Ciência sem Fronteiras (Science Without Borders) - 1st Technical Report

August 19, 2015

1 INTRODUCTION AND PHD PROJECT CONTEXT

The research work performed in this project is closely aligned with the ALEXANDRIA and iCrawl projects at the L3S Research Center - Leibniz Universität Hannover[1]. The L3S Research Center, the host institution of the PhD project, is one of the leading institutions focusing on basic and applied research in the field of Web Science.

The ALEXANDRIA project [2] aims at creating a foundation for temporal retrieval, exploration and analytics in Web archives. Web archives are a rich source of data since they reflect rapid evolution of the digital world. The interest in using the data stored within Web archives has been observed by researchers from different areas, such as history, sociology and law [5]. For example, the discussion in [1] indicates that Web archives are an important source for communication and media history, as well as within historiography in general. In the context of Web archives, unique Web data collections available within the ALEXANDRIA project represent a great opportunity to create new methods to explore and analyse the large scale data within Web archives. It also provides contacts with high level research scientists that has been crucial to the development of the PhD project. While ALEXANDRIA focuses on already existing data collections within Web archives, the iCrawl project[3] aims at creation of such collections efficiently on demand, using Web and Social Web data. In particular, the role of Social Media to achieve better temporal and topical coherence of the resulting collections is in the main research focus of the project.

In this reporting period, the student worked on the development of novel semantic analytics techniques to enable efficient light-weight access to large scale Web and Social Web data collections available in the context of the ALEXANDRIA and the iCrawl projects. In the following sections, we present a detailed description regarding the PhD work developed at the L3S during this period. Then we present the work plan for the second year. Table 1.1 illustrates the detailed information about the PhD candidate at the L3S Research Center.

[1] The L3S Research Center: `www.L3S.de`

[2] The ALEXANDRIA project: `http://alexandria-project.eu/`

[3] `http://icrawl.l3s.uni-hannover.de/`

Name	Tarcísio Souza Costa
Process number	201836/2014-9
Report number	1
Area	Computer Science
Sub-area	Computer Science
Institution	Instituto Federal do Maranhão - Brasil
Department	Professional Education
Professional Address	Pacas Street 5, Pinheiro Campus, Pinheiro - Maranhão, Brasil
E-mail	souza@l3s.de | tarcisio@ifma.edu.br
Research Project	ALEXANDRIA
Title	Master
Supervisor	Prof. Dr. Wolfgang Nejdl | nejdl@l3s.de
Advisor/Mentor	Dr. Elena Demidova | demidova@l3s.de
Period of Evaluation	October, 2014→July, 2015

Table 1.1: Student description

2 DESCRIPTION OF RESEARCH ACTIVITIES

2.1 Research and development activities

In the first year of the PhD project, the student has been involved in some of the key projects at L3S that focus on analytics of the Social Web and Web content, including the iCrawl and the ALEXANDRIA projects.

2.1.1 R&D activities in the context of the iCrawl project

In this time period, the goal of the iCrawl project was to study how Social Web signals can be used to support focused collection of fresh Web content relevant to the current events, such as an Ebola outbreak, or a military conflict in Ukraine, in an effective and efficient way. In iCrawl, Social Web signals, in this case Twitter messages, have been used to guide the Web crawler towards the newly created pages [4]. In this initial phase, the student studied the literature related to the problems of focused crawling and Social Web analytics. He also became familiar with the software developed in the context of the iCrawl project as well as with the Big Data research infrastructure of the L3S Research Center. Also in this period, the student started to learn new Big Data technologies, such as Hadoop MapReduce, Hive and Pig Latin.

Large scale Web data collections, like those created in the iCrawl project, or, more generally, data contained within Web archives, require efficient access techniques that can enable to efficiently locate relevant Social Web and Web content stored within the archive to enable further research. The goal of providing efficient light-weight access to large scale Web collections lead the student to joining the ALEXANDRIA project at L3S later on as this project explicitly focuses on providing efficient entity-centric access to large scale Web collections. An important question on the interconnection of the both projects is how to use current Web to discover resources within Web archives.

A URL discovery algorithm: In order to provide efficient URL-based access to the Web archive data in the ALEXANDRIA context, the student developed a URL discovery algorithm that maps a set of URLs from the current Web to the corresponding URLs from the Web archive. In this period

the student investigated several methods to optimize the URL discovery algorithm using Big Data technologies, such as the join optimization using Hive tables and an integration with Apache HBase.

2.1.2 R&D activities in the context of the ALEXANDRIA project

The student became familiar with the work plan of the ALEXANDRIA project and joined Work Package 1 (Evolution-Aware Entity-Based Enrichment and Indexing), where he focused on creation of novel efficient light-weight methods for entity-centric access to Web archives. In this context, he became familiar with the state-of-the-art methods and tools in the field of Named Entity Recognition (NER), and other information extraction tools.

A light-weight full-text URL-based index: As full-text indexing of large scale Web archive data is computationally very expensive, the student developed a light-weight full-text index that only takes the URLs into account. To this extent, the student has written an analyzer for correct pre-processing and parsing of URLs to detect keywords. At the same period the student also learned a distributed indexing technology, named Elastic Search. The index was developed and made available in Kibana (a user-friendly Elastic Search graphical user interface). The student developed categorization of URL search results by domain and mime type, performed index optimization such as setting some fields as analyzed by the search engine as well as refinement of search engine results to retrieve documents in certain periods of time. The index was further improved by inserting mime type, domain and improvements related to temporal expressions automatically extracted from the URLs, also changes in the main structure were performed. This index should be used as a starting point for the iCrawl to find relevant documents within the archive.

Named entity extraction from URLs: Further semantic analytics of the URLs included named entity extraction from URLs. The paper describing this work was accepted for publication at the 1st International Keystone Conference [1] at the beginning of July 2015. In this paper, the student analysed a subset of URLs, named as Popular German Web. Such a subset is related to popular domains according to Alexa ranking[2]. In this paper we measure the quantity of URLs that are alive (the status code starts with "2") per year and per domain, also counting in which domains categories they are included, as Sports, Games, Computes and so on. We also extracted temporal expressions and entities using only the information within URLs. We continue working on this topic to further improve the results, for instance by adding better pre-processing and post-filtering procedures in the URLs before applying state of the art techniques to extract entities, as Stanford Named Entity Recognition (NER) [3]. Using those procedures we reach a precision up to 85% in our dataset. Further improvements such as evaluating precision and filtering on a sample of distinct URLs, instead of captures and presentation of a more detailed description of how the precision was measured are currently being performed by the student as a part of the camera-ready preparation and to address the review comments.

[1] http://www.keystone-cost.eu/ikc2015/about.php

[2] http://www.alexa.com

[3] http://nlp.stanford.edu/software/CRF-NER.shtml

2.1.3 Artificial Bee Clustering Search (ABCS) algorithm

In addition to the research activities and publications in the ALEXANDRIA project, during this reporting period the student received an acceptance message from an extended version of a paper to be published in a journal, the Artificial Bee Clustering Search (ABCS) algorithm [2]. After the final version submitted, the journal paper was published [3].

2.2 Integration in the L3S research environment

At the beginning of the first year at the L3S Research Center, the student joined the team of the iCrawl project lead by Dr. Elena Demidova, who was appointed as the student advisor by Prof. Wolfgang Nejdl. Later on, the student joined a larger group of researchers working on the ALEXANDRIA project.

A part of the student integration in the L3S research environment is the regular meetings with the advisor (once a week), and with Prof. Nejdl (once a month) as well as regular participation in the iCrawl and ALEXANDRIA project meetings. These meetings are very helpful for the student to obtain feedback on his research work, in particular with respect to further improvement of the results and promising research directions for future work.

During this reporting period, the student made two presentations of the research and development results in the ALEXANDRIA meetings: The student presented the URL index using Kibana interface during the ALEXANDRIA meeting on March 5th, 2015. The paper accepted for the 1st International Keystone Conference was presented by the student on June 18th, 2015. These presentations provided valuable feedback for the student work from a larger group of the L3S researchers involved in the project.

2.3 Community services

The student worked as a sub-reviewer under supervision of Dr. Thomas Risse for the Joint Conference on Digital Libraries 2015 (JCDL 2015). That was a good experience since the paper student reviewed was related to the PhD work and was very useful to improve the student knowledge in the field.

2.4 Courses

In winter semester, the student followed the course “Foundations of Information Retrieval" including lectures and exercises offered by Prof. Nejdl. This lecture provides a solid foundation to better understand Information Retrieval concepts and terminology used in the context of the ALEXANDRIA project and also to become familiar with the typical evaluation methods.

In summer semester, the student participated in the course “Advanced Methods of Information Retrieval” conducted by Dr. Elena Demidova. In this course Information Retrieval specific methods were presented from the practical point of view, including implementation exercises and challenging research questions during the lectures. This course was very useful for the student to become familiar with a diversity of relevant topics in the field as well as open research directions relevant to the doctoral study.

At the moment there are no grades for these courses, that is the reason the doctorate transcript will not be provided, only in next annual report. The student will take the information retrieval exams still in current month and on September.

2.5 Accepted publications

Title: Semantic URL Analytics to Support Efficient Annotation of Large Scale Web Archives.
Authors: Tarcísio Souza, Elena Demidova, Thomas Risse, Helge Holzmann, Gerhard Gossen and Julian Szymanski.
To appear in: Proceedings of the 1st International Keystone Conference, Coimbra, Portugal, September 2015.
Abstract: "Long-term Web archives comprise Web documents gathered over longer time periods and can easily reach hundreds of terabytes in size. Semantic annotations such as named entities can facilitate intelligent access to the Web archive data. However, the annotation of the entire archive content on this scale is often infeasible. The most efficient way to access the documents within Web archives is provided through their URLs, which are typically stored in dedicated index files. The URLs of the archived Web documents can contain semantic information and can offer an efficient way to obtain initial semantic annotations for the archived documents. In this paper, we analyse the applicability of semantic analysis techniques such as named entity extraction to the URLs in a Web archive. We evaluate the precision of the named entity extraction from the URLs in the Popular German Web dataset and analyse the proportion of the archived URLs from 1,444 popular domains in the time interval from 2000 to 2012 to which these techniques are applicable. Our results demonstrate that named entity recognition can be successfully applied to a large number of URLs in our Web archive and provide a good starting point to efficiently annotate large scale collections of Web documents."

3 CONCLUSION AND OUTLOOK

During the first reporting period, the student worked on the problems of light-weight semantic annotation of large web collections in the context of the ALEXANDRIA and the iCrawl projects. The student learned state-of-the-art Big Data processing technologies, developed novel approaches for Web data analytics, and published the results.

The first period at the L3S Research Center was a really rich experience, since the student had an opportunity to study different research areas, learn new technologies as well as working with top-level researchers. During this period the student could find a well defined working approach implemented by the L3S, where regular meetings with the supervisor (one per month) and the advisor (one per week) were performed, which was very useful to improve the student knowledge, also helping him to improve the PhD work that has been developed in last months.

In the second year, the student will continue research in the context of the ALEXANDRIA and the iCrawl projects. In particular, he will further develop and evaluate semantic analysis techniques for URLs, such as terms, named entities and temporal expressions extraction and correlation. Furthermore, he will develop novel methods to identify interconnections between such semantic information spread across URLs to enable light-weight URL-based sub-collection extraction of information related to events from large unfocused Web data collections. In this context the student will rely on machine learning and statistical techniques as well as soft computing - based algorithms.

BIBLIOGRAPHY

[1] Niels Brügger. Probing a nation's web sphere: A new approach to web history and a new kind of historical source, 2014.

[2] Tarcísio Souza Costa and Alexandre César Muniz de Oliveira. Artificial bee clustering search. In Ignacio Rojas, Gonzalo Joya Caparrós, and Joan Cabestany, editors, *IWANN (2)*, volume 7903 of *Lecture Notes in Computer Science*, pages 20–27. Springer, 2013.

[3] Tarcísio Souza Costa and Alexandre Cesar Muniz de Oliveira. Artificial bee and differential evolution improved by clustering search on continuous domain optimization. *Soft Computing*, pages 1–12, 2014.

[4] Gerhard Gossen, Elena Demidova, and Thomas Risse. icrawl: Improving the freshness of web collections by integrating social web and focused web crawling. In Paul Logasa Bogen II, Suzie Allard, Holly Mercer, Micah Beck, Sally Jo Cunningham, Dion Hoe-Lian Goh, and Geneva Henry, editors, *JCDL*, pages 75–84. ACM, 2015.

[5] Thomas Risse, Elena Demidova, and Gerhard Gossen. What do you want to collect from the web? In *Proc. of the Building Web Observatories Workshop (BWOW) 2014*, 2014.