Espelho da prova de proficiência em Língua Inglesa – Ciências Exatas

SCRIPTLATTES: AN OPEN-SOURCE KNOWLEDGE EXTRACTION SYSTEM FROM THE LATTES PLATFORM[1]

Jesús Pascual Mena-Chalco, Roberto Marcondes Cesar Junior Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo – USP

Abstract

The Lattes platform is the major scientific information system maintained by the National Council for Scientific and Technological Development (CNPq). This platform allows to manage the curricular information of researchers and institutions working in Brazil based on the so called Lattes Curriculum. However, the public information is individually available for each researcher, not providing the automatic creation of reports of several scientific productions for research groups. It is thus difficult to extract and to summarize useful knowledge for medium to large size groups of researchers. This paper describes the design, implementation and experiences with scriptLattes: an open-source system to create academic reports of groups based on curricula of the Lattes Database. The scriptLattes system is composed by the following modules: (a) data selection, (b) data preprocessing, (c) redundancy treatment, (d) collaboration graph generation among group members, (e) research map generation based on geographical information, and (f) automatic report creation of bibliographical, technical and artistic production, and academic supervisions. The system has been extensively tested for a large variety of research groups of Brazilian institutions, and the generated reports have shown an alternative to easily extract knowledge from data in the context of Lattes platform. The source code, usage instructions and examples are available at http://scriptlattes.sourceforge.net/.

Keywords: academic production report, Lattes platform, knowledge discovery.

Introduction

Knowledge extraction and visualization from large datasets is an important research topic in computer science with strong potential impact in all scientific fields. The research on this topic tipically involves the treatment of large datasets which can not be processed and understood by human experts due to its volume, diversity and complexity. Techniques from datamining, summarization, visualization, network modeling and high-performance computing are often brought together in order to solve problems of this nature. The present work is focused on the summarization and visualization of scientific reports obtained from the Brazilian Lattes platform.

The Conselho Nacional de Desenvolvimento Científico e Tecnológico (Brazilian National Council for Scientific and Technological Development - CNPq), makes efforts to integrate the Curricula of people associated to Brazilian scientific communities, in a

[1] Journal of the Brazilian Computer Society, 2009; 15(4):31-39

curricular information system denominated Lattes. For this reason, the so-called “Lattes Curriculum” is considered a national standard of information about the scientific and academic accomplishments of students, professors, researchers and professionals involved in science and technology in general.

The Lattes curriculum is used for academic evaluation because it represents the history of scientific, academic and professional activities. It is hence a rich and powerful database that presents inumerous potential applications (scientific, technological, economical, etc.) The Lattes curriculum, in HTML format as available in the CNPq site, displays information only in a personal way, i.e., the registered information is individually associated to each person. This characteristic does not easily provide a way to figure out the bibliographical, technical or artistic productions of a given group, such as a research group, professors of an academic department or members of a Brazilian institution.

Currently, most of the Brazilian academic institutions usually explore the Lattes curricula in order to elaborate reports about scientific productions, supervisions, and projects of research groups related with these institutions, as well as to evaluate the graduate programs in Brazil. The reports are typically created by manually-assisted analysis of the Lattes curriculum data of each member of the group in order to obtain a complete digest of all scientific productions, supervisions and projects of the group. It is important to note that, despite having structured information, this procedure is very cumbersome and time consuming, being highly susceptible to errors caused by the manual treatment.

There are some interesting bibliometric questions that may be answered about a group just based on the respective Lattes curricula: a) How many bibliographical, technical or artistic productions were elaborated? b) What is the profile (i.e. proportion of publication) of the different types of bibliographical productions? c) How is the regularity and the evolution of the publications along the years? d) How is the collaboration/cooperation among researchers? e) How many thesis and dissertations have been concluded? f)What is the geographic distribution of the researchers? g) What is the scientific formation influence of the considered researchers?

The scriptLattes, an open-source system, was designed to provide answers to the above questions through automatically created reports. Given a group of researchers registered in the Lattes platform, the scriptLattes download their Lattes curricula from the CNPq site, extract the information of interest, eliminate the redundant scientific productions and create reports about the production, reports of academic supervisions as well as the collaboration graph and the research map from the members of the group. We believe that the introduced system is a useful tool to easily extract knowledge about the Lattes curricula of a group. This knowledge may be used to explore, identify or validate patterns of academic activities, thus bringing bibliometric information about a group of interest.

This open-source system runs on a PC with GNU/Linux using Perl modules and basic structures of programming languages. The scriptLattes is a project registered in the Free Software Competence Center at University of São Paulo, being hosted at SourceForge. To the best of our knowledge, the system is the first to be widely used in several Brazilian academic groups, including to the University of São Paulo (USP), the State of São Paulo Research Foundation (FAPESP), and the Agency for Agro-business Technology in the State of São Paulo (APTA), for instance. The system was successfully tested with at least 300 research groups of Brazilian institutions.

Therefore, the present paper describes a system that allows scientific data summarization from a strucutred database of curricula vitae, i.e. the Lattes platform. In this context, the paper describes a new system that allows the extraction of useful summarized knowledge from large set of data, a task that would be too difficult (in many cases, impossible) to be performed manually. In order to produce such a system, different solutions and algorithms have been adopted or proposed and implemented, being described in the paper. Possible applications of the system are also discussed. The paper's contributions show how Computer Science tools (developed data-structures, algorithms, visualization techniques and software system) solve an important knowledge extraction problem from large datasets. The relevance of the paper's contributions relies in the context of knowledge extraction and visualization systems that deal with possibly large datasets, an important Computer Science research topic.

The remaining of the paper is organized as follows: Section 2 discusses some important background references followed by Section 3, which describes the modules of the proposed system. Some results illustrating the use of scriptLattes, as well as a form to explore the obtained information, are described in Section 4. Finally, the conclusions and future directions are summarized in Section 5.

Background

There has been intensive growing attention to the problem of extracting meaningful summarized knowledge from large volumes of data. Papers under different perspectives (knowledge extraction, e-Science, data intensive paradigm, etc.) have appeared in most important Computer Science forums and discussed the main involved aspects, from computational tools to applications. For instance, Kouzes and collaborators describe some of the typical architecture components of a knowledge extraction systems. Some of these components inspired the tools implemented in the scriptLattes. An interest resource is the Digging into Data website, which presents a National Science Foundation (among other funding agencies) call for applications to extract knowledge based on data-driven inquiry of large sets of books (mainly on Social Sciences and Humanities). The book discusses how knowledge extraction plays a central role in modern scientific fields such as environment, health and scholarly communication. Different problems, systems and computational solutions are presented by the authors in this edited book. An important field of application for such techniques is the analysis of scholar data such as publications, supervisions, collaboration and scholar influence. The analysis of academic data is carried out in different levels, from the case where students and researchers are exploring collaborations and supervisors to institutional levels of academic assessment of whole departments. Many works have been devoted to the analysis of co-authorship and collaborations from papers databases. Many of the issues in this research appear from the low degree of structure of the data as well as the ambiguities often present. Another important issue is the analysis of the networks themselves.

It is worth noting that co-autorship is not the only approach to create scientific networks. For instance, textmining helps the generation of paper networks allowing clustering and hierarchical analyis of large datasets of papers based on subject. In all such cases, visualization is of utmost importance in order to produce good interfaces to allow the user to understand the summarized data. Visualization of networks plays a central role in helping the user to understand and to interact with data, mainly because large volumes are tipically involved.

Modules of the System

The input of the system is composed by an ASCII list of Lattes curriculum's IDs in conjunction with the time period of each member of the group to be analyzed, i.e., the years where each member has been associated to the group (research group, institute, department, university, etc.) The ID of a Lattes curriculum is a number of 16 algarisms associated to each person registered in the Lattes platform, being easily obtained from the Lattes curriculum. Therefore, the IDs are commonly used in the request of a given curriculum (...)

Data selection

This module allows to download the Lattes curricula, in HTML format, from the site of the Lattes platform. In order to facilitate the comparison between strings, the curricula were normalized to use the same characters codification.

The curricula are downloaded in HTML format because they are publicly available only as plain HTML. Public users of the Lattes platform do not have access either to the Lattes database or to data in XML format. As a result, special attention is devoted to extract the information about the scientific productions, as explained in the data preprocessing module.

Com base no texto “**SCRIPTLATTES: AN OPEN-SOURCE KNOWLEDGE EXTRACTION SYSTEM FROM THE LATTES PLATFORM”**, responda às questões de 1a 5.

Questão 01 (2,0)

Com base no texto, responda às seguintes questões:

a) Qual o objetivo do artigo?

Espera-se que o (a) candidato (a) consiga entender que o objetivo da pesquisa é descrever o design, a implementação e experiência com scriptLattes – um sistema de código aberto para criar relatórios acadêmicos de grupos de pesquisa com base nos currículos da base de dados da Plataforma Lattes.

b) Quais os módulos que compõem o sistema do scriptLattes?

Espera-se que o (a) candidato (a) entenda que as partes são os seguintes: a) seleção de dados, pré-processamento de dados, c) tratamento de redundância, d) geração de gráfico de colaboração entre os membros do grupo, e) pesquisa geração de mapas com base em informações geográficas, f) criação automática de relatórios bibliográficos, técnicos e artísticos, produção e supervisão acadêmica.

Questão 02

De acordo com o texto, responda às seguintes questões:

a) Quais são as palavras-chave do artigo?

Espera-se que o (a) candidato (a) compreenda que as palavras-chave são: a) relatório de produção acadêmica, b) plataforma Lattes, c) descoberta de conhecimento.

b) Qual o resultado do estudo?

Espera-se que o (a) candidato (a) perceba que o sistema foi amplamente testado com uma variedade de grupo de pesquisa de instituições brasileiras, os relatórios gerados evidenciam a facilidade de extração de dados da Plataforma Lattes.

Questão 03

De acordo com o texto, responda às seguintes questões:

a) Quais técnicas são utilizadas no tratamento de grandes conjuntos de dados na pesquisa em ciência da computação?
Espera-se que o (a) candidato (a) infira que o tratamento de grandes conjuntos de dados se dá por meio mineração de dados, resumo, visualização, modelagem de rede e computação de alto desempenho.

b) Por que o currículo Lattes pode ser utilizado como objeto de avaliação acadêmica no Brasil?
Espera-se que o (a) candidato (a) compreenda que o currículo Lattes pode ser usado como objeto de avaliação acadêmica porque representa a história científica, acadêmica e descreve atividades profissionais de pesquisadores brasileiros.

Questão 04

Conforme a leitura do texto, responda às seguintes perguntas:

a) Como o artigo foi organizado?
Espera-se que o (a) candidato (a) entenda que o artigo está dividido em seções. A primeira é a introdução. A segunda seção discute algumas referências teóricas. A terceira descreve os módulos do sistema proposto. A quarta, alguns resultados são apresentados ilustrando o uso do script lattes, bem como a maneira de explorar as informações obtidas. Por fim, na última seção, as conclusões e futuros direcionamentos são resumidos.

b) Qual o papel da visualização de redes no auxílio dos usuários?
Espera-se que o (a) candidato (a) compreenda a visualização de redes desempenham um papel central em ajudar o usuário a compreender e interagir com os dados.

Questão 05

De acordo com o artigo, o que o módulo seleção de dados permite realizar?

Espera-se que o (a) candidato (a) responda que o módulo de seleção de dados permite baixar os currículos Lattes, em formato HTML, do site da plataforma Lattes. Para facilitar a comparação entre as cadeias de caracteres, os currículos foram normalizados para usar a mesma codificação de caracteres. Os currículos são baixados em formato HTML porque eles estão publicamente disponíveis apenas como HTML simples.