

Search Engine for DBLP

- with Whoosh

Authors:

Piccinni Marco

Tortoli Daniele

Obiettivi del progetto

- ◈ Realizzazione di un sistema di ricerca full-text che consenta di effettuare ricerche nella bibliografia di DBLP ordinando i risultati secondo un modello di ranking.
- ◈ Il sistema è quindi responsabile di due aspetti:
 1. Creazione e gestione degli indici a partire dal file XML di DBLP.
 2. Supporto per ricerche full-text in base al linguaggio proposto.

Sintassi del linguaggio di ricerca



Fasi di sviluppo



Analisi e parsing del file XML



Modellazione degli indici



Parsing della query utente



Algoritmi di ricerca



Gestione dei risultati



Analisi XML

Tipi di documenti

Venue

- ◇ Raccolta di pubblicazioni
- ◇ Può essere:
 - ◇ Conferenza (*proceedings*)
 - ◇ Libro (*book*)
 - ◇ Rivista (*journal*)

Pubblication

- ◇ Il documento dello studio pubblicato
- ◇ Può essere:
 - ◇ Articolo (*article*)
 - ◇ *incollection*
 - ◇ Intervento (*inproceedings*)
 - ◇ Tesi di dottorato (*ph. Thesis*)
 - ◇ Tesi di laurea (*master thesis*)



Analisi XML

Riferimenti tra pubblicazioni e raccolte

inproceedings
article



proceedings

```
<inproceedings key="conf/dsd/WurstDHZSCBB19" mdate="2019-10-23">
  <author>Marko Bertogna</author>
  <author>Paolo Burgio</author>
  <title>
    System Performance Modelling of Heterogeneous HW Platforms: An Automated Driving Case Study.
  </title>
  <crossref>conf/dsd/2019</crossref>
  ...
</inproceedings>
```

incollection



book

```
<incollection key="series/acvpr/DenmanHFS17" mdate="2019-08-20">
  <author>Simon Denman</author>
  <title>
    Locating People in Surveillance Video Using Soft Biometric Traits.
  </title>
  <crossref>series/acvpr/TC2017</crossref>
  ...
</incollection>
```

article



journal

```
<article key="journals/nature/Beckwith13" mdate="2019-11-28">
  <author>Christopher I. Beckwith</author>
  <title>History of science: Science spun on the Silk Road.</title>
  <year>2013</year>
  <volume>502</volume>
  <journal>Nature</journal>
  ...
  <url>db/journals/nature/nature502.html#Beckwith13</url>
</article>
```



Parsing XML

Il primo passo per la creazione del programma è stata l'analisi del file XML

Due approcci possibili:

1. DOM (Tree Based)
2. SAX (Event Based)

DOM (Document Object Model)

- ◇ Crea la struttura in memoria.
- ◇ Veloce ma oneroso in termini di risorse.

SAX (Simple Api for XML)

- ◇ Esegue azioni in seguito a degli eventi.
- ◇ Più lento ma richiede meno memoria.

SAX si è rivelata la scelta ottimale per le macchine utilizzate.



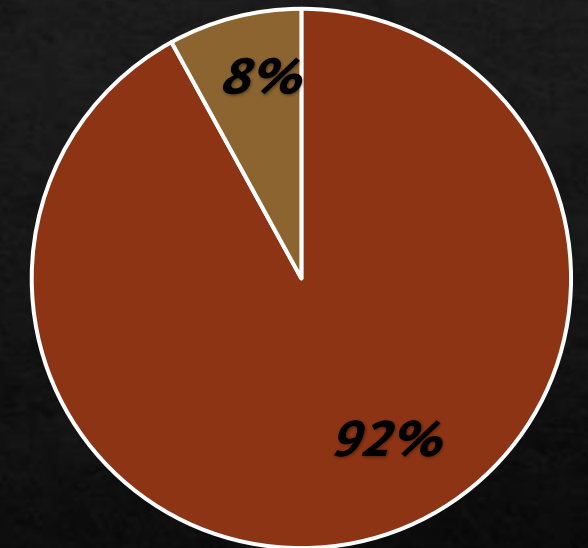
Parallelizzazione

In seguito a test si è deciso di parallelizzare il parsing e la creazione degli indici, tramite Whoosh.

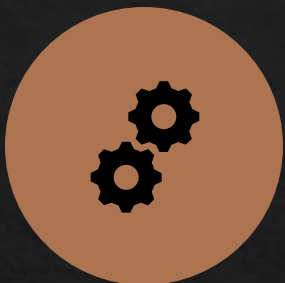
◆ Dati

- ◆ ~17 minuti per eseguire il parsing in parallelo.
- ◆ ~19 minuti in sequenziale.
- ◆ Il carico di lavoro è completamente sbilanciato.
- ◆ Le risorse sono distribuite uniformemente tra i processi, limitando l'utilizzo della RAM disponibile all'80%

Carico di lavoro



■ *pubblicazione* ■ *venue*



Modellazione degli indici

◈ Indici creati durante la fase di parsing (parallelo)

◈ 3 fasi di processing:

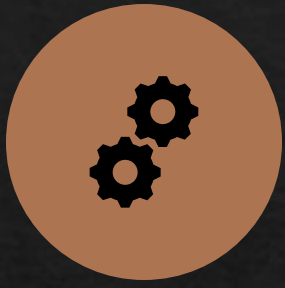
1. Pubblicazioni
2. Collezioni
3. Aggiunta riviste (trovate nella fase 1) alle Collezioni



Caso particolare: Journal

- ◇ Se durante il parsing delle pubblicazioni viene trovato l'attributo *journal*:
 1. Controllo univocità *key* per evitare duplicazione
 2. Creazione di *key* = *<nome_journal>/year/volume/number*
 3. *Journal* salvato su file
 4. Aggiornata *crossref* della relativa pubblicazione

- ◇ Aggiunti all'indice *venue* dopo la sua creazione



Caso particolare: Journal

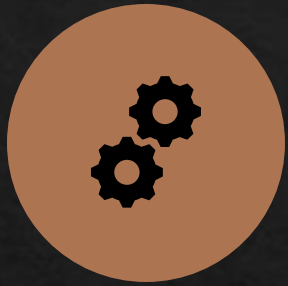
Scelta della chiave

PRO

◇ Riferimenti più specifici

CONTRO

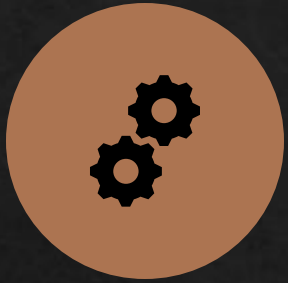
◇ Numero di riviste più elevato



Caso particolare: Journal

Esempi di key

GTELaboratoriesIncorporated/1989/TM-0149-06-89-165/~GTE Laboratories Incorporated-1989-TM-0149-06-89-165~~db/journals/gtelab/index.html~
GTELaboratoriesIncorporated/1995/TR-0310-11-95-165/~GTE Laboratories Incorporated-1995-TR-0310-11-95-165~~db/journals/gtelab/index.html~
GTELaboratoriesIncorporated/1991/TR-0146-06-91-165/~GTE Laboratories Incorporated-1991-TR-0146-06-91-165~~db/journals/gtelab/index.html~
GTELaboratoriesIncorporated/1993/TR-0231-08-93-165/~GTE Laboratories Incorporated-1993-TR-0231-08-93-165~~db/journals/gtelab/index.html~
GTELaboratoriesIncorporated/1993/TR-0244-12-93-165/~GTE Laboratories Incorporated-1993-TR-0244-12-93-165~~db/journals/gtelab/index.html~
GTELaboratoriesIncorporated/1991/TR-0169-12-91-165/~GTE Laboratories Incorporated-1991-TR-0169-12-91-165~~db/journals/gtelab/index.html~
GTELaboratoriesIncorporated/1990/TM-0332-11-90-165/~GTE Laboratories Incorporated-1990-TM-0332-11-90-165~~db/journals/gtelab/index.html~
GTELaboratoriesIncorporated/1988/TM-0014-06-88-165/~GTE Laboratories Incorporated-1988-TM-0014-06-88-165~~db/journals/gtelab/index.html~
GTELaboratoriesIncorporated/1993/TR-0236-09-93-165/~GTE Laboratories Incorporated-1993-TR-0236-09-93-165~~db/journals/gtelab/index.html~
UniversityofCaliforniaatBerkeley/1979/UCB/ERLM79/28/~University of California at Berkeley-1979-UCB/ERL M79/28~~~
ANSIX3H2/1990/X3H2-90-412/~ANSI X3H2-1990-X3H2-90-412~~db/conf/x3h2/index.html~
ANSIX2H2/1991/DBL:KAW-006X3H2-91-133rev1/~ANSI X2H2-1991-DBL:KAW-006 X3H2-91-133rev1~~db/conf/x3h2/index.html~
ANSIX3H2/1990/X3H2-90-292/~ANSI X3H2-1990-X3H2-90-292~~db/conf/x3h2/index.html~
ANSIX3H2/1991/DBL:ARL-029X3H2-91-083rev1/~ANSI X3H2-1991-DBL:ARL-029 X3H2-91-083rev1~~db/conf/x3h2/index.html~
ANSIX3H2/1992/X3H2-92-062/~ANSI X3H2-1992-X3H2-92-062~~db/conf/x3h2/index.html~
IWBSReport/1991/191/~IWBS Report-1991-191~~~
LILOG-Report/1988/59/~LILOG-Report-1988-59~~~
LILOG-Report/1987/15/~LILOG-Report-1987-15~~~



Schema degli indici

Publication

pubtype = TEXT(stored = True)

key = STORED

author = TEXT(stored = True)

title = TEXT(stored = True)

pages = STORED

year = TEXT(stored = True)

journal = STORED

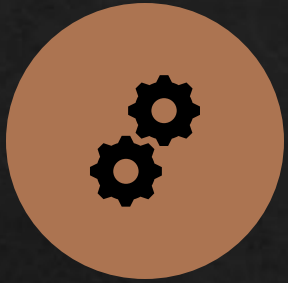
volume = STORED

number = STORED

url = STORED

ee = STORED

crossref = ID(stored = True)



Schema degli indici

Venue

pubtype = TEXT(stored = True)

key = ID(stored = True)

author = STORED

title = TEXT(stored = True)

journal = STORED

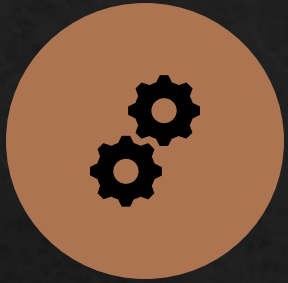
publisher = TEXT(stored = True)

url = STORED

ee = STORED

year = STORED

isbn = STORED



Analisi degli indici

Dati

Publications: 4.786.307

Venue: 60.507

Venue con
journals: 461.164



Parsing delle interrogazioni

1. Si spezza l'interrogazione dell'utente dividendo le ricerche come da sintassi
2. Le singole interrogazioni sono adattate secondo il *Whoosh Query Language*
3. Le query sono collegate tra loro in OR

Casi particolari:

- ◇ *Search_pattern* senza *element_field* sono cercati in entrambi gli indici
- ◇ *Phrasal retrieval* ammessa secondo il formato: "*search_pattern*"
- ◇ Ricerche meno specifiche: *Prefix* e *suffix* (* , ?)



Struttura dei documenti

I risultati della ricerca sono una lista di documenti, ognuno con una struttura ben definita:

```
doc = {'key': <string>, 'score': <float>, 'ven': <dict>,  
      'pub': <list>, 'alternative': <list>}
```

```
'pub' = {'author': <string>, 'crossref': <string>,  
        'ee': <string>, 'journal': <string>, 'key': <string>,  
        'number': <string>, 'pages': <string>,  
        'pubtype': <string>, 'title': <string>, 'url': <string>,  
        'volume': <string>, 'year': <string>, 'o_score': <float>}
```

```
'ven' = {'author': <string>, 'ee': <string>,  
        'isbn': <string>, 'key': <string>,  
        'publisher': <string>, 'pubtype': <string>,  
        'journal': <string>, 'title': <string>, 'url': <string>,  
        'year': <string>, 'o_score': <float>}
```



Algoritmi di ricerca

Okapi BM25F

Frequenza

Altre opzioni: Fuzzy-term



Algoritmi di ricerca

OKAPI BM25F

Best Match 25 Model with Extension to Multiple Weighted Fields

- ◇ Estensione del modello BM25
- ◇ Basato sui modelli probabilistici e su *tf-idf*
- ◇ Applicabile a documenti strutturati consistenti di campi multipli

L'idea generale è che alcuni campi abbiano più importanza di altri

Un match del titolo con la query utente si prevede essere più rilevante che un match con una parte di testo!

- ◇ Ne esistono diverse implementazione open source: Whoosh usa quella Okapi



Algoritmi di ricerca

Frequency

- ◇ Il peso di un termine che occorre in un documento è semplicemente proporzionale alla frequenza del termine
- ◇ Implementato da Whoosh sotto la classe *whoosh.scoring.Frequency*
- ◇ Spesso usato come fattore di peso nelle ricerche di *information retrieval*



Algoritmi di ricerca

Fuzzy-term

- ◆ Trova i documenti contenenti parole simili al termine dato
- ◆ Opzione implementata da *Whoosh* sotto la classe:

```
class.whoosh.query.FuzzyTerm(fieldname, text, boost=1.0,  
maxdist=1, prefixlength=1, constantscore=True)
```



Threshold

◆ Permette di restituire all'utente un pool di risultati più specifico!

◆ Viene invocato solo quando viene effettuata una ricerca su entrambi gli indici.

(a) publications

| score | crossref |
|-------------|----------------|
| pub score 1 | pub crossref 1 |
| pub score 2 | pub crossref 2 |
| pub score 3 | pub crossref 3 |
| pub score 4 | pub crossref 4 |
| ... | ... |
| pub score N | pub crossref N |

(b) venue

| key | score |
|-------------|---------------|
| venue key 1 | venue score 1 |
| venue key 2 | venue score 2 |
| venue key 3 | venue score 3 |
| venue key 4 | venue score 4 |
| ... | ... |
| venue key N | venue score N |

(c) threshold

| Threshold: score |
|------------------|
| sum_score |
| sum_score1 |
| sum_score2 |
| sum_score3 |
| |
| sum_scoreN |

if *Threshold* < *sum_score1*
l'algoritmo termina



Threshold

- ◆ Elementi ordinati per punteggio decrescente
- ◆ Cerca se tra i risultati di publications e venue esiste qualche collegamento, cercando la crossref della prima pub nelle key delle venue (e viceversa)
- ◆ Se trova corrispondenza, si crea un unico elemento che ha come score la somma degli score dei due elementi e come key quella di publication

(a) publications

| score | crossref |
|-------------|----------------|
| pub score 1 | pub crossref 1 |
| pub score 2 | pub crossref 2 |
| pub score 3 | pub crossref 3 |
| pub score 4 | pub crossref 4 |
| ... | ... |
| pub score N | pub crossref N |

(b) venue

| key | score |
|-------------|---------------|
| venue key 1 | venue score 1 |
| venue key 2 | venue score 2 |
| venue key 3 | venue score 3 |
| venue key 4 | venue score 4 |
| ... | ... |
| venue key N | venue score N |





Threshold

(c) threshold

- ◇ *score* è dato dalla somma degli score della pubblicazione e della *venue* scelta.

| Threshold: score |
|------------------|
| sum_score |
| sum_score1 |
| sum_score2 |
| sum_score3 |
| |
| sum_scoreN |

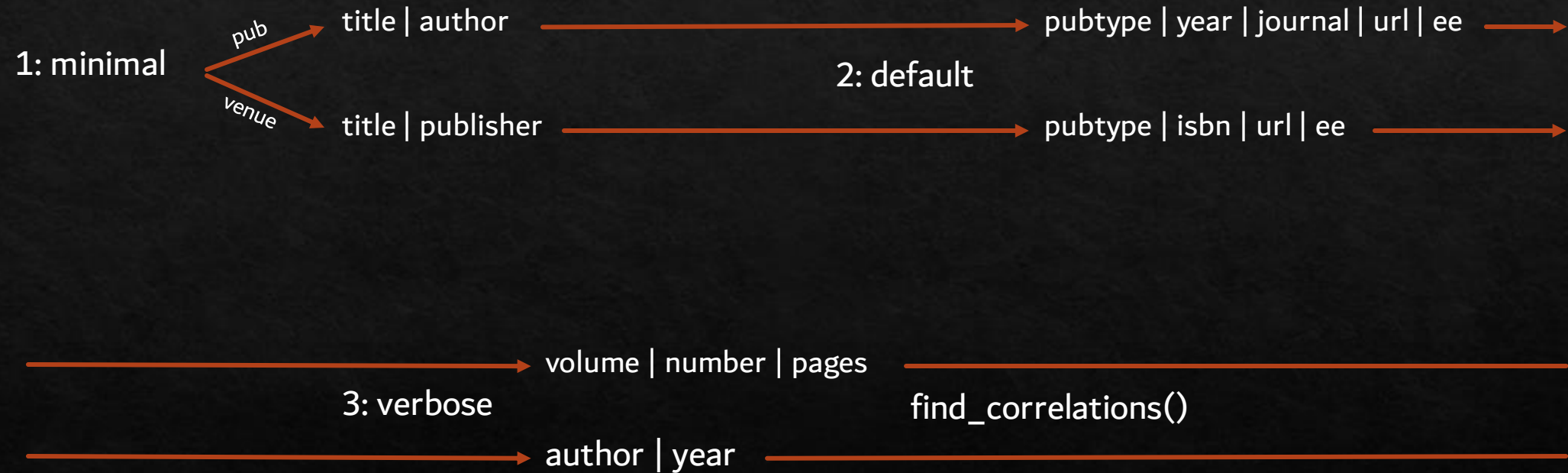
if *Threshold* < *sum_score1*
l'algoritmo termina

- ◇ Questo processo si itera finche' non si giunge a verificare la condizione!!!



Gestione dei risultati

3 livelli di output:





Gestione dei risultati

Pub:

Query:

1) score: 29.16489

Publication

Type: inproceedings

Title: A Statistical Mechanics Approach to Immigrant Integration in Emilia Romagna (Italy).

Author: Francesco De Pretis, Cecilia Vernia

Year: 2014

Link: <https://dblp.uni-trier.de/db/conf/complenet/complenet2014.html#PretisV14>

Alternative link: https://doi.org/10.1007/978-3-319-05401-8_6

publication.author:"Francesco De Pretis"

Venue:

1) score: 10.43794

Venue

Type: book

Title: The physics of quantum information: quantum cryptography, quantum teleportation, quantum computation.

Publisher: Springer

ISBN: 3540667784

Alternative link: <http://www.worldcat.org/oclc/43919627>

venue.title:quantum



Gestione dei risultati

Venue con pub(s) rilevanti:

Query:

1) score: 65.72163

Venue

Type: proceedings

Title: Proceedings of the VLDB 2019 PhD Workshop, co-located with the 45th International Conference on Very Large Databases (VLDB 2019), Los Angeles, California, USA, August 26-30, 2019.

Publisher: CEUR-WS.org

Link: <https://dblp.uni-trier.de/db/conf/vldb/phd2019.html>

Alternative link: <http://ceur-ws.org/Vol-2399>
<http://nbn-resolving.de/urn:nbn:de:0074-2399-4>

Relevant Publications

Type: inproceedings

Title: Database Systems 2.0.

Author: Johannes Gehrke

Year: 2019

Link: <https://dblp.uni-trier.de/db/conf/vldb/phd2019.html#Gehrke19>

Alternative link: <http://ceur-ws.org/Vol-2399/keynote1.pdf>

Type: inproceedings

Title: Structured Data Meets News.

Author: Cong Yu 0001

Year: 2019

Link: <https://dblp.uni-trier.de/db/conf/vldb/phd2019.html#Yu19>

Alternative link: <http://ceur-ws.org/Vol-2399/keynote2.pdf>

inproceedings.title: "Database Systems 2.0"
inproceedings.title: "Structured Data Meets News"
venue:VLDB



Gestione dei risultati

Pub con venue rilevante:

```
2)    score: 26.72775
      Publication
      Type: article
      Title: Computer Science Education.
      Author: Raymond Lister
      Year: 2008
      Journal: Computer Science Education
      Link: https://dblp.uni-trier.de/db/journals/csedu/csedu18.html#Lister08
      Alternative link: https://doi.org/10.1080/08993400802172449

      In Relevant Venue
      Title: Computer Science Education
      Year: 2008
```

Query:

```
publication.title:"computer vision"
venue:image
```



Gestione dei risultati

find_correlations()

Pub con venue non rilevante

```
3)    score: 17.72883
      Publication
      Type: article
      Title: Computer Vision.
      Author: Azriel Rosenfeld
      Year: 1988
      Journal: Advances in Computers
      Volume: 27
      Pages: 265-308
      Link: https://dblp.uni-trier.de/db/journals/ac/ac27.html#Rosenfeld88
      Alternative link: https://doi.org/10.1016/S0065-2458\(08\)60261-2

      In Venue
      Title: Advances in Computers
      Year: 1988
      Link: https://dblp.uni-trier.de/db/journals/ac/ac27.html
      Alternative link: https://doi.org/10.1016/S0065-2458\(08\)60256-9
```

Query:

```
publication.title:"computer vision"
venue:3D
```

Data una *pub* mostra a quale *venue* appartiene se questa non è stata precedentemente accoppiata!



Gestione dei risultati

find_correlations()

Venue con pubs non rilevanti:

Query:

2) score: 12.12054

Venue

Type: journal

Title: VLDB J.

Year: 2010

Link: <https://dblp.uni-trier.de/db/journals/vldb/vldb19.html>

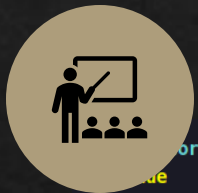
Alternative link: <https://doi.org/10.1007/s00778-009-0159-9>

Pubs Included

- Schema mapping and query translation in heterogeneous P2P XML databases.
- Continuous spatial assignment of moving users.
- Revisiting the cube lifecycle in the presence of hierarchies.
- Understanding the meaning of a shifted sky: a general framework on extending skyline query.
- Threshold-based probabilistic top-
- Continuous authentication on relational streams.
- A framework for testing DBMS features.

venue:VLDB

Data una *venue* mostra alcune delle *pubs* in esse contenute, se questa non è già stata precedentemente accoppiata



Gestione dei risultati

find_correlations()

Venue con *pubs* rilevanti e incluse

Query:

computer science

Data una *venue* con *pubs* rilevanti mostra alcune pubblicazioni incluse

ore: 44.97706

ue

Type: journal

Title: ACM Comm. Computer Algebra

Year: 2010

Link: <https://dblp.uni-trier.de/db/journals/cca/cca44.html>

Alternative link: <https://doi.org/10.1145/1940475.1940487>

Relevant Publications

Type: article

Title: Symbolic computation software composability protocol and its implementations.

Author: The SCIENCE project

Year: 2010

Journal: ACM Comm. Computer Algebra

Volume: 44

Number: 3/4

Pages: 210-212

Link: <https://dblp.uni-trier.de/db/journals/cca/cca44.html#project10>

Alternative link: <https://doi.org/10.1145/1940475.1940522>

Type: article

Title: SymGrid-Par: parallel orchestration of symbolic computation systems.

Author: The SCIENCE project

Year: 2010

Journal: ACM Comm. Computer Algebra

Volume: 44

Number: 3/4

Pages: 213-216

Link: <https://dblp.uni-trier.de/db/journals/cca/cca44.html#project10a>

Alternative link: <https://doi.org/10.1145/1940475.1940523>

Pubs Included

- transalpyne: a language for automatic transposition.
- Nullspace computation over rational function fields for symbolic summation.
- Sturm root counting using chebyshev expansion.
- Solving linear recurrence equations.
- A fast recursive algorithm for computing cyclotomic polynomials.
- A FPGA implementation of Chen's algorithm.

Menu e opzioni

MAIN MENU

1. Make a search.
2. Change settings.
3. Print active settings.
4. Exit.

Type your choice:

> 3

Options:

Ranking: *bm25f*

Results limit: *10*

Fuzzy: *False*

Output level: *2*

oppure

frequency

n

True

1

3