Identifying Diachronic Topic-Based Research Communities by Clustering Shared Research Trajectories

Francesco Osborne, Giuseppe Scavo, and Enrico Motta

Knowledge Media Institute, The Open University, MK7 6AA, Milton Keynes, UK
{f.osborne,g.scavo,e.motta}@open.ac.uk

Abstract. Communities of academic authors are usually identified by means of standard community detection algorithms, which exploit 'static' relations, such as co-authorship or citation networks. In contrast with these approaches, here we focus on *diachronic topic-based communities* –i.e., communities of people who appear to work on semantically related topics at the same time. These communities are interesting because their analysis allows us to make sense of the dynamics of the research world –e.g., migration of researchers from one topic to another, new communities being spawn by older ones, communities splitting, merging, ceasing to exist, etc. To this purpose, we are interested in developing clustering methods that are able to handle correctly the dynamic aspects of topic-based community formation, prioritizing the relationship between researchers who appear to follow the same *research trajectories*. We thus present a novel approach called *Temporal Semantic Topic-Based Clustering (TST)*, which exploits a novel metric for clustering researchers according to their research trajectories, defined as distributions of *semantic topics* over time. The approach has been evaluated through an empirical study involving 25 experts from the Semantic Web and Human-Computer Interaction areas. The evaluation shows that TST exhibits a performance comparable to the one achieved by human experts.

Keywords: #eswc2014Osborne, Community Detection, Scholarly Data, Scholarly Ontologies, Semantic Technologies, Clustering, Similarity Metrics, Fuzzy C-Means.

1 Introduction

Communities of academic authors are usually identified by using standard community detection algorithms, which typically exploit co-authorship or citation graphs [1]. However, an interesting type of community, which has received much less attention in the literature [2], is formed by the set of researchers who, at a given time, are working on the same topic. Obviously, this type of *topic-based community* has a degree of overlap with co-authorship and citation communities; nonetheless it provides a distinct way of identifying groups of related researchers. Co-authorship communities can certainly be seen as examples of topic-based communities, however one does not need to co-author with another researcher in order to be part of the same

V. Presutti et al. (Eds.): ESWC 2014, LNCS 8465, pp. 114–129, 2014.

topic-based community. Hence, co-authorship networks only provide an incomplete view of a topic-based community. In addition co-authorship relations can span different topics, hence providing a noisy mechanism to identify a topic-based community. An analogous argument applies to the use of citation networks to identify topic-based communities: on the one hand citations may cut across different topics and on the other hand there is no guarantee that people working on the same topic actually cite each other. Hence, citation networks also define poor approximations of topic-based communities.

Topic-based communities are interesting because their analysis allows us to make sense of the dynamics of the research world –e.g., migration of researchers from one topic to another, new communities being spawn by older ones, communities (and therefore associated topics) splitting, merging, ceasing to exist, etc. More precisely, the formal identification and characterization over time of topic-based communities allows us to give an extensional computational treatment of a topic (or set of topics), say T, in terms of all the researchers and publications related to T at a given time. Thus, we can then measure precisely the size of the topic, its scientific impact (in terms of a variety of academic impact measures), its evolution, relations between topics in terms of overlap of researchers, migrations across topics, etc. In the rest of the paper we will use the term *temporal topic-based community* to refer to this type of communities.

In this paper we propose a novel approach to identifying temporal topic-based communities, called *Temporal Semantic Topic-Based Clustering (TST).* TST exploits a novel metric, called *ATTS* (*Adjusted Temporal Topic Similarity*), which measures the similarity between *research trajectories*. These are in turn defined as distributions of *semantically-characterized topics* over time –i.e., topics structured in terms of semantic relationships, such as *skos:broaderGeneric* or *relatedEquivalent* [3]. Thus, TST is able to detect *diachronic* groups of authors with similar behavior over a period of time.

An important aspect of TST is that, in contrast with methods which rely on co-authorship or citation networks, it does not require a complete graph of relations between community members. Hence, it can also be used in non-academic contexts, where such relations are typically not available. In addition, we characterize temporal topic-based communities as fuzzy clusters and as a result each author is then associated with a set of membership values, which express the degree of work done for different communities. Hence, this model naturally handles both the common situation in which an author contributes to more than one community and also the situation in which a community is defined in terms of multiple dynamic topics over time –e.g., the community of all researchers who worked in Knowledge Acquisition during the 90s and then worked primarily on the Semantic Web during the 00s.

Our approach increases the granularity of the representation of the research environment and makes it possible to discover interesting dynamics. For example, we can highlight the behaviour of groups of researchers reacting to a mutation in the scientific environment, such as the introduction of a new technology (e.g., Mobile Devices), a new vision (e.g., Semantic Web), or a grant on a particular theme (e.g., Smart Cities). We can also get interesting insights into the 'DNA' of specific communities. For example, a topic-centred analysis of Semantic Web (SW) researchers over time reveals that the authors with a World-Wide-Web (WWW) background, who joined the SW research area in the first years of this century, were by and large the ones who progressed the Linked Data topic at the end of the decade.