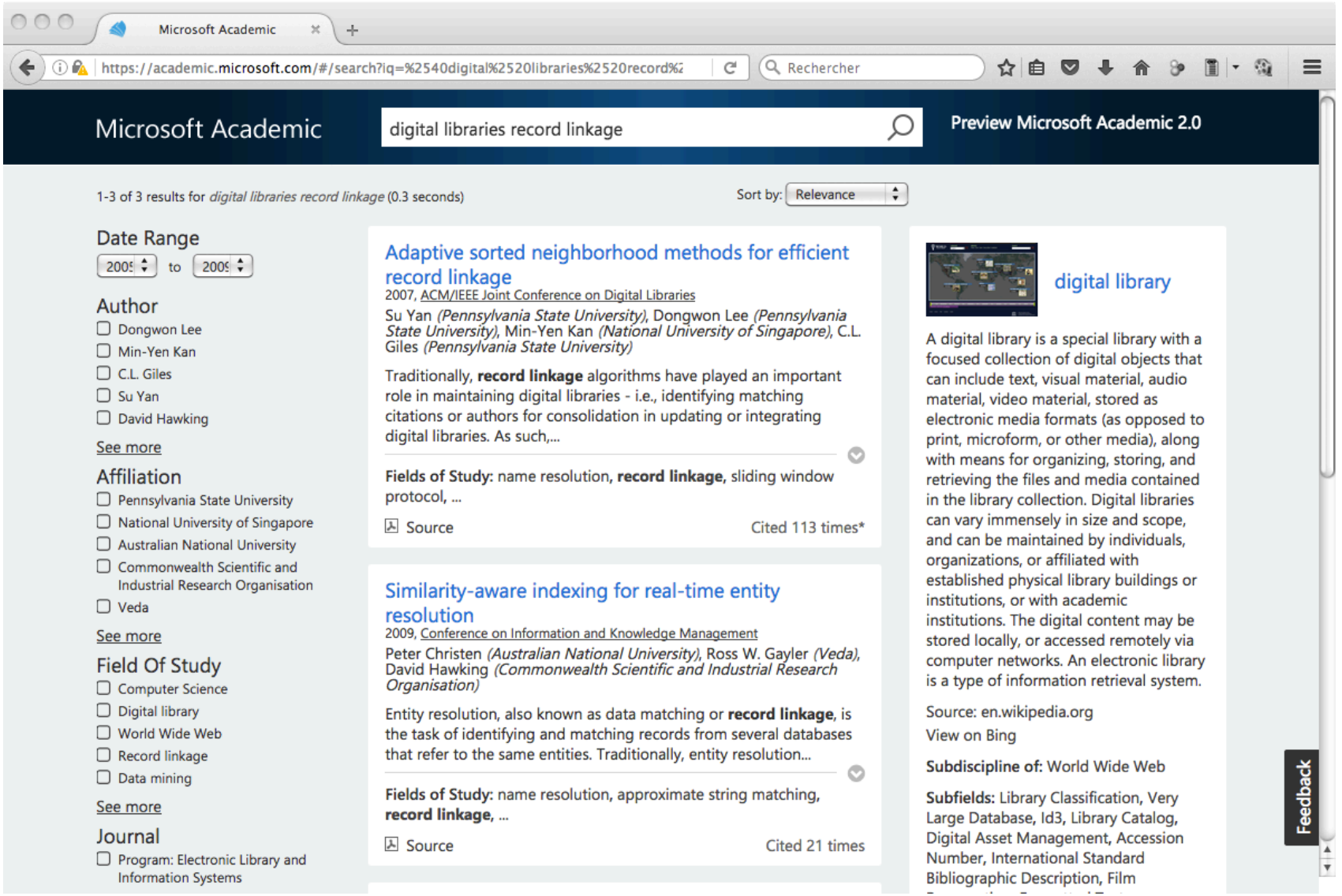
citations, références, co-citations, regroupements

panier, profil

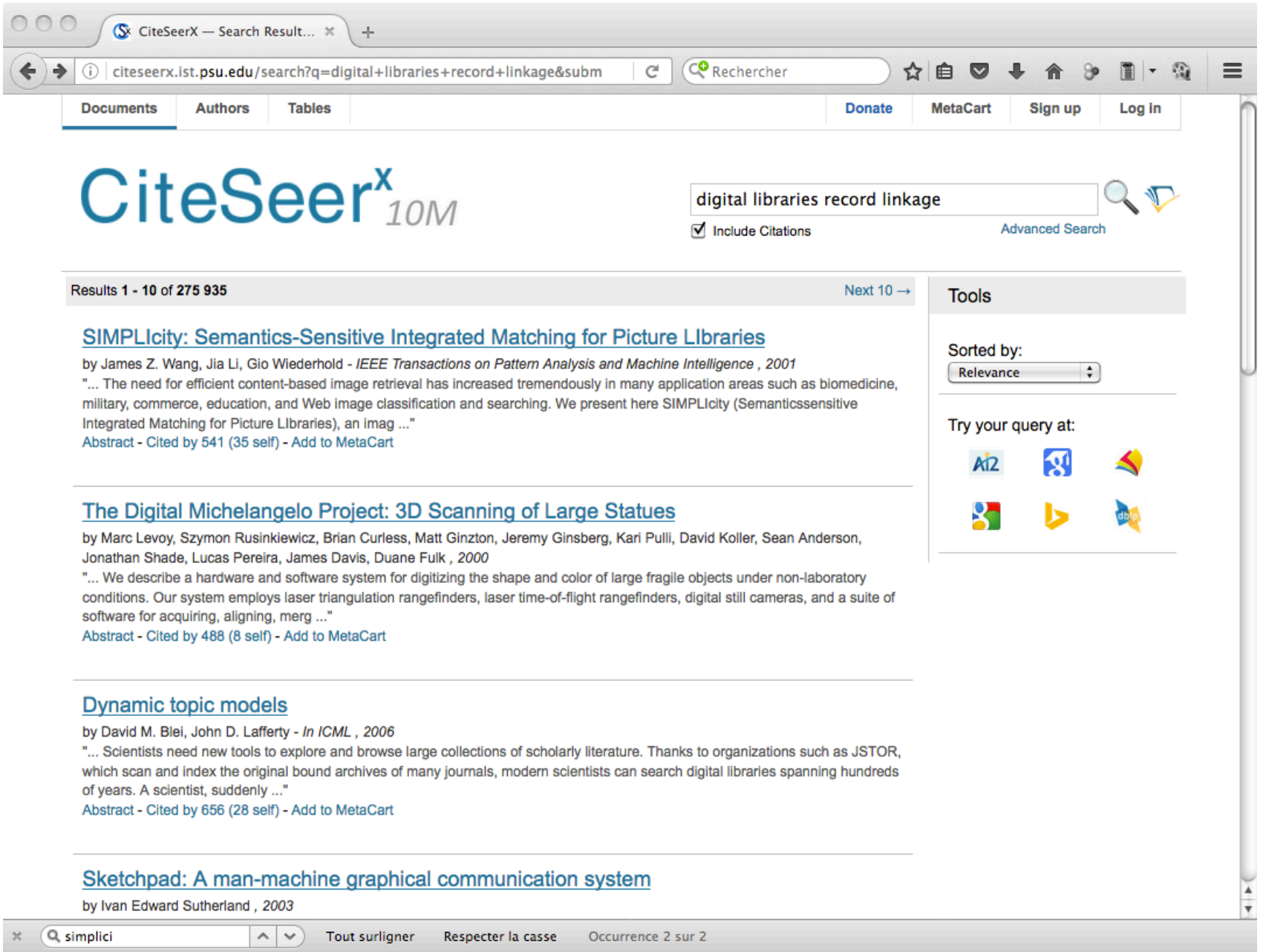
Google Scholar



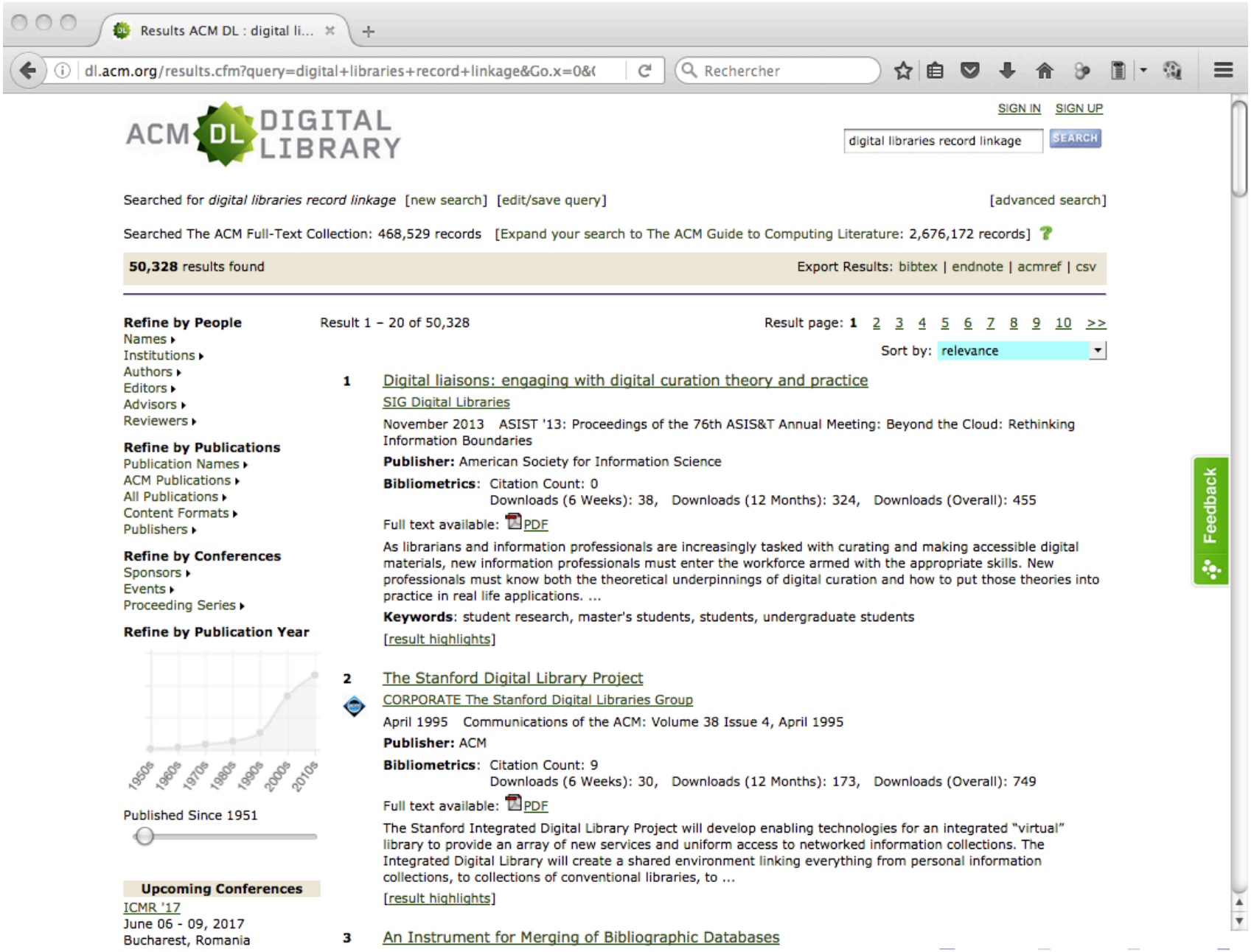
Microsoft Academic



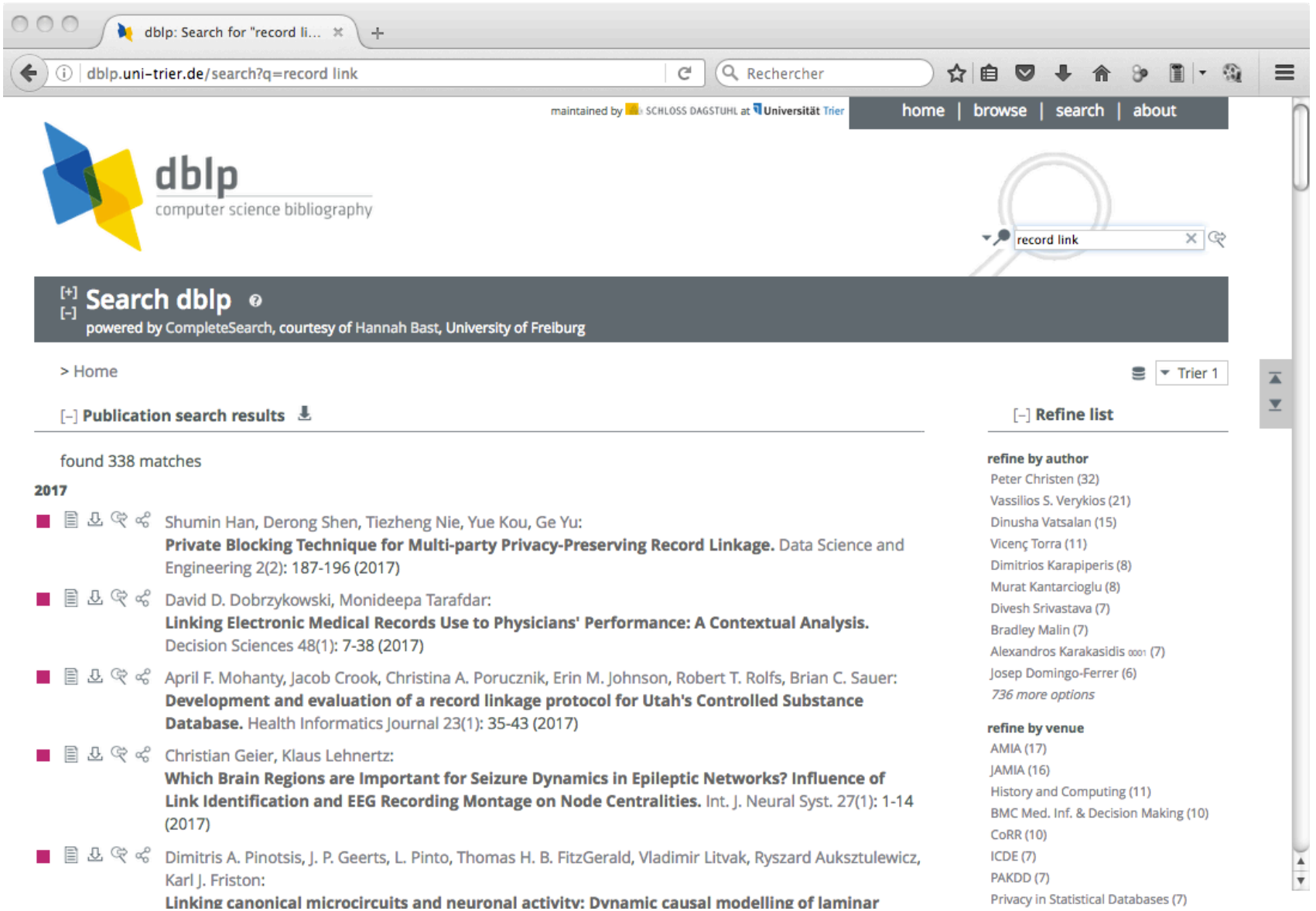
CiteSeer^X


CiteSeerX — Search Result...
citeseerx.ist.psu.edu/search?q=digital+libraries+record+linkage&subm
Rechercher
Documents
Authors
Tables
Donate
MetaCart
Sign up
Log in
CiteSeerX 10M
digital libraries record linkage
Include Citations
Advanced Search
Results 1 - 10 of 275 935
Next 10 →
SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture LIbraries
by James Z. Wang, Jia Li, Gio Wiederhold - IEEE Transactions on Pattern Analysis and Machine Intelligence , 2001
"... The need for efficient content-based image retrieval has increased tremendously in many application areas such as biomedicine, military, commerce, education, and Web image classification and searching. We present here SIMPLIcity (Semanticssensitive Integrated Matching for Picture LIbraries), an imag ..."
Abstract - Cited by 541 (35 self) - Add to MetaCart
The Digital Michelangelo Project: 3D Scanning of Large Statues
by Marc Levoy, Szymon Rusinkiewicz, Brian Curless, Matt Ginzton, Jeremy Ginsberg, Kari Pulli, David Koller, Sean Anderson, Jonathan Shade, Lucas Pereira, James Davis, Duane Fulk , 2000
"... We describe a hardware and software system for digitizing the shape and color of large fragile objects under non-laboratory conditions. Our system employs laser triangulation rangefinders, laser time-of-flight rangefinders, digital still cameras, and a suite of software for acquiring, aligning, merg ..."
Abstract - Cited by 488 (8 self) - Add to MetaCart
Dynamic topic models
by David M. Blei, John D. Lafferty - In ICML , 2006
"... Scientists need new tools to explore and browse large collections of scholarly literature. Thanks to organizations such as JSTOR, which scan and index the original bound archives of many journals, modern scientists can search digital libraries spanning hundreds of years. A scientist, suddenly ..."
Abstract - Cited by 656 (28 self) - Add to MetaCart
Sketchpad: A man-machine graphical communication system
by Ivan Edward Sutherland , 2003
Tools
Sorted by:
Relevance
Try your query at:
simplici
Tout surligner
Respecter la casse
Occurrence 2 sur 2

ACM DL


Results ACM DL : digital li...
dl.acm.org/results.cfm?query=digital+libraries+record+linkage&Go.x=0&(
Rechercher
ACM DIGITAL LIBRARY
SIGN IN SIGN UP
digital libraries record linkage
SEARCH
Searched for digital libraries record linkage [new search] [edit/save query]
[advanced search]
Searched The ACM Full-Text Collection: 468,529 records [Expand your search to The ACM Guide to Computing Literature: 2,676,172 records]
50,328 results found
Export Results: bibtex | endnote | acmref | csv
Refine by People
Names
Institutions
Authors
Editors
Advisors
Reviewers
Refine by Publications
Publication Names
ACM Publications
All Publications
Content Formats
Publishers
Refine by Conferences
Sponsors
Events
Proceeding Series
Refine by Publication Year
1950s 1960s 1970s 1980s 1990s 2000s 2010s
Published Since 1951
Upcoming Conferences
ICMR '17
June 06 - 09, 2017
Bucharest, Romania
Result 1 – 20 of 50,328
Result page: 1 2 3 4 5 6 7 8 9 10 >>
Sort by: relevance
1 Digital liaisons: engaging with digital curation theory and practice
SIG Digital Libraries
November 2013 ASIST '13: Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries
Publisher: American Society for Information Science
Bibliometrics: Citation Count: 0
Downloads (6 Weeks): 38, Downloads (12 Months): 324, Downloads (Overall): 455
Full text available: PDF
As librarians and information professionals are increasingly tasked with curating and making accessible digital materials, new information professionals must enter the workforce armed with the appropriate skills. New professionals must know both the theoretical underpinnings of digital curation and how to put those theories into practice in real life applications. ...
Keywords: student research, master's students, students, undergraduate students
[result highlights]
2 The Stanford Digital Library Project
CORPORATE The Stanford Digital Libraries Group
April 1995 Communications of the ACM: Volume 38 Issue 4, April 1995
Publisher: ACM
Bibliometrics: Citation Count: 9
Downloads (6 Weeks): 30, Downloads (12 Months): 173, Downloads (Overall): 749
Full text available: PDF
The Stanford Integrated Digital Library Project will develop enabling technologies for an integrated "virtual" library to provide an array of new services and uniform access to networked information collections. The Integrated Digital Library will create a shared environment linking everything from personal information collections, to collections of conventional libraries, to ...
[result highlights]
3 An Instrument for Merging of Bibliographic Databases
Feedback

dblp


dblp: Search for "record li...
dblp.uni-trier.de/search?q=record link
Rechercher
maintained by SCHLOSS DAGSTUHL at Universität Trier
home | browse | search | about
dblp
computer science bibliography
record link
Search dblp
powered by CompleteSearch, courtesy of Hannah Bast, University of Freiburg
> Home
Trier 1
[–] Publication search results
found 338 matches
2017
Shumin Han, Derong Shen, Tiezheng Nie, Yue Kou, Ge Yu:
Private Blocking Technique for Multi-party Privacy-Preserving Record Linkage. Data Science and Engineering 2(2): 187-196 (2017)
David D. Dobrzykowski, Monideepa Tarafdar:
Linking Electronic Medical Records Use to Physicians' Performance: A Contextual Analysis. Decision Sciences 48(1): 7-38 (2017)
April F. Mohanty, Jacob Crook, Christina A. Porucznik, Erin M. Johnson, Robert T. Rolfs, Brian C. Sauer:
Development and evaluation of a record linkage protocol for Utah's Controlled Substance Database. Health Informatics Journal 23(1): 35-43 (2017)
Christian Geier, Klaus Lehnertz:
Which Brain Regions are Important for Seizure Dynamics in Epileptic Networks? Influence of Link Identification and EEG Recording Montage on Node Centralities. Int. J. Neural Syst. 27(1): 1-14 (2017)
Dimitris A. Pinotsis, J. P. Geerts, L. Pinto, Thomas H. B. FitzGerald, Vladimir Litvak, Ryszard Auksztulewicz, Karl J. Friston:
Linking canonical microcircuits and neuronal activity: Dynamic causal modelling of laminar
[–] Refine list
refine by author
Peter Christen (32)
Vassilios S. Verykios (21)
Dinusha Vatsalan (15)
Vicenç Torra (11)
Dimitrios Karapiperis (8)
Murat Kantarcioglu (8)
Divesh Srivastava (7)
Bradley Malin (7)
Alexandros Karakasidis 0001 (7)
Josep Domingo-Ferrer (6)
736 more options
refine by venue
AMIA (17)
JAMIA (16)
History and Computing (11)
BMC Med. Inf. & Decision Making (10)
CoRR (10)
ICDE (7)
PAKDD (7)
Privacy in Statistical Databases (7)

État de l'art

État de l'art: travaux académiques

Retour sur les besoins du néophyte:

- il ne connaît pas les articles de référence sur le domaine
- « domaine »: recherche d'information thématique
- « de référence »: utilisation des liens
  - de citations [Salton, 1963]
  - de co-autorat, co-citations [Beel, 2016]
  - composantes relationnelles dans les calculs de score
  - recommandation: basée sur les usages
  - travail proche [Raamkumar, 2017]

État de l'art: évaluation en recherche d'information

- Évaluation en recherche d'information
  - disponibilité de données
  - disponibilité des logiciels des méthodes à comparer
    ◇ en particulier pour les modèles de référence
  - mesures d'évaluation
  - campagnes d'évaluation
    ◇ besoins d'information (requêtes) et jugements de pertinence
    ◇ pas disponible ni pour NeoTEX, ni plus généralement pour les BN

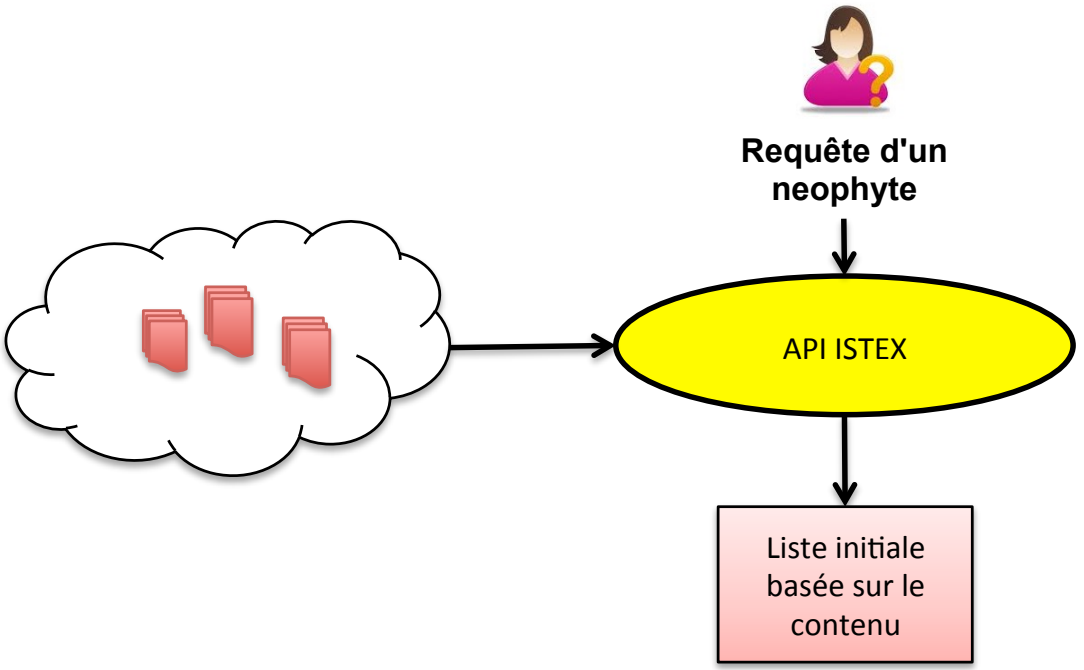
Conclusion de l'état de l'art

Objectifs	BN	travaux académiques
Besoins des néophytes	non	non
Utilisation de données relationnelles	oui	oui
Évaluation	?	pas de collection de test

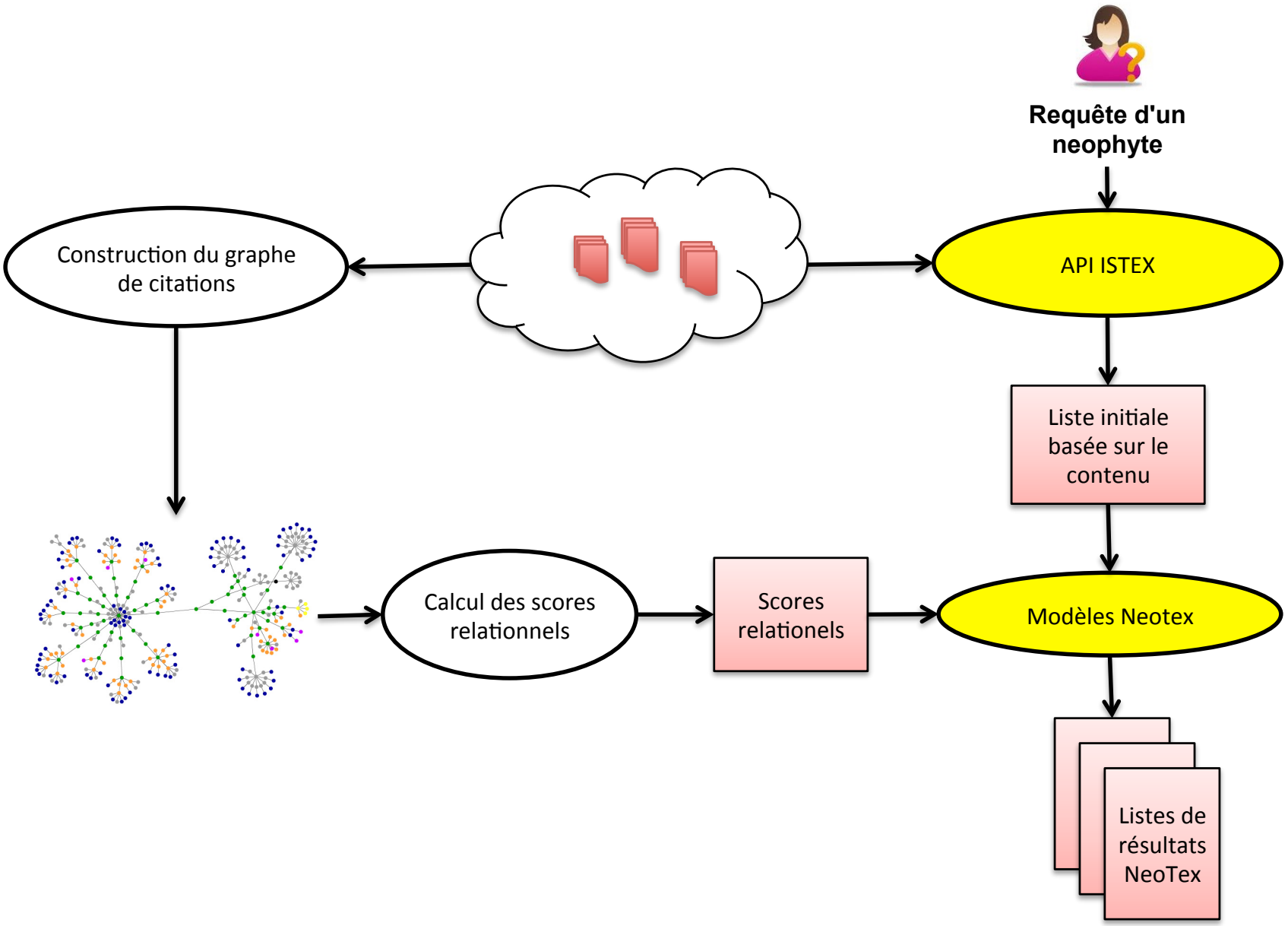
- **Contributions**:
  - création de vérités terrains
  - construction de modèles de référence

Système NeoTEX

Modèle de RI


Requête d'un neophyte
API ISTEX
Liste initiale basée sur le contenu

Système NeoTEX



Le graphe (orienté) de citation

- nœuds
  - les documents ISTEX
  - les documents cités par les documents ISTEX
- arcs: lien (citations, références)
- **Contribution**: unification des nœuds
  - documents citants (ISTEX): identifiant ISTEX, méta-informations
  - documents cités: méta-informations bruitées

Les scores relationnels

- degré entrant: S_I
- degré sortant: S_O
- PageRank: S_{PR}

- en perspective: d'autres mesures de centralité

Les modèles

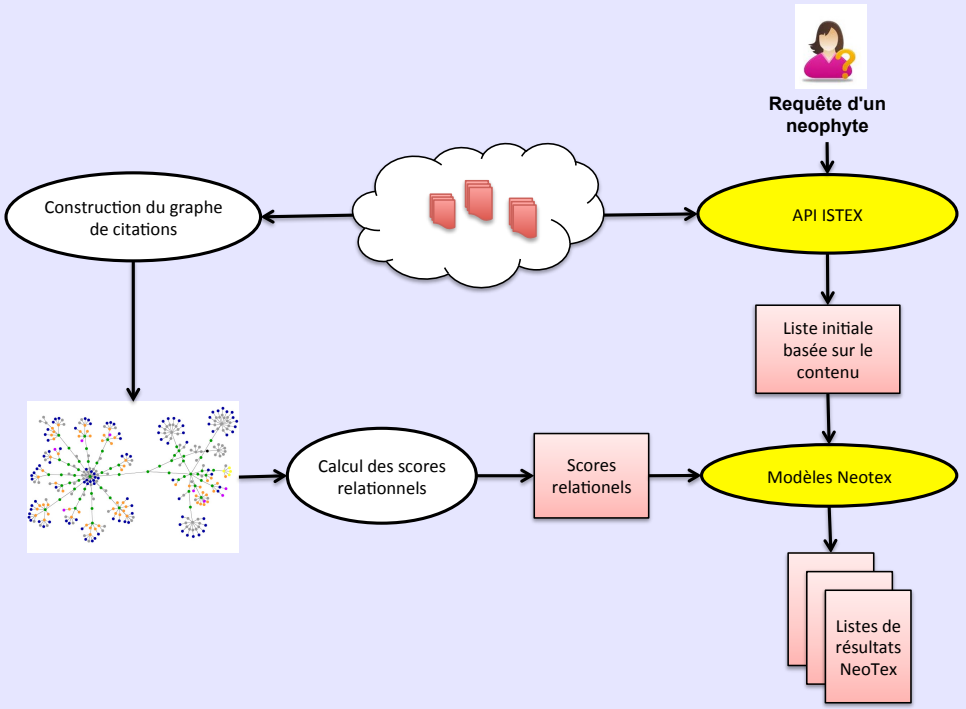
- liste L des 1000 premiers documents classés selon leur score de contenu (S_c)
- les listes reclassées selon tel ou tel score (S_I, S_O ou S_{PR})
- les listes agrégées
- **Contribution**: les listes basées sur l'apprentissage automatique

Contributions

- les besoins d'information et les requêtes
  - issues de 25 thèses en Informatique soutenues en 2006
  - requêtes construites manuellement à partir des titres, résumés, et mots-clés des thèses
- les vérités terrains (les jugements de pertinence)
  - **thèses** la section « références » des 25 thèses
  - **manuel** jugement d'un expert
  - **réputation** nombre de citations en 2016
  - **thèses** $\cup$ **réputation**

Plan

- Construction du graphe
- Les modèles, en particulier ceux qui utilisent de l'apprentissage
- Les expériences et leurs résultats



Construction du graphe: les données

Extrait des données téléchargées au format Json:

```
{
  "corpusName": "elsevier",
  "author": ...
  "title": "Nuclear antigens in the HeLa cell cycle"...
  "refBibs": [
        "title": "Multiplicaiton and division in Mammalian  Cells",
            ...
             ]
}
...
{
   ...
  "title": "Multiplication and Division in Mammalian Cells",
   ...
}
...
```

7 316 816	titres citants
117 946 803	titres cités
125 263 619	titres

Volume de données: 9,6 Gio (1950–2005, titre de plus de 6

Le but

```
0          000 000 100 0 source world christian trends demographic futures for chris
1          000 000 clinical perspectives 593 adults with cystic fibrosis meeting the
2          000 000 miles of fluid evaluation in city bus automatic transmission
...        [...]
3 486      0956 a study of the conditions and mechanisms of the diphenylamine reacti
...        [...]
108 310    1956 a study of the conditions and mechanism of the diphenylamine reactio
...        [...]
108 313    1956 a study of the mechanism of the diphenylamine reaction for the color
...        [...]
46 852 687 zzzv and aaaa2 v a decade later he spoke of ashmole as my honoured friend
```

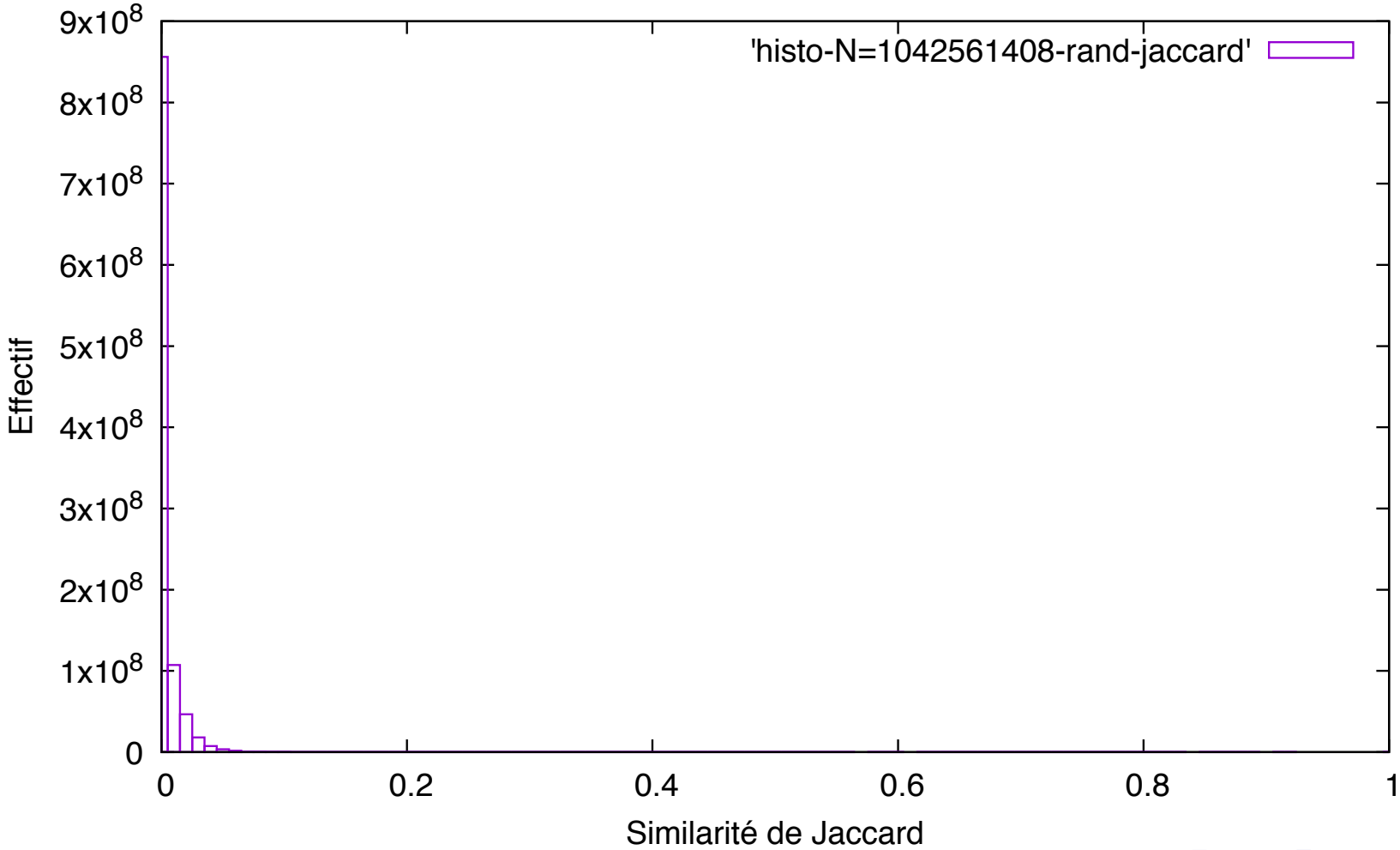
Le moyen. . . qui n'en est pas un

Le but: reconnaître les titres qui se ressemblent

- une première étape, de complexité linéaire ($O(n)$)
  - normaliser
  - c'est-à-dire remplacer les caractères non alphanumériques par des espaces, conversions en minuscules, compactage des espaces
- une deuxième étape, de complexité pseudo-linéaire ($O(n \log n)$)
  - pour trouver les duplicats exacts
  - résultat: 46 852 688 titres normalisés uniques
- une troisième étape, de complexité quadratique ($O(n^2)$)
  - comparer chaque paire selon une similarité (Jaccard, Levenshtein, etc.)
  - temps à raison de 20–30 μs la comparaison: un millénaire
- et choisir un seuil de similarité

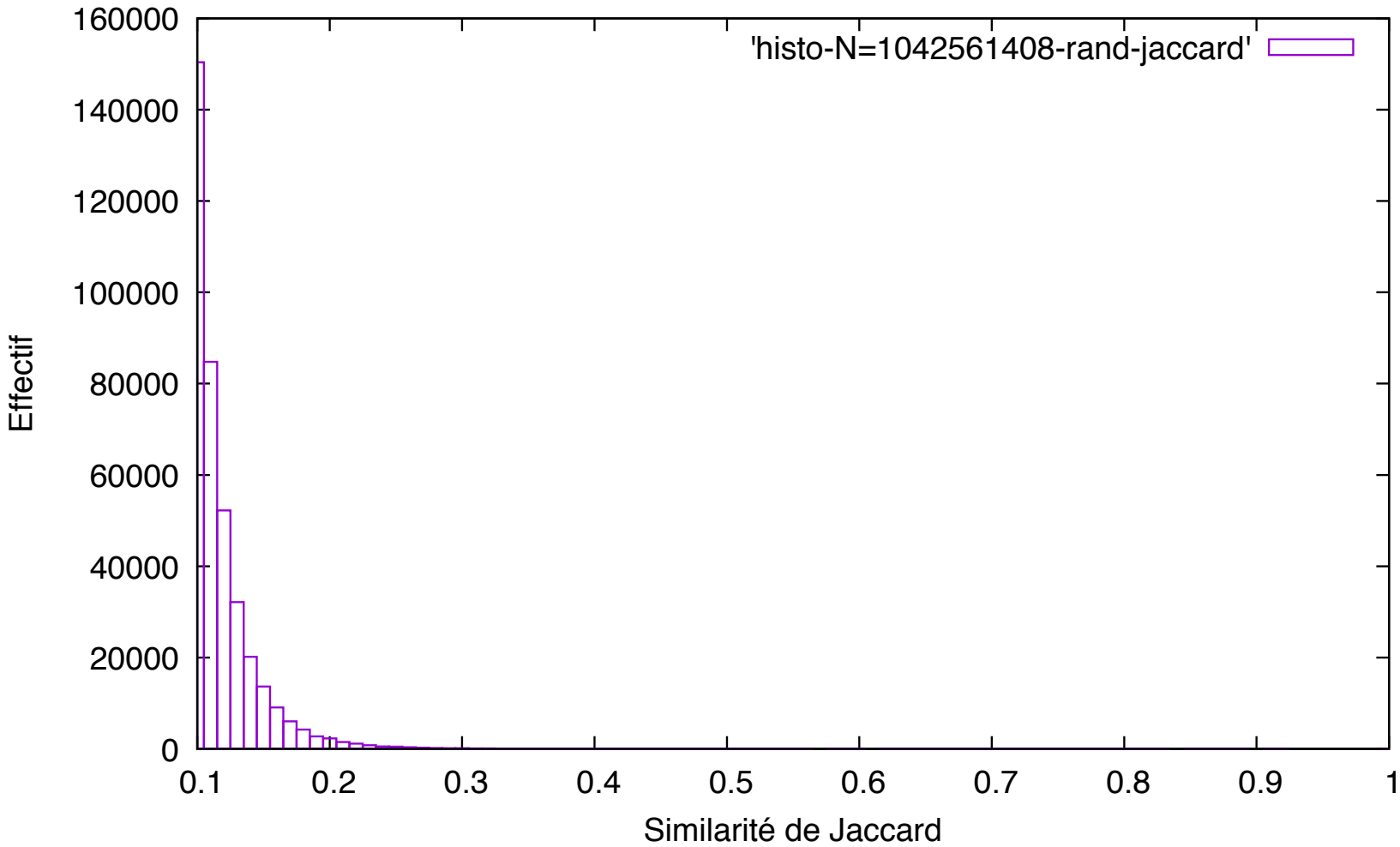
Histogramme similarité de Jaccard: échantillon aléatoire

Histogramme de similarité de Jaccard sur un échantillon de
1 milliard de paires tirées au hasard
850 millions entre 0 et 1/100, et 100 millions entre 1/100 et 2/100



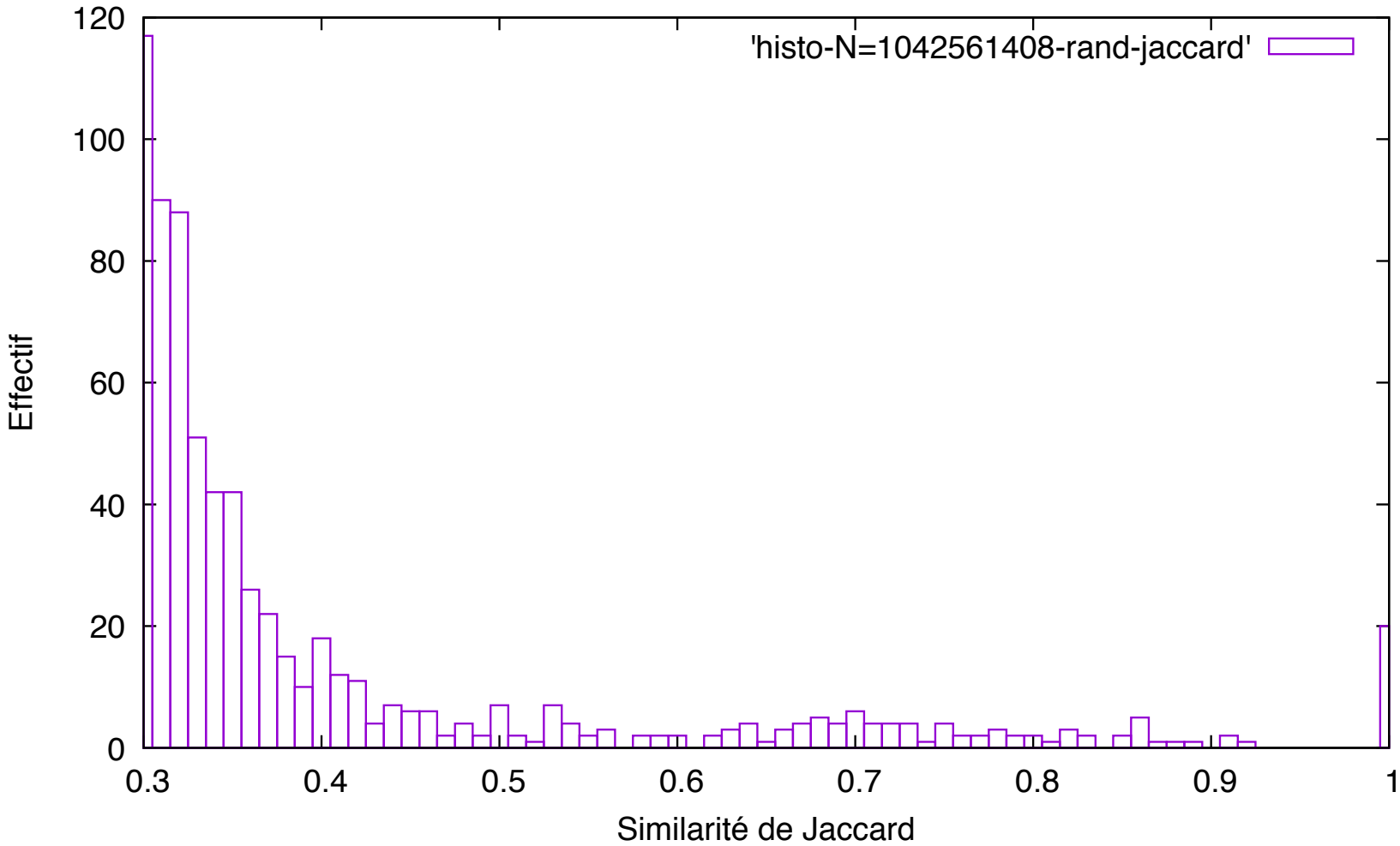
Histogramme similarité de Jaccard ≥ 0.1: échantillon aléatoire

Zoom de l'histogramme de similarité de Jaccard avec $s \geq 1/10$



Histogramme similarité de Jaccard ≥ 0.3: échantillon aléatoire

Zoom de l'histogramme de similarité de Jaccard avec $s \geq 3/10$



Le moyen

Le *hashage* sensible à la localité: *Locality sensitive hashing*
[1998-ACM Th. of computing-Indyk Motwani]

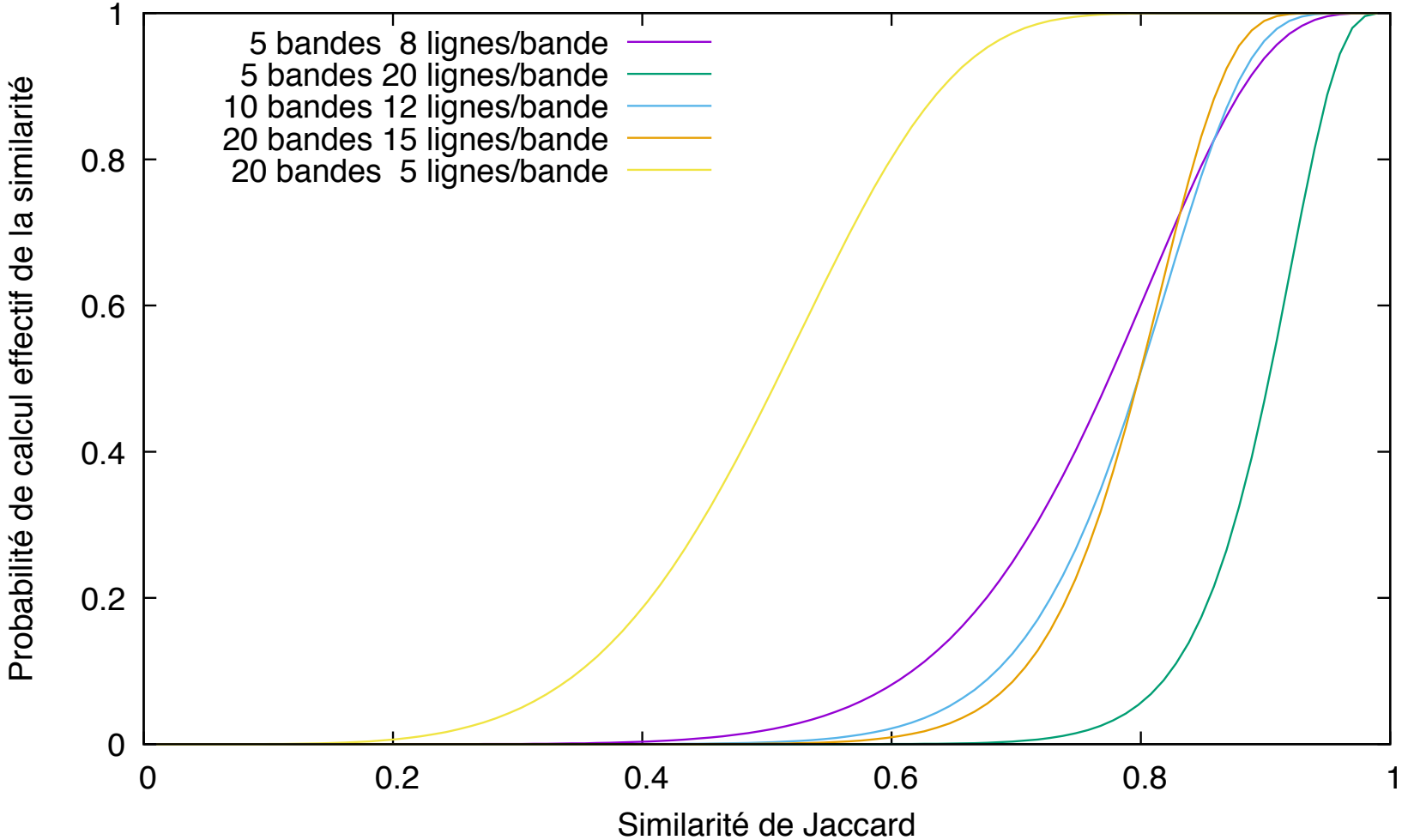
[1999-VLDB-Gionis Indyk Motwani]

- prendre r « extraits » des objets à comparer
- si ces r « extraits » sont les mêmes (*hashage* classique) pour deux objets, c'est un indice qu'il sont peut-être proches et il faudra les comparer
- le faire b fois (*bande*)
- les « extraits » doivent être choisis en cohérence avec une mesure de similarité
- pour la similarité de Jaccard: le *hashage par minimum* (*Min Hashing*)

  [1998-ACM Th. of computing-Broder Charikar Frieze Mitzenmacher]
- probabilité que deux objets de similarité x partagent leurs r « extraits » dans au moins une des b bandes:

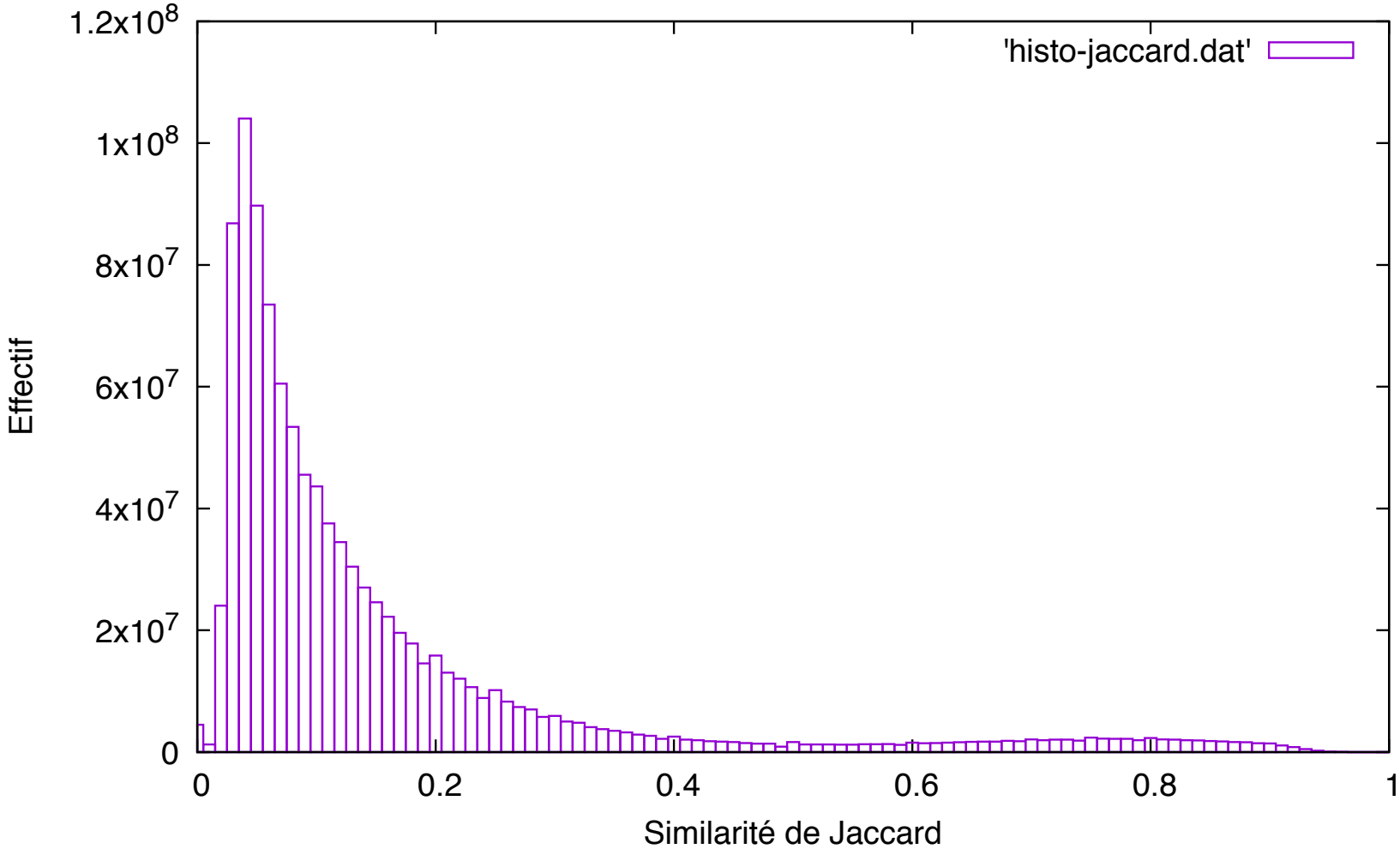
$$1 - (1 - x^r)^b$$

Choix du nombre de bandes b et de lignes par bande r



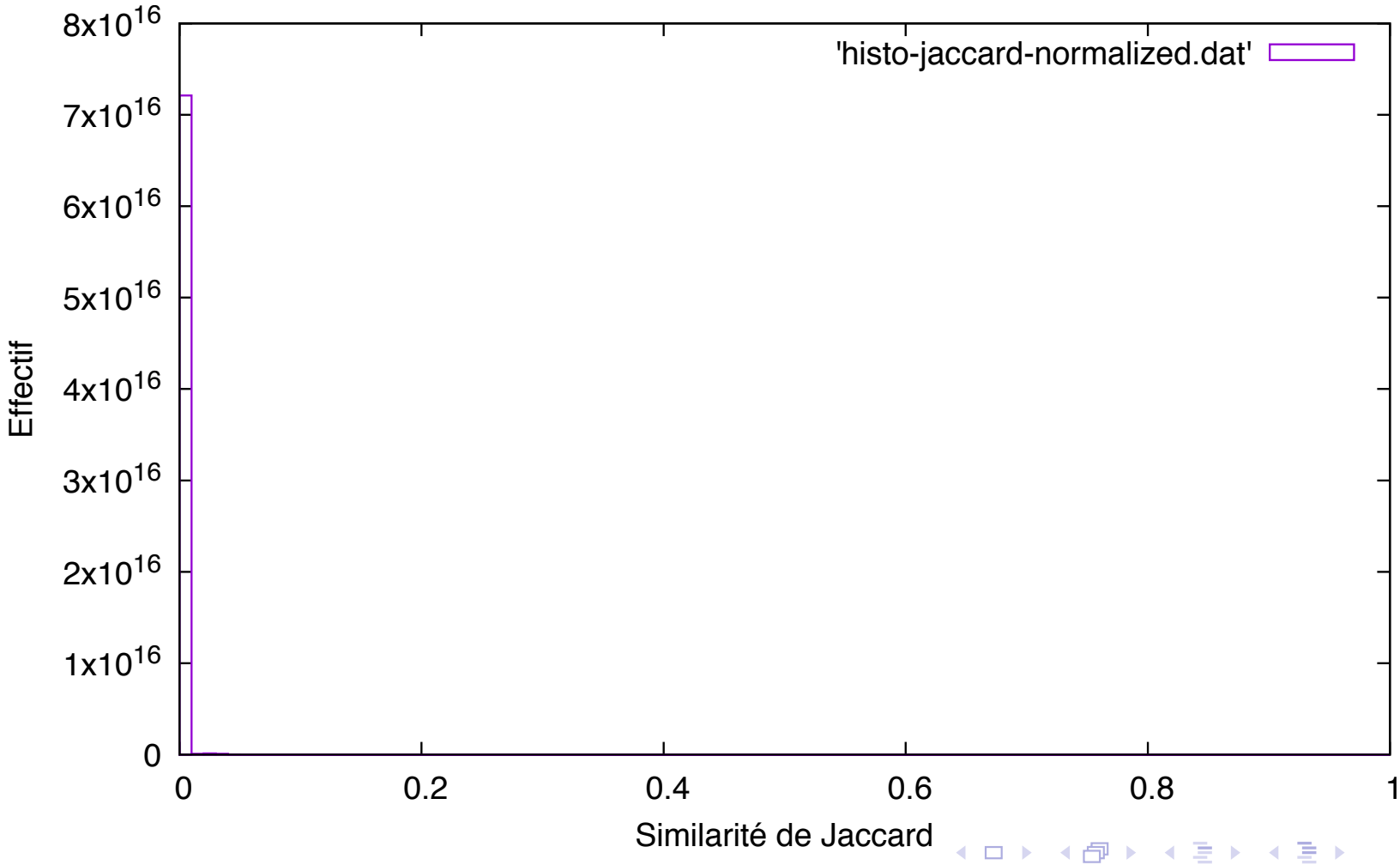
Histogramme similarité de Jaccard: échantillon LSH

Histogramme de similarité de Jaccard sur les paires sélectionnées par le *hashage* sensible à la localité.



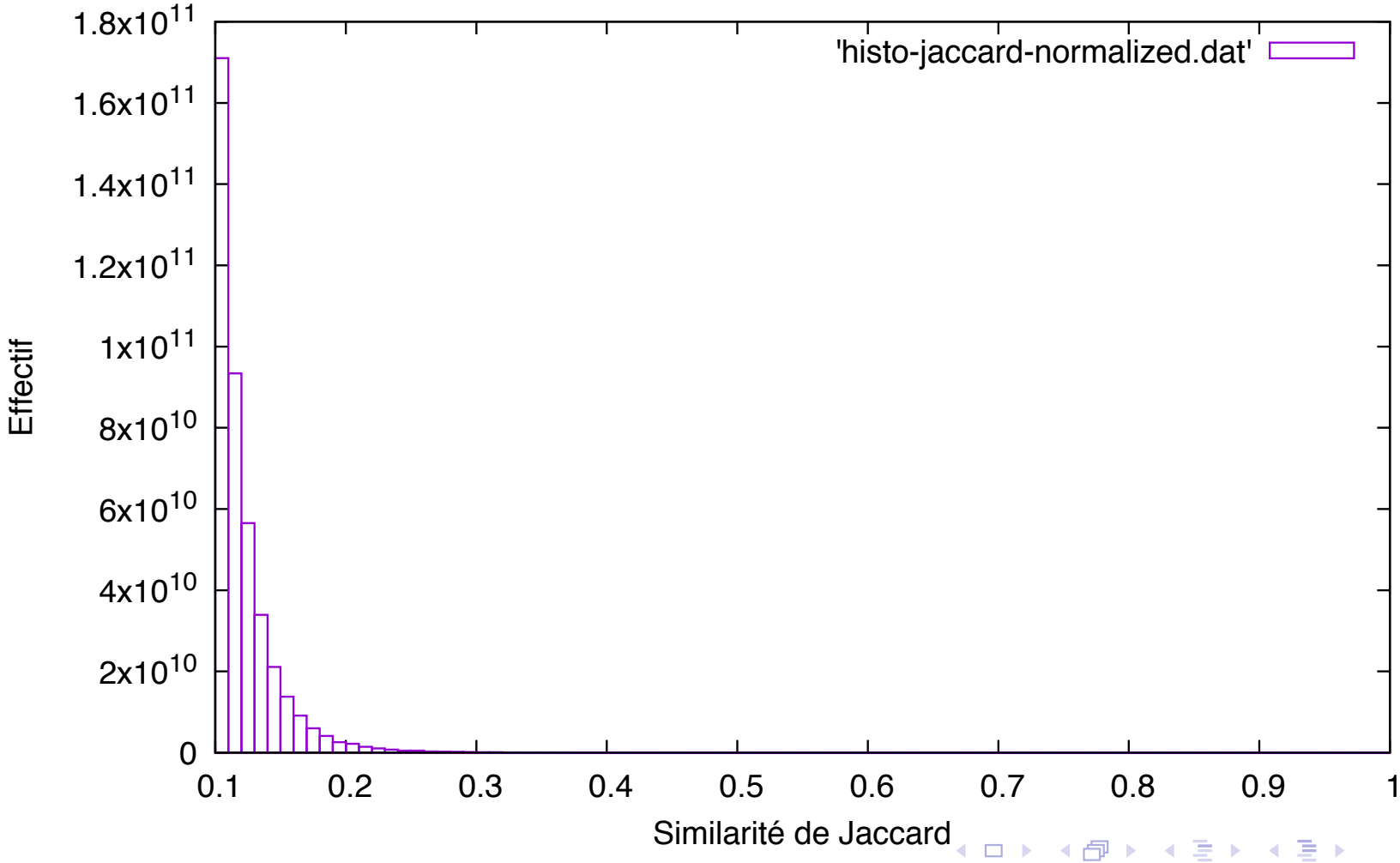
Histogramme similarité de Jaccard: interpolation de l'échantillon LSH

Histogramme de similarité de Jaccard sur l'interpolation à partir des paires sélectionnées par le *hashage* sensible à la localité.



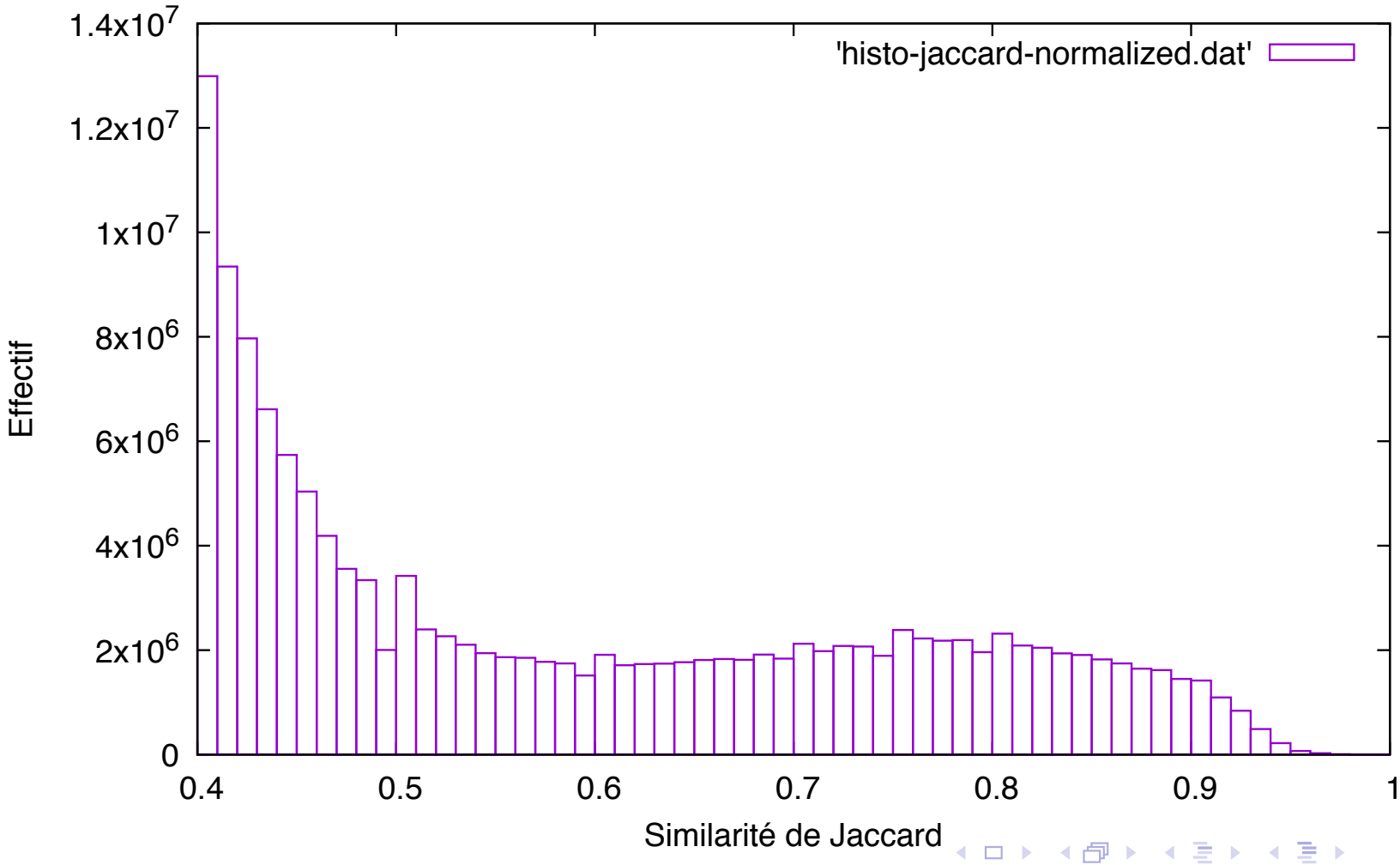
Histogramme similarité de Jaccard ≥ 0.1: interpolation de l'échantillon LSH

Histogramme de similarité de Jaccard sur l'interpolation à partir des paires sélectionnées par le *hashage* sensible à la localité.



Histogramme similarité de Jaccard ≥ 0.4: interpolation de l'échantillon LSH

Histogramme de similarité de Jaccard sur l'interpolation à partir des paires sélectionnées par le *hashage* sensible à la localité.



Application du seuil et recherche des composantes connexes

- seuillage: ne sont conservées parmi les paires candidates que celles de similarité supérieure au seuil choisi (0,85 dans notre expérience)
- recherche des composantes connexes pour obtenir une partition ce qui peut amener à mettre dans un même sous-ensemble deux éléments qui ont une similarité de Jaccard inférieure au seuil choisi.

Effectif	taille de la composante connexe
3 522 922	2
1 236 552	3
576 094	4
309 439	5
181 799	6
114 913	7
76 098	8
52 425	9
37 719	10
...	...

Effectif	taille de la composante connexe
...	...
1	647
1	709
1	728
1	818
1	880
1	933
1	981
1	1394
1	2077

Un extrait d'une composante connexe

```
    Dd8c: 0951 protein measurement with the folin phenol reagent
D12f7914: i951 protein measurement with the folin phenol reagent
D153df04: i protein measurement with the folin phenol reagent
D1e9c0d4: protein measurement with the folin protein phenol reagent
D1e9d686: protein protein measurement with the folin phenol reagent
 D405429: a protein measurement with the folin phenol reagent
D12fac0d: i a all protein measurement with the folin phenol reagent
D154347c: i rotein measurement with the folin phenol reagent
D159a4b0: j protein measurement with the folin phenol reagent
D171334d: l protein measurement with the folin phenol reagent
D1e99d7a: protein in measurement with the folin phenol reagent
D1e9bdec: protein measuremennt with the folin phenol reagent
D1e9be32: protein measurement reagent with the folin phenol reagent
D1e9bed8: protein measurement with ent with the folin phenol reagent
D1e9bfdb: protein measurement with the folin in phenol reagent
D1e9bfdc: protein measurement with the folini phenol reagent
D1e9bfe0: protein measurement with the folinn phenol reagent
D1e9bfee: protein measurement with the folin phenenol reagent
D1e9c002: protein measurement with the folin phenol phenol reagent
D1e9c02a: protein measurement with the folin phenol reagent
D1e9c085: protein measurement with the folin phenol reagent protein measurement
          with the folin phenol reagent
D1e9c0f5: protein measurement with the foli phenol reagent
D1e9c13d: protein measurement with the protein folin phenol reagent
```

Pour 47 millions de titres sans duplicats exacts

construction des 6-grammes

3,8 Gio 56 Gio 15 min

***hashage* des 6-grammes**

56 Gio 31 Gio 14 min

tri

31 Gio 31 Gio 16 min mémoire centrale:. . .

***hashage* par minimum et *hashage* sensible à la localité**

31 Gio 7 Gio 30 min mémoire centrale: 18 Gio

Recherche des candidats: même *hash* dans la même bande (tri)

7 Gio 7 Gio 5 min

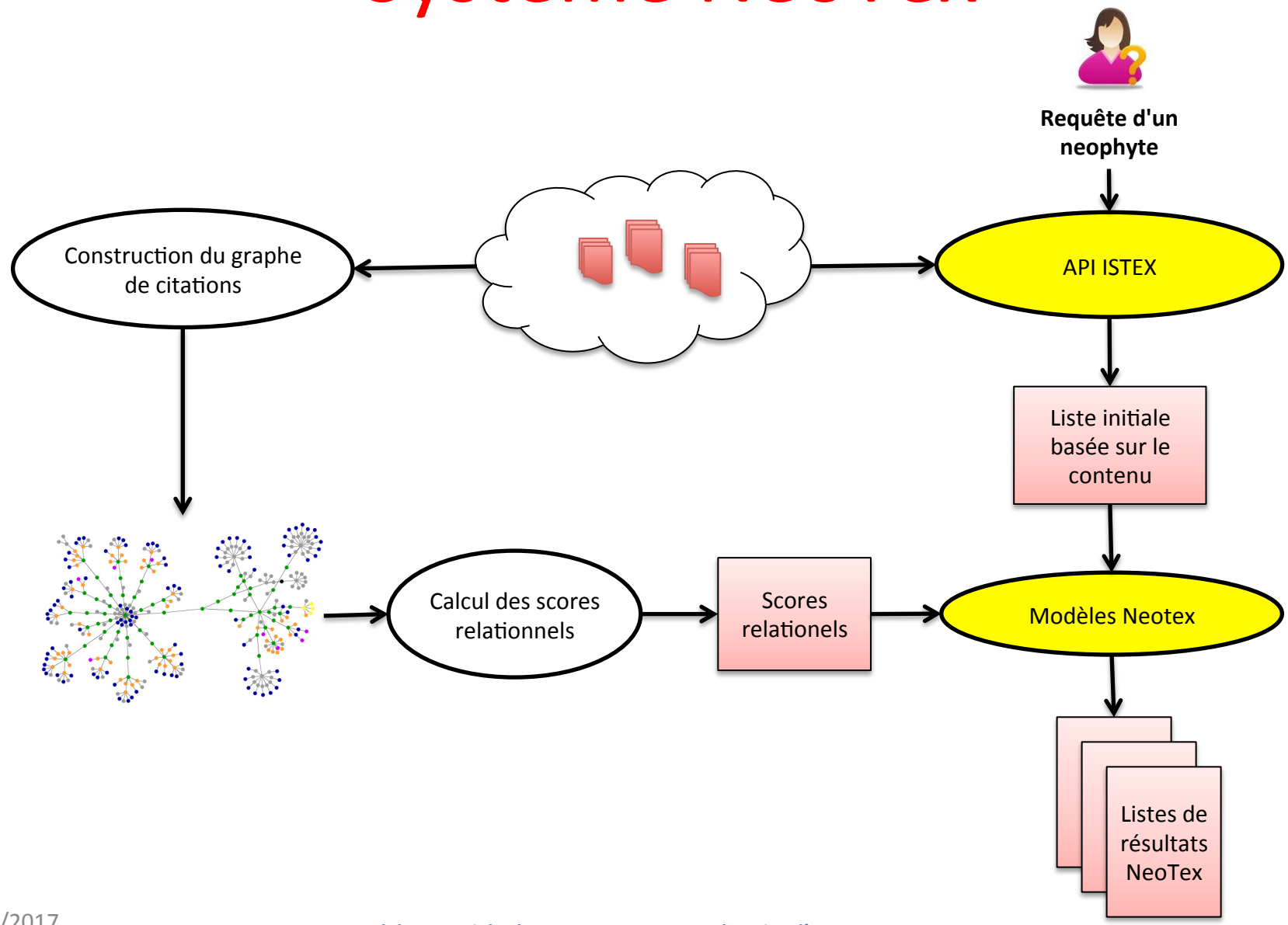
Calcul des similarités de Jaccard des paires candidates, seuillage

7 Gio 4 Gio 7 h 26 min

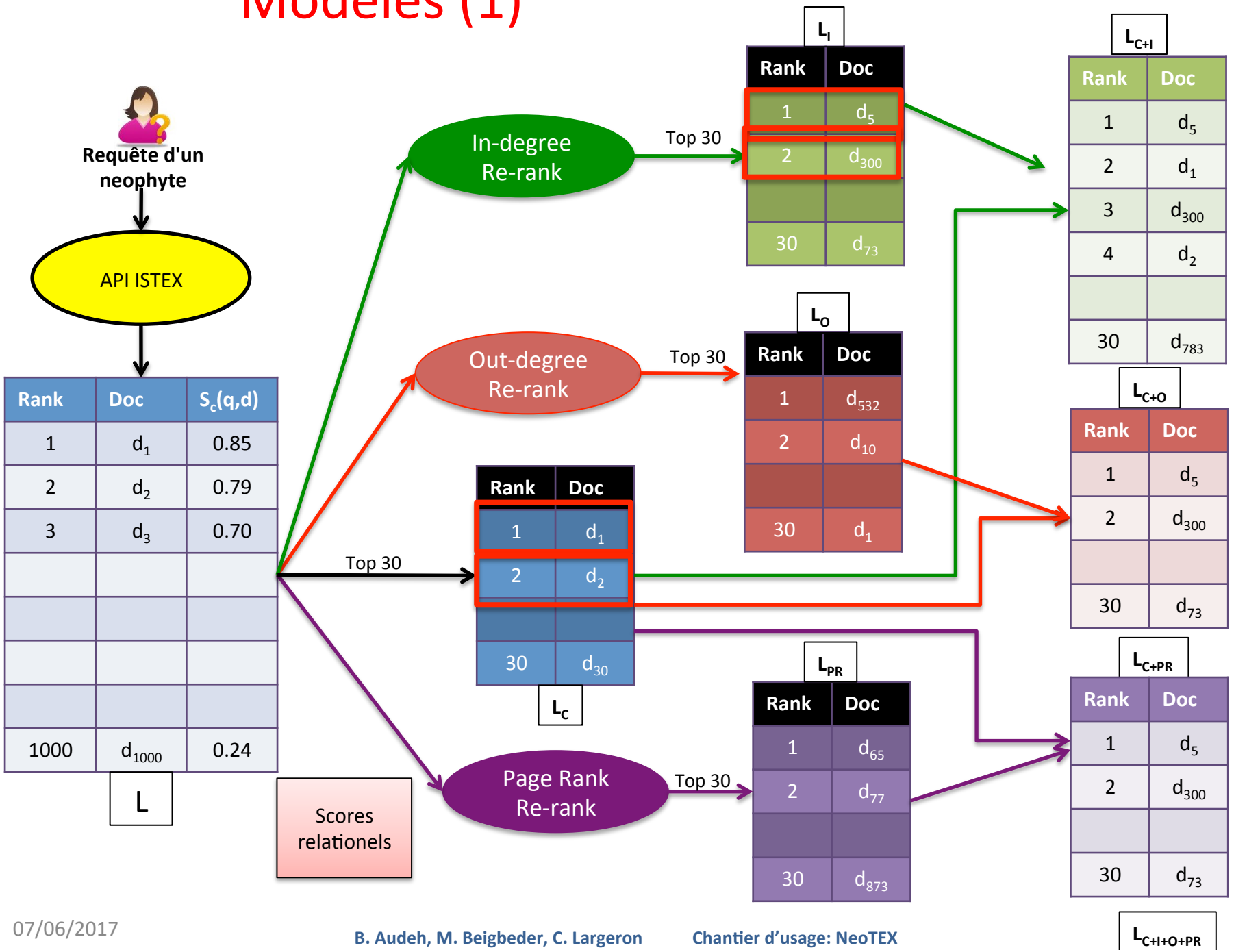
Calcul des composantes connexes

4 Gio 0,2 Gio 6 min

Système NeoTex



Modèles (1)


Requête d'un neophyte
API ISTEX
Rank Doc Sc(q,d)
1 d1 0.85
2 d2 0.79
3 d3 0.70
1000 d1000 0.24
L
Top 30
In-degree Re-rank
Out-degree Re-rank
Page Rank Re-rank
Scores relationels
Rank Doc
1 d1
2 d2
30 d30
LC
Top 30
LI
Rank Doc
1 d5
2 d300
30 d73
Top 30
LO
Rank Doc
1 d532
2 d10
30 d1
Top 30
LPR
Rank Doc
1 d65
2 d77
30 d873
LC+I
Rank Doc
1 d5
2 d1
3 d300
4 d2
30 d783
LC+O
Rank Doc
1 d5
2 d300
30 d73
LC+PR
Rank Doc
1 d5
2 d300
30 d73
LC+I+O+PR

Apprentissage supervisé

Objectif

Prédire si un document *d* est pertinent pour une requête *q*

$$\hat{E}(q,d) \in \{0,1\}$$

Formalisation du modèle

- Étape 1: Apprentissage

A partir d'un ensemble de couples (*q,d*), où chaque couple est décrit par les scores

- $S_c(q,d)$: score de pertinence (ISTEX API)
- $S_I(d)$: in-degree du document d
- $S_O(d)$: out-degree du document d
- $S_{PR}(d)$: pagerank du document d

et par le vérité terrain

- $E(q,d) \in \{0,1\}$

construire un model pour prédire

→ $\hat{E}(q,d) \in \{0,1\}$

Formalisation du modèle

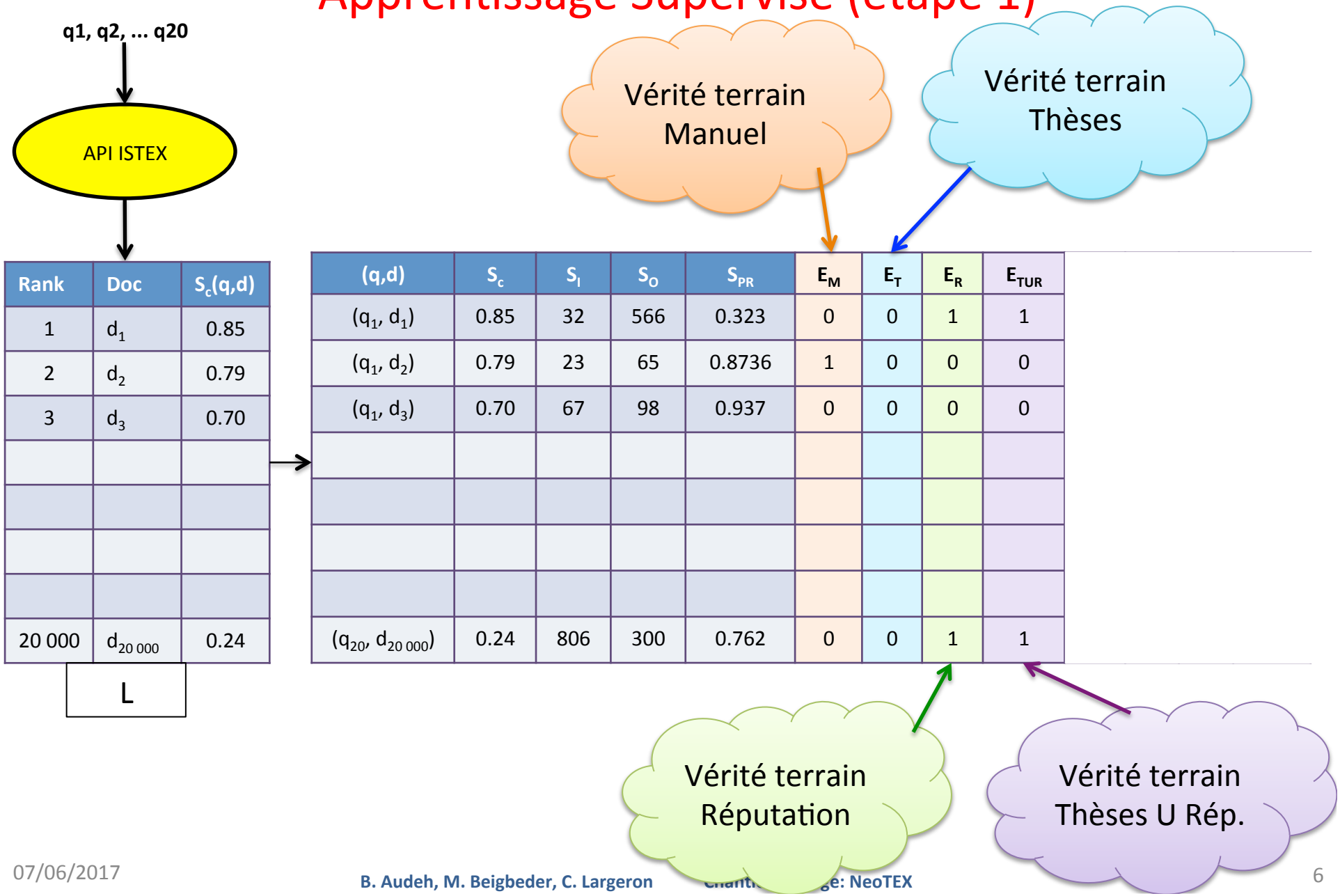
- Étape 2: Utilisation du modèle

Pour une nouvelle requête, prédire la pertinence d'un document pour lequel on dispose des scores

- Trois méthodes d'apprentissage :
  - Arbres de décision (AD)
  - Random Forest (RF)
  - SVM

Modèles (2)

Apprentissage Supervisé (étape 1)



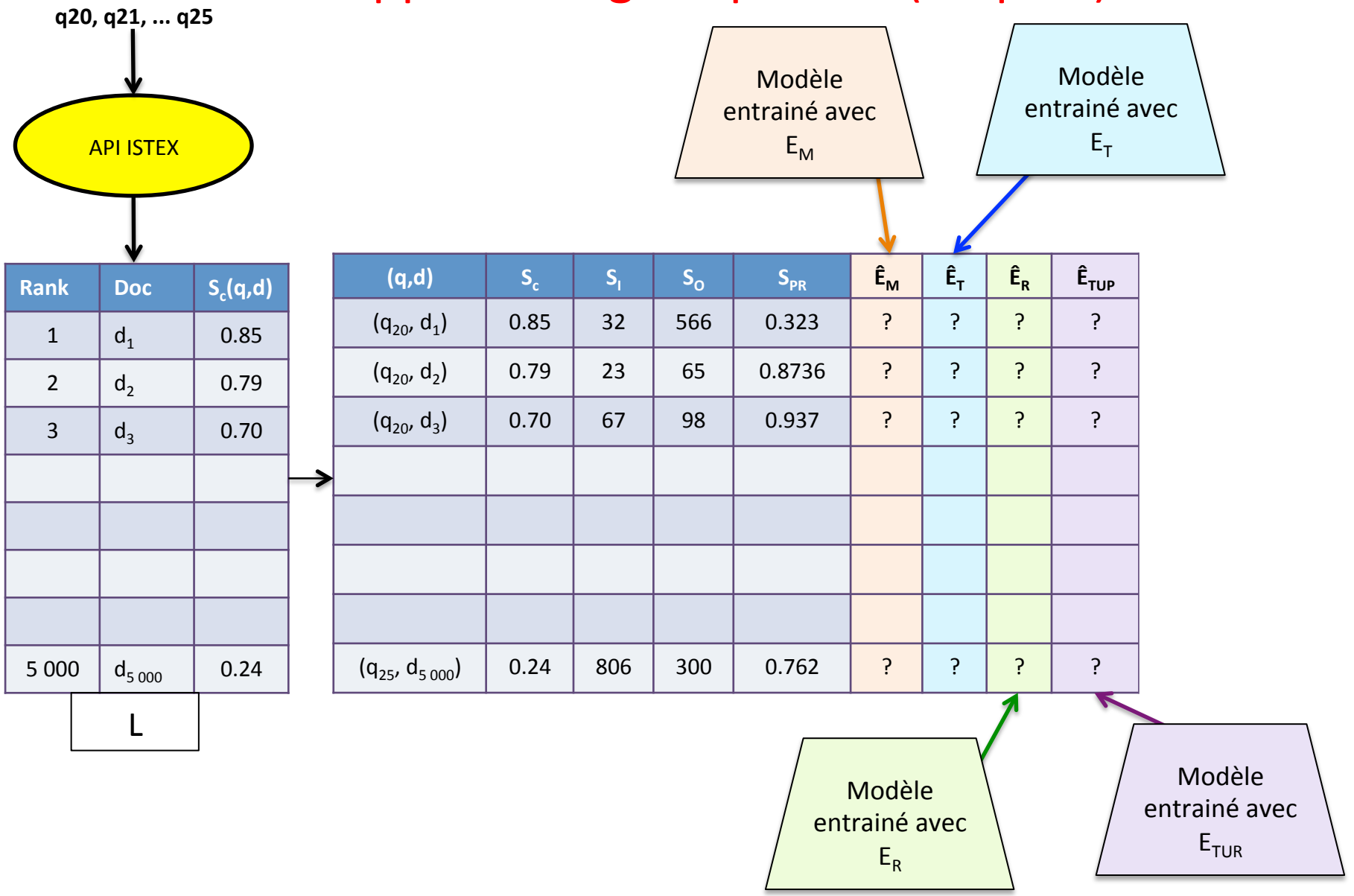
Rank	Doc	$S_c(q,d)$
1	d_1	0.85
2	d_2	0.79
3	d_3	0.70
20 000	$d_{20\,000}$	0.24

L

(q,d)	S_c	S_I	S_O	S_{PR}	E_M	E_T	E_R	E_{TUR}
(q_1, d_1)	0.85	32	566	0.323	0	0	1	1
(q_1, d_2)	0.79	23	65	0.8736	1	0	0	0
(q_1, d_3)	0.70	67	98	0.937	0	0	0	0
$(q_{20}, d_{20\,000})$	0.24	806	300	0.762	0	0	1	1

Models (2)

Apprentissage Supervisé (étape 2)



Rank	Doc	$S_c(q,d)$
1	d_1	0.85
2	d_2	0.79
3	d_3	0.70
5 000	$d_{5\,000}$	0.24

(q,d)	S_c	S_I	S_O	S_{PR}	$\hat{E}_M$	$\hat{E}_T$	$\hat{E}_R$	$\hat{E}_{TUP}$
(q_{20}, d_1)	0.85	32	566	0.323	?	?	?	?
(q_{20}, d_2)	0.79	23	65	0.8736	?	?	?	?
(q_{20}, d_3)	0.70	67	98	0.937	?	?	?	?
$(q_{25}, d_{5\,000})$	0.24	806	300	0.762	?	?	?	?

Modèles (2)
Apprentissage Supervisé (étape 2)

Option 1: Filterage Booléen

q

ISTEX API

Rank	Doc	$S_c(q,d)$
1	d_1	0.85
2	d_2	0.79
3	d_3	0.70
1000	d_{1000}	0.24

L

(q,d)	S_c	S_I	S_O	S_{PR}
(q1, d_1)	0.85	32	566	0.323
~~(q1, d_2)~~	~~0.79~~	~~23~~	~~65~~	~~0.8736~~
(q1, d_3)	0.70	67	98	0.937
~~(q1, d_{1000})~~	~~0.24~~	~~806~~	~~300~~	~~0.762~~

Top 30 où $\hat{E}_M$=1

Résultat final

Rank	Doc
1	d_1
2	d_3
3	d_{900}
30	d_{50}

$ML\_B\_E_M$

Modèle entrainé avec E_M

Booléan

Modèles (2)
Apprentissage Supervisé (étape 2)

Option 2: Reclassement par probabilité

q

ISTEX API

Rank	Doc	$S_c(q,d)$
1	d_1	0.85
2	d_2	0.79
3	d_3	0.70
1000	d_{1000}	0.24

L

(q,d)	S_c	S_I	S_O	S_{PR}
($q1$, d_1)	0.85	32	566	0.323
($q1$, d_2)	0.79	23	65	0.8736
($q1$, d_3)	0.70	67	98	0.937
($q1$, d_{1000})	0.24	806	300	0.762

Arbre de décision entrainé avec E_M

Top 30

triés par $P^{\wedge}_M$ puis par S_C

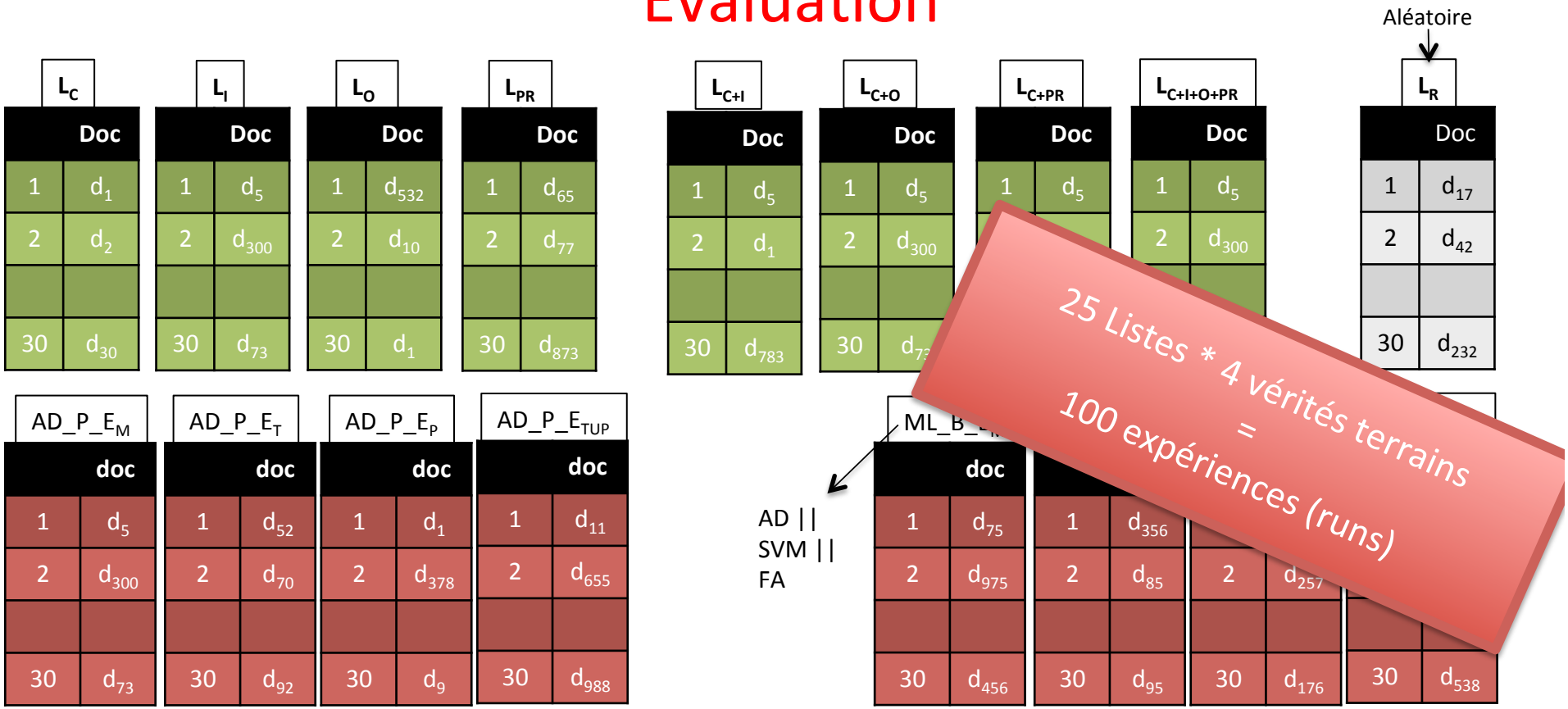
Résultat final

Rank	Doc
1	d_3
2	d_1
3	d_{1000}
30	d_2

$AD\_P\_E_M$

Probabilité

Evaluation



Trec_eval

Doc	Rel
d_1	0
d_2	1

Vérité terrain Manuel

Doc	Rel
d_1	0
d_2	0

Vérité terrain Thèses

Doc	Rel
d_1	1
d_2	1

Vérité terrain Réputation

Doc	Rel
d_1	1
d_2	1

Vérité terrain Thèse U Rép.

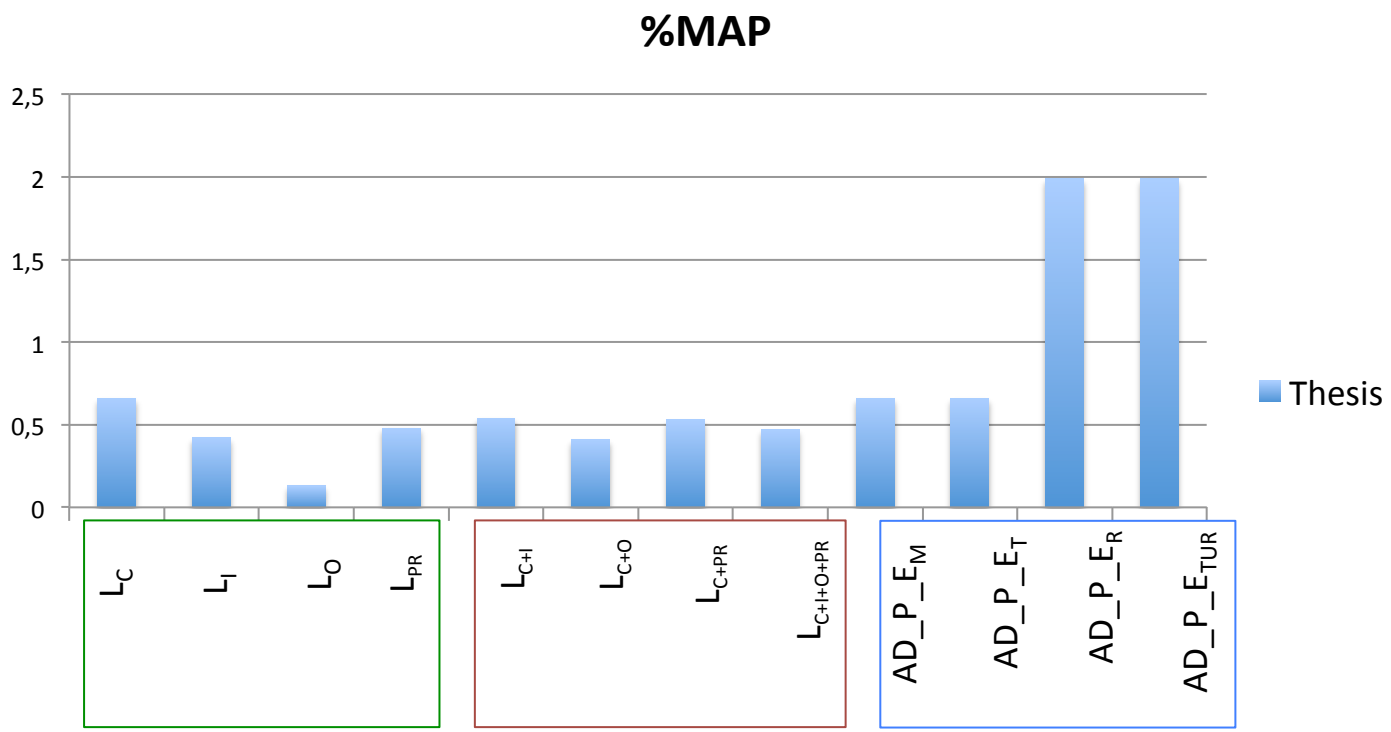
Evaluation

- Data
  - Métadonnées d'ISTEX de 1950 à 2005 (18 millions docs)
  - Requêtes: 25 sujets de thèse (Informatique 2006)
- Apprentissage suppervisé
  - Modèles testés
    - Arbre de décision
    - SVM
    - Forêts aléatoires
  - Échantillons
    - 20 requêtes d'apprentissage
    - 5 requêtes de test
    - Validation croisée (5-fold)

Résultats

F-measure	Thèses	MAN	Réputation
L_C	1,71	17,40	4,40
L_R	-	0,76	5,30
L_I	1,93	6,65	**37,73**
L_O	1,07	2,56	28,40
L_P	2,36	6,65	36,00
L_{C+I}	**2,79**	**23,52**	30,27
L_{C+O}	1,50	18,92	24,27
L_{C+PR}	2,36	23,52	28,67
$L_{C+I+O+PR}$	2,57	23,52	35,33
$AD\_P\_E_M$	1,71	17,40	4,40
$AD\_P\_E_T$	1,71	17,40	4,40
$AD\_P\_E_R$	2,36	22,76	23,60
$AD\_P\_E_{TUR}$	2,36	22,76	23,73
$RF\_B\_E_R$			28,27
$RF\_B\_E_{TUR}$			28,27

Résultats



Pistes d'amélioration par rapport au problème d'échantillons déséquilibrés :

- procédure de rééquilibrage des classes
- introduire des coûts d'erreur (pénaliser le faux négatif)

Conclusion

- NeoTex pour ISTEX : de nouvelles fonctionnalités pour l'utilisateur
  - Graphe de citations: permet d'associer à chaque document: nombre de références, de citations et pagerank **grace à la collection intégrant les références et à Grobid**
  - Fournir à un utilisateur des listes d'articles de référence selon plusieurs points de vue

Conclusion

- NeoTex pour ISTEX : l'articulation de la collection ISTEX avec DBLP pour produire de nouveaux attributs prédictifs
  - Impact factor:
    - calcul basé sur le graphe, comparaison avec WOS -> forte corrélation
  - Popularité des auteurs (DBLP):
    - Nb de publis, co-auteurs, ancienneté
  - Multi-disciplinarité:
    - Catégories WOS d'ISTEX
    - Autres ressources (to do ISTEX2 cf Gully)
  - *Spécialisation/vulgarisation (to do: collaboration ISTEX2 cf Cuxac)*
    - *Ressource sémantique*
    - *Topic Model, LDA*

Perspectives

- Evaluation du système NeoTex
  - Choix du domaine
    - Actuellement: Informatique
    - En cours: Science de la vie
- Graphe
  - Construction et évaluation du graphe
    - choix des paramètres pour le LSH
    - appliquer la méthode sur toute la collection ISTEX (< une semaine)
    - lier avec les identifiants ISTEX (ISTEX2: cf. CILLEX)
    - visualiser le graphe de voisinage (références/citations) d'un document donné pour permettre une recherche itérative par l'utilisateur (NeoTex2?)
    - construire le graphe d'auteurs (très difficile-ISTEX2)
  - Lier à d'autres collections: par exemple dblp, ACM pour l'informatique

Merci

Questions?