Génération et détection automatique de faux articles scientifiques

Ateliers de l'information

Cyril Labbé

Université Joseph Fourier - LIG

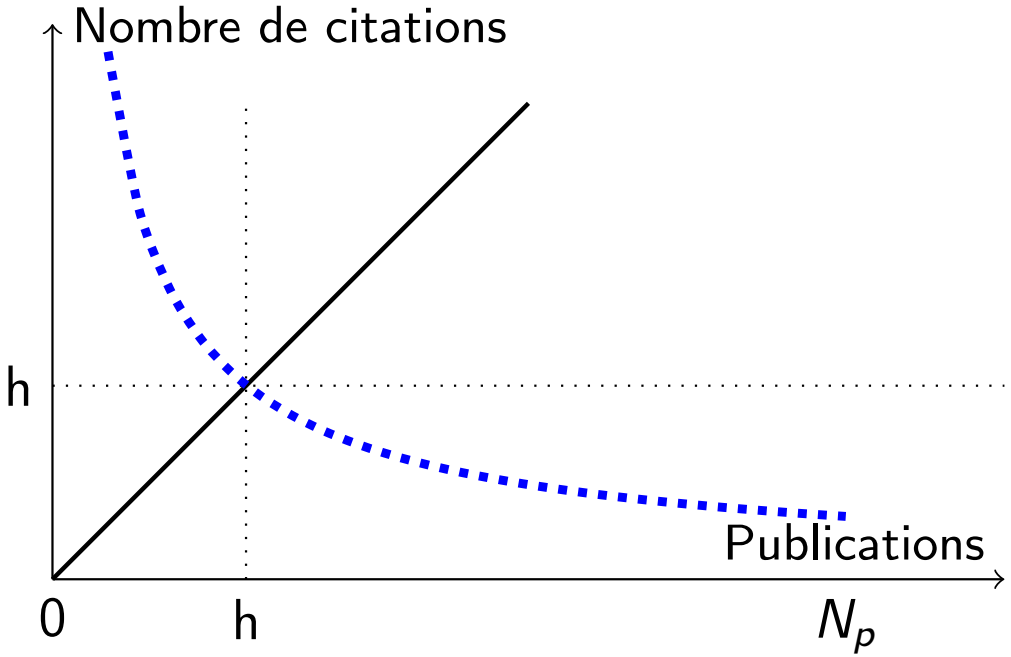
12 Janvier 2015

1. Préliminaires
   - Scientométrie
   - SCIgen une grammaire probabiliste hors contexte
2. Ike Antkare, one of the great starts in the scientific firmament
3. Detection de papiers SCIgen
   - Google Search
   - Classification automatique

Classement des scientifiques et des journaux

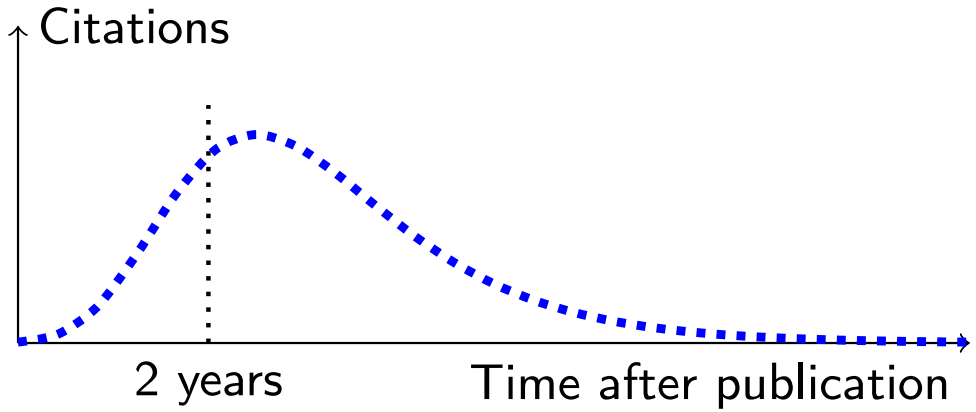
Definition (h-index [Hirsch, 2005])

A scientist has index h if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have $\leq h$ citations each.



Definition (Impact Factor)

Average number of citations to papers published by the journal over the last two years. Computed since 1975.



Compter les citations.

Outils payants.

- Fournis par les maisons d'édition (Elsevier, Thomson reuters);
- A partir des catalogues (ACM, IEEE, Springer, Elsevier);
- Sélection stricte (évaluation par les pairs).

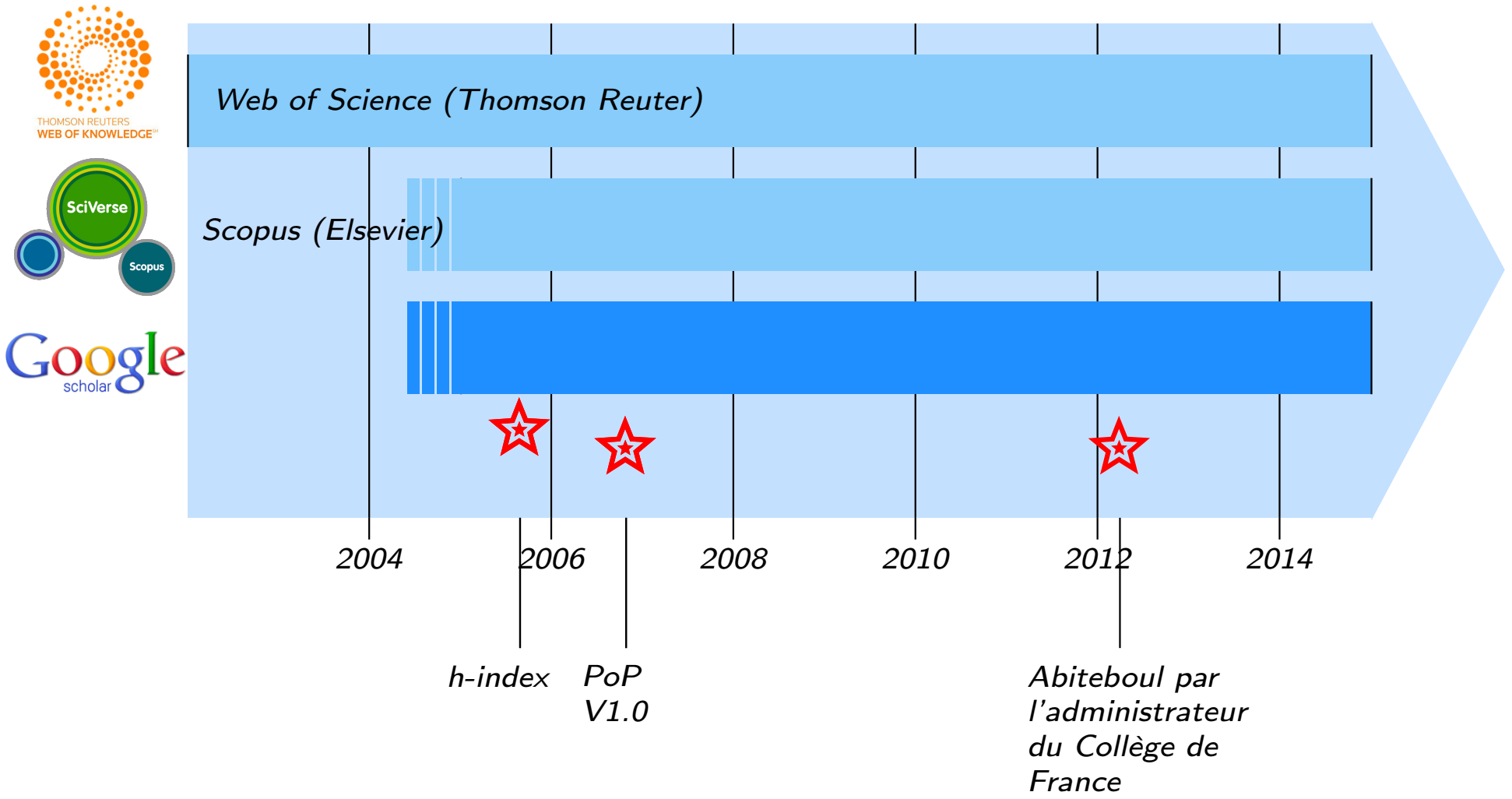
Outils gratuits :

- Google Scholar, CiteSeerX,...
- Parcours du web / de catalogues / ajoutées par les utilisateurs;
- Média sociaux (Google+, Scholarometer, Microsoft Academics...).

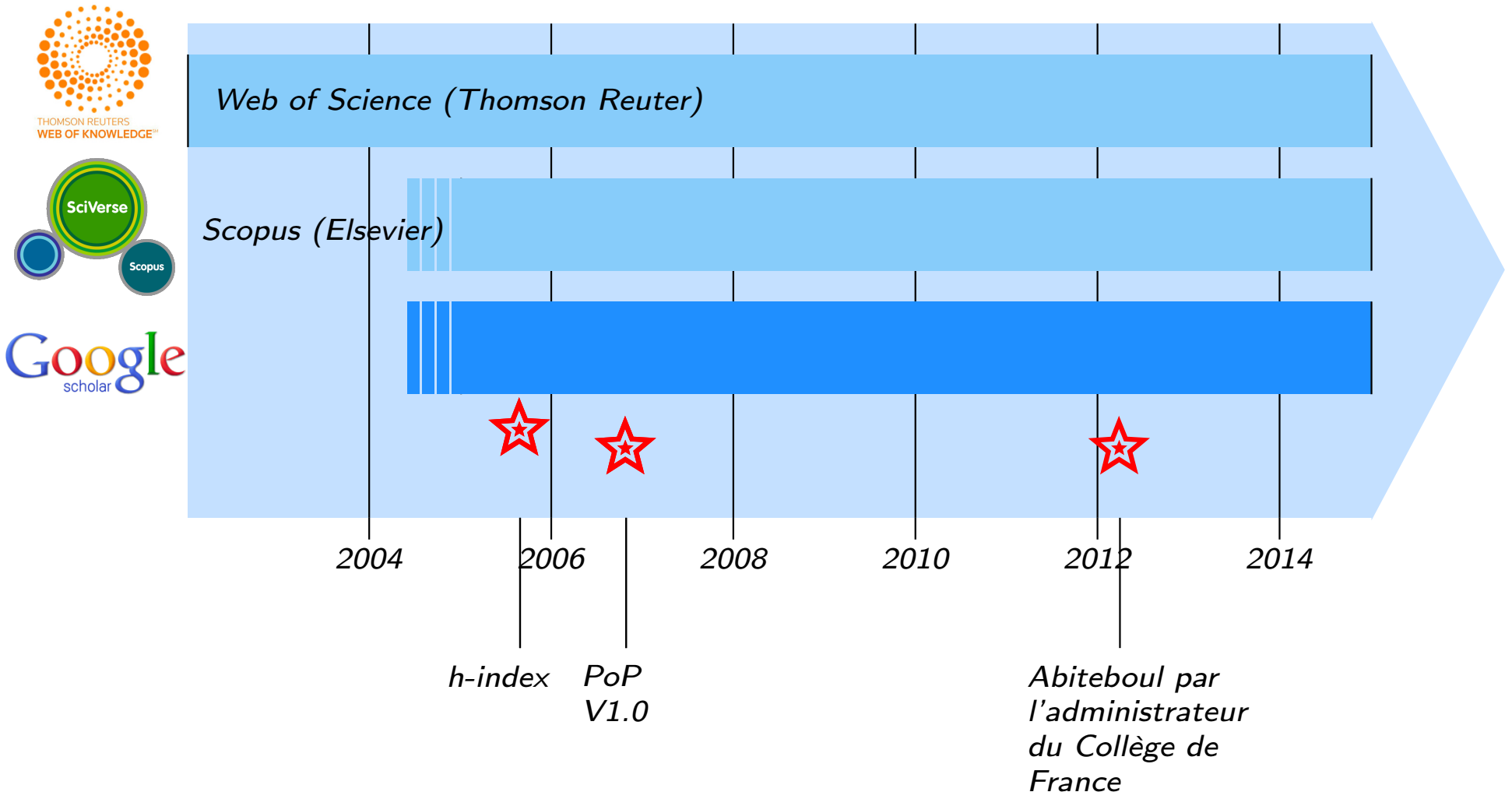
Outils gratuits pour calculer les indicateurs

Publish or Perish; Scholarometer; Microsoft Academics; Google+;

Chronos


THOMSON REUTERS
WEB OF KNOWLEDGE
SciVerse
Scopus
Google
scholar
Web of Science (Thomson Reuter)
Scopus (Elsevier)
2004
2006
2008
2010
2012
2014
h-index
PoP
V1.0
Abiteboul par
l'administrateur
du Collège de
France

Chronos



Outils de génération de publications.

Grammaire probabiliste hors contexte

Ensemble de symboles

- Non terminaux $\mathcal{N} = \{\mathcal{SP}, \mathcal{S}, \mathcal{V}, \mathcal{P}\}$,
- Terminaux $\Sigma = \{".", \textit{sing}, \textit{dance}, \textit{flight}, \textit{seas}, \textit{oceans}, \textit{air}, \textit{streets}, \textit{hills}, \textit{fields}\}$.

Set of rules $\mathcal{R}_i$

$\mathcal{R}_1$:	$\mathcal{SP}$	$\longrightarrow$	$\mathcal{S}.$	$p(\mathcal{R}_1)=1$	
$\mathcal{R}_2$:	$\mathcal{S}$	$\longrightarrow$	*We shall* $\mathcal{V}$ *in the* $\mathcal{P}$	$p(\mathcal{R}_2)=1/4$	
$\mathcal{R}_4$:	$\mathcal{S}$	$\longrightarrow$	*We shall* $\mathcal{V}$ *in the* $\mathcal{P}$ *and in the* $\mathcal{P}$, $\mathcal{S}$	$p(\mathcal{R}_4)=1/4$	
$\mathcal{R}_3$:	$\mathcal{S}$	$\longrightarrow$	$\mathcal{S}, \mathcal{S}$	$p(\mathcal{R}_3)=1/2$	
$\mathcal{R}_{5..7}$:	$\mathcal{V}$	$\longrightarrow$	$\textit{sing} \mid \textit{dance} \mid \textit{flight}$	$p(\mathcal{R}_i)=1/3$	$i=5..7$
$\mathcal{R}_{8..13}$:	$\mathcal{P}$	$\longrightarrow$	$\textit{seas} \mid \textit{oceans} \mid \textit{air} \mid \textit{streets} \mid \textit{hills} \mid \textit{fields}$	$p(\mathcal{R}_i)=1/6$	$i=8..13$

Exemple :

s : We shall sing in the air and in the hills, We shall dance in the fields.

$p(s) = \prod_j p(\mathcal{R}_j)$

Grammaire probabiliste hors contexte

Ensemble de symboles

- Non terminaux $\mathcal{N} = \{\mathcal{SP}, \mathcal{S}, \mathcal{V}, \mathcal{P}\}$,
- Terminaux $\Sigma = \{".", \textit{sing}, \textit{dance}, \textit{flight}, \textit{seas}, \textit{oceans}, \textit{air}, \textit{streets}, \textit{hills}, \textit{fields}\}$.

Set of rules $\mathcal{R}_i$

$\mathcal{R}_1$:	$\mathcal{SP}$	$\longrightarrow$	$\mathcal{S}.$	$p(\mathcal{R}_1)=1$	
$\mathcal{R}_2$:	$\mathcal{S}$	$\longrightarrow$	*We shall* $\mathcal{V}$ *in the* $\mathcal{P}$	$p(\mathcal{R}_2)=1/4$	*Non−zero*
$\mathcal{R}_4$:	$\mathcal{S}$	$\longrightarrow$	*We shall* $\mathcal{V}$ *in the* $\mathcal{P}$ *and in the* $\mathcal{P}$, $\mathcal{S}$	$p(\mathcal{R}_4)=1/4$	*probability*
$\mathcal{R}_3$:	$\mathcal{S}$	$\longrightarrow$	$\mathcal{S}, \mathcal{S}$	$p(\mathcal{R}_3)=1/2$	*to* ∞
$\mathcal{R}_{5..7}$:	$\mathcal{V}$	$\longrightarrow$	$\textit{sing} \vert \textit{dance} \vert \textit{flight}$	$p(\mathcal{R}_i)=1/3$	$i=5..7$
$\mathcal{R}_{8..13}$:	$\mathcal{P}$	$\longrightarrow$	$\textit{seas} \vert \textit{oceans} \vert \textit{air} \vert \textit{streets} \vert \textit{hills} \vert \textit{fields}$	$p(\mathcal{R}_i)=1/6$	$i=8..13$

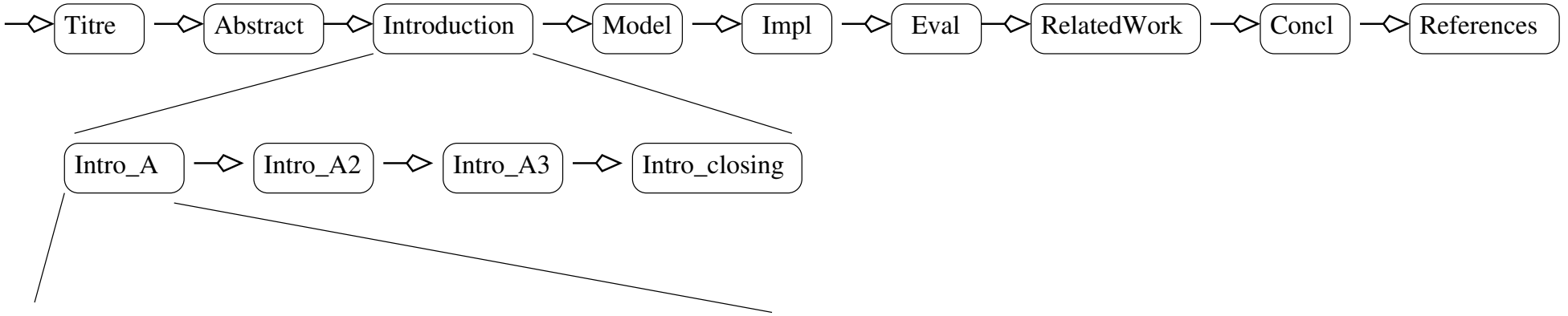
Exemple :

s : We shall sing in the air and in the hills, We shall dance in the fields.

$p(s) = \prod_j p(\mathcal{R}_j)$

SCIgen 2005 by J. Stribling, M. Krohn & D. Aguayo

... maximize amusement, rather than coherence ...



$\mathcal{I}ntro\_A \longrightarrow$ Many SCI_PEOPLE would agree that, had it not been for SCI_GENERIC_NOUN, ...

$\mathcal{I}ntro\_A \longrightarrow$ In recent years, much research has been devoted to the SCI_ACT; , ...

$\mathcal{I}ntro\_A \longrightarrow$ SCI_THING_MOD and SCI_THING_MOD, while SCI_ADJ in theory, have not until...

$\mathcal{I}ntro\_A \longrightarrow$ The SCI_ACT is a SCI_ADJSCI_PROBLEM.

$\mathcal{I}ntro\_A \longrightarrow$ The SCI_ACT has SCI_VERBEDSCI_THING_MOD, and current trends...

$\mathcal{I}ntro\_A \longrightarrow$ The implications of SCI_BUZZWORD_ADJ SCI_BUZZWORD_NOUN have...

$\ldots \longrightarrow \ldots$

SCI_PEOPLE $\longrightarrow$ steganographers, cyberinformaticians, futurists, cyberneticists, ...

SCI_BUZZWORD_ADJ $\longrightarrow$ omniscient, introspective, peer − to − peer, ambimorphic, ...

Rooter: A Methodology for the Typical Unification of Access Points and Redundancy

Jeremy Stribling, Daniel Aguayo and Maxwell Krohn

Abstract

Many physicists would agree that, had it not been for congestion control, the evaluation of web browsers might never have occurred. In fact, few hackers worldwide would disagree with the essential unification of voice-over-IP and public-private key pair. In order to solve this riddle, we confirm that SMPs can be made stochastic, cacheable, and interposable.

The rest of this paper is organized as follows. For starters, we motivate the need for fiber-optic cables. We place our work in context with the prior work in this area. To address this obstacle, we disprove that even though the much-tauted autonomous algorithm for the construction of digital-to-analog converters by Jones [10] is NP-complete, object-oriented languages can be made signed, decentralized, and

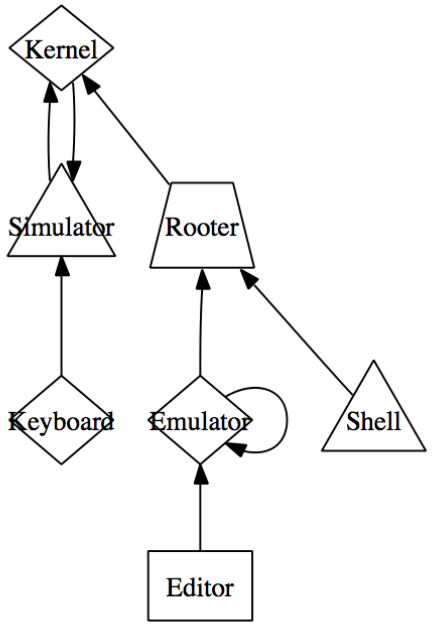


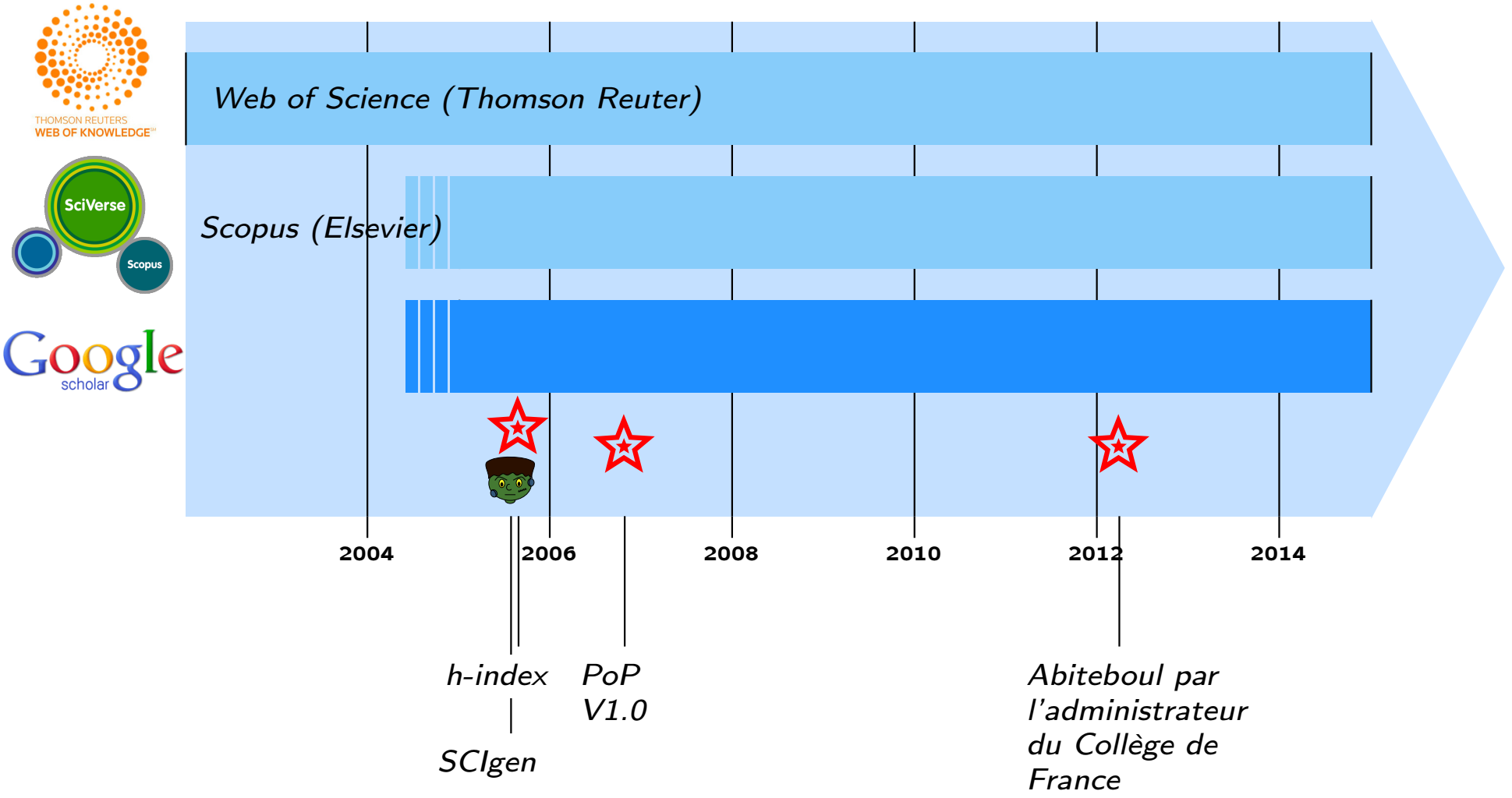
Fig. 2. The schematic used by our methodology.

References

[1] S. Abiteboul, Y. Huang and V. Ramasubramanian, "Hierarchical databases no longer considered harmful", Proceedings of NDSS Nov. 2005, pp. 22-28.

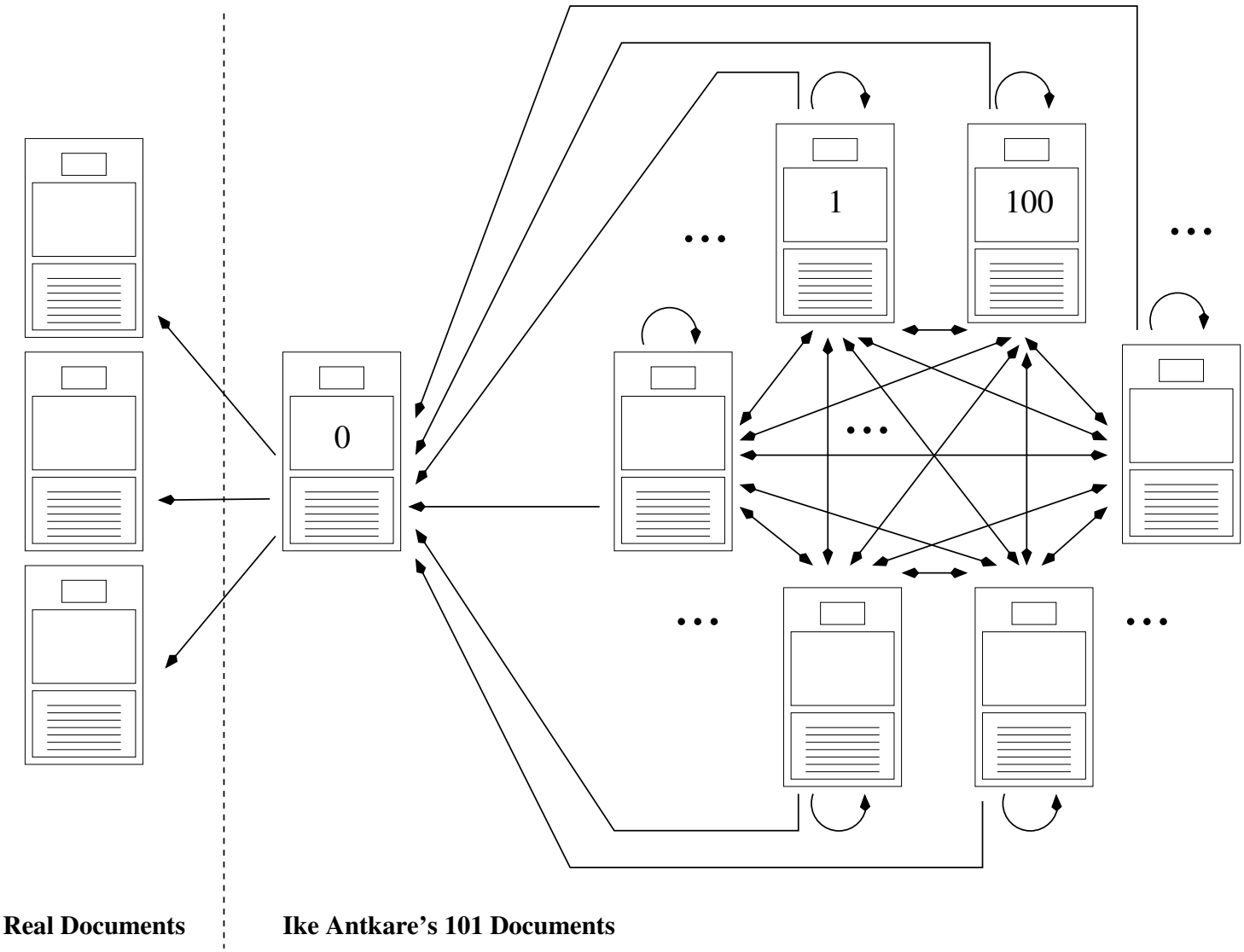
[2] O. Dahl, D. Johnson and R. Turing, "A. Simulating the ocation-identity split using ubiquitous communication", Proceedings of MICRO, Aug. 2006, pp.34-38.

Chronos


THOMSON REUTERS
WEB OF KNOWLEDGE
SciVerse
Scopus
Google
scholar
Web of Science (Thomson Reuter)
Scopus (Elsevier)
2004
2006
2008
2010
2012
2014
h-index
SCIgen
PoP
V1.0
Abiteboul par
l'administrateur
du Collège de
France

Une *ferme de citations* [Labbé, 2010]

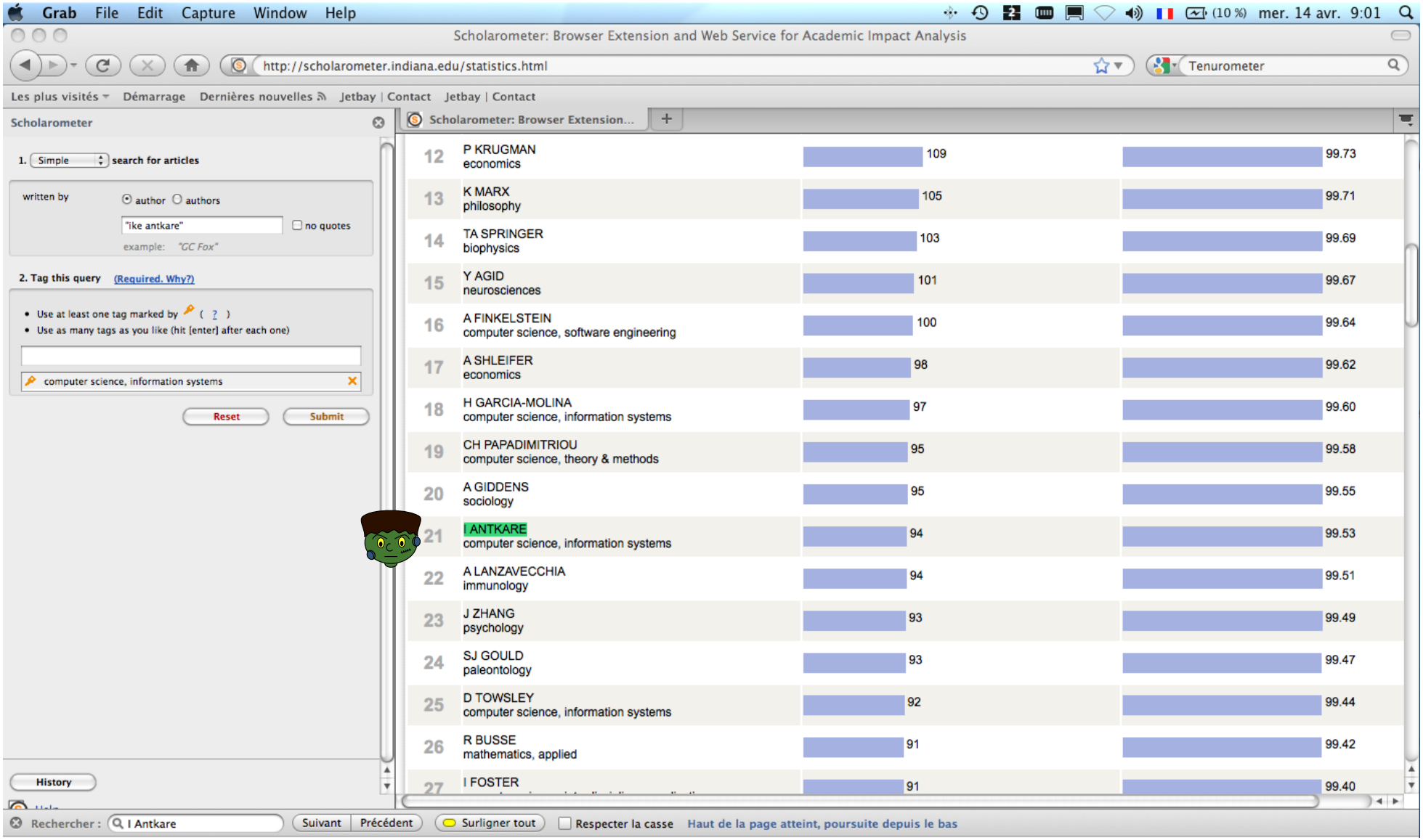
SCIgen modifié



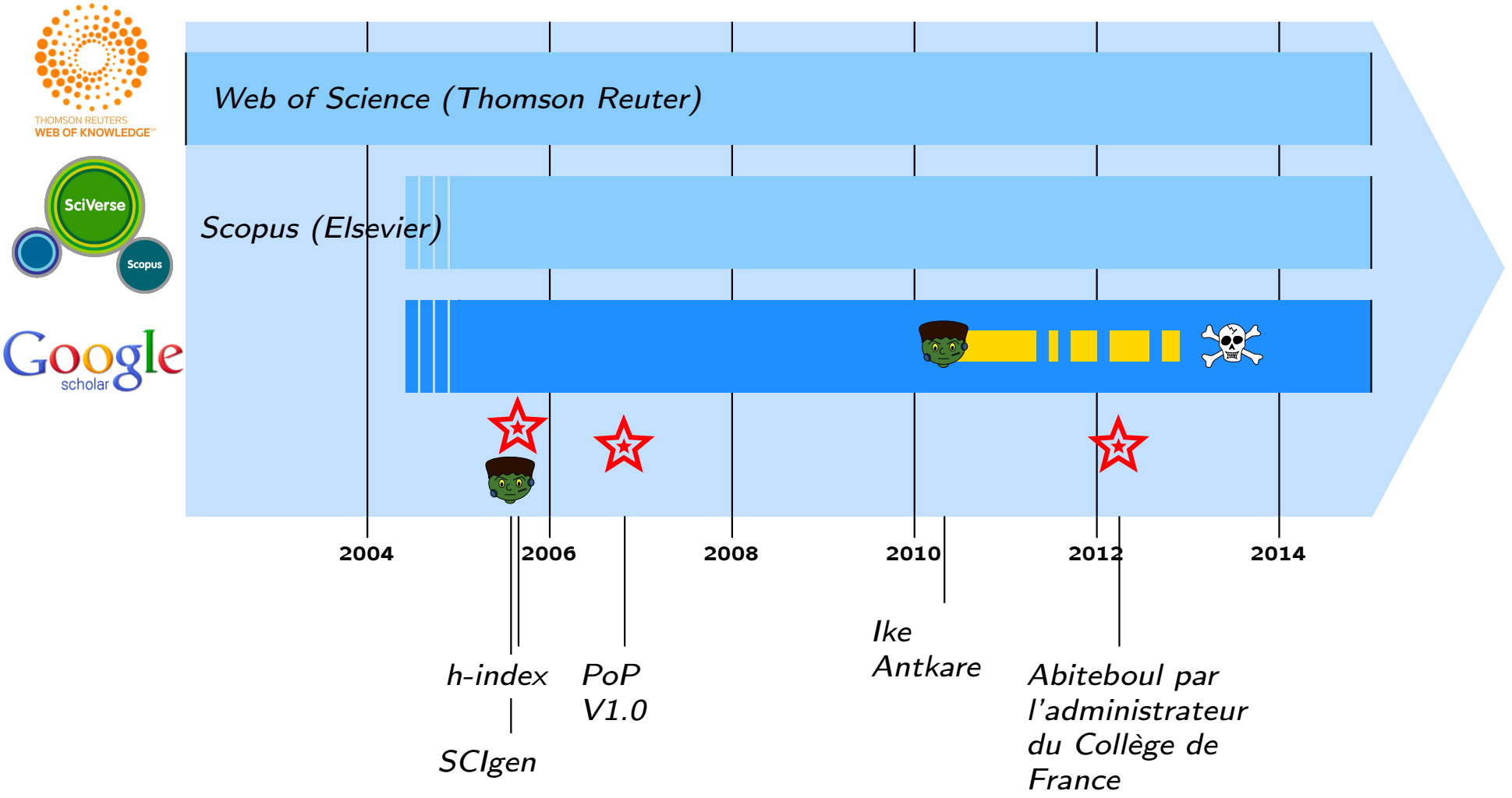
Ike Antkare h-index

Google scholar Scholarometer

[Labbé, 2010]



Chronos


THOMSON REUTERS
WEB OF KNOWLEDGE
SciVerse
Scopus
Google
scholar
Web of Science (Thomson Reuter)
Scopus (Elsevier)
2004
2006
2008
2010
2012
2014
h-index
SCIgen
PoP
V1.0
Ike
Antkare
Abiteboul par
l'administrateur
du Collège de
France

Get cited or Perish

Conclusion

	Complétude	Exactitude	Robustesse
Google Scholar (gratuit)	Bonne	Assez bonne	Spamable
WoK / Scopus (payant)	incomplète	Sans erreur	Excellente

Un scientifique ne fraudera jamais ainsi...

Get cited or Perish

Conclusion

	Complétude	Exactitude	Robustesse
Google Scholar (gratuit)	Bonne	Assez bonne	Spamable
WoK / Scopus (payant)	incomplète	Sans erreur	Excellente

Un scientifique ne fraudera jamais ainsi...

1. Préliminaires
   - Scientométrie
   - SCIgen une grammaire probabiliste hors contexte
2. Ike Antkare, one of the great starts in the scientific firmament
3. Detection de papiers SCIgen
   - Google Search
   - Classification automatique

Recherche de phrases et *More Like This* IEEE

http://www.computer.org

```
Many SCI_PEOPLE would agree that, had it not been for SCI_GENERIC_NOUN, ...
In recent years, much research has been devoted to the SCI_ACT; ...
SCI_THING_MOD and SCI_THING_MOD, while SCI_ADJ in theory, have not until ...
The SCI_ACT has SCI_VERBEDSCI_THING_MOD, and current trends ...
The implications of SCI_BUZZWORD_ADJ SCI_BUZZWORD_NOUN have ...
```

Recherche de phrases et *More Like This* IEEE

http://www.computer.org

```
Many SCI_PEOPLE would agree that, had it not been for SCI_GENERIC_NOUN, ...
In recent years, much research has been devoted to the SCI_ACT; ...
SCI_THING_MOD and SCI_THING_MOD, while SCI_ADJ in theory, have not until ...
The SCI_ACT has SCI_VERBEDSCI_THING_MOD, and current trends ...
The implications of SCI_BUZZWORD_ADJ SCI_BUZZWORD_NOUN have ...
```

Recherche de phrases et *More Like This* IEEE

http://www.computer.org

```
Many SCI_PEOPLE would agree that, had it not been for SCI_GENERIC_NOUN, ...
In recent years, much research has been devoted to the SCI_ACT; ...
SCI_THING_MOD and SCI_THING_MOD, while SCI_ADJ in theory, have not until ...
The SCI_ACT has SCI_VERBEDSCI_THING_MOD, and current trends ...
The implications of SCI_BUZZWORD_ADJ SCI_BUZZWORD_NOUN have ...
```

Corpus name	Downloaded from	Years	Type of papers	Number of papers	Acceptance rate	Corpus size
MLT	IEEE	2008	various	122	NA	122

Identification des faux :

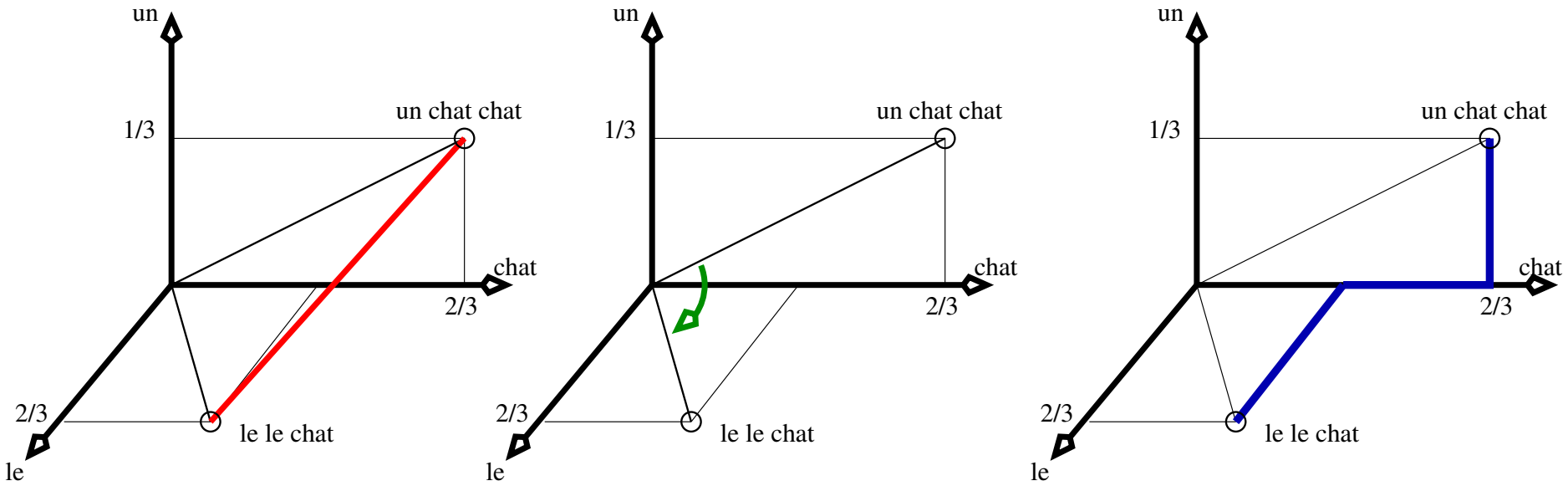
Corpus name	Downloaded from	Years	Type of papers	Number of papers	Acceptance rate	Corpus size
MLT	IEEE ieee.org	2008 2010	various	122	NA	122
Corpus Z	Conf. Web Site	2010	Track 1 Track 2 Track 3 Demo	58 33 36 32	18.4% 16.1% 36%	153
Ike	SCIgen	2009-2010	-	100	100%	100

- Extraction du texte à partir du pdf
- Calcule de la matrice des distances (sur les textes bruts) et construction d'un dendrogramme

Distance intertextuelle: [Labbé and Labbé, 2006]

A: {le le chat} $(\frac{1}{3}, \frac{2}{3}, \frac{0}{3})$

B: {un chat chat } $(\frac{2}{3}, \frac{0}{3}, \frac{1}{3})$



Distance intertextuelle : $D_{(A,B)} = \frac{1}{2} \sum_{i \in (A \cup B)} |f_{i,A} - f_{i,B}| = \frac{2}{3}$

Interprétation:

- $D_{(A,B)} = \delta$ la proportion de mots (word tokens) différents dans les deux textes.

Regroupement Hiérarchique [Labbé and Labbé, 2013]

$$D_{(I,J)} = \frac{1}{|I||J|}\left(\sum_{i\in I}\sum_{j\in J} D_{(i,j)} + D_{(i,j)}\right)$$

	I	J
I	0	0.45
J	0.45	0

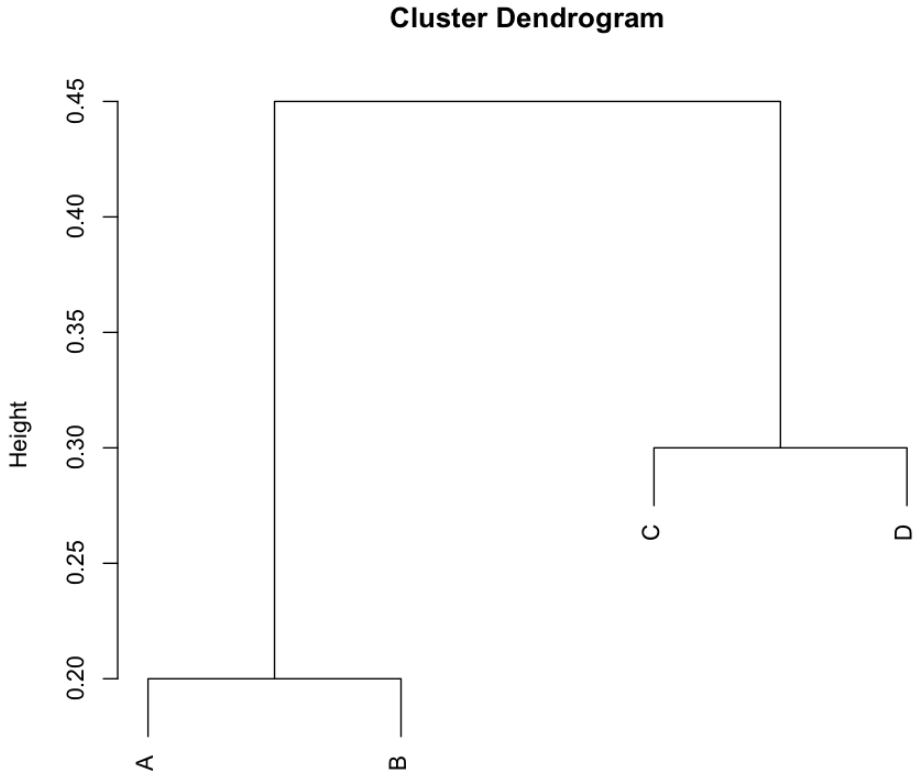
C et D forment le groupe J

$$D_{(I,x)} = \frac{1}{2}(D_{(A,x)} + D_{(B,x)})$$

	I	C	D
I	0	0.35	0.55
C	0.35	0	0.3
D	0.55	0.3	0

A et B forment le groupe I

	A	B	C	D
A	0	0.2	0.3	0.5
B	0.2	0	0.4	0.6
C	0.3	0.4	0	0.3
D	0.5	0.6	0.3	0

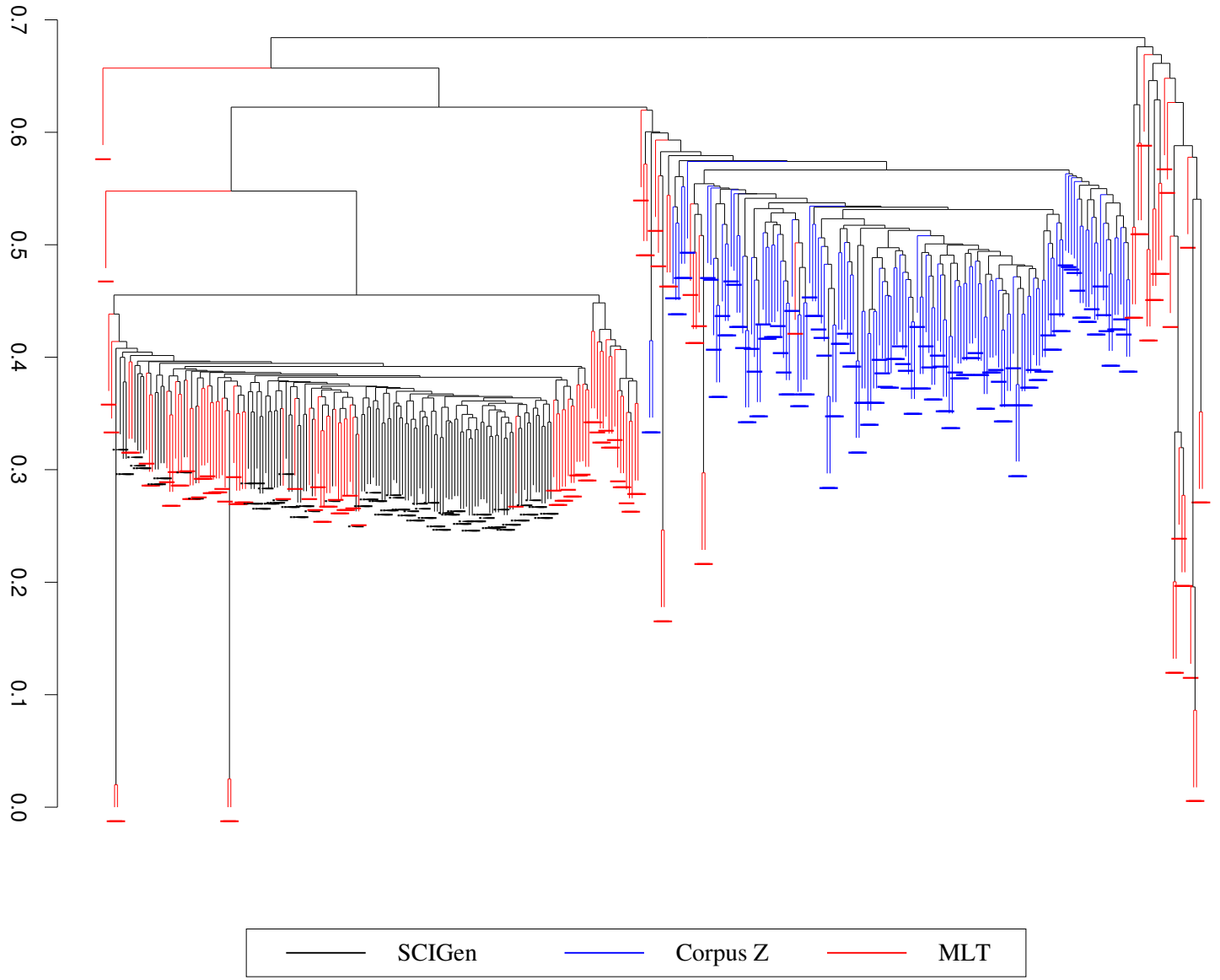


Identification des faux :

Corpus name	Downloaded from	Years	Type of papers	Number of papers	Acceptance rate	Corpus size
MLT	IEEE ieee.org	2008 2010	various	122	NA	122
Corpus Z	Conf. Web Site	2010	Track 1 Track 2 Track 3 Demo	58 33 36 32	18.4% 16.1% 36%	153
Ike	SCIgen	2009-2010	-	100	100%	100

- Extraction du texte à partir du pdf
- Calcule de la matrice des distances (sur les textes bruts) et construction d'un dendrogramme

Dendrogramme (Z, MTL, Ike)



Détection de SCIgen : méthode proposée http://scigendetection.imag.fr

Corpus	Downloaded	Years	Field	Corpus size
arXiv[1]	arxiv.org	08–10	Computer Science	15338
MLT	ieee.org	08–10	Computer Science	122
SCIgen-Origin	Original SCIgen	–	Computer Science	236
SCIgen-Physics	Modified SCIgen	–	Physics	414

Soit

- t un texte à tester.
- δ_t^{Fake} la distance entre t et le SCIgen le plus proche

Si $\delta_t^{Fake} < \delta_{Seuil}$

- Alors une provenance SCIgen doit sérieusement être considérée (misclass. risk $< 10^{-5}$).
- Sinon ($\delta_t^{Fake} > \delta_{Seuil}$) une origine non-SCIgen doit être considérée.

[1] open repository for scholarly papers

Site web de détection

http://scigendetection.imag.fr

Site de démonstration pour l'article [Labbé and Labbé, 2013]

- Input : *MyConf.zip* contenant des fichiers pdf
- Output : la classe (SCIgen/non-SCIgen) de chaque pdf, dendrogramme, doublons,...
- Utilisation *en production*.

Hindawi

Springer

INSTICC

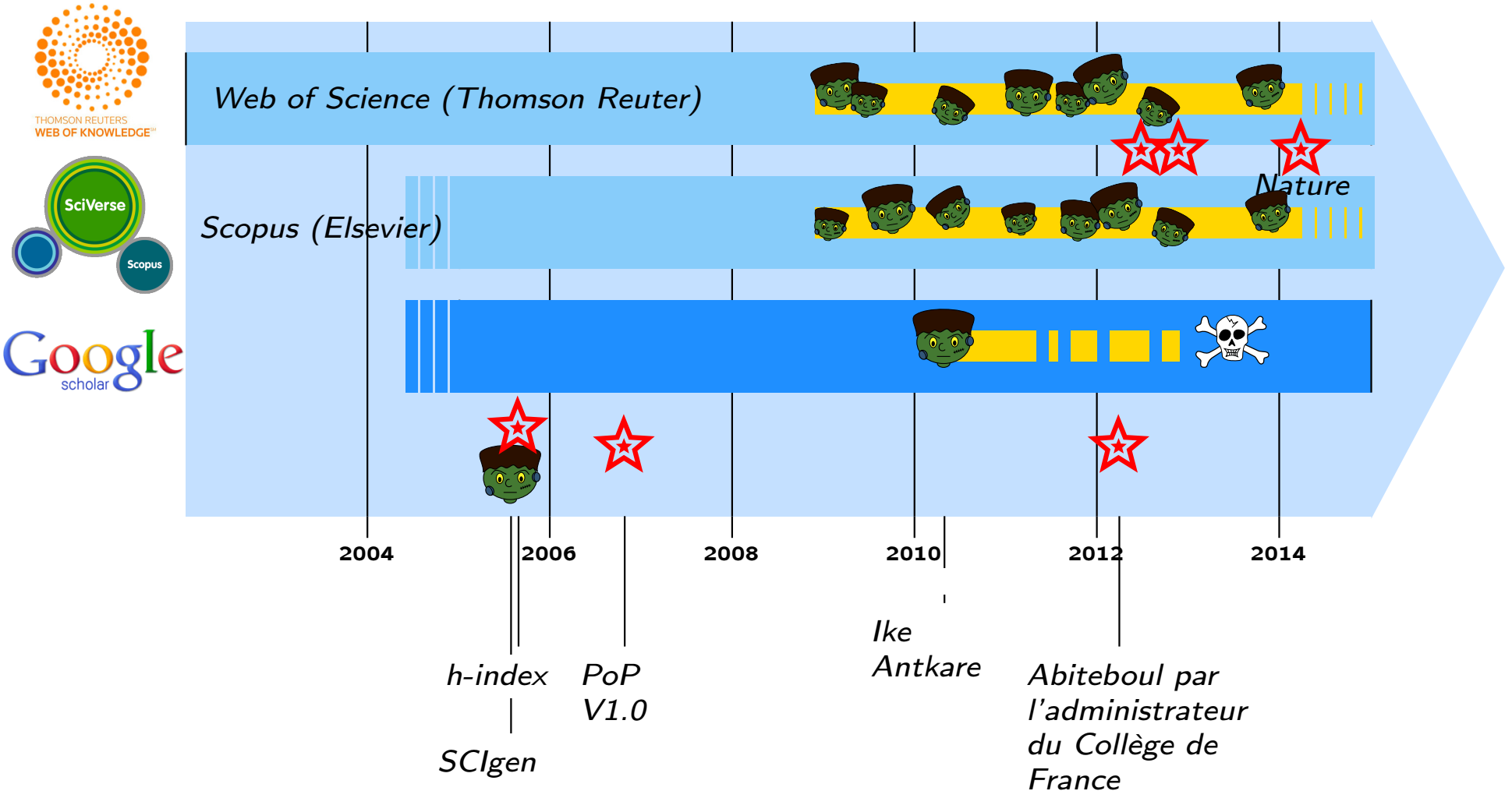
Utilisation automatique (depuis février 2014) :

- Nombre d'archives soumises $>$ 51000 (nombre d'articles testés $>$ 100000)

Utilisation curative :

- Détection de SCIgen déjà parus : 120 IEEE, 16 Springer, 3 Hal

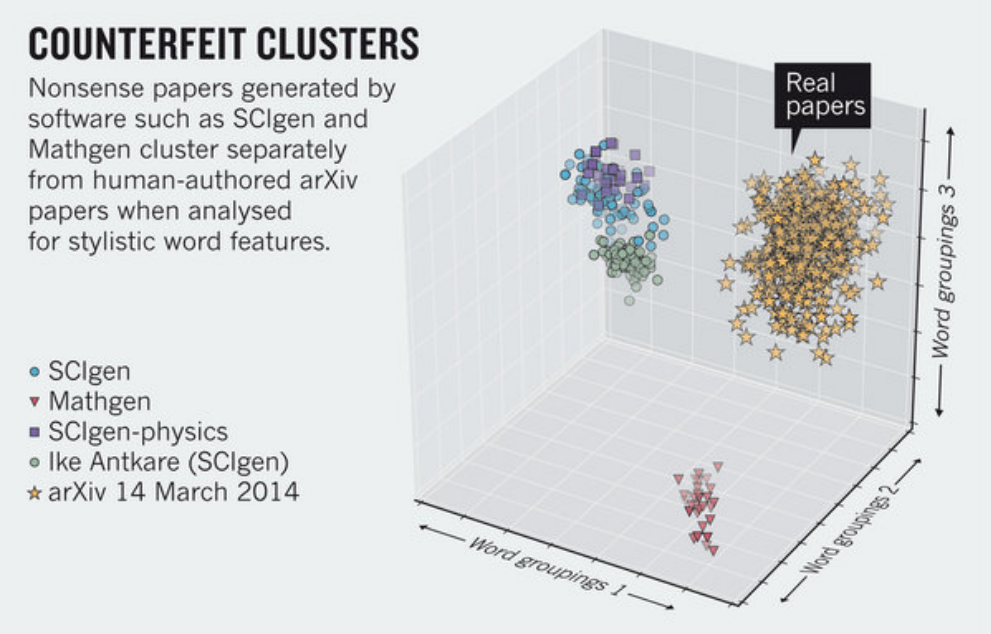
Scopus, Wok,...


THOMSON REUTERS
WEB OF KNOWLEDGE
SciVerse
Scopus
Google
scholar
Web of Science (Thomson Reuter)
Scopus (Elsevier)
Nature
2004
2006
2008
2010
2012
2014
h-index
SCIgen
PoP
V1.0
Ike
Antkare
Abiteboul par
l'administrateur
du Collège de
France

Travaux dérivés et connexes

- Spoofing [Beel and Gipp, 2010, Lopez-Cozar et al., 2012], Academic optim. [Beel et al., 2010];
- Detecting methods: Bib. based [Xiong and Huang, 2009], Compression [Dalkilic et al., 2006], ad-hoc dist. [Lavoie and Krishnamoorthy, 2010], Phrase search [Springer, 2014], Structural distances between texts [Fahrenberg et al., 2014].

Pas de SCIgen dans arXiv (Computer Science)



- Image borrowed from [Ginsparg, 2014];
- PCA, mots vides uniquement.
- Calibrage de la méthode à l'aide des corpus.

Conclusion et travaux à venir

Procédures, modèles et habitudes de publications

- Pourquoi de faux papiers ont été acceptés, publiés et ... vendus.
- Publication traditionnelles *vs* accès publique.
- Diffusion du savoir: mieux et moins... où autant que possible.

Les règles de gestion aveugles...

- ...incitent à la fraude : saucissonnage, plagiat, données trafiquées,...

Détection automatique de nouveaux générateurs

- Grammaires : trouver des groupes dense dans de grandes populations.
- Etudier d'autres sortes de générateurs (modèle de langue).

Le web aujourd'hui

- Extraction/détection/génération automatique du savoir.
- Comment séparer le bon grain de ivraie...

Merci !

Beel, J. and Gipp, B. (2010).
Academic search engine spam and google scholar's resilience against it.
Journal of Electronic Publishing, 13(3).

Beel, J., Gipp, B., and Wilde, E. (2010).
Academic search engine optimization (aseo).
Journal of scholarly publishing, 41(2):176–190.

Dalkilic, M. M., Clark, W. T., Costello, J. C., and Radivojac, P. (2006).
Using compression to identify classes of inauthentic texts.
In *Proceedings of the 2006 SIAM Conference on Data Mining*.

Fahrenberg, U., Biondi, F., Corre, K., Jégourel, C., Kongshøj, S., and Legay, A. (2014).
Measuring structural distances between texts.
CoRR, abs/1403.4024.

Ginsparg, P. (2014).
Automated screening: Arxiv screens spot fake papers.
Nature, 508(7494):44–44.

Hirsch, J. E. (2005).
An index to quantify an individual's scientific research output.
Proceedings of the National Academy of Science, 102:16569–16572.

Labbé, C. (2010).
Ike antkare, one of the great stars in the scientific firmament.
International Society for Scientometrics and Informetrics Newsletter, 6(2):48–52.

Labbé, C. and Labbé, D. (2006).
A tool for literary studies. intertextual distance and tree classification.
Literary and Linguistic Computing, 21(3):311–326.

Labbé, C. and Labbé, D. (2013).
Duplicate and fake publications in the scientific literature: how many scigen papers in computer science?
Scientometrics, 94(1):379–396.

Lavoie, A. and Krishnamoorthy, M. (2010).
Algorithmic Detection of Computer Generated Text.
ArXiv e-prints.

Lopez-Cozar, E. D., Robinson-García, N., and Torres-Salinas, D. (2012).
Manipulating google scholar citations and google scholar metrics: Simple, easy and tempting.
arXiv preprint arXiv:1212.0638.

Xiong, J. and Huang, T. (2009).
An effective method to identify machine automatically generated paper.
In *Knowledge Engineering and Software Engineering, 2009. KESE '09. Pacific-Asia Conference on*, pages 101–102.