11th World Congress on Computational Mechanics (WCCM XI)
5th European Conference on Computational Mechanics (ECCM V)
6th European Conference on Computational Fluid Dynamics (ECFD VI)
July 20 - 25, 2014, Barcelona, Spain

An Efficient Process for Extraction and Identification in Scientific Collaboration Networks

*** Thiago M. R. Dias[1], Gray F. Moita[1]**

[1] CEFET-MG - Federal Center for Technological Education of Minas Gerais, Av. Amazonas, 7675, Nova Gameleira, 30510-000, Belo Horizonte, MG, Brazil, thiagomagela@gmail.com and gray@dppg.cefetmg.br

Key Words: *Extraction and data integration, Information Retrieval, Scientific Collaboration.*

1. Introduction

In the scientific domain, an example of a social network is the scientific collaboration network that is observed as a graph in which the vertices correspond to the authors of specific scientific publications and the lines or edges correspond to co-authorship relationship. In this type of network, the edges mayor may not be weighted. The addition of weight represents the number of joint papers in which the authors connected by the edge under analysis have participated.

For Stroele et al. [1], scientific social networks are specific types of social networks that represent social interactions originating in the academic environment. These interactions usually occurthrough the publication of scientific articles, academic guidelines and the development of research projects. Various goals can lead to the study of scientific collaboration networks, such as recommendation of new collaborators, intensifying the collaboration, ranking of groups or individuals, oridentifying groups and their characteristics.

This work presents an efficient process for identifying collaborations in large scientific databases for the modelling and characterisation of networks wherein the vertices represent the authors and the edges represent papers co-producedby two or more authors.

2. Development

Data from the CNPq Lattes Platform were used to produce this work. The Lattes Platform was conceived to integrate the information systems of Brazilian federal agencies, optimising the Science and Technology (S&T) management process from the standpoint of both the user as well as promotion agencies and institutions of education and research. [2]

The selection of the Lattes Platform for extraction is related to it being extremely rich.This is becauseit deals with the integration of scientific data, both curricular and institutional in the S&T field, recording academic data and scientific production from researchers and institutions, allowing the researchers them selves to update the information. Currently, the Lattes Platform includes approximately 3 million recorded curricula vitae.

For each article title recorded in each of the platform's curricula vitae, an algorithm was developed to identify scientific collaborationthat produces various types of processing. Figure 1.

```
Identification-Collaboration
    1   n ← number of articles author
    2   for i ← 1 to n
    3      x ← string[i]     // x is article title [i]
    4      x ← stopword[x]    // removes token without semantic value
    5      x ← normalization[x]    // remove whitespace and accentuation
    6      x ← lowercase[x]
    7      if hash[x] in dictionary   // checks whether x is in the dictionary
    8         dictionary[x] ← idauthor
    9      else dictionary ← x, idauthor
```

Figure 1. Algorithm for identification of collaboration

As shown by the algorithm in Figure 1, each title in a given publication undergoes a transformation designed to obtain the same title without words with any semantic value, without any kind of accentmarks and without spaces. Next, the entire text is standardised in lowercase characters and the resulting string is transformed into a key.

As a result of this transformation stage, a comparison is performed to verify whether this key already exists in the directory used to identify collaborations. Should the key already exist in the directory, the author's identifier in the curriculum vitae under analysis is inserted in the key position; otherwise, the key and the identifier are inserted in the abovementioned directory.

3. Results

As a result of adopting the algorithm, a directory is generated consistingof keys which are the transformed titles of the articles and the identifiers of the authors. Given this, it is possible to generate the entire collaboration network of the whole database by inserting clicks in the collaboration graph by juxtaposition.

The greatest advantage of adopting this technique is with regards to its computational cost. Asonly one comparison for each article title is performed, it is possible to generate the collaboration network at a linear cost ($\theta(n)$), thusallowing the construction of the network in a much more interesting timeframe than with algorithms that work with cross-validation at a polynomial cost ($\theta(n^2)$).

REFERENCES

[1] V. Stroele, G. Zimbrão and J. M. Souza, Análise de Redes Sociais Científicas: Modelagem Multi-relacional. In: I Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 2012, Curitiba, PR, Brazil.

[2] T. M. R. Dias, G. F. Moita, T. Moreira, L. Santos and P. M. Dias, Modelagem e Caracterização de Redes Científicas: Um Estudo Sobre a Plataforma Lattes. In: II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 2013, Maceio, AL, Brazil