Towards automatic discovery of co-authorship networks in the Brazilian academic areas

Jesús P. Mena-Chalco and Roberto M. Cesar Junior
{jmena,cesar}@vision.ime.usp.br

Department of Computer Science
Institute of Mathematics and Statistics
University of São Paulo

http://scriptlattes.sourceforge.net/

12/05/2011
Outline

1 Motivation

2 Data Collection

3 The system
   - Data selection
   - Data extraction
   - Redundancy treatment
   - Co-authorship networks generation
   - AuthorRank computation

4 Experiments and results

5 Conclusions
Outline

1 Motivation

2 Data Collection

3 The system

4 Experiments and results

5 Conclusions
Motivation

- **Data mining** is a process of extracting large volumes of data, commonly used to identify or reveal possible **relationships between instances** of the treated elements.

- The extraction of scientific data, identification of bibliometric patterns, and effective modeling and visualization of **networks of interaction** among co-authors are relevant topics in **e-Science**, Bibliometrics, and Scientometrics.

- In recent years, there has been growing interest in such topics because of the **knowledge discovery process** that may be implemented from the treatment of datasets available in the **repositories of scientific production**.
Curriculum Lattes

(semi-structured information)
An open-source knowledge extraction system from the Lattes platform for analysing individuals Lattes curricula.
Outline

1 Motivation
2 Data Collection
3 The system
4 Experiments and results
5 Conclusions
Data collection

- The data used in our work is from the Lattes platform on May 5th, 2011 and contains about 1.3 million curricula.

- We have created several procedures in order to crawl curricula, in HTML format, by accessing the public web-search interface provided by the Lattes platform.

- Several web-query were executed for the purpose of access to large Lattes curricula dataset.

- In summary, it was performed eighty queries to the Lattes platform, each one associated to an knowledge domain.
Outline

1 Motivation

2 Data Collection

3 The system
   ■ Data selection
   ■ Data extraction
   ■ Redundancy treatment
   ■ Co-authorship networks generation
   ■ AuthorRank computation

4 Experiments and results

5 Conclusions
A high level view of the system

Academic area, production's type, and period time

- Data selection
  - Lattes curricula
- Data extraction
  - list of productions
- Redundancy treatment
  - list of filtered productions
- Co-authorship networks generation
  - network with normalized weights
- AuthorRank computation
  - AuthorRank measures
A. Data selection

- This module allows to select the **Lattes curricula**, in HTML format, from the local database.

- Only the **subset** of researchers Lattes curricula is chosen in accordance to the academic (expertise) areas **informed by the user**.

- Public users of the Lattes platform **do not have access** to the complete **Lattes database**. For this reason, special attention is given to extract the information about the scientific productions.
B. Data extraction

- In this module, a deterministic HTML Parser is used to automatically extract all information of each researcher.

- The designed parser allows to extract the list of academic productions limited by the years indicated in the time period.

- A computational challenge for the system is processing the data in HTML format, where the constituent parts of the academic productions (e.g. publication title, the journal name of publication, number, volume, pages, year) are presented without any indication of division/separation.
### (i) Bibliographical production

- Articles in scientific journals
- Book published/organized
- Book chapter published
- Articles in newspapers/magazines
- Complete works published in proceedings of conferences
- Expanded summary published in proceedings of conferences
- Summary published in proceedings of conferences
- Articles accepted for publication
- Presentations of work
- Other kinds of bibliographical production

### (ii) Technical production

- Patented or registered software
- Not patented or registered software
- Technological products
- Techniques or process
- Technical works
- Other kinds of technical production

### (iii) Artistic productions

- Artistic/cultural production
C. Redundancy treatment

- It is common that some academic productions are made in co-authorship with one or more researches of the selected group in the Data selection module. Thus, the same academic production may be being registered in each Lattes curriculum of the corresponding co-authors.

- Two or more researchers are considered co-authors if there is a common production between them.

- Due to inconsistencies in filling information in the Lattes curriculum (e.g. typos), the comparison of any two productions is made through an inexact matching procedure between the title of the academic productions.
C. Redundancy treatment (cont.)

**Author 1**

**Author 2**

**Co-authors list:**
HIRATA JR., R. ; BRUN, M. ; BARRERA, J. ; DOUGHERTY, E. R.
HIRATA JR, Roberto ; BARRERA, Junior ; BRUN, Marcel ; DOUGHERTY, E R

**Titles:**
aperture filters: theory, application, and multiresolution analysis
aperture filters: theory, application and multi-resolution analysis
**D. Co-authorship networks generation**

- The designed system uses a *graph* to represent the **co-authorship among researchers** based on scientific productions.

- Each researcher is represented by a *node*. An *edge* is created between a pair of nodes whenever a common production of the corresponding researchers is detected.

- Our work is focused on creating different co-authorship networks:
  - Networks without weights
  - Networks with un-normalized weights
  - Networks with co-authorship frequencies
  - Networks with normalized weights
D. Co-authorship networks generation (cont.)

### Scientific production

- Production 1 written by the authors A1 and A2
- Production 2 written by the authors A1, A2 and A3
- Production 3 written by the authors A1 and A4

### Exclusivity
- Production 1: $A1-A2 = 1$
- Production 2: $A1-A2 = 0.5$
- Production 2: $A1-A3 = 0.5$
- Production 2: $A2-A3 = 0.5$
- Production 3: $A1-A4 = 1$

<table>
<thead>
<tr>
<th>Author</th>
<th>Productions</th>
<th>Degree</th>
<th>AuthorRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>3</td>
<td>3</td>
<td>1.648</td>
</tr>
<tr>
<td>A2</td>
<td>2</td>
<td>2</td>
<td>1.114</td>
</tr>
<tr>
<td>A3</td>
<td>1</td>
<td>2</td>
<td>0.620</td>
</tr>
<tr>
<td>A4</td>
<td>1</td>
<td>1</td>
<td>0.617</td>
</tr>
</tbody>
</table>

### Co-authorship networks

- **Network without weights**
- **Network with un-normalized weights**
- **Network with co-authorship frequencies**
- **Network with normalized weights**
E. AuthorRank computation

- The AuthorRank (AR) of an author $i$, which has $n$ co-authors, is given as follows:

$$AR^{(t)}(i) = (1 - d) + d \sum_{j=1}^{n} (AR^{(t-1)}(j) \times w_{j,i})$$  \hspace{1cm} (1)$$

where $w_{j,i}$ corresponds to the normalized weight between author $j$ and author $i$, and $d$ is the damping factor.

- The researchers with more collaborative status have the highest AuthorRank values.
E. AuthorRank computation (cont.)

Scientific production

- Production 1 written by the authors A1 and A2
- Production 2 written by the authors A1, A2 and A3
- Production 3 written by the authors A1 and A4

<table>
<thead>
<tr>
<th>Exclusivity</th>
<th>Author</th>
<th>Productions</th>
<th>Degree</th>
<th>AuthorRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production 1: A1-A2 = 1</td>
<td>A1</td>
<td>3</td>
<td>3</td>
<td>1.648</td>
</tr>
<tr>
<td>Production 2: A1-A2 = 0.5</td>
<td>A2</td>
<td>2</td>
<td>2</td>
<td>1.114</td>
</tr>
<tr>
<td>Production 2: A1-A3 = 0.5</td>
<td>A3</td>
<td>1</td>
<td>2</td>
<td>0.620</td>
</tr>
<tr>
<td>Production 3: A1-A4 = 1</td>
<td>A4</td>
<td>1</td>
<td>1</td>
<td>0.617</td>
</tr>
</tbody>
</table>

Co-authorship networks

(a) network without weights
(b) network with un-normalized weights
(c) network with co-authorship frequencies
(d) network with normalized weights
Outline

1 Motivation
2 Data Collection
3 The system
4 Experiments and results
5 Conclusions
Experiments and results

- Researchers from **four domains** have been considered in the present study:

<table>
<thead>
<tr>
<th>Area</th>
<th>Researchers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer Science</td>
<td>6 014</td>
</tr>
<tr>
<td>Mathematics</td>
<td>5 008</td>
</tr>
<tr>
<td>Physics</td>
<td>6 410</td>
</tr>
<tr>
<td>Statistics</td>
<td>2 274</td>
</tr>
</tbody>
</table>

- The Data selection module has been configured to consider only **researchers with a doctoral degree** (at least).

- In order to standardize the analysis of automatic generation of co-authorship networks, only the **journal papers** published between **2000 and 2010** were considered.
Experiments and results

Computer Science

degree

authorRank
Experiments and results

degree

authorRank

Mathematics
Experiments and results

degree

authorRank

Physics
Experiments and results

degree

authorRank

Statistics
Experiments and results

Co-authorship network belonging to researchers associated to four Brazilian academic areas
In our experiment, we computed the Pearson’s correlation coefficient of 0.719 (p-value < 2.2e-16) between the degree values and the AuthorRank measures.

In fact, both measures indicate **how related is a researcher** in the academic community.
Top 10 researchers w.r.t. degree values and AuthorRank measures

<table>
<thead>
<tr>
<th>Researcher</th>
<th>Degree</th>
<th>AuthorRank</th>
<th>Papers</th>
<th>Area(s)</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>A6201</td>
<td>126</td>
<td>7.829</td>
<td>173</td>
<td>P</td>
<td>1B</td>
</tr>
<tr>
<td>A14631</td>
<td>104</td>
<td>13.716</td>
<td>265</td>
<td>CS, P</td>
<td>1A</td>
</tr>
<tr>
<td>A2763</td>
<td>98</td>
<td>8.095</td>
<td>154</td>
<td>P</td>
<td>1A</td>
</tr>
<tr>
<td>A12704</td>
<td>98</td>
<td>7.459</td>
<td>151</td>
<td>P</td>
<td>1A</td>
</tr>
<tr>
<td>A1839</td>
<td>96</td>
<td>5.389</td>
<td>135</td>
<td>P</td>
<td>1C</td>
</tr>
<tr>
<td>A260</td>
<td>91</td>
<td>4.328</td>
<td>85</td>
<td>P</td>
<td>1D</td>
</tr>
<tr>
<td>A3599</td>
<td>89</td>
<td>6.507</td>
<td>122</td>
<td>P</td>
<td>1C</td>
</tr>
<tr>
<td>A14752</td>
<td>88</td>
<td>8.118</td>
<td>142</td>
<td>P</td>
<td>1B</td>
</tr>
<tr>
<td>A9190</td>
<td>87</td>
<td>2.740</td>
<td>51</td>
<td>P</td>
<td>1C</td>
</tr>
<tr>
<td>A11411</td>
<td>84</td>
<td>8.062</td>
<td>189</td>
<td>P</td>
<td>1A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Researcher</th>
<th>AuthorRank</th>
<th>Degree</th>
<th>Papers</th>
<th>Area(s)</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>A14631</td>
<td>13.716</td>
<td>104</td>
<td>265</td>
<td>CS, P</td>
<td>1A</td>
</tr>
<tr>
<td>A4600</td>
<td>9.706</td>
<td>41</td>
<td>239</td>
<td>CS, S</td>
<td>1B</td>
</tr>
<tr>
<td>A9220</td>
<td>9.442</td>
<td>48</td>
<td>148</td>
<td>CS, M, P</td>
<td>1B</td>
</tr>
<tr>
<td>A15939</td>
<td>8.332</td>
<td>41</td>
<td>102</td>
<td>M</td>
<td>1A</td>
</tr>
<tr>
<td>A14752</td>
<td>8.118</td>
<td>88</td>
<td>142</td>
<td>P</td>
<td>1B</td>
</tr>
<tr>
<td>A2763</td>
<td>8.095</td>
<td>98</td>
<td>154</td>
<td>P</td>
<td>1A</td>
</tr>
<tr>
<td>A11411</td>
<td>8.062</td>
<td>84</td>
<td>189</td>
<td>P</td>
<td>1A</td>
</tr>
<tr>
<td>A14054</td>
<td>7.950</td>
<td>50</td>
<td>143</td>
<td>P</td>
<td>1A</td>
</tr>
<tr>
<td>A16431</td>
<td>7.888</td>
<td>35</td>
<td>70</td>
<td>CS</td>
<td>1C</td>
</tr>
<tr>
<td>A6201</td>
<td>7.829</td>
<td>126</td>
<td>173</td>
<td>P</td>
<td>1B</td>
</tr>
</tbody>
</table>
Outline

1 Motivation

2 Data Collection

3 The system

4 Experiments and results

5 Conclusions
Conclusions

- We proposed an automatic system for generation of large-scale co-authorship networks, based on Lattes curricula belongings to researchers associated to specific academic areas.

- Our work is the first study to automatically examine Brazilian co-authorship networks in such large-scale.

- We observe a significant correlation between the node measures such as the degree value and the AuthorRank measure.

- The system can be used for exploring, identifying and validating patterns of scientific activities, thus bringing bibliometric and/or scientometric information about a group of interest.
Towards automatic discovery of co-authorship networks in the Brazilian academic areas

Jesús P. Mena-Chalco and Roberto M. Cesar Junior
{jmena,cesar}@vision.ime.usp.br

Department of Computer Science
Institute of Mathematics and Statistics
University of São Paulo

http://scriptlattes.sourceforge.net/

12/05/2011