

BibTeX Bibliography Index Maker: Notes

Ramon Xuriguera

1 BibTeX

Aspectes del format BibTeX a tenir en compte:

- Com podem distingir entre diferents tipus d'entrada (article, book, inproceedings, etc.) a partir del fitxer?
- Format dels noms. Un nom consisteix de diferents parts: First, von, Last, Jr. El token *von* o *de la* cal posar-los en minúscules. Per tal que el nom es reconegui, cal que tingui el format: von Last, Jr, First. D'aquesta manera, si hi ha més d'un cognom no passa res.
- Caràcters Unicode entre claus per poder ser utilitzats correctament amb l'estil *alpha*. Per exemple: `Jos{\'\{e}}`
- Per prevenir que BibTeX canviï un text a minúscules, cal posar el text entre claus.
- Si hi ha massa autors, truncar la llista amb *et al.*
- Utilitzar abreviatures de tres lletres per als mesos
- Utilitzar el camp **key** per a organitzacions amb un nom llarg, de manera que s'utilitzin les inicials de l'organització al fer una cita.

2 Extracció del contingut dels fitxers PDF

2.1 Software

xPDF proporciona eines executables des de la línia de comandes per extreure text i altres elements dels fitxers PDF. Es distribueixen binaris de la utilitat tant per Windows com per Linux (que també funcionen per MAC OS). El principal motiu pel qual hem escollit xPDF és la qualitat dels resultats, en especial, el fet que no separa els paràgrafs en diferents línies i que obté el text segons l'ordre de lectura i no l'ordre en que es troben en el document (e.g. dues columnes). També serà útil la possibilitat d'extreure les metadades del fitxer de forma fàcil.

Altres opcions que s'han tingut en compte:

- PyPDF
- PDFMiner
- PDFBox

2.2 Alguns punts a considerar

A continuació es comenten alguns punts sobre els quals cal prendre decisions:

- Caràcters Unicode:
Tan el programari seleccionat com la resta d'alternatives no permeten extreure de forma correcta els caràcters unicode. En la majoria dels casos, els articles estaran publicats en anglès i només trobarem caràcters especials en els noms dels autors, universitats, etc.
 - Lletres amb signes diacrítics: Si aconseguim obtenir la lletra, no hi ha problema perquè els cercadors permeten buscar amb l'equivalent ASCII. Altrament, caldrà idear alguna manera de buscar. Idees:
 - * Cerques inexactes amb caràcters comodí: Google permet utilitzar el caràcter *, però Google Scholar no.
 - * Google Fight: fer cerques per diferents possibilitats dels caràcters que ens manquen i decidir segons els resultats.
 - Altres caràcters Unicode: Els podem ometre ja que només estem interessats en cercar text.
- Text en negreta:
Si realment ens interessa obtenir el text que en el fitxer es troba en negreta, podem utilitzar l'eina `pdftohtml`, que generarà etiquetes ``. A priori sembla que no ho necessitem.
- Símbols de puntuació:
Els podem ometre ja que les cerques exactes de Google Scholar retorna els mateixos resultats tant si s'inclouen a la consulta com no. Per exemple, la consulta `cluster or NoW(Network of Workstations)`. retorna els mateixos resultats que `cluster or NoW Network of Workstations`.

2.3 Procediment

1. L'usuari indica un directori
2. L'aplicació obté la llista de tots els documents del directori (i subdirectoris)
3. Per cada document:
 - (a) Extreu les metadades del fitxer, en cas que en tingui
 - (b) Extreu el contingut en forma de text o HTML
 - (c) Utilitza la informació per cercar la referència Bibtex a Internet
 - (d) Executa una sèrie de tests (a establir) per comprovar la correctesa de les dades obtingudes
 - Si es passen els tests, s'afegeix la referència al fitxer BibTex o a la base de dades de JabRef
 - En cas de dubte, indica a l'usuari que hauria de revisar la referència
 - En cas de no poder obtenir cap tipus d'informació, n'informa a l'usuari

2.4 Idees

Algunes idees sobre alguns dels problemes:

- Caràcters Unicode:
El software que ens permet extreure el contingut dels fitxers PDF no treballa bé amb els caràcters Unicode. Com que la majoria de cercadors permeten cercar amb ASCII, podem ometre i obtenir les dades correctes d'Internet.
- Reconeixement de les parts d'un document:
Com es pot veure a l'apèndix A, no hi ha gaires elements en comú entre les capçaleres dels documents. L'element que sí que es repeteix en gairebé tots ells és l'*abstract*. La solució proposada és utilitzar part d'aquest resum per tal de cercar a quin article correspon cada fitxer. Agafant un número prou elevat de paraules consecutives del resum (a la pràctica, unes 7-8), els motors de cerca limiten la cerca a només resultats sobre l'article. Exemples de cerques: *"we discuss the use of boundary methods"*, *critical juncture with regard to HPC* o *Consider a strongly connected directed weighted*

3 Obtenció de referències

DBLP++

DBLP++ proporciona un servei web que amplia la funcionalitat de l'API de DBLP. Permet cercar per paraules clau. DBLP++ ofereix el fitxer WSDL necessari per generar les classes en Java o Python.

Portal ACM

Es poden obtenir construint les URLs adequades.

CiteSeerX

OAIHarvester en Java

Arxiv.org

<http://arxiv.org/help/api/index>

3.1 Algunes opcions

- Utilitzar un cercador web (o *Google Scholar*):
Similar a cercar a totes les bases de dades a la vegada. El problema passa a ser com obtenir les referències de la multitud de pàgines diferents (veure apèndix B), en les quals hi pot haver el codi Bibtex entre les etiquetes HTML, o bé algun enllaç o acció Javascript (la cosa es complica) que ens hi porti.
- APIs o serveis web de les bases de dades més importants:
Es tractaria de tenir una sèrie de classes implementant una interfície. Per cada nova base de dades, caldria
- Reaprofitar els *imports* de JabRef:
Aquesta opció sorgeix una mica de la idea anterior, però no està exempta de problemes. Per què no aprofitem els *imports* ja desenvolupats a Jabref? A mesura que se'n van afegint, el nostre plug-in els podria anar utilitzant.
Punts problemàtics: no totes les bases de dades permeten buscar a partir del resum dels articles, ens hem d'assegurar que un plug-in pot accedir a les classes d'un altre.

4 JabRef

Java, llicència: LGPL.

És possible crear plug-ins amb *Java Plug-in Framework*. L'última versió d'aquest framework és de fa més de dos anys (pre OSGi), actualment es troba en estat *frozen*. En el fòrum i el tracker de *JabRef* no hi ha cap missatge en el que es discuteixin canvis sobre aquest tema.

Extension-points permesos:

- **ImportFormat** Add importers to JabRef accessible from the 'Import into ... database'.
- **EntryFetcher** Add access to databases like Citeseer or Medline to the Web Search menu.
- **ExportFormatTemplate** Add a template based export like the ones accessible using the Manage Custom Exports.
- **ExportFormat** Add an export filter to JabRef's export dialog, that is more complicated than the simple template based one.
- **ExportFormatProvider** A more powerful way to add export formats to JabRef.
- **LayoutFormatter** Add formatters that can be used in the layout based exporters.
- **SidePanePlugin** Add a side pane component that can do any kinds of operations. The panel is accessed from a Plugins menu in JabRef's main window.

5 Llenguatge

La idea seria desenvolupar una aplicació de línia de comandes en *Jython*, (Python sobre la JVM). Pel que fa al plug-in de JabRef, si és possible també podria ser interessant utilitzar Jython.

El toolkit NLTK pot ser útil a l'hora de tractar els textos amb Python. Actualment hi ha alguns problemes per executar-lo sobre Jython, però sembla ser que hi ha *workarounds*.

A Exemples de text extret

A continuació es mostren algunes capçaleres d'articles i el resultat obtingut a l'extreure'n el text:

- **Text 1**

- PDF:

- Text:

arXiv:0912.0725v1 [astro-ph.CO] 3 Dec 2009

Lorentz symmetry violation , dark matter and dark energy

Luis Gonzalez-Mestresa

a

LAPP, Universit de Savoie , CNRS/IN2P3, B.P. 110, 74941 Annecy
-le-Vieux Cedex, France e

Lorentz symmetry violation, dark matter and dark energy

Luis Gonzalez-Mestres^a

^aLAPP, Université de Savoie, CNRS/IN2P3, B.P. 110, 74941 Annecy-le-Vieux Cedex, France

Taking into account the experimental results of the HiRes and AUGER collaborations, the present status of bounds on Lorentz symmetry violation (LSV) patterns is discussed. Although significant constraints will emerge, a wide range of models and values of parameters will still be left open. Cosmological implications of allowed LSV patterns are discussed focusing on the origin of our Universe, the cosmological constant, dark matter and dark energy. Superbradyons (superluminal preons) may be the actual constituents of vacuum and of standard particles, and form equally a cosmological sea leading to new forms of dark matter and dark energy.

1. Patterns of Lorentz symmetry violation particle. For $p \gg mc$, one has:

Taking into account the experimental results of the HiRes and AUGER collaborations, the present status of bounds on Lorentz symmetry violation (LSV) patterns is discussed. Although significant constraints will emerge, a wide range of models and values of parameters will still be left open. Cosmological implications of allowed LSV patterns are discussed focusing on the origin of our Universe, the cosmological constant, dark matter and dark energy. Superbradyons (superluminal preons) may be the actual constituents of vacuum and of standard particles, and form equally a cosmological sea leading to new forms of dark matter and dark energy.

1. Patterns of Lorentz symmetry violation A formulation of Planck-scale Lorentz symmetry violation (LSV) testable in ultra-high energy cosmic-ray (UHECR)

– HTML:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN
" ><HTML>
<HEAD>
<TITLE></TITLE>
</HEAD>
<BODY>
<A name=1></a>Lorentz symmetry violation, dark matter and
dark energy<br>
Luis Gonzalez-Mestres<br>
aLAPP, Université de Savoie, CNRS/IN2P3, B.P. 110, 74941
Annecy-le-Vieux Cedex, France<br>
Taking into account the experimental results of the HiRes and
AUGER collaborations, the present status of<br>
bounds on Lorentz symmetry violation (LSV) patterns is
discussed. Although significant constraints will emerge,<br>
<br>a wide range of models and values of parameters will
still be left open. Cosmological implications of allowed<
```

LSV patterns are discussed focusing on the origin of our Universe, the cosmological constant, dark matter and dark energy. Superbradyons (superluminal preons) may be the actual constituents of vacuum and of standard particles, and form equally a cosmological sea leading to new forms of dark matter and dark energy.

1. Patterns of Lorentz symmetry violation

• Text 2

– PDF:



Available online at www.sciencedirect.com



Journal of Computer and System Sciences 74 (2008) 775–795



www.elsevier.com/locate/jcss

Compact roundtrip routing with topology-independent node names

Marta Arias^{a,1}, Lenore J. Cowen^{b,2}, Kofi A. Laing^{b,*,3}

^a Center for Computational Learning Systems, Columbia University, New York, NY 10115, USA

^b Department of Computer Science, Tufts University, Medford, MA 02155, USA

Received 9 November 2004; received in revised form 24 January 2007

Available online 14 September 2007

Abstract

Consider a strongly connected directed weighted network with n nodes. This paper presents compact roundtrip routing schemes with $\tilde{O}(\sqrt{n})$ sized local tables⁴ and stretch 6 for any strongly connected directed network with arbitrary edge weights. A scheme with local tables of size $\tilde{O}(\epsilon^{-1}n^{2/k})$ and stretch $\min((2^{k/2}-1)(k+\epsilon), 8k^2+4k-4)$, for any $\epsilon > 0$ is also presented in the case where edge weights are restricted to be polynomially-sized. Both results are for the topology-independent node-name model.

– Text:

Journal of Computer and System Sciences 74 (2008) 775–795 www.elsevier.com/locate/jcss

Compact roundtrip routing with topology-independent node names

Marta Arias^{a,1}, Lenore J. Cowen^{b,2}, Kofi A. Laing^{b,*,3}

^a Center for Computational Learning Systems, Columbia University, New York, NY 10115, USA ^b Department of Computer Science, Tufts University, Medford, MA 02155, USA

Received 9 November 2004; received in revised form 24 January 2007 Available online 14 September 2007

Abstract Consider a strongly connected directed weighted network with n nodes. This paper presents compact roundtrip routing schemes with $O(\sqrt{n})$ sized local tables⁴

and stretch 6 for any strongly connected directed network with arbitrary edge weights. A scheme with local tables of size $O(-1 n^2/k)$ and stretch $\min((2k/2 - 1)(k +), 8k^2 + 4k - 4)$, for any > 0 is also presented in the case where edge weights are restricted to be polynomially-sized

– HTML:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN
">
<HTML>
<HEAD>
<TITLE>doi:10.1016/j.jcss.2007.09.001</TITLE>
<META http-equiv="Content-Type" content="text/html; charset=
ISO-8859-1">
<META name="generator" content="pdftohtml 0.36">
<META name="date" content="2008-05-20T09:39:49+00:00">
</HEAD>
<BODY vlink="blue" link="blue">
<A name=1</a>Journal of Computer and System Sciences 74
(2008) 775-795<br>
www.elsevier.com/locate/jcss<br>
Compact roundtrip routing with topology-independent node
names<br>
Marta Arias a,1, Lenore J. Cowen b,2, Kofi A. Laing b,,3<br>
a <i>Center for Computational Learning Systems, Columbia
University, New York, NY 10115, USA</i><br>
b <i>Department of Computer Science, Tufts University,
Medford, MA 02155, USA</i><br>
Received 9 November 2004; received in revised form 24 January
2007<br>
Available online 14 September 2007<br>
<b>Abstract</b><br>
Consider a strongly connected directed weighted network with
n nodes. This paper presents compact roundtrip routing
schemes<br>
<br>
with  $\tilde{O}(n)$  sized local tables<br>
and stretch 6 for any strongly
connected directed network with arbitrary edge weights. A
scheme<br>
with local tables of size  $\tilde{O}(n^2/k)$ 
and stretch  $\min((2k/2 - 1)(k + ), 8k^2 + 4k - 4)$ ,
for any  $> 0$  is also presented in the<br>
```

- Text 3

– PDF:

– Text:

Enhancing Prediction on Non-dedicated Clusters^{*}

Joseph Ll. L rida¹, F. Solsona¹, F. Gin ¹, J.R. Garc a²,
M. Hanzich², and P. Hern ndez²

¹ Departamento de Inform tica e Ingenier a Industrial, Universitat de Lleida, Spain
{jlerida,francesc,sisco}@diei.udl.cat

² Departamento de Arquitectura y Sistemas Operativos,
Universitat Aut noma de Barcelona, Spain
{jrgarcia,mauricio,porfidio.hernandez}@aomail.uab.es

Abstract. In this paper, we present a scheduling scheme to estimate the turnaround time of parallel jobs on a heterogeneous and non-dedicated cluster or NoW(Network of Workstations). This scheme is based on an analytical prediction model that establishes the processing and communication slowdown of the execution times of the jobs based on the cluster nodes and links powerful and occupancy. Preservation of the local application responsiveness is also a goal.

Enhancing Prediction on Non-dedicated Clusters

Joseph Ll. L rida¹, F. Solsona¹, F. Gin ¹, J.R. Garc a², e
e i a M. Hanzich², and P. Hern ndez²

¹

Departamento de Inform tica e Ingeniera Industrial,
Universitat de Lleida, Spain a i {jlerida,francesc,sisco}
@diei.udl.cat ² Departamento de Arquitectura y Sistemas
Operativos, Universitat Aut noma de Barcelona, Spain o {
jrgarcia,mauricio,porfidio.hernandez}@aomail.uab.es

Abstract. In this paper, we present a scheduling scheme to estimate the turnaround time of parallel jobs on a heterogeneous and non-dedicated cluster or NoW(Network of Workstations). This scheme is based on an analytical prediction model that establishes the processing and communication slowdown of the execution times of the jobs based on the cluster nodes and links powerful and occupancy.

– HTML:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN
">
<HTML>
<HEAD>
<TITLE>Title</TITLE>
<META http-equiv="Content-Type" content="text/html; charset=
ISO-8859-1">
<META name="generator" content="pdftohtml 0.36">
```


<META name="author" content="Author">
 <META name="keywords" content="">
 <META name="date" content="2008-08-19T14:04:20+00:00">
 <META name="subject" content="Subject">
 </HEAD>
 <BODY bgcolor="#A0A0A0" vlink="blue" link="blue">
 <A name=1Enhancing Prediction on Non-dedicated
 Clusters

 Joseph Ll. L´erida1, F. Solsona1, F. Gin´e1, J.R. Garc´ia2,

 M. Hanzich2, and P. Hern´andez2

 1 Departamento de Inform´atica e Ingenier´ia Industrial ,
 Universitat de Lleida , Spain

 {jlerida ,francesc ,sisco}@diei.udl.cat

 2 Departamento de Arquitectura y Sistemas Operativos ,

 Universitat Aut´onoma de Barcelona , Spain


 {jrgarcia ,mauricio ,porfidio.hernandez}@aomail.uab.es

 Abstract. In this paper, we present a scheduling
 scheme to estimate the
turnaround time of parallel jobs
 on a heterogeneous and non-dedicated cluster
or NoW(
 Network of Workstations). This scheme is based on an
 analytical pre-
diction model that establishes the
 processing and communication slowdown of
the execution
 times of the jobs based on the cluster nodes and links
 powerful and
occupancy.

B Exemples de les diferents bases de dades on-line

The screenshot shows the CiteSeerX website interface. At the top, there is a navigation bar with links: Home, Statistics, About, Bulletin, Submit Documents, Feedback, MetaCart, and Sign in to MyCiteSeerX. Below this, the CiteSeerX logo is displayed. A search bar is located on the right side of the header. The main content area features a document titled "Feature Set Evaluation and Robust Neural Networks using Boundary Methods" by J.L. Sancho, William E. Pierson, Batu Ulug, Stanley C. Ahalt, and A. R. Figueiras-vidal. The document is categorized under "Feature Set Evaluation and Robust Neural Networks using Boundary Methods". There are buttons for "Summary", "Related Documents", and "Version History". A "DOWNLOAD:" link is provided with the URL "http://er4www.eng.ohio-state.edu/ips/Publications/". A "CACHED:" link is also present. Below the document title, there are buttons for "Add to Collection", "Correct Errors", and "Monitor Changes". The "Abstract:" section contains the following text: "In this paper we discuss the use of Boundary Methods (BM) for distribution analysis. We view these methods as tools which can be used to extract useful information from sample distributions. We believe that the information thus extracted has utility for a number of applications, but in particular we discuss the use of BM as a new mechanism to Feature Set Evaluation (FSE) and as an aid to constructing robust and efficient Neural Networks (NN) to solve classification problems. In the first case, BM can establish which feature set is most appropriate for classification. We demonstrate experimentally that the derived ranking is consistent with alternative ranking techniques based on Bayes error (ffl), showing the theoretical relationship between Overlap Sum (OS), the BM measure of class separability, and ffl. Next, we investigate complexity issues associated with using BMs for FSE and compare with other techniques used for FSE. Finally, BM are used as Sample Seleccion (SS) mechanism to tra...". The "Citations" section lists three references: "2189 Neural networks - a comprehensive foundation - Haykin - 1994", "2046 Learning internal representations by error propagation - Rumelhart, Hinton, et al. - 1986", and "1777 Introduction to statistical pattern recognition (2nd ed) - Fukunaga - 1990". On the right side, there is a "POPULAR TAGS" section with a "Add a tag:" input field and a "Submit" button. Below this, it states "No tags have been applied to this document." There is also a "BIBTEX | ADD TO METACART" section containing a BibTeX entry: "@MISC{Sancho_featureset, author = {J.L. Sancho and William E. Pierson and Batu Ulug and Stanley C. Ahalt and A. R. Figueiras-vidal}, title = {Feature Set Evaluation and Robust Neural Networks using Boundary Methods}, year = {} }".


Figura 1: Referència BibTex a la mateixa pàgina


[Subscribe \(Full Service\)](#) [Register \(Limited Service, Free\)](#) [Login](#)

Search: ☐ The ACM Digital Library ☒ The Guide

[THE GUIDE TO COMPUTING LITERATURE](#) [Feedback](#)

Toward a High Performance Computing Economy

Full text  (19 KB)

Source Conference on High Performance Networking and Computing [archive](#)
 Proceedings of the 2004 ACM/IEEE conference on Supercomputing [table of contents](#)
 Page: 61
 Year of Publication: 2004
 ISBN: 0-7695-2153-3

Author [Stan Ahalt](#) Ohio Supercomputer Center

Sponsor [SIGARCH](#): ACM Special Interest Group on Computer Architecture

Publisher IEEE Computer Society Washington, DC, USA


Bibliometrics Downloads (6 Weeks): 1, Downloads (12 Months): 18, Citation Count: 0

Tools and Actions: [Review this Article](#) Aquest enllaç executa codi Javascript per obrir un pop-up amb la referència BibTeX
[Save this Article to a Binder](#) Display Formats: [BibTeX](#) [EndNote](#) [ACM Ref](#)

DOI Bookmark: [10.1109/SC.2004.59](#)

Collaborative Colleagues:
 Stan Ahalt: [colleagues](#)

Figura 2: Enllaç amb codi Javascript per obtenir la referència

 CERN Document Server

[CDS](#) [Indico](#) [Library](#) [Bulletin](#) [EDMS](#)

[Search](#) [Submit](#) [Help](#) [Your CDS](#) [login](#)

[Home](#) > [Articles & Preprints](#) > [Preprints](#) > Record#1226949: Warm gas phase chemistry as possible origin of high HDO/H₂O ratios in hot and dense gases: application to inner protoplanetary discs

[Information](#) [References](#) [Discussion](#) [Fulltext](#)

Preprint


Report number arXiv:0912.0701


Title Warm gas phase chemistry as possible origin of high HDO/H₂O ratios in hot and dense gases: application to inner protoplanetary discs

Author(s) [Thi, Wing-Fai](#) (Institute for Astronomy, University of Edinburgh, UK) ; [Woitke, Peter](#) (UK Astronomical Technology Centre, Edinburgh) ; [Kamp, Inge](#) (Kapteyn Astronomical Institute, Groningen, The Netherlands)

Abstract Full Text (PDF) ... discriminate between the different origins of water on Earth.

Record created 2009-12-04, last modified 2009-12-05 [Similar records](#)

[external link:](#)  Preprint

Rate this document:

 (Not yet reviewed)

Add to personal basket
 Export as [BibTeX](#), [MARC](#), [MARCXML](#), [DC](#), [EndNote](#), [NLM](#)
Enllaç a una altra pàgina HTML amb el contingut BibTeX

Figura 3: Enllaç a una altra pàgina HTML