Topic modeling and hypergraph mining to analyze the EGC conference history

Adrien Guille*, Edmundo-Pavel Soriano-Morales*, Ciprian-Octavian Truica**

*Laboratoire ERIC, Université Lumière Lyon 2
{adrien.guille;edmundo.soriano-morales}@univ-lyon2.fr
**Computer Science dept., University Politehnica of Bucharest
ciprian.truica@cs.pub.ro

Abstract. Each year the EGC conference gathers researchers and practitioners from the knowledge discovery and management domain to present their latest advances. This year's edition features an open challenge that encourages participants to leverage the EGC rich anthology which spans from 2004 to 2015. The ultimate goal is to highlight the dynamics of the conference history and to try to get a glimpse of the coming years. In this context, we first describe our methodology for inferring latent topics that pervade this corpus using non-negative matrix factorization. Based on the discovered topics and other properties of the articles (e.g., authors, affiliations) we shed light on interesting facts on both the topical and collaborative structures of the EGC society. Secondly, we employ a hypergraph itemset extraction process to discover existent but latent relations between authors or between topics. We also propose topic-author and author-author recommendations with a content-based approach. Lastly, we describe a Web interface for browsing this collection of articles complemented with the discovered knowledge.

1 Introduction

In this article, we describe work done in the context of the first edition of the EGC challenge, which ultimate goal is to highlight the dynamics of the conference history and to try to get a glimpse of the coming years.

Dataset Participants of the challenge are provided with the descriptions of 1935 articles published by RNTI, out of which 1041 are articles presented at the EGC conference. Each article is described by several fields: year, title, abstract (potentially missing), list of authors, and a URL pointing at the first page of the article (potentially missing or unreachable). When it is possible, we enrich the descriptions of the articles with (i) the language detected from their abstracts and (ii) the authors' affiliations. Language detection relies on a naive Bayes classifier, trained on Wikipedia articles covering French and English, using n-grams of characters as features (Cavnar and Trenkle, 1994). The identification of authors' affiliations relies on regular expressions to match e-mail addresses in the content of the first page of the articles, assuming