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 el text. Només converteix a text pla, però no separa els diferents paràgrafs en línies diferents. Permet obtenir el resultat en diferents codificacions de caràcters, però no conserva el text correcte. És a dir, continua tenint problemes per extreure accents, etc. Amb l'eina `pdftohtml` podem obtenir informació d'algunes de les paraules en negreta, però separa cada línia amb una etiqueta br.
- PDFBox Llibreria escrita en Java i publicada sota la llicència *Apache License v2.0*. Actualment es troba a la incubadora d'Apache. Permet obtenir el text i les metadades d'un fitxer. Separa els resultats per línia segons es troben en el document, no per paràgrafs.

  Podem obtenir les següents dades del fitxer sense haver-ne d'extreure el text: número de pàgines, títol, autor, assumpte i paraules clau. Si aquests camps no s'han omplert al generar el fitxer, estaran en blanc.

2.2 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

2.3 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"*, *çritical juncture with regard to HPC"* o *Çonsider a strongly connected directed weighted"*

3 Obtenció de referències

DBLP++
DBLP++ proporciona un servei web que amplia molt 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, 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 podría 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 (per motius de consistència).

El toolkit NLTK pot ser bastant 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:

Lorentz symmetry violation, dark matter and dark energy

Luis Gonzalez-Mestres[a]

[a]LAPP, 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:

  - 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

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-Mestresa<br>
aLAPP, Universite 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>a wide range of models and values of parameters will
    still be left open. Cosmological implications of allowed<
    br>LSV patterns are discussed focusing on the origin of
    our Universe, the cosmological constant, dark matter and<
    br>dark energy. Superbradyons (superluminal preons) may be
    the actual constituents of vacuum and of standard<br>
    particles, and form equally a cosmological sea leading to
    new forms of dark matter and dark energy.<br>
1. Patterns of Lorentz symmetry violation<br>
```

- **Text 2**
  - PDF:

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

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

JOURNAL OF COMPUTER AND SYSTEM SCIENCES

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 tables[4] 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
<br/>
Abstract Consider a strongly connected directed weighted
    network with n nodes. This paper presents compact
    roundtrip routing schemes  with O( 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 n2/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 ˜<br>
O( n) sized local tables4 and stretch 6 for any strongly
    connected directed network with arbitrary edge weights. A
    scheme<br>
with local tables of size ˜<br>
O( −1n2/k) and stretch min((2k/2 − 1)(k + ), 8k2 + 4k − 4),
    for any<br>
&gt; 0 is also presented in the<br>
```

- **Text 3**
  - PDF:

Enhancing Prediction on Non-dedicated Clusters*

Joseph Ll. Lérida[1], F. Solsona[1], F. Giné[1], J.R. García[2],
M. Hanzich[2], and P. Hernández[2]

[1] Departamento de Informática e Ingeniería Industrial, Universitat de Lleida, Spain
{jlerida,francesc,sisco}@diei.udl.cat
[2] Departamento de Arquitectura y Sistemas Operativos,
Universitat Autònoma 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.

  - Text:

```
Enhancing Prediction on Non−dedicated Clusters
Joseph Ll. L rida1 , F. Solsona1 , F. Gin 1 , J.R. Garca2 , 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=1></a><b>Enhancing Prediction on Non-dedicated
    Clusters</b><br>
Joseph Ll. L´erida1, F. Solsona1, F. Gin´e1, J.R. Garc´ia2,<
    br>
M. Hanzich2, and P. Hern´andez2<br>
1 Departamento de Inform´atica e Ingenier´ia Industrial,
    Universitat de Lleida, Spain<br>
{jlerida,francesc,sisco}@diei.udl.cat<br>
2 Departamento de Arquitectura y Sistemas Operativos,<br>
Universitat Aut‘onoma de Barcelona, Spain<br>
{jrgarcia,mauricio,porfidio.hernandez}@aomail.uab.es<br>
<b>Abstract. </b>In this paper, we present a scheduling
    scheme to estimate the<br>turnaround time of parallel jobs
     on a heterogeneous and non-dedicated cluster<br>or NoW(
    Network of Workstations). This scheme is based on an
    analytical pre-<br>diction model that establishes the
    processing and communication slowdown of<br>the execution
    times of the jobs based on the cluster nodes and links
    powerful and<br>occupancy.
```