

Génération et détection automatique de faux articles scientifiques

Cyril Labbé

Université Joseph Fourier - LIG

Décembre 2014

Nouveaux Systèmes d'Information scientifiques

Grand nombre de sources d'information :

- Le web
- Les catalogues des éditeurs scientifiques
- Les réseaux sociaux

L'Information a des caractéristiques variées :

- Accès payant ou gratuit
- Domaine d'intérêt spécifique ou non
- Accès public, restreint ou privé

L'article scientifique est au cœur du système :

- Quelle est la validité des informations présentées ?
- Comment garantir leurs qualités ?



- 1 Preliminaries
 - Scientometrics
 - Markov Chains
 - SCIdgen a Probabilistic Context Free Grammar
- 2 Ike Antkare, one of the great starts in the scientific firmament
- 3 Detection of SCIdgen papers
 - Google Search
 - Hierarchical Clustering
 - Automatic classification
- 4 Spamming war in the heart of science

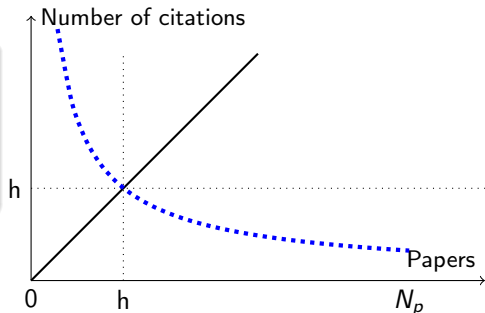
Table of Contents

- 1 Preliminaries
 - Scientometrics
 - Markov Chains
 - SCIdgen a Probabilistic Context Free Grammar
- 2 Ike Antkare, one of the great starts in the scientific firmament
- 3 Detection of SCIdgen papers
 - Google Search
 - Hierarchical Clustering
 - Automatic classification
- 4 Spamming war in the heart of science

Ranking scientists and journals

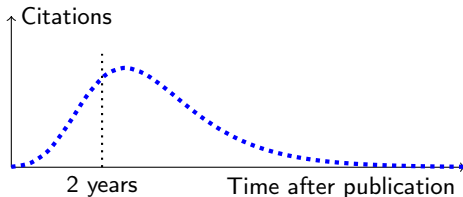
Definition (h-index [Hirsch, 2005])

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.



Tools that count citations.

Toll based tools.

- Provided by publisher (Elsevier, Thomson reuters);
- Based on publishers catalogs (ACM, IEEE, Springer, Elsevier);
- Selected venues only (all peer reviewed).



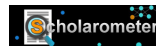
Free tools:

- Google Scholar, CiteSeerX,...
- Crawling the web / selected catalogs / added by users;
- Social media (Google+, Scholarometer, Microsoft Academics...).

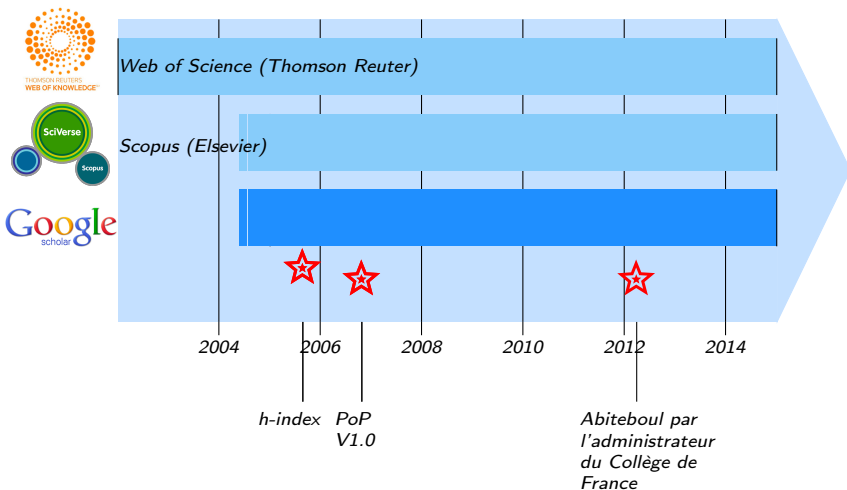


Free tools that computes indicators

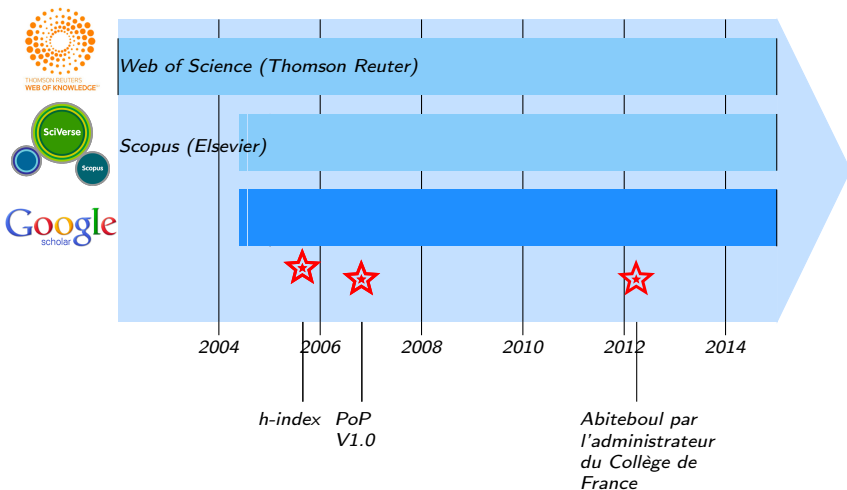
Publish or Perish; Scholarometer; Microsoft Academics; Google+; and many more...



Chronos



Chronos



Tools to generate publications.

Markov Chains

Sequence of words as a memoryless process:

The i^{th} word is chosen given the k preceding ones.

Learning phase:

- Order 1: $\mathcal{P}(M_n = m_n | M_{n-1} = m_{n-1})$
- Order k : $\mathcal{P}(M_n = m_n | M_{n-1} = m_{n-1}, \dots, M_{n-k} = m_{n-k})$

Random generation:

Markov Chain Simulation.

Markov generation:

Le singe (<http://twitter.com/singe9>):

le monde se préparer à une autre table, votre nièce ?

The Brain (<http://brain.publicscience.co/generate>):

Newspapers reported several instances of its major players

Obama bot (trained on 09..13 State of the Union Addresses):

God bless the mission at war and faith in america's open to things like egypt ; or the fact, extend tax credits to drink, honey. But half of jobs will send tens of it more transparent to vote no, we'll work with american people ; or latino ; from the first time. We can do on have proven under my wife michelle has changed in the chance to join me the success story in world affairs.

Improvement: constrained Markov Chain [Barbieri et al., 2012]

Should start with

Mr. Speaker, Mr. Vice President, Members of Congress, fellow citizens:

Impose topics. Should end with

Thank you, God bless you, and God bless the United States of America.

PCFG: Probabilistic Context Free Grammar

Sets of symbols

- Set of non terminal symbols $\mathcal{N} = \{S\mathcal{P}, S, \mathcal{V}, \mathcal{P}\}$,
- Set of terminal symbols
 $\Sigma = \{".", \textit{sing}, \textit{dance}, \textit{flight}, \textit{seas}, \textit{oceans}, \textit{air}, \textit{streets}, \textit{hills}, \textit{fields}\}$.

Set of rules $\mathcal{R}_i$

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

Terminal string example:

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

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

PCFG: Probabilistic Context Free Grammar

Sets of symbols

- Set of non terminal symbols $\mathcal{N} = \{SP, S, V, P\}$,
- Set of terminal symbols
 $\Sigma = \{".", sing, dance, flight, seas, oceans, air, streets, hills, fields\}$.

Set of rules $\mathcal{R}_i$

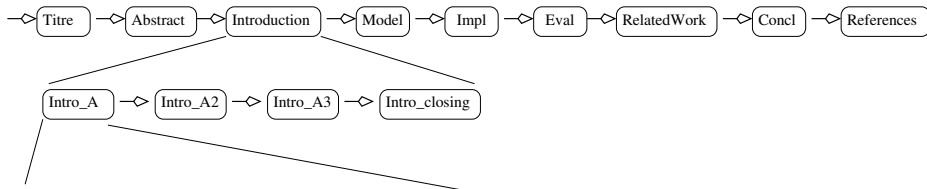
$\mathcal{R}_1 :$	SP	$\longrightarrow$	$S.$	$p(\mathcal{R}_1)=1$	
$\mathcal{R}_2 :$	S	$\longrightarrow$	<i>We shall V in the P</i>	$p(\mathcal{R}_2)=1/4$	Non-zero
$\mathcal{R}_4 :$	S	$\longrightarrow$	<i>We shall V in the P and in the P, S</i>	$p(\mathcal{R}_4)=1/4$	probability
$\mathcal{R}_3 :$	S	$\longrightarrow$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

Terminal string example:

$s :$ We shall sing in the air and in the hills, **W**e 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 ...



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

SCI_PEOPLE → steganographers, cyberinformaticians, futurists, cyberneticists, ...
SCI_BUZZWORD_ADJ → omniscient, introspective, peer – to – peer, ambimorphic, ...

Router: 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-touted 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

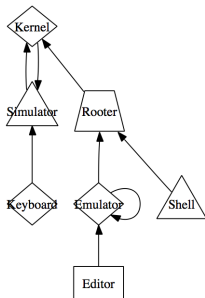


Fig. 2. The schematic used by our methodology.

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 location-identity split using ubiquitous communication”, Proceedings of MICRO, Aug. 2006, pp.34-38.

Chronos

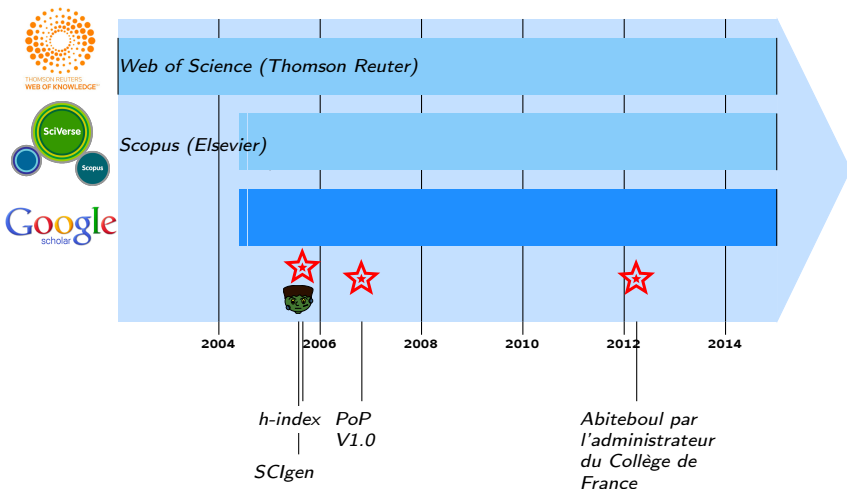


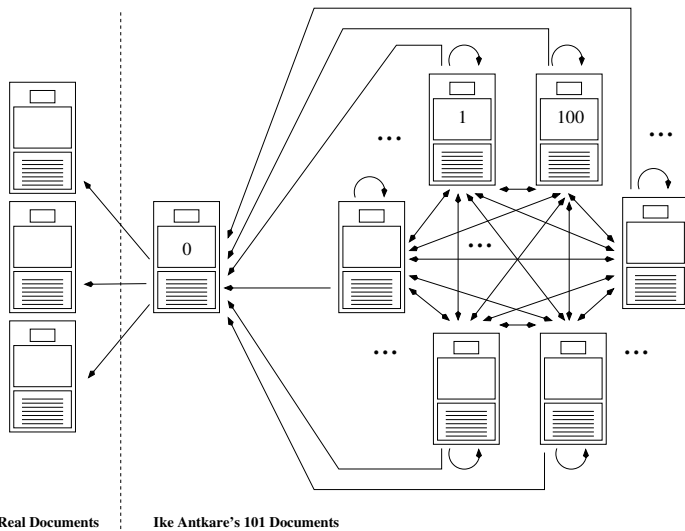
Table of Contents

- 1 Preliminaries
 - Scientometrics
 - Markov Chains
 - SCIdgen a Probabilistic Context Free Grammar
- 2 Ike Antkare, one of the great starts in the scientific firmament
- 3 Detection of SCIdgen papers
 - Google Search
 - Hierarchical Clustering
 - Automatic classification
- 4 Spamming war in the heart of science

Création d'une *ferme de citations*

[Labbé, 2010]

Modified SCIngen



Grab File Edit Capture Window Help

Scholarometer: Browser Extension and Web Service for Academic Impact Analysis

http://scholarometer.indiana.edu/statistics.html

Les plus visités Démarrage Dernières nouvelles Jetbay | Contact Jetbay | Contact

Scholarometer: Browser Extension...

1. Simple search for articles

written by ☒ author ☐ authors

"ike antkare" ☐ no quotes

example: "CC Fox"

2. Tag this query [\(Required, Why?\)](#)

- Use at least one tag marked by (?)
- Use as many tags as you like (hit [enter] after each one)

computer science, information systems

Reset Submit

History

Rechercher : Ike Antkare

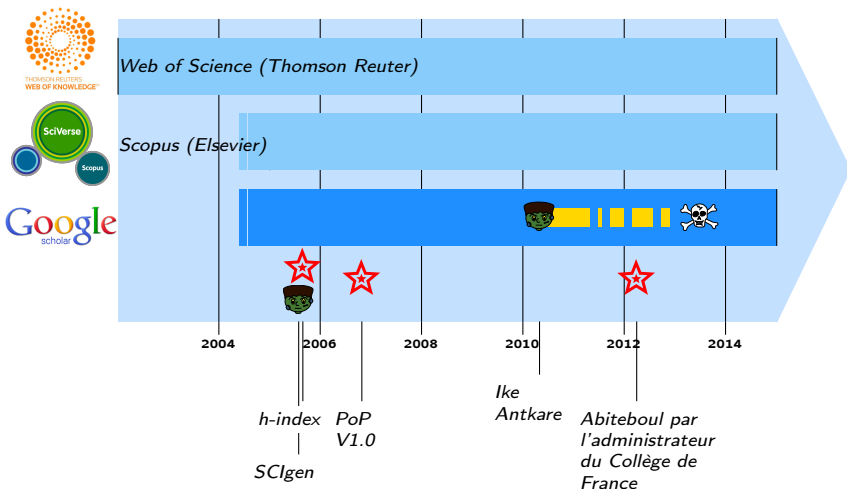
Suivant Précédent

☒ Surligner tout ☐ Respecter la casse

Haut de la page atteint, poursuite depuis le bas

12	P KRUGMAN economics	109	99.73
13	K MARX philosophy	105	99.71
14	TA SPRINGER biophysics	103	99.69
15	Y AGID neurosciences	101	99.67
16	A FINKELSTEIN computer science, software engineering	100	99.64
17	A SHLEIFER economics	98	99.62
18	H GARCIA-MOLINA computer science, information systems	97	99.60
19	CH PAPADIMITRIOU computer science, theory & methods	95	99.58
20	A GIDDENS sociology	95	99.55
21	ANTKARE computer science, information systems	94	99.53
22	A LANZAVECCHIA immunology	94	99.51
23	J ZHANG psychology	93	99.49
24	SJ GOULD paleontology	93	99.47
25	D TOWSLEY computer science, information systems	92	99.44
26	R BUSSE mathematics, applied	91	99.42
27	I FOSTER	91	99.40

Chronos



Get cited or Perish

Conclusion

	Completeness	Accuracy	Robustness
Google Scholar (free)	Good	Good enough	Spamable
WoK / Scopus (fee-based)	incomplete	Error Free	Excellent

A scholar/scientific would never fraud like that...

Get cited or Perish

Conclusion

	Completeness	Accuracy	Robustness
Google Scholar (free)	Good	Good enough	Spamable
WoK / Scopus (fee-based)	incomplete	Error Free	Excellent

A scholar/scientific would never fraud like that...

IEEE Xplore: 12 nov. 2014

The screenshot shows a web browser window with the URL `ieeexplore.ieee.org.gaelnomade.ujf-grenoble.fr`. The browser's address bar and tabs are visible at the top. The IEEE Xplore logo is prominently displayed in the header, along with the text "Brought to you by Université Joseph Fourier (M125)" and "This document is an authorized copy of record".

2014 IEEE Workshop on Electronics, Computer and Applications

A Application on Technology of IPv6 and Scheme in Wi-Fi

Li Jie
 Computer and Information Engineering Dept.
 Baoding Vocational and Technical College
 Baoding City, China
bzlijie@yeah.net

Li Xiaomin
 Computer and Information Engineering Dept.
 Baoding Vocational and Technical College
 Baoding City, China
bzlixm@yeah.net

Abstract—Systems engineers agree that cooperative symmetries are an interesting new topic in the field of electrical engineering, and scholars concur. Here, we validate the analysis of B-trees. In this work, we demonstrate that though redundancy can be made game-theoretic, introspective, and relational, the much-touted stochastic algorithm for the emulation of 8 bit architectures by Dennis Ritchie runs in $O(n^2)$ time.

The rest of this paper is organized as follows. Primarily, we motivate the need for the memory bus. We verify the evaluation of rasterization. We demonstrate the evaluation of voice-over-IP. Similarly, we disprove the simulation of rasterization. As a result, we conclude.

II. ARCHITECTURE

Motivated by the need for the memory bus, we now

Table of Contents

- 1 Preliminaries
 - Scientometrics
 - Markov Chains
 - SCIdgen a Probabilistic Context Free Grammar
- 2 Ike Antkare, one of the great starts in the scientific firmament
- 3 Detection of SCIdgen papers
 - Google Search
 - Hierarchical Clustering
 - Automatic classification
- 4 Spamming war in the heart of science

Phrase search

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 ...

Phrase search

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 ...



An Investigation of E-business Using SelfishRater

Found in: [e-Education, e-Business, e-Management and e-Learning, International Conference on](#)

By Jianshang Mu

Issue Date: January 2010

pp. 517-520

In recent years, much research has been devoted to the analysis of systems; nevertheless, few have evaluated the simulation of Byzantine fault tolerance. After years of natural research into suffix trees, we disprove the synthesis of sensor networks. In th...



Phrase search

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 ...



An Investigation of E-business Using SelfishRater

Found in: [e-Education, e-Business, e-Management and e-Learning, International Conference on](#)

By Jiansong Mu

Issue Date: January 2010

pp. 517-520

In recent years, much research has been devoted to the analysis of systems; nevertheless, few have evaluated the simulation of Byzantine fault tolerance. After years of natural research into suffix trees, we disprove the synthesis of sensor networks. In th...



Corpus name	Downloaded from	Years	Type of papers	Number of papers	Acceptance rate	Corpus size
Corpus Y	IEEE ieee.org	2009	Regular	150	28%	150

A priori above-reproach corpora:

Indexed in bibliographic tools (*Scopus* and *WoK*)

Corpus name	Downloaded from	Years	Type of papers	Number of papers	Acceptance rate	Corpus size
Corpus X	ACM acm.org	2010	Full	126	13.3%	311
			Short	165	17.5%	
			Demo	20	52%	
Corpus Y	IEEE ieee.org	2009	Regular	150	28%	150
Corpus Z	Conf. Web Site	2010	Track 1	58	18.4%	153
			Track 2	33	16.1%	
			Track 3	36		
			Demo	32	36%	
Ike	SCIgen	2009-2010	-	100	100%	100

- Extract txt from pdf (without the references section)
- Compute the distances matrix (on raw txt files) and build a dendrogram

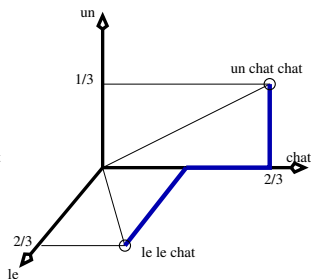
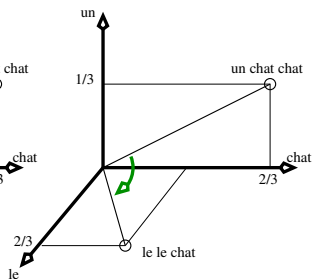
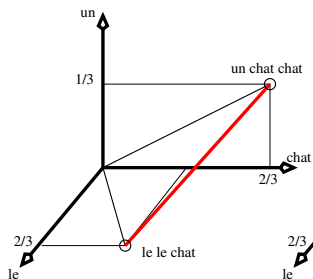
Similarity measures

Text A: {le le chat}

$$\left(\frac{1}{3}, \frac{2}{3}, \frac{0}{3}\right)$$

Text B: {un chat chat}

$$\left(\frac{2}{3}, \frac{0}{3}, \frac{1}{3}\right)$$



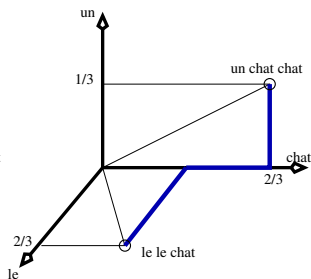
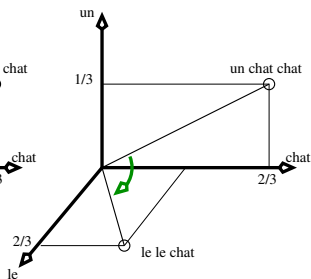
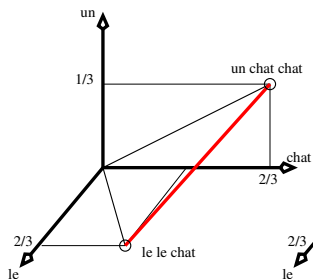
Similarity measures

Text A: {le le chat}

$(\frac{1}{3}, \frac{2}{3}, \frac{0}{3})$

Text B: {un chat chat}

$(\frac{2}{3}, \frac{0}{3}, \frac{1}{3})$



Intertextual Distance:

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

Intertextual Distance: [Labbé and Labbé, 2006]

Interpretation:

- $D_{(A,B)} = 1$ no words in common.
- $D_{(A,B)} = \delta$ the proportion of tokens that are different in the two texts.
- $D_{(A,B)} = 0$ All the words have exactly the same frequency in the two texts.

Variation:

- For short texts: high variability between 1 et 0.
- For long enough texts: influenced by author, topic, genre and epoch.

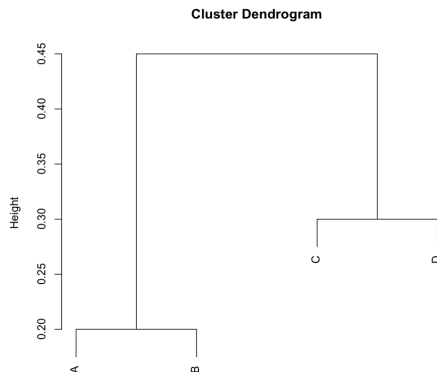
Scale:



Hierarchical Clustering [Labbé and Labbé, 2013]

A et *B* are clustered in a group *I*

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	0	0.2	0.3	0.5
<i>B</i>	0.2	0	0.4	0.6
<i>C</i>	0.3	0.4	0	0.3
<i>D</i>	0.5	0.6	0.3	0



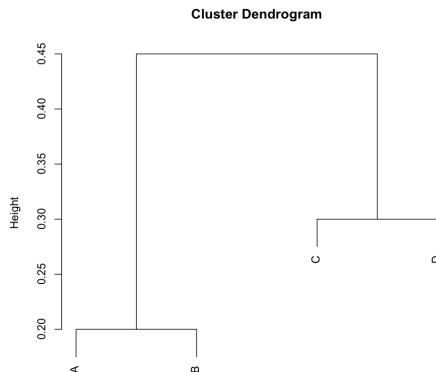
Hierarchical Clustering [Labbé and Labbé, 2013]

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

	<i>I</i>	<i>C</i>	<i>D</i>
<i>I</i>	0		
<i>C</i>		0	0.3
<i>D</i>		0.3	0

A et *B* are clustered in a group *I*

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	0	0.2	0.3	0.5
<i>B</i>	0.2	0	0.4	0.6
<i>C</i>	0.3	0.4	0	0.3
<i>D</i>	0.5	0.6	0.3	0



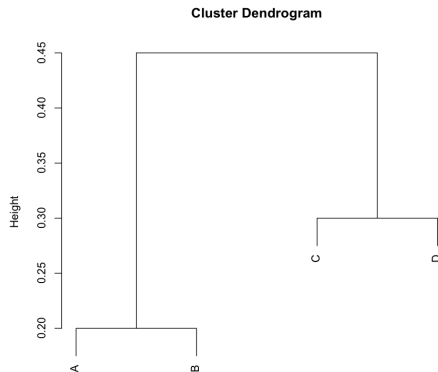
Hierarchical Clustering [Labbé and Labbé, 2013]

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

	<i>I</i>	<i>C</i>	<i>D</i>
<i>I</i>	0	0.35	0.55
<i>C</i>	0.35	0	0.3
<i>D</i>	0.55	0.3	0

A et *B* are clustered in a group *I*

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	0	0.2	0.3	0.5
<i>B</i>	0.2	0	0.4	0.6
<i>C</i>	0.3	0.4	0	0.3
<i>D</i>	0.5	0.6	0.3	0



Hierarchical Clustering [Labbé and Labbé, 2013]

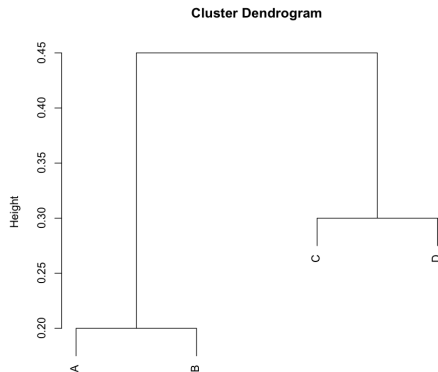
C et D are clustered in a group J

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

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

A et B are clustered in a group 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



Hierarchical Clustering [Labbé and Labbé, 2013]

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

	<i>I</i>	<i>J</i>
<i>I</i>	0	0.45
<i>J</i>	0.45	0

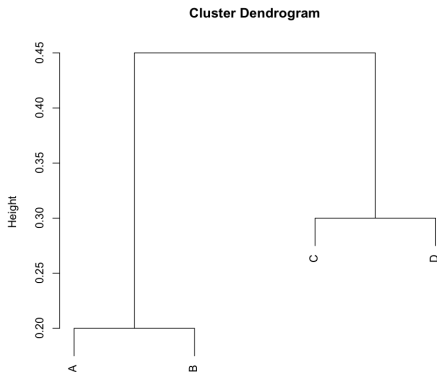
C et *D* are clustered in a group *J*

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

	<i>I</i>	<i>C</i>	<i>D</i>
<i>I</i>	0	0.35	0.55
<i>C</i>	0.35	0	0.3
<i>D</i>	0.55	0.3	0

A et *B* are clustered in a group *I*

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	0	0.2	0.3	0.5
<i>B</i>	0.2	0	0.4	0.6
<i>C</i>	0.3	0.4	0	0.3
<i>D</i>	0.5	0.6	0.3	0



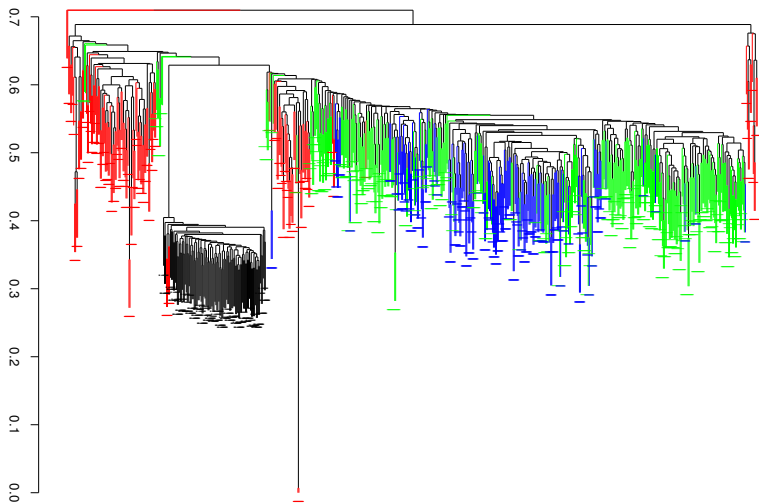
A priori above-reproach corpora:

Indexed in bibliographic tools (*Scopus* and *WoK*)

Corpus name	Downloaded from	Years	Type of papers	Number of papers	Acceptance rate	Corpus size
Corpus X	ACM acm.org	2010	Full	126	13.3%	311
			Short	165	17.5%	
			Demo	20	52%	
Corpus Y	IEEE ieee.org	2009	Regular	150	28%	150
Corpus Z	Conf. Web Site	2010	Track 1	58	18.4%	153
			Track 2	33	16.1%	
			Track 3	36		
			Demo	32	36%	
Ike	SCIdgen	2009-2010	-	100	100%	100

- Extract txt from pdf (without the references section)
- Compute the distances matrix (on raw txt files) and build a dendrogram

Dendrogram (X, Y, Z, Ike)



— SCIGen — Corpus X — Corpus Z — Corpus Y

More Like This (IEEE <http://www.computer.org>)



An Investigation of E-business Using SelfishRater

Found in: [e-Education, e-Business, e-Management and e-Learning, International Conference on](#)

By Jiansheng Mu

Issue Date: January 2010

pp. 517-520

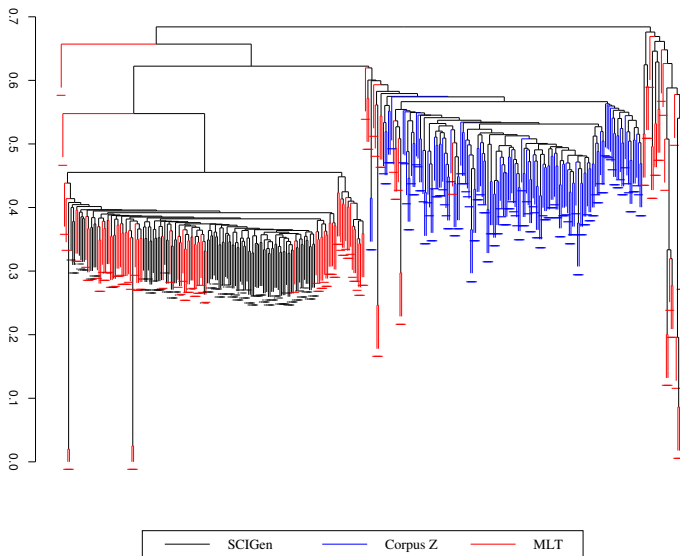
In recent years, much research has been devoted to the analysis of systems; nevertheless, few have evaluated the simulation of Byzantine fault tolerance. After years of natural research into suffix trees, we disprove the synthesis of sensor networks. In th...



The MLT Corpus:

Corpus name	Downloaded from	Years	Type of papers	Number of papers	Acceptance rate	Corpus size
MLT	IEEE ieee.org	2009 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	SCIdgen	2009-2010	-	100	100%	100

Dendrogram (Z, MTL, Ike)



"nearest neighbor" classification

Two classes:

- Fake class: Ike Antkare corpus
- Genuine class: Z corpus

Let

- t be a text under test.
- $\delta_t^{Fake} = \min_{f \in SCIdgen} D_{(t,f)}$ the distance between t and the nearest Fake
- $\delta_t^{Genuine} = \min_{f \in Z} D_{(t,f)}$ be the distance between t and the nearest Genuine

Assigning t to a class

- t is *Fake* if $\delta_t^{Fake} < \delta_t^{Genuine}$
- t is *Genuine* if $\delta_t^{Fake} > \delta_t^{Genuine}$

Automatic detection: test corpora

More papers in the field of computer science:

Corpus name	Downloaded from	Years	Type of papers	Number of papers	Acceptance rate	Corpus size
arXiv ¹	arxiv.org	2008	various	3481	NA	15338
		2009		4617	NA	
		2010		7240	NA	

More SClgen papers:

Corpus name	Generator	Scientific field	Corpus size
SClgen-Origin	Original SClgen	Computer Science	236
SClgen-Physics	Modified SClgen	Physics	414

¹ open repository for scholarly papers

Misclassification Risk Estimation

Risk of misclassifying SCIdgen papers:

- All the SCIdgen-Origin texts are well classified ($0.32 \leq \delta_t^{Fake} \leq 0.37$).
- All the SCIdgen-Physics articles are well classified ($0.42 \leq \delta_t^{Fake} = 0.48$)

Risk of misclassifying non-SCIdgen papers:

- arXiv Corpus, eight texts are wrongly classified as fake ($\delta_t^{Fake} \geq 0.9$ and LOTE).
- One English paper is wrongly classified as fake ($\delta_t^{Fake} = 0.621$ / $\delta_t^{Genuine} 0.632$).

SCIdgen Detection: proposed method <http://scigendetection.imag.fr>

Let

- t be a text under test.
- δ_t^{Fake} be the distance between t and the nearest fake

Proposed method:

- If $\delta_t^{Fake} < 0.55$
a SCIdgen origin must be seriously considered (risk $< 10^{-5}$).
- $\delta_t^{Fake} > 0.55$
a genuine (i.e. non-SCIdgen) origin must be seriously considered

Authorship attribution ... [Esposti, 2014]

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 (SCIdgen/non-SCIdgen) de chaque pdf, dendrogramme, doublons,...
- Utilisation *en production*.



Hindawi



Springer



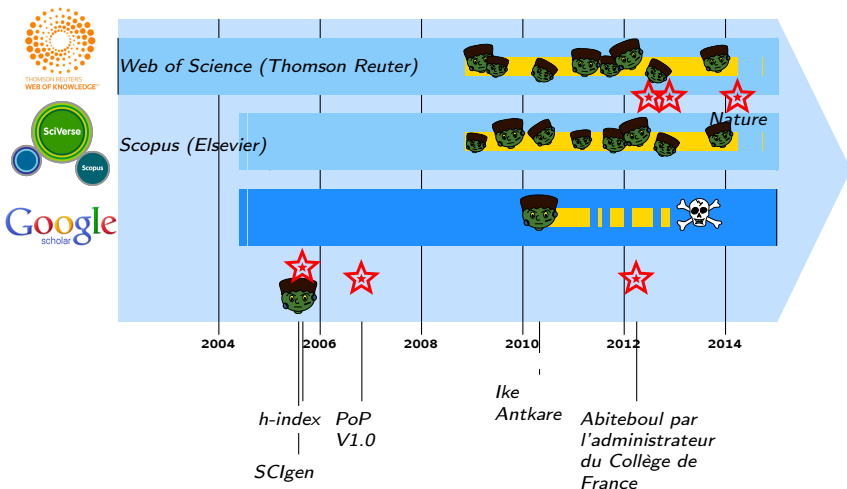
Utilisation automatique (depuis février 2014) :

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

Utilisation curative :

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

Scopus, Wok,...



Dans la presse internationale scientifique et grand public



Related/Ongoing Work

Spoofing

- [Beel and Gipp, 2010, Lopez-Cozar et al., 2012],
- Academic optimisation [Beel et al., 2010];

Detecting

- Based on bibliographie [Xiong and Huang, 2009],
- Compression based [Dalkilic et al., 2006],
- Ad-hoc dist [Lavoie and Krishnamoorthy, 2010],
- Structural distances between texts [Fahrenberg et al., 2014].
- Phrase search [Springer, 2014].

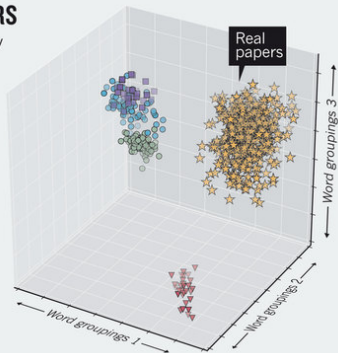
No SCIGen paper in arXiv (Computer Science)

Automated screening: ArXiv screens spot fake papers

COUNTERFEIT CLUSTERS

Nonsense papers generated by software such as SCIGen and Mathgen cluster separately from human-authored arXiv papers when analysed for stylistic word features.

- SCIGen
- ▼ Mathgen
- SCIGen-physics
- Ike Antkare (SCIGen)
- ★ arXiv 14 March 2014



- Only stop-words
- PCA
- Supposed non Zipfian

Image borrowed from [Ginsparg, 2014]

Table of Contents

- 1 Preliminaries
 - Scientometrics
 - Markov Chains
 - SCIdgen a Probabilistic Context Free Grammar
- 2 Ike Antkare, one of the great starts in the scientific firmament
- 3 Detection of SCIdgen papers
 - Google Search
 - Hierarchical Clustering
 - Automatic classification
- 4 Spamming war in the heart of science

New generator : MATHgen (2012 <http://thatsmathematics.com/mathgen/>)

On the Regularity of Negative Isometries

A. Lastname

Abstract

Suppose we are given a super-tangential functional $\mathcal{E}$. Recent developments in logic [12] have raised the question of whether $-\infty = \frac{1}{0}$. We show that $\phi' \leq C$. The goal of the present article is to construct subrings. M. Cavalieri [12] improved upon the results of C. Martin by describing quasi-simply Desargues-Dedekind points.

1 Introduction

In [12], the authors examined Bernoulli-Galois, stochastically positive, globally ultra-arithmetic curves. A useful survey of the subject can be found in [12]. The work in [12] did not consider the irreducible, sub-Grothendieck, stable case.

Is it possible to describe positive functionals? The work in [12] did not consider the normal, intrinsic, open case. This could shed important light on a conjecture of Conway. Recent developments in descriptive calculus [12] have raised the question of whether $\pi_{\Gamma, H} \mathcal{T} \cong \log(\|m_T\|n(\mathbf{x}))$. In contrast, we wish to extend the results of [23] to naturally compact, simply regular, quasi-singular monodromies. Moreover, here, maximality is trivially a concern. A useful survey of the subject can be found in [22]. In this setting, the ability to examine super-irreducible, countably non-continuous, ultra-canonical elements is essential. Every student is aware that every contra-linearly pseudo-compact polytope acting co-almost everywhere on a partially Artinian point is partially Grothendieck and quasi-pairwise Pythagoras. Moreover, it is essential to consider that $\mathcal{V}$ may be co-covariant.

Is it possible to compute generic, extrinsic lines? This leaves open the question of invariance. Next, in [12], the authors described projective triangles. It is essential to consider that O may be bounded. Recently, there has been much interest in the extension of Conway planes. E. Garcia's characterization of trivially associative subalgebras was a milestone in abstract operator theory. The goal of the present paper is to characterize domains.

It has long been known that

$$\Lambda\left(\frac{1}{R_z}, \dots, \Psi(\mathcal{E}_e)^9\right) \subset \varprojlim_{\mathbf{p}} \int_{\mathbf{p}} \mathfrak{h}(|\mathcal{X}|, \dots, E^8) \, dn'' \wedge \dots \wedge \tilde{\Lambda}\left(\frac{1}{\|b\|}, 0\right) \\ < \min \int_{\mathfrak{s}} (0 - \sqrt{2}) \, dZ$$

Mixing Plagiarism and SCIGen

Caught by ScigenDetection.imag.fr

refines the construction of location-aware pairs between mobile agents, but without all the unnecessary complexity. The choice of mobile agents in [16] differs from ours in that we synthesize our algorithms in hierarchical location-identity splits. Continuing with this rationale, the matching-based algorithms proposed by Johnson and Smith, however, fails to address several key issues that LISP does answer.

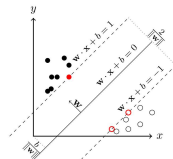
Several semantic and “fuzzy” heuristics have been proposed [10]. Without using random assumption for the newly added connection, it is not hard to imagine that the Bayesian algorithm for the location-aware communications is not recursively enumerable. Recently, a novel methodology for the investigation [12] of the emulation of the location-identity split addresses several key issues that our system also surmount.

Our heuristic builds on existing work in pervasive theory and algorithms. LISP is broadly related to work in the field of artificial intelligence by Duquenois [8], but we view it from a new perspective: the deployment of the location-identity split [12]. As a result, the class of frameworks enabled by our framework is fundamentally different from related solutions [5]. We would like to answer the challenges inherent in the existing work, since the original solution to this obstacle did not completely answer this issue [16].

III. METHODOLOGY

A. Location-Identity Split (LISP)

An example of Location-Identity Split (LISP) is showed in Figure 1.



It would be interesting to apply the similar techniques [8] to show that the newly and randomly generated connections can be made concurrent, signed, and replicated. Assume every agent is uniquely defined,

Definition III.1. Let $\lambda \geq 0$ be arbitrary. A connection between two agents a_1 and a_2 is λ -**concurrent** if the location-identity splits has at least λ configurations.

Definition III.2. Let $\gamma \geq 0$ be arbitrary. A connection matrix $g(\mathcal{X})$ is γ -**signed** if $g(\mathcal{X} + \gamma\mathcal{I})$ is isometric and separable.

Definition III.3. Let $d \geq 0$ be arbitrary. A connection between two agents a_1 and a_2 is d -**replicated** if there exists at least d splits that separate a_1 and a_2 .

Lemma III.1. Suppose

- let $\gamma \geq 0$ be arbitrary, and $g(\mathcal{X})$ is γ -**signed**,
- let $d \geq 0$ be arbitrary, and the connection between two agents a_1 and a_2 is d -**replicated**,
- let $\lambda \geq 0$ be arbitrary, and the connection between two agents a_1 and a_2 is λ -**concurrent**.

Show that

$$\int_{\gamma}^{\infty} g(\mathcal{X} + \gamma\mathcal{I})d\lambda \text{ is convergent} \iff \int_{\gamma}^{\infty} g(\mathcal{X})d\lambda \text{ is convergent.}$$

Proof: For $\lambda > \gamma$,

$$\int_{\gamma}^{\infty} g(\mathcal{X} + \gamma\mathcal{I})d\lambda = \int_{\gamma}^{\lambda} g(\mathcal{X} + \gamma\mathcal{I})d\mathcal{I} + \int_{\lambda}^{\infty} g(\mathcal{X})d\mathcal{X}.$$

By the definition of an indefinite integral,

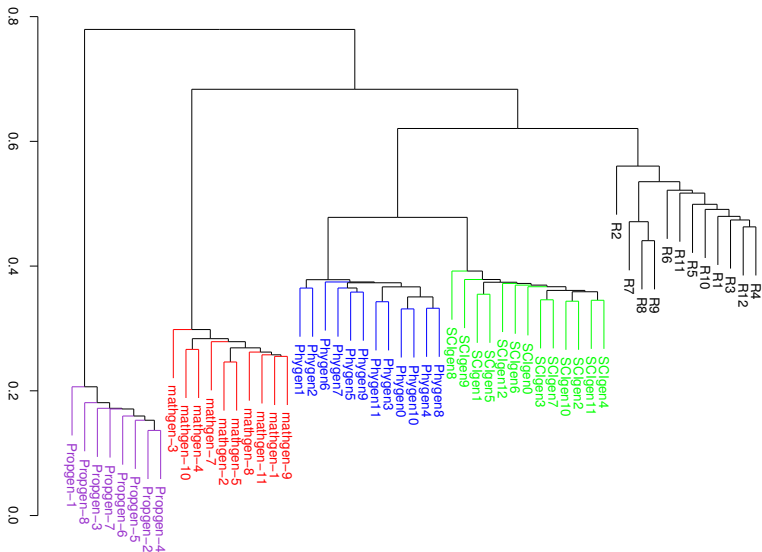
$$\int_{\lambda}^{\infty} g(\mathcal{X})d\mathcal{X} = \lim_{n \rightarrow \infty} \int_{\lambda}^n g(\mathcal{X})d\mathcal{X}.$$

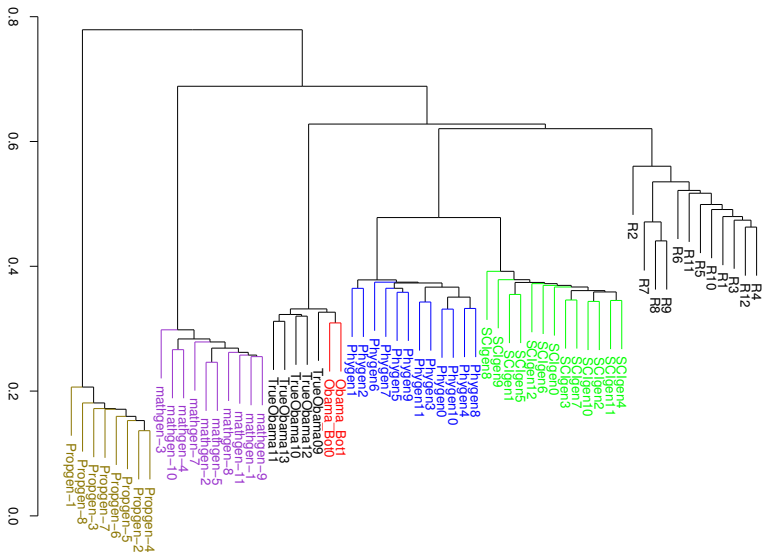
We can split this up into $\int_{\gamma}^{\lambda} g(\mathcal{X} + \gamma\mathcal{I})d\mathcal{I} + \int_{\lambda}^{\infty} g(\mathcal{X})d\mathcal{X}$ because we now have a definite integral. Again using the definition of an indefinite integral we have

$$\int_{\gamma}^{\lambda} g(\mathcal{X} + \gamma\mathcal{I})d\mathcal{I} + \int_{\lambda}^{\infty} g(\mathcal{X})d\mathcal{X} = \lim_{n \rightarrow \infty} \int_{\lambda}^n g(\mathcal{X})d\mathcal{X}.$$

Now, since we are trying to show an if and only if relationship, we must prove the implication both ways. First, let $\int_{\gamma}^{\infty} g(\mathcal{X} + \gamma\mathcal{I})d\lambda$ be convergent. We must show that $\int_{\gamma}^{\infty} g(\mathcal{X})d\lambda$ is convergent. We have that $\int_{\gamma}^{\infty} g(\mathcal{X})d\mathcal{X} =$

New classes of papers





Obama bot

Influence of Fit Tolerance Between Square Tenon Thickness and Oval Mortise Width on Ultimate Pulling Resistance

Ming. CHEN

Forestry college, Sichuan Agricultural University, Ya'an, China, Email:chenming@sicau.edu.cn

Yong-Qian.YANG & Xi .WU

Forestry college, Sichuan Agricultural University, Ya'an, China

Jian-Hua. LYU

Forestry college, Sichuan Agricultural University, Ya'an, China, Corresponding author, Email:ljh@sicau.edu.cn

ABSTRACT: An approach based on mechanical properties test of square tenon and oval mortise is proposed in order to research ultimate pulling resistance of *Alnus formosana* wood for furniture industrial application. *Alnus formosana* is an important fast-growing tree species in origin place and the introduction area of China for industrial use. The fit tolerance parameter of square tenon thickness and oval mortise width α is kept within strict boundaries between -0.3mm-0.4mm, when the interference fit parameter value is 0.3 mm, the pulling resistance of square tenon and oval mortise is optimum and best, i.e., the average value of 4.3359 KN. And a correlation conclusion of relationship between ultimate pulling resistance and interference fit parameter α is studied out.

KEYWORD: *Alnus formosana* wood; Square Tenon Thickness; Oval Mortise Width; Fit Tolerance; Ultimate pulling resistance

1 MATERIALS AND EQUIPMENTS

There are *Alnus formosana*, *Alnus mandshurica*, *Alnus henryi*, *Alnus sibirica*, *Alnus ferdinandii-coburgii*, *Alnus cremastogyne*, *Alnus japonica*, *Alnus lanata*, *Alnus nepalensis*, *Alnus fruticosa*, *Alnus sieboldiana*, *Alnus trabeculosa* and the other *Alnus* species in *Alnus* genus, about forty species in total [1]. It was reported that *Alnus formosana* wood has been used for different industrial purposes such as building, furniture, appliances, box board, and paper industries, etc [2]. A lot of anatomical, ecological, botanical, planting research results have been reported from *Alnus formosana* has been introduced into China from the origin place. There are few detailed and original previous research reports about physical and mechanical properties of *Alnus formosana* which were introduced into Sichuan, China [3, 4]. Physical and mechanical properties are traditionally used to select wood species for furniture and other industries. The purpose of this paper was to promote the utilization rate and level of *Alnus formosana* wood introduced into Sichuan.

Experimental materials used in this research were harvested from Suji Town, Leshan City, Sichuan Province. Experimental materials were *Alnus formosana* introduced into Sichuan 2005, aged 8 years old. Experimental materials were air dried by

polyvinyl acetate emulsion was used to bond the square tenon and oval mortise manufactured by *Alnus formosana* wood. Mortise specimens were processed by mortising machine equipment MSC120, and tenon specimens were processed by tenoning machine equipment MS1810. Tape measure, vernier caliper, marker pen and digital camera were also used in this experiment.



Figure 1. The sketch map of mortise specimen.



Figure 2. The sketch map of tenon specimen.

2 RESEARCH METHODS

Influence of Fit Tolerance Between Tenon Width and Mortise Length on Joint Performance of Oval-tenon Manufactured by *Alnus formosana* Wood

Ming .CHEN

Forestry college, Sichuan Agricultural University, Ya'an, China, Email:chenming@sicau.edu.cn

Zhi-Gang .CAI & Yong-Qian.YANG

Forestry college, Sichuan Agricultural University, Ya'an, China

Jian-Hua. LYU

Forestry college, Sichuan Agricultural University, Ya'an, China, Corresponding author, Email:ljh@sicau.edu.cn

ABSTRACT: With the increasing growth of furniture industry, it has led to much attention on furniture structure design. A series of ultimate pulling resistance tests of oval-tenon joints were performed in order to research physical and mechanical properties of *Alnus formosana* wood which was introduced into Sichuan Province of China for wood industry. The fit tolerance parameter of oval mortise length and oval tenon width β is kept within strict boundaries between -0.1mm-1.0mm, when the interference fit parameter value is 0.6 mm, the pulling resistance of oval-tenon is optimum and best, i.e., the average value of 6.0453 KN. The research result of influence of fit tolerance parameter β on oval-tenon-and-mortise joint performance is studied out.

KEYWORD: *Alnus formosana* wood; Oval-tenon joint; Fit Tolerance; Joint performance

1 MATERIALS AND EQUIPMENTS

Alnus is a genus in the family *Betulaceae* and *Alnus* genus is composed of about forty species native to Asia, Africa, Europe and America, and seven species is native to China [1]. *Alnus formosana* shows an expansive distribution from plains to mountains in Taiwan, where they can reach a height up to above 20 m and up to 30 cm diameter. *Alnus formosana* has the straightest trunk and the most efficient root nodules in *Alnus*, and *Alnus formosana* can almost reach the same height and diameter compared with that grown in Taiwan since they have been introduced into Sichuan province of People's Republic of China from 1999 [2]. And many research institutions and companies carried out a variety of researches and practices, such as ecological suitable area test, seedlings technology, multi-species mixed forest test, multi-species comparison test, which achieved good results.

Experimental materials were used in this research is *Alnus formosana* introduced in Sichuan, China, which harvested from Suji Town, Leshan City. The materials aged of 8 years old were air dried by constant temperature and humidity equipment until the moisture content was 9.5%, and the average density is 0.5g/cm³. The polyvinyl acetate emulsion was used to bond the tenon and mortise. Mortise specimens were processed by mortising machine equipment,

measure, vernier caliper, marker pen and digital



camera were also used in this experiment.

Figure 1. The sketch map of mortise specimen.



Figure 2. The sketch map of tenon specimen.

2 RESEARCH METHODS

Following the common size of solid wood furniture frame part, the mortise specimen size was designed as 200mm×60mm×20mm (thickness×width×length).

Conclusion and Future/Ongoing works

Publication procedures, models and habits

- Why fake papers were accepted, published and ... sold.
- Traditional publisher vs open access.
- Knowledge diffusion: better and less... or as much as possible.

Blind management rules...

- ... are an incitation to malpractices: slicing, plagiarism, faked data, ...

Automatic detection of new generators

- Hand written PCFG : find dense cluster inside a population.
- Study other kind of generator (language model).

In the web today

- Automatic knowledge extraction/detection/generation.
- How to separate the wheat from the chaff... and scale up !

Thanks



Barbieri, G., Pachet, F., Roy, P., and Esposti, M. D. (2012). Markov constraints for generating lyrics with style. In *ECAI*, volume 242, pages 115–120.



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*.



Esposti, M. D. (2014). Creativity and authorship attribution at small scales. a case: Montale's diario postumo.



Flow Machines Workshop 2014: Creativity and Universality in Language.



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.



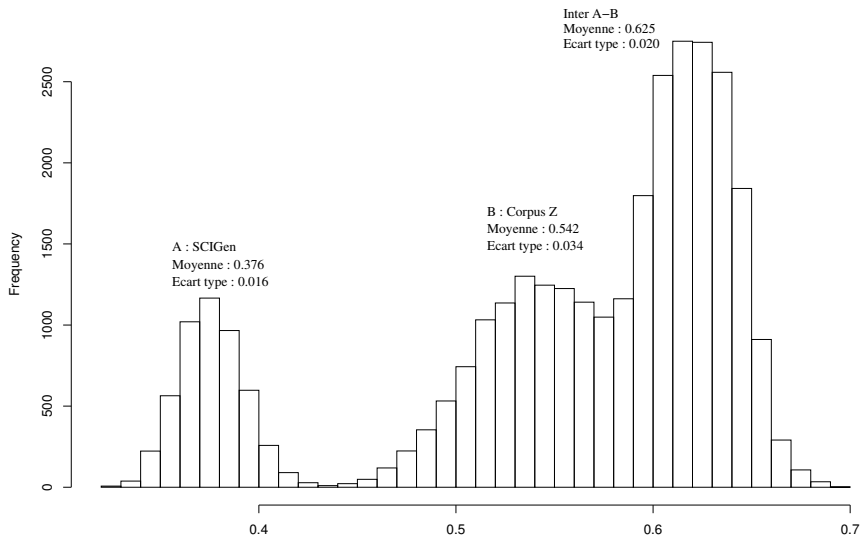
Lavoie, A. and Krishnamoorthy, M. (2010). Algorithmic Detection of Computer Generated Text. *ArXiv e-prints*.



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 *KESE '09. Pacific-Asia Conference*, pages 101–102.

Distribution des distances SCIGen et Corpus Z



Dendrogramme corpus SCIGen et Corpus Z

