Building a Research Social Network from an Individual Perspective

Alberto H. F. Laender, Mirella M. Moro, Marcos André Gonçalves, Clodoveu A. Davis Jr., Altigran S. da Silva‡, Allan J. C. Silva, Carolina A. S. Bigonha, Daniel Hasan Dalip, Eduardo M. Barbosa, Eli Cortez‡, Peterson S. Procópio Jr., Rafael Odon de Alencar, Thiago N. C. Cardoso, Thiago Salles

Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
{laender,mirella,mgoncalv,clodoveu,allan,carolb,hasan,emb,peterson,odon,thiagon,tsalles}@dcc.ufmg.br

‡Universidade Federal do Amazonas, Manaus, Brazil
{alti,eccv}@dcc.ufam.edu.br

ABSTRACT

In this poster paper, we present an overview of CiênciaBrasil, a research social network involving researchers within the Brazilian INCT program. We describe its architecture and the solutions adopted for data collection, extraction, and deduplication, and for materializing and visualizing the network.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

General Terms

Design, Management, Measurement

Keywords

Social Networks, Research Collaboration, CiênciaBrasil

1 Introduction

In July 2008, the Brazilian National Council for Scientific and Technological Development - CNPq, lauched a nationwide research program called National Institutes of Science and Technology (or simply INCT)[1]. Among the 126 institutes approved, INWeb (INCT for the Web) focuses on cutting edge research and technology on Web-related topics. One of its projects, *CiênciaBrasil* - The Brazilian Portal of Science and Technology[2], consists of a research social network that allows studying science and technology indicators (e.g., knowledge production, research collaboration, human resources formation, technology transfer, etc.) involving Brazilian researchers within the INCT program.

Building and analyzing research networks beget challenges that go beyond those from regular social networks [3]. The data is usually not provided by the actors (researchers) themselves but must be gathered and extracted from various sources (including the Web). Also, the relationships among those actors are implicitly embedded in the data and must be derived by processing it. Even if the data can be correctly obtained, issues related to data duplication and (name) ambiguity, which can occur since data may be collected from several heterogeneous sources, need to be dealt with to guarantee a minimum of quality in the construction of a network. Besides, issues associated with the visualization, navigation and analysis of such networks must be addressed.

In Brazil, *Lattes*[3] is a Web platform made available by CNPq for storing, managing and searching for curricula vitae of researchers involved with Brazilian institutions. Lattes allows researchers (from undergrads to seniors) to inform all their academic achievements and has been recognized as one of "the cleanest researcher databases in existence" [2]. With such organized data, Lattes became the natural source for building CiênciaBrasil. Indeed, such a standardized platform provides a large, individually centered repository of scientific and educational information, from which it is possible to put together research groups and associations for building a fairly reliable research social network covering several knowledge fields. However, obtaining and organizing data from Lattes is per se a challenge: researchers upload their information by filling out forms with predefined fields and all the data is published on the Web in a free format. Therefore, we must deal with typos and duplicated information (e.g., same paper on different curricula) as well as with complex data extraction issues.

In this poster paper, we introduce CiênciaBrasil and provide an overview of our solutions for some of the challenges involved. We also illustrate its potential as a platform for large scale academic research evaluation and analysis.

2 Architecture and Basic Features

Figure 1 illustrates the architecture we conceived for building and maintaining CiênciaBrasil with data from Lattes. The architecture supports the execution of four major processes: *crawling*, *data extraction*, *data deduplication* and *network materialization*.

Crawling. Before crawling Lattes, for each INCT we manually input in the portal repository (a relational database) general data (name, principal investigator, institution, etc.) and the list of Lattes URLs for its participating researchers (arrow 1). The list of URLs is then fed to the crawling process (arrow 2) for collecting the respective curricula vitae. The crawling process applies an asynchronous concurrent download strategy to increase throughput and help keep track of download errors to avoid any possible data loss. After the crawling, the gathered data is available in pure HTML format (arrow 3), which will be further decomposed in finer-grained structures called *information blocks*.

[1]http://www.cnpq.br/programas/inct/_apresentacao
[2]http://pbct.inweb.org.br

[3]http://lattes.cnpq.br/english

JCDL'11, June 13–17, 2011, Ottawa, Ontario, Canada.
ACM 978-1-4503-0744-4/11/06.

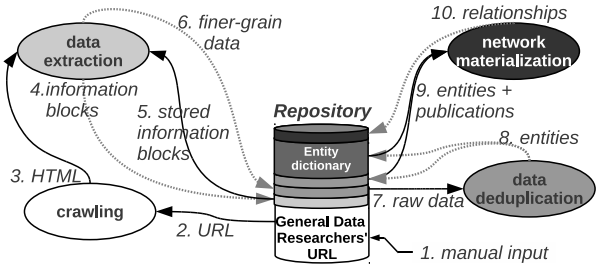

Figure 1: CiênciaBrasil architecture

Data Extraction. The data extraction process involves two steps. First, we extract information blocks from the HTML pages, and then we extract specific fine-grain data from the information blocks, such as publication metadata (e.g., title, venue, etc.). In the first step, despite the standardized structure of the Lattes curricula vitae, there are no explicit delimiters to guide the extraction process. For example, some unrelated fields are grouped by the same HTML markup and others are freely formatted fields that need additional parsing (e.g., the researchers' address). To address this, we have crafted a number of regular expressions for discovering and grouping related information into blocks. Extracted blocks are then stored in the repository for further processing (arrow 4). In the second step, we retrieve the previously stored information blocks (arrow 5) and extract from them detailed data (arrow 6) to build the research social network. Specifically, the CiênciaBrasil network is based on coauthorship derived from publication metadata.

Data Deduplication. As CiênciaBrasil processes data from individual curricula vitae (arrow 7), each coauthored publication is likely to be listed in more than one curriculum vitae. However, not all of these occurrences are exact duplicates, as typos or abbreviations may be introduced while filling up them. To avoid duplicates interfering with derived statistics and further analysis, we deduplicate the publication entries by using an unsupervised heuristics-based method [1]. Furthermore, for better performance, the data deduplication process is parallelized using a map-reduce framework. The unique entries are then stored back into the repository, as part of the Entity Dictionary (arrow 8).

Network Materialization. For building the network, we derive coauthorship relationships based on the existence of common publications in the researchers' curricula vitae (arrows 9 and 10). An effective way to visualize such a network is through graphs. Specifically, an author graph is built with the perspective of an individual author, as in Figure 2. This way, each researcher that has collaborated with such an author is represented by a circle and positioned in a horizontal line. An arc represents a collaboration between those two researchers and its thickness represents the strength of the collaboration (number of publications with both of them as coauthors). An INCT graph shows collaborations between researchers from a same INCT, as in Figure 3, and includes three types of arc: black arcs represent collaborations that existed but stopped after the project started in 2008, blue arcs represent collaborations intensified after 2088 and orange arcs represent new collaborations.

Basic Features. CiênciaBrasil front-end was developed as a web application in Django, a Python framework for agile web development. Basic features include an initial page with general information about the project, browsing and visualization functions for exploring researchers and INCTs, and a general keyword search box that results a list of both INCTs and researchers whose names approximately match the given keywords.

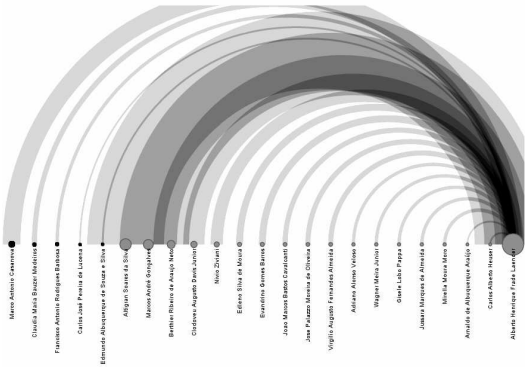

Figure 2: Author graph example

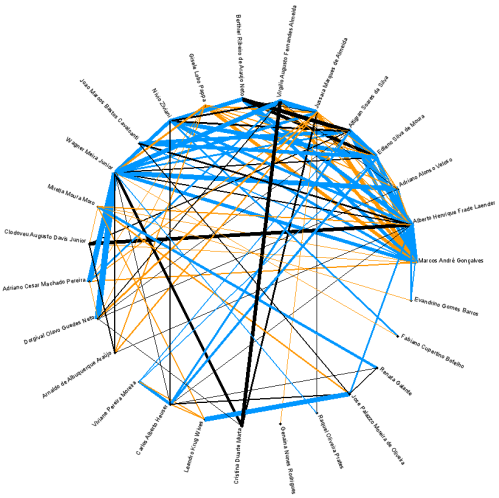

Figure 3: INCT graph example

3 Conclusion

We presented an overview of CiênciaBrasil, a research social network involving researchers within the Brazilian INCT program. It has been built from an individual perspective using the CNPq Lattes platform as data source for researchers' information. At the moment, this implies redoing all steps shown in Figure 1 every time we need to update the repository. Thus, as future work, we must devise a more efficient way to keep the repository up-to-date with Lattes. We also intend to provide citation information based on data available on other sources.

Acknowledgements. This work is partially supported by INWeb (MCT/CNPq grant 57.3871/2008-6) and by the authors' individual grants and scholarships from CNPq, CAPES and FAPEMIG.

4 References

[1] E. N. Borges et al. An unsupervised heuristic-based approach for bibliographic metadata deduplication. *Information Processing & Management*, 2011. (In press, available online)

[2] J. Lane. Let's make science metrics more scientific. *Nature*, 464(7288):488–489, March 2010.

[3] J. Tang et.al. ArnetMiner: Extraction and Mining of Academic Social Networks. In *Procs. of KDD*, pages 990–998, 2008.