PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

Researchers Exchange Report
Using Saffron for Measuring the Impact of Evaluation Campaigns

NUIG – RSLIS

10/15/2012

SEVENTH FRAMEWORK PROGRAMME

Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191

Document Information

Report title:	Using Saffron for measuring the impact of evaluation campaigns
Researcher Exchange date:	10/15/2012
Visitor(s):	Georgeta Bordea (NUIG)
Host(s):	Birger Larsen (RSLIS)
Preparation date:	10/30/2012
Author(s):	Georgeta Bordea, Birger Larsen, Toine Bogers

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
1.0	11/06/2012	Draft	NUIG	Circulated to all partners

Table of Contents

Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191

1 Introduction

The purpose of this visit was to present the CLEF instance of Saffron[1], set up by NUIG as part of task 3.6 from Work Package (WP) 3, and to discuss possible directions for joint research between WP3 and WP6. In PROMISE, NUIG is mainly concerned with developing methods for matching experimental data with underlying publications, extracted topics and experts in different research areas. These goals are investigated using Saffron, a system that provides insights in a research community or organization by analysing its main topics of investigation and the experts associated with these topics [1]. The results of the topical analysis performed by Saffron on CLEF working notes are a promising resource for assessing scholarly impact of scientific campaigns. One of the directions that were discussed during this visit is the identification of research communities. The identification of research communities that cite publications from a scientific campaign would allow a more in depth analysis of impact.

2 Conducted Work

The conducted work focused on using Saffron as a tool for impact analysis (Section 2.1). Also discussed as part of this visit was a collaboration in the area of expert search in scientific communities (Section 2.2).

2.1 Impact Analysis of the CLEF campaign

A first step in measuring the impact of a research activity is to collect a dataset about its publications and the citations they receive. Several sources can be considered including commercial products (e.g., Web of Science, Scopus) or freely available services such as Google Scholar. At the same time, a considerable number of datasets about publications are available as Linked Open Data, but only a limited number of these datasets include citation information. In addition to academic publications and citation networks, we need access to full content publications or abstracts in order to apply Saffron. During this visit we analysed suitable datasets for analysing the impact of CLEF on other research communities.

2.1.1 Open datasets for impact analysis

A promising dataset that has a wide coverage of the Computer Science domain is CiteSeer[2]. This dataset is appropriate for the purpose of our investigation as it gives access to full content papers and it provides citation information that is needed for assessing scholarly impact. A preliminary investigation revealed that a search with the keyword “CLEF” in CiteSeer returns about 2000 publications and about 2700 results including citations, while a more restrictive search using both the “CLEF” keyword and the “retrieval” keyword returns about 1300 publications and about 1400 results including citations. Another reason that would recommend CiteSeer is that NUIG has access to a

[1] The CLEF instance of Saffron: http://saffron.deri.ie/clef
[2] CiteSeer: http://citeseerx.ist.psu.edu

Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191

full dump of CiteSeer from 2010. Two other open datasets which fulfil our requirements are GESIS and ACM. RSLIS mentioned that it has access to a dump of arXiv.org of more than 300 k publications with automatically identified citation links.

2.1.2 Identification of research communities

The identification of citing research communities will allow us to analyse connections between CLEF and other communities and areas of research. Knowledge transfer and uptake enabled by a scientific campaign can be estimated by identifying influential topical areas picked up by other communities. By analysing the connections between research areas at different time slices, we will attempt to identify newly created connections between previously disconnected research areas. Georgeta Bordea presented the CLEF and the Information Retrieval (IR) instances of Saffron as well as an automatically built generalisation hierarchy of IR concepts. It was agreed that Saffron can be used to identify and analyse citing communities. Another direction of research, which was identified during the course of this visit, is to improve the existing algorithm for hierarchy construction by identifying main clusters in the citation graph

2.1.3 Citation context analysis

Another direction of research, briefly discussed during this visit, was the analysis of citation context. An appropriate dataset is not currently available and should be acquired through crowdsourcing. This dataset will allow us to investigate how much context is needed as well as different ways in which papers are cited (e.g., *agree, disagree, use*), similar to valence and subjectivity in sentiment analysis.

2.2 Expert search

2.2.1 Workshop committee members dataset

NUIG and RSLIS independently gathered two datasets on workshop committee members from research areas including Semantic Web, Computational Linguistics, Information Retrieval, Digital Libraries and Recommender Systems. It was considered opportune to align these datasets and publish a joint work that introduces a new dataset for Expert Search. The workshops dataset will be a valuable resource for evaluation, as it provides a gold standard for expert finding and a silver standard for expert profiling. This dataset has the advantage to be readily available and easy to collect but it is expected that expert profiling results will have a low upper bound due to incompleteness and partial matching of expertise topics. Another limitation of the dataset is its subjectivity, as committee members can be selected based on their generic expertise not necessarily on their interest on the specific area of the workshop. A corpus of publications will be provided for each research area and we will release standard performance results for two baselines (Saffron, a traditional IR approach) along with relevance assessments.

2.2.2 Expert Recommendation

Typically, expert recommendation is done using different measures of similarity between people [2]. Finding experts with complementary skills is an alternative direction for expert recommendation that did not receive much attention. We will investigate this approach on a dataset about research projects from the University of Tilburg.

3 References

[1] Georgeta Bordea, Sabrina Kirrane, Paul Buitelaar, Bianca Pereira. Expertise Mining for Enterprise Content Management. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* 2012.

[2] Wongchokprasitti, Chirayu and Brusilovsky, Peter and Parra-Santander, Denis (2010) Conference Navigator 2.0: Community-Based Recommendation for Academic Conferences. In: Workshop on Social Reminder Systems (SRS '10), 7 February 2010, Hong Kong, China.