A Hybrid Semantic Approach to Building Dynamic Maps of Research Communities

Francesco Osborne, Giuseppe Scavo, Enrico Motta

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

Abstract. In earlier papers we characterised the notion of *diachronic topic-based communities* –i.e., communities of people who work on semantically related topics at the same time. These communities are important to enable topic-centred analyses of the dynamics of the research world. In this paper we present an innovative algorithm, called Research Communities Map Builder (RCMB), which is able to automatically link diachronic topic-based communities over subsequent time intervals to identify significant events. These include topic shifts within a research community; the appearance and fading of a community; communities splitting, merging, spawning other communities; etc. The output of our algorithm is a map of research communities, annotated with the detected events, which provides a concise visual representation of the dynamics of a research area. In contrast with existing approaches, RCMB enables a much more fine-grained understanding of the evolution of research communities, with respect to both the granularity of the events and the granularity of the topics. This improved understanding can, for example, inform the research strategies of funders and researchers alike. We illustrate our approach with two case studies, highlighting the main communities and events that characterized the World Wide Web and Semantic Web areas in the 2000 – 2010 decade.

Keywords: Semantic Web, Community Detection, Change Detection, Trend Detection, Pattern Recognition, Data Mining, Scholarly Data.

1 Introduction

Understanding the dynamics of research communities is a challenging and important task. As Yan et al [1] point out, there is a need for "better tools for identifying emergent trends and the development of new scholarly communities." In particular, we need to be able i) to capture effectively the relationships between authors and *research topics*, ii) to learn how these evolve over time, and iii) to understand how research communities interact with each other, trading topics and researchers.

Current approaches to these tasks, e.g., [1,2,3,4,5], suffer from a number of limitations. First, they model research topics simply as keywords, failing to capture the semantic relationships that can exist between topics and thus reducing the opportunities for a sophisticated modelling of topic spaces [6]. They also tend to focus on 'static snapshots' of research communities [2,4,5] and do not take into account the dynamic aspects associated with the evolution of the *research trajectories* of authors (i.e., which topics they work on over time). As we know, communities are dynamic entities that may merge, split and evolve by incorporating new topics and

visions [1,7,8]. By analysing the trajectories of different authors and identifying similarities between them it becomes possible to go beyond such static snapshots, to develop models of the evolution of topic-centred research communities. This approach makes also possible to estimate the effects of specific external events (e.g., grants, new technologies, historical events) on the research dynamics.

In Osborne et al. [9] we took a first step to addressing the aforementioned limitations by presenting Temporal Semantic Topic-Based Clustering (TST), an algorithm which identifies *temporal topic-based communities*, i.e., communities grouping authors who share similar research trajectories.

In this paper we present an innovative algorithm, called Research Communities Map Builder (RCMB), which is able to automatically link diachronic topic-based communities over subsequent time intervals to identify significant events. These include topic shifts within a community; the appearance and fading of a community; communities splitting, merging, spawning other communities; etc. The output of RCMB is a map of research communities, annotated with the detected events, which provides a concise visual representation of the dynamics of a research area. RCMB is integrated in Rexplore [10], a system that capitalises on innovative integrated solutions in semantic technologies and data mining to support users in exploring and making sense of scholarly data.

These research maps can be important for a variety of tasks and users. For example, they can help research managers to understand which are the main research communities in a field, in which directions they are evolving, and how they interact, to inform critical decisions on funding and recruitment policies. Editors can use research maps to ensure timely editorial decisions, e.g., by detecting emerging themes for special issues. Researchers can use them for making sense of new trends and identifying appropriate collaborators. Moreover, by analysing the history of research communities it is possible to learn good practices. For example, we can investigate how scientific communities adapt and cooperate to implement visions into concrete technologies (as the Semantic Web community did with Linked Data) and try to replicate successful stories. Finally, discovering how groups of researchers are influenced by technological breakthroughs, grants or visions, and how they interact to generate new ideas and technologies, can allow us to predict future research trends.

We will illustrate our approach with two case studies, highlighting the main communities and events that characterized the World Wide Web (WWW) and Semantic Web (SW) areas in the 2000 – 2010 decade.

The rest of the paper is organized as follows. In Section 2, we discuss the limitations of current approaches to identifying the dynamics of research communities. Section 3 describes RCMB in detail. In Section 4 we illustrate our case studies, highlighting key community trends within the two aforementioned research areas. In Section 5 we evaluate our approach and in Section 6 we conclude by restating the main contributions of this work and outlining future directions of research.

2 State of the Art

Community detection is a popular topic in computer science and has been addressed using a variety of techniques. Most of this work deals with graphs (e.g., co-authorship networks) and their topological structure rather than exploiting the topic similarity of

the members. Current approaches to community detection are usually classified according to the strategy they employ [5], as either heuristic [11] or optimization-based methods [12]. However, from the point of view of capturing coherent topic-centric research communities, both fail to correctly cover all people who work in a research area (lack of recall) and also typically include individuals interested in different topics (lack of precision).

More recently, there has been a growing interest in topic-based communities [1], i.e., communities generated on the basis of topic similarities between individuals, rather than on their connections in a graph. These communities can in fact be used to describe the evolution of topics in a research area [7]. For example, Upham et al. [3] presented an analysis of "knowledge communities", defined as intellectually cohesive, organic inter-organizational forms. Similarly to our approach, they detect communities by using a clustering scheme that produces dynamic clusters over a timeline. However, they focus mainly on the citation graph and on language-level similarities between publications to identify communities. In contrast with this approach, we focus instead on identifying the research trajectories shared by groups of authors. Zhao et al. [5] identify topic-based communities by analyzing co-authorship networks. Their method lacks however the temporal dimension, which is needed for detecting research trends. Other techniques focus on identifying dynamic trends within an individual topic. For example, Racherla and Hu [13] identify topic communities by exploiting a topic similarity matrix and assigning a predefined research topic to each document and author. This approach however fails to take into account the fact that most research communities cannot be described by a single topic, being most often characterized by a distribution of related topics [7].

Topic models, which compute the similarity of entities according to shared keywords, can also be seen as a form of community detection. These methods usually rely on the detection of latent topics for capturing semantic dependencies between keywords and exploit Probabilistic Latent Semantic Indexing (pLSI) [14] or Latent Dirichlet Allocation [15]. For example the Author-Conference-Topic (ACT) model [16] treats authors as probability distributions over topics, conferences and journals extracted by means of an unsupervised learning technique. Our approach is instead based on a semantic characterization of research topics, which are connected by three types of semantic relations [6].

RCMB integrates TST [9], which adopts a Fuzzy C-Means (FCM) [17] clustering algorithm that exploits a novel similarity metric on a vector of *semantic topics*. Similarly, Yan et al. [4] use a K-Means algorithm to examine a paper-to-paper network based on shared word relations, while Van Eck and Waltman [2] exploit modularity-based clustering techniques for detecting research communities. Differently from these approaches, we focus on authors' research trajectories rather than on papers, as our aim is to identify diachronic communities and their dynamics, rather than simply detecting the existence of a community in a certain time interval. Another important feature of RCMB is its ability of linking together communities over time intervals. Similarly, Yan at al [1] use cosine similarity over the topic distributions and empirically set thresholds to detect links between communities. RCMB however brings this idea further by taking into account also the migrations of authors from one community to another and finding automatically the relevant thresholds by minimizing an evaluation function with the Nelder-Mead algorithm [18]. In addition, RCMB relies on the chi-square test and a sliding window algorithm

to detect significant changes in the topic distribution of a community. Similar statistical approaches can be found in the change detection research area, such as scene change detection [19].

3 The Research Communities Map Builder

In this section we will describe in detail the Research Communities Map Builder algorithm, which generates a map of the main communities within a given topic.
Most methods for detecting topic-based communities [1,2,4,7] use keywords as proxies for research topics. Hence, pairs of obviously related terms, e.g., "Semantic Web" and "Linked Data", are considered as unrelated entities. We instead rely on the Klink algorithm [6] to address this issue and generate an ontology describing a structured set of semantic topics. Specifically, Klink is able to infer three kinds of semantic relationships between keywords, which are *skos:broaderGeneric* (topic, T_1 is a sub-topic of topic T_2), *relatedEquivalent* (two topics are alternative names for the same research area) and *contributesTo (*research in topic T_1 is an important contribution to research in topic T_2*)*. Hence, the topic vectors representing the publications are modified by labelling with topic T_1 any publication tagged with topic T_2, if T_2 is sub-topic of T_1 or equivalent to T_1.

RCMB takes as input i) a research area (e.g., Semantic Web); ii) a time interval (e.g., 2000-2006); iii) a granularity index (e.g., 3), which will determine the dimension of the intervals to be processed (e.g., 2000-2002, 2002-2004, 2004-2006); iv) a number of authors *active* in the research area, associated with their publications; and v) a knowledge base, automatically populated by Klink, specifying the semantic topics and their relationships. We define as active in a research area, say *R*, any author who has at least one publication a year in *R*, during the time interval considered.

The output is a map of the research communities within the research area in question, describing for each year of the time interval the existing communities, the topic shifts (i.e., changes in some of the major topics which characterize a community); and a number of other significant events that alter the research landscape, such as the splitting of a community. RCMB relies on statistical methods to infer topic shifts and on heuristic rules to detect six kinds of important events. More formally, RCMB is described by the following pseudo code:

```
Function RCMB (main_topic, time_interval, granularity_index, authors, topic_kb) {
   // Split the time interval in multiple intervals with a length equal to the granularity
   // and the last year of an interval being the first year of the subsequent interval
   time_intervals = splitTimeInterval(time_interval, granularity);
   // Extract the semantic topic vectors from the publication, see Sec. 3.1
   stv = semanticTopicExtraction(authors, topic_kb);
   // Weights each topic accordingly to its similarity with main_topic, see Sec. 3.1
   wstv = topicWeighting(stv, main_topic);
foreach (time_intervals as t) {
      // Clusters the authors in fuzzy communities using FCM, see Sec. 3.2
      initial_centroids = substractiveMethod(wstv[t]);
      communities[t] = FCM(wstv[t], initial_centroids);
   }
   // Infer the shifts of interest in the returned communities, see Section 3.3
```

```
shifts = inferTopicsShifts(communities);
// Links communities in different time intervals, see Sec. 3.4
s_links = estimateStrongLinks(communities, shifts);
w_links = estimateWeakLinks(communities, shifts);
// Detect additional shifts of interest, see Section 3.3
shifts = inferAdditionalTopicsShifts(shifts, communities, s_links);
// Detect key events, see Sec. 3.5
events = inferSignificantEvents(communities, s_links, w_links, shifts );
map = buildResearchMap(communities, s_links, w_links, shifts, events);
return map; }
```

In the following sections we will discuss in details the steps of the algorithm. The first three steps (described in Section 3.1 and 3.2) are actually a slight modification of the TST algorithm [9], thus they will be explained only briefly here.

3.1 Semantic Topic Integration and Topic Vector Weighing

In this step we extract the semantic topics from the publications to represent each author as a semantic topic distribution over subsequent years. We do so by summing the semantic topic vectors of an author's publications in each year. Then, for each pair of topics sharing a *contributesTo*(T_1,T_2) relationship in the topic vector, we assign to T_2 a fraction of the publications in T_1 according to the formula:

$$CT(T) = \sum_{i=1}^{n} P\big(T \big| ct(i,T)\big)^{\varphi}$$

where $ct(i,T)$ indicates the set of topics associated with the *i-th* publication that is in a *contributesTo* relationship with *T*. $P\big(T\big|ct(i,T)\big)$ is the probability that a paper with such a set of topics is also explicitly associated with area *T* (or with a topic having a *broaderGeneric* or *relatedEquivalent* relationship with *T*) at the publication date of the *i-th* paper. The summation is carried out over the number *n* of publications that are not already associated with *T* but have at least one topic in a *contributesTo* relationship with *T*.

To privilege the communities strongly related to the research area given as input to RCMB (the "main topic"), we weigh each topic according to its relationship with the main topic. Given a topic *T*, the weight *W(T)* is calculated as follows:

$$W(T) = 1 + k\frac{C(T)}{S(T)}$$

where *C(T)* is the number of co-occurrences of topic *T* with the main topic in the selected time interval; *S(T)* is the number of total occurrences of the topic *T* in the selected time interval, and *k* is an arbitrary constant that can be tuned to amplify the effect of the weight on the system.

3.2 Fuzzy Clustering Based on Temporal Topic Similarity

We employ a Fuzzy C-means algorithm (FCM) [17] over the *weighted semantic topic vectors* in the timeframe to compute the clusters. We use a fuzzy clustering technique since researchers are often members of multiple communities.

For clustering authors according to their shared topic trajectory over the years we use a similarity metric called *ATTS* (*Adjusted Temporal Topic Similarity*). ATTS is computed by averaging the cosine similarities of the semantic topic vectors over progressively smaller intervals of time. We first define the *TTS* (*Temporal Topic Similarity)* between author *A* and author *B* in the interval t_1-t_2 as:

$$TTS(A,B,t_1,t_2) = \sum_{i=0}^{\lfloor log_2(t_2-t_1)\rfloor} \frac{\left[\left(\sum_{j=0}^{2^i-1} \mathrm{TS}\left(A,B,t_1+\left\lfloor\frac{j\cdot(t_2-t_1)}{2^i}\right\rfloor, t_1+\left\lfloor\frac{(j+1)\cdot(t_2-t_1)}{2^i}\right\rfloor\right)\right)/2^i\right]}{\lfloor log_2(t_2-t_1)\rfloor}$$

where *TS*(A,B,t_1, t_2) = cos($\sum_{i=t_1}^{t_2} \hat{a}_i$, $\sum_{i=x}^{t_2} \hat{b}_i$), $\hat{a}_i$ and $\hat{b}_i$ are the topic vectors of the two authors in the i-th year and *cos(s,t)* is the cosine similarity.

To account for those cases in which an author may not be present in all the years of the timeframe, we define the *adjusted temporal topic similarity* ATTS as:

$$ATTS(A,B,t_1,t_2) = TTS(A,B,t_1,t_2) \cdot \frac{I_s^\gamma}{I_s^\gamma + I_{ns}^\gamma} + P \cdot \frac{I_{ns}^\gamma}{I_s^\gamma + I_{ns}^\gamma},$$

where I_s is the number of years in which both authors were active, I_{ns} is the remaining number of years, *P* is a constant equal to the average TTS of *n* random couples of authors in the given timeframe (*n=500* in the prototype), and γ>1 a parameter for weighing their relationship (γ= 2 in the prototype).

Since ATTS is a similarity metric, while a FCM needs a distance, we use as norm the inverse of the ATTS minus 1.

In order to choose the candidate centroids for the FCM algorithm we use the subtractive clustering method [20], which estimates the initial centroids by assigning a potential to each individual in the dataset according to its number of neighbours. FCM returns a list of cluster centroids and a partition matrix where each author is associated with its degree of membership to each cluster. The returned centroids are actually the topic vectors of the communities in each year, allowing for an easy detection of shifts and changes throughout the years. By summing the vectors over the years and selecting the topics with the highest values, it is possible to label communities in terms of their most significant topics.

3.3 Detecting Topic Shifts

One of the main advantages of using communities characterized by an evolving distribution of topics is that we can study the dynamic changes in the research interests of the community. In this section we describe an automatic method which detects that a certain community underwent a topic shift and outputs an explanation. As shown in the pseudo code, RCMB infers community topic shifts in two occasions. The first time is before linking the communities (*inferTopicsShift* function), because these initial topic shifts will be used to estimate the threshold value for the linking process (see Section 3.4). The procedure is then run again on the linked communities (*inferAdditionalTopicsShifts* function), to allow the detection of topic shifts across the given time intervals.

We define a *topic shift* as a statistically significant change in the topic distribution of a community which occurred in a certain time interval. We compute the null hypothesis *p* that the difference between two topic distributions of the same community in two subsequent time t_1 and t_2 are due to random fluctuations by using

the *chi-square test*; if $p \leq 0.05$, the two distributions are considered different and a topic shift is detected in the time interval [t_1-t_2].
After detecting the topic shift it is also important to detect which topics were the main protagonists of this event. We do so by applying the chi-square test to the topic distributions, excluding each time a different topic, and selecting the topic whose absence yields the bigger increment in the *p* value. If after excluding this topic we still obtain $p \leq 0.05$, the procedure is repeated to select another topic, until p > 0.05.

In most cases the topic shift does not happen abruptly in one year, but it is a gradual process. Thus we implemented a sliding window algorithm that checks for a topic shift by comparing the initial topic distribution in time t with the topic distributions in time *t+1*, *t+2*… *t+n* (n=10 in the prototype). The same procedure is repeated after incrementing *t*, until $t = m-1$, where *m* is the total number of years in the investigated interval.

This algorithm may output different topic shifts that point to the same real life phenomenon (e.g., the growth of topic T in community C at the beginning of 2000), but we want to select only the ones that would be more significant for a human user. We assume that a human would prefer a sharp shift over a slow one and would prefer to recognize it as soon as possible: hence, if two or more topic shifts share part of the same time interval (e.g., 2000-2002 and 2000-2003) we keep only the shortest one, and if they have the same length, we keep the oldest one.

3.4 Linking Communities

We will now discuss the RCMB approach to linking a number of communities on a timeline, and automatically detecting their interactions. The inputs are the temporal topic-based communities detected in different time intervals, with the last year of an interval being the first year of the subsequent interval –e.g., 2000-2002, 2002-2004 and so on. The topic vector of the year in common between two successive intervals is then used to compute the *community similarity* (*CS*) between each couple of communities. CS, which assesses the possibility that two communities in different time intervals C and D are actually the same one, is computed as:

$$CS(C,D) = \left(cos(\widehat{c_y}, \widehat{d_y}) + aut(\mathrm{C},\mathrm{D})\right)/2$$

where C and D are two communities in subsequent time intervals, *y* is the year they have in common (e.g., last year of *C* and first year of *D*), $cos(\widehat{c_y}, \widehat{d_y})$ is the cosine similarity between the communities topic distribution in year *y*, and $aut(C,D)$ is the percentage of authors which moved from *C* to *D*. *CS(C,D)* varies between 0 and 1, with a high value pointing to the fact that *C* and *D* denote the same community in different intervals.

As in [1], the CS(C,D) measure is used to detect two different links between communities. We define two thresholds, t_s and t_w, with $t_s > t_w$. When $CS(C,D) \geq t_s$ we infer a *strong link*, when instead $t_w \leq CS(C,D) < t_s$ we infer a *weak link*.

The *strong link* is defined as a link that connects the same community in subsequent timeframes –e.g., the link that connects the "Ontology" community of SW in the 2004-2006 interval with the "Ontology" community of SW in 2006-2008.

The *weak link* is defined as the link that connects community C_1 with community C_2 in a subsequent timeframe, if C_1 has an impact over C_2 in terms of migrating authors and/or topics. For example a weak link can be established between the

"Intelligent Agent" community and the "Semantic Web Services" community, if a number of authors from the first one flow into the second.

Estimating the correct value for t_s and t_w is vital for generating a legitimate and useful map of research communities. We determine those values by using the Nelder-Mead algorithm [18], which is a derivative-free optimization method. The Nelder-Mead algorithm is used to solve parameter estimation problems where the function values are uncertain or subject to noise. It works by defining a simplex, whose vertices represent the parameters to be found, and then performing a sequence of geometrical transformations on it, aimed at decreasing an evaluation function. Hence, we need an evaluation function that will estimate the quality of the links.

It is possible to use as evaluation function the F-Measure between the links in a generated map and those from a human generated gold standard. However it is not trivial to find the right group of experts and it may be argued that the threshold values obtained within a certain research area may not be feasible to be reused for a different one. We thus propose a different solution, which uses a number of preferable properties that a map should possess to be practical and readable.

The properties on which we focus are:

- Ideally the communities should not appear intermittently, since it would be unrealistic to assume that a community keep disappearing and reappearing.
- Ideally the communities that fade out should contribute to other communities; it would be odd if both the authors and the topics suddenly disappear from a map without any explanation. More formally this means that communities on the verge of disappearing should have at least one exiting weak link.
- Ideally the communities that receive a weak link should observe a topic shift, since they are being modified by the inflow of another community.
- The definition of strong link entails that two strong links should not 1) fork from a community or 2) merge into two different communities.

From these preferable properties we derive the Map Evaluation Function (MEF):

$$MEF = (I\, w_I + D\, w_D + NTS\, w_{NTS} + SLF\, w_{SFL} + SLM\, w_{SFM}) - (WL + SL)$$

where, *I* is the number of times a community disappeared and then reappeared in a subsequent year, *D* the number of times a community disappears without any exiting link in the previous time interval, *NTS* the number of times a weak link does not produce a topic shift, *SLF* the number of strong links forking, and *SLM* the number of strong links merging. *WL* and *SL* define the ratio between the sum of all weak/strong links and the average number of communities in an interval; the purpose of these components is to avoid that the best solution would yield a low or null number of links. The components of this function are attuned by means of weights (i.e., $w_I, w_{SFL}, w_{SFM}, w_D, w_{NTS}$) that can be set empirically or learnt by maximizing, using the Nelder-Mead algorithm, the F-Measure between the links of the generated community and the links of a human crafted gold standard.

We take the minimum values of t_s and t_w that minimize the MEF.

3.5 Detecting Other Key Events

The two kinds of links, which connect the communities in subsequent years, are also exploited for automatically detecting six other kinds of typical dynamics. These events are detected by applying the following heuristic rules:

1. If a community has no strong links with any precedent interval communities, we detect an event of **type 1**, i.e., the **appearance** of a community.
2. If a community has no strong links with any subsequent interval communities, we detect an event of **type 2**, i.e., the **fading** of a community.
3. If two or more communities are linked to one community in the subsequent interval and one of the inlinks is a strong link, we detect an event of **type 3.A**, i.e., the **assimilation** of one or more communities into the community *C* characterized by the strong link. If the communities fade after the event, they are labelled as **absorbed by *C***, else they are labelled as **contributing to C**.
4. If two or more communities are linked to one community in the subsequent interval and none of the inlinks is a strong link, we detect an event of **type 3.B**, i.e., the **merging** of two or more communities in a new community C. If the communities fade after the event, they are labelled as **merged in C**, else they are labelled as **contributing to C**.
5. If a community is linked to more than one community in the subsequent interval and one of the links is a strong one we detect an event of **type 4.A**, i.e., the **forking** of one or more communities out of the community characterized by the strong link.
6. If a community is linked to more than one community in the subsequent interval and none of the links is a strong one we detect an event of **type 4.B**, i.e., the **splitting** of a community into multiple communities.

It should be noted that, from the definition of strong link, it follows that we should never have more than one of them in 3.A and 4.A events. However, the opposite may actually happen in practice and the conflict is then solved by considering as strong link only the one with the highest *CS*.

The final output of RCMB is a data structure that we call 'map of research communities', describing the detected communities within a certain topic, the links, the topic shifts and the other detected events. A map of research communities can be used as a knowledge base and visualized in a variety of ways. For example Rexplore [10] currently implements a view in which communities are represented as nodes in a graph and linked to the related authors, affiliations and countries. In this paper we will instead use a timeline based visualization (see Figures 1 and 2), which is particularly useful for highlighting the evolution and interactions of research communities.

4 Mapping the Dynamics of WWW and SW

In this section we discuss the most interesting events that we detected by applying RCMB to two research areas: World Wide Web (WWW) and Semantic Web (SW). For each of these areas we will highlight the main topic-based communities and how they evolved in the time interval 2000-2010. Since an in-depth analysis of all significant dynamics and patterns discovered in these two areas would take too much space, here we will provide broad overviews of these areas and we will elaborate only on the most interesting automatically detected patterns. We will support our analysis by evidencing the most significant changes in the topic distributions of the communities and the flow of authors between them. We will also show examples of *representative authors*. These are defined as members of a community who fit well the community profile (i.e., they are very similar to the community centroid) and have

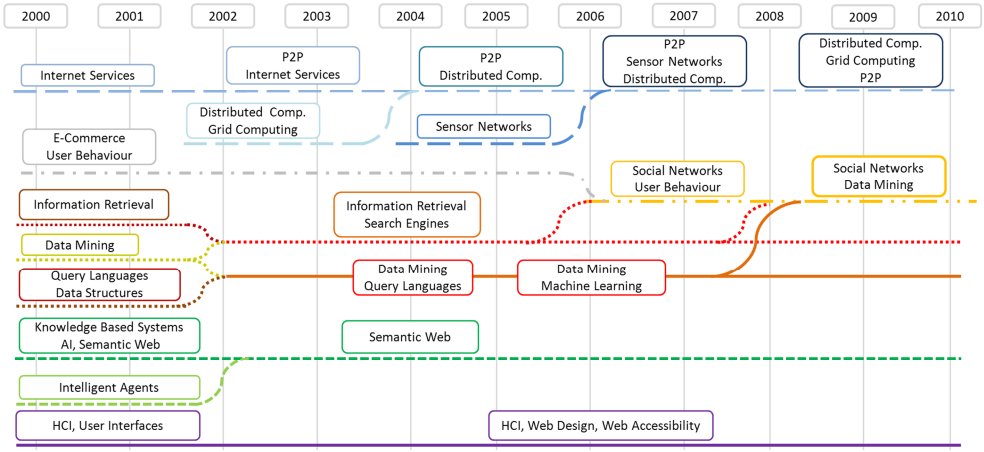

Figure 1. The main research communities in WWW and their trajectories.

an H-Index of at least 10. For readability, we will label some communities only with their one or two most significant topics.

Our study was based on a dataset built from data retrieved by means of the API provided by Microsoft Academic Search (http://academic.research.microsoft.com). In particular, we first retrieved authors and papers labelled with WWW and SW or with their first 150 co-occurring topics. By running the Klink algorithm over these keywords we detected about 700 semantic relationships between them. We then run RCMB on WWW and SW in the 2000-2010 time interval with a granularity of 3. The average number of authors selected in each year was 932 for WWW and 646 for Semantic Web.

4.1 Modelling the Dynamics of the World Wide Web Research Area

WWW is a major area of computer science and contains a variety of communities continuously renewed by the introduction of new technologies and topics. Figure 1 shows the main communities and trends in the WWW area in the timeframe 2000-2010. For year 2000, the RCMB algorithm yielded eight main communities: “Information Retrieval” (210 authors), “HCI, User Interface” (178), “E-Commerce, User Behaviour” (167), “Intelligent Agents” (150), “Query Languages, Data Structures” (149), “Data Mining” (142), “Internet Services, Mobile Agents, Middleware” (90) and “Semantic Web, Artificial Intelligence” (86).
It should be noted that the number of authors does not represent the total of authors who happen to work in some of these topics, but only those who 1) have at least one publication per year in the WWW area and 2) share a common research trajectory with the other authors of the community, i.e. have a similar topic distribution in the same years[1].

We will now examine more closely some of the trends within the WWW area. As shown in Figure 1, in 2002 some authors from the “Data Mining” community (about

[1] Since RCMB returns fuzzy clusters, for the sake of this analysis we assigned each author to the communities for which she/he had a membership score > 0.4.

55% of them) merged with the "Query Languages, Data Structures" community, while others (43%) flew into the "Information Retrieval" community. The merging of "Data Mining" and "Query Languages, Data Structures" in 2002 gave birth to a novel community that counted 245 authors with a hybrid topic distribution, since it was composed by authors who appeared to be interested both in "Query Languages" and in "Data Mining". Moreover, since 2004, the topic distribution of the "Data Mining" community continued to evolve, being progressively enriched by topics representing "Machine Learning" approaches, such as "Neural Networks" and "Genetic Algorithms". Also the "Internet Services, Mobile Agents, Peer To Peer" community went through some interesting changes. In fact, after 2002, the "Peer to Peer" component became prominent in this community, growing from 11% to almost 50%.

It can be hypothesised that this phenomenon was linked to the emergence in 2001-2002 of influential web applications relevant to the Peer To Peer area, such as BitTorrent, Kademlia and Gnutella. Indeed the behaviour of this community appears to be very sensitive to technological shifts: in 2004 it merged with the "Distributed Computing, Grid Computing" community, and in 2006 with "Sensor Networks", finally becoming the "Distributed Computing, Grid Computing, Peer To Peer" community.

The main transformation in the 2000-2010 timeframe for the WWW research area is however the emergence of the "Social Networks" community in 2006, the same year in which Facebook went public outside USA. According to RCMB, the "Social Networks" community drew its authors mainly from the "E-Commerce, User Behaviour" (31%), "Information Retrieval" (27%), "Data Mining" (18%) and "Semantic Web" (18%) communities. The topic distribution of "Social Networks" in 2006 includes: "Social Network" 52%, "User Behaviour" 18%, "Complex Networks" 14%, "E-Commerce" 12% and "Knowledge Based Systems" 10%. It should be noted that, since we are using Klink, the main components tend to be high-level topics, which implicitly include the lower level ones. Thus "Social Networks" actually includes other areas, such as "Social Web", "Social Media", "Collaborative Networks" etc. It is possible to examine each of these components, but here we will instead focus on the big picture. In 2008 there is again a strong flow of authors from the "Data Mining" community to the "Social Networks" community (32%). This migration causes a topic shift in the community, which becomes more focused on "Data Mining", "Information Retrieval" and "Machine Learning" at the expense of "User Behaviour" and "Complex Network". In fact, Social Networks became a major domain for researchers in Data Mining in these years.

4.2 Modelling the Dynamics of the Semantic Web Research Area

The Semantic Web area is a particularly interesting one to be analysed in the timeframe 2000-2010, since it became officially recognised only in 2000. It is thus possible to study how authors from other areas gradually enriched the Semantic Web environment and how the main communities within the Semantic Web area were created. Figure 2 provides an overview of trends in the Semantic Web. At the beginning of 2000, we detect 3 main communities: "Knowledge Based Systems, AI, Ontology" (77 authors), "Intelligent Agents, Software Agents, Multi Agent Systems" (47), "WWW, Data Models, RDF" (61). These communities are composed by the initial authors setting up the Semantic Web area and indeed represent the three main

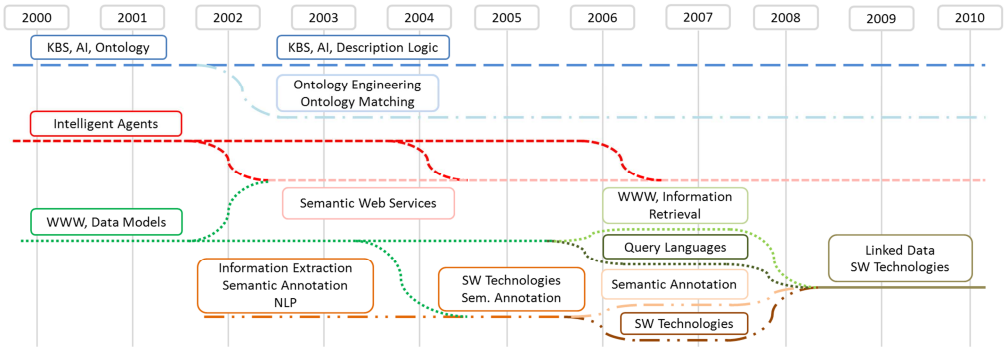

Figure 2. The main research communities in SW and their trajectories.

classes of experts needed to generate this novel area according to the vision of Tim Berners Lee et al. [21]. In the following years, these original groups will interact and merge with each other to create new communities.

Topics	*Rep. Authors*	*Topics*	*Rep. Authors*
2002-2004: KBS, AI, Description Logic (212)		**2002-2004: Semantic Web Services** (149)	
KBS: 26 % AI: 22 % Des. Logic: 16 % Ontology: 6 % Logic Prog. : 5 %	F. Van Harmelen, P. Patel-Schneider, I. Horrocks, S. Bechhofer, J. Heflin, P. Hayes, D. McGuinness, C. Welty.	SW Services: 30 % WS Composition : 15 % WS Discovery: 9 % WWW: 7 % SW Tech.: 6 %	B. Norton, K. Verma, A. Sheth, N. Srinivasan, S. Kumar Agarwal, U. Keller, J. Miller, C. Patel, S. McIlraith.
2002-2004: O. Alignment., O. Engin. (116)		**2002-2004: Inf. Extraction, Sem. Ann., NLP** (63)	
Ontologies: 21 % KBS: 11 % O. Alignment: 10% O. Engineering: 7% Sem. Matching: 5%	S. Castano, N. Silva, J. Euzenat, A. Ferrara, Y. Kalfoglou, M. Ehrig, S. Montanelli, J. M. Rocha.	Inf. Extraction: 26 % Sem. Annotation: 18 % Natural Language: 18 % KBS: 7 % Machine Learning: 5 %	Y. Wilks, B. Popov, H. Cunningham, T. Declerck, A. Dingli, M. Vargas-Vera, K. Bontcheva, F. Ciravegna.

Table 1. Novel Semantic Web communities in 2002-2004.

We can see in Table 2 that by 2002-2004 the scenario has changed and we now have four new communities. Specifically, in 2002 the “Knowledge Based Systems” community split into two different branches: about 22% of the authors migrated to “Ontology Engineering, Ontology Matching” (116 authors) and 58% to “Knowledge Based Systems, AI, Description Logic” (212 authors), which in 2004 would create the OWL standard. In the same year, some authors from the “WWW, Data Models” and the “Intelligent Agents” communities came together to build the “Semantic Web Services” community. According to RCMB, the authors from “Intelligent Agents” kept flowing to “Semantic Web Services” in the 2002-2006 interval, and eventually “Intelligent Agents” went under the critical mass to be considered as a main community for the SW. This insight is actually confirmed by a 2007 editorial titled “Where are all the Intelligent Agents?” [22] in which Jim Hendler, one of the founders of the SW area, observed that the role of agent research in SW at that time was not as strong as envisaged in the original 2001 vision.

The fourth novel community to appear in 2002 is the “Information Extraction, Semantic Annotation, Natural Language Processing” community, which has a hybrid composition, acquiring authors from “Knowledge Based Systems, AI, Description Logic” (18%), “WWW, Data Models” (14%), and “Intelligent Agents” (12%). In 2004 this community is enriched again by authors from the “WWW, Data Models” community and becomes the “Semantic Web Technologies, Semantic Annotation”

community. In 2006, this community splits into “Semantic Web Technologies, World Wide Web, Ontologies” (51% of the authors) and in “Semantic Annotation, Knowledge Based Systems, Semantic Wikis” (38% of the authors). The former appears to focus primarily on the design and use of ontologies and SW technologies, while the latter focuses on the task of annotating the World Wide Web. One of their most significant outputs was DBpedia [23], the RDF based version of Wikipedia (http://www.wikipedia.org), which today provides a widely used information source in the Linked Data environment. In the meantime, the original “WWW, Data Models” community followed a parallel track and also split in 2006: 30% of the authors became part of the novel “WWW, Information Retrieval” community while 32% became part of the “Query Languages, Data Models” community. “WWW, Information Retrieval” focused on methods to exploit semantic markups for Information Retrieval on the WWW, whereas “Query Languages, Data Models” gave birth to SPARQL [24], the W3C candidate recommendation query language for RDF.

By 2008 these four communities had crafted a multitude of RDF resources (e.g., DBpedia), a query language (SPARQL) and a variety of approaches for exploiting Semantic Web Technologies. Soon after, they all merged into the “Linked Data, Semantic Web Technologies, Information Retrieval” community, focusing on the key task of building a large-scale Semantic Web. Indeed, about 48% of the authors from the original “WWW, Data Model” community and 52% of the authors from the 2004 “Information Extraction” community eventually flowed into the “Linked Data” community. These trends provide a good example of how a community of researchers started with a vision that was ahead of its time, split into sub-communities, which tackled different elements of the problem, and then merged again years later, once the key components of the vision had become established.

5 Evaluation

In this section we compare the RCMB algorithm with other approaches and show how the different techniques introduced in Section 3 incrementally improve the identification of well-formed research communities.

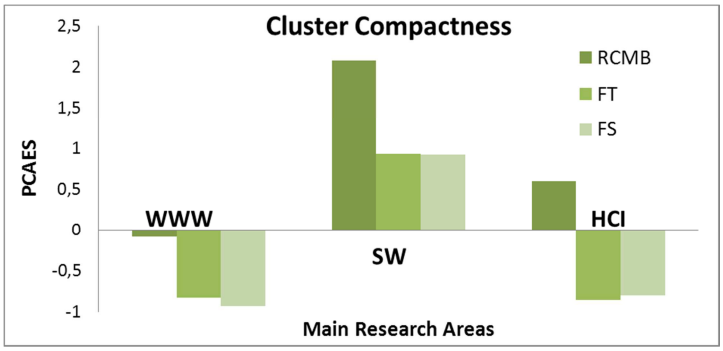

Figure 3. PCEAS of the cluster sets returned by RCMB, FT and FS.

In particular, we tested four different methods: 1) Fuzzy C-Mean (FCM) using cosine similarity on regular keywords (**FC**), 2) FCM using cosine similarity on semantic topics (**FS**), 3) FCM using TTS on semantic topics (**FT)** and 4) FCM using TTS on weighted semantic topics (**RCMB)**.

We ran these approaches on WWW, SW and Human Computer Interaction research areas in the timeframe 2000-2010 and compared the detected clusters. As shown by Figure 3 in all cases RCMB performs much better than the other two approaches in term of cluster compactness, yielding a PCAES [25] 129±41% higher than FT and FS. According to the chi-square test, the differences between RCMB and either FT or FS are statistically significant (p=0.03), thus we can say that the semantic techniques introduced in this paper allow finding much more compact clusters.

Table 3 describes the communities detected by the approaches in the WWW area. The left panel shows that RCMB works also better in term of community identified, while the other two algorithms miss some important communities such as the "HCI" and the "Social Network" ones. Table 4 shows a comparison between the main topics of the WWW communities according to both RCMB and Fuzzy C-Mean using the cosine similarity. The absence of semantics has an evident negative impact both on the clusters and on their topic distribution. The first three clusters found by **FC** may probably be mapped to the "Information Retrieval", "Semantic Web" and "Data Mining" clusters detected by RCMB, but the others appear to be very noisy.

Clusters	RCMB	FT	FS
Inf. Retrieval, Search Eng.	Y	Y	Y
Semantic Web, AI	Y	Y	Y
P2P, Dis. Computing	Y	Y	Y
Data Mining, Inf. Retrieval	Y	Y	Y
Social Networks	Y	Y	N
E-Commerce	Y	Y	N
HCI, Design Process	Y	N	N
Data Structures, Data Models	N	N	Y

RCMB	Baseline Fuzzy C-Mean
Information Retrieval, Search Engines, Web Search	Search Engine, Web Search, Digital Libraries, Information
Semantic Web, AI, KBS, Intelligent Agents	SW, Linked Data, Semantic Web Services
Data Mining, Information Retrieval	Data Mining, Data Streams, lotus japonicus
Social Networks, Complex Networks, User Behavior	Astrophysics Data System, Xml Document, Query Language
Peer To Peer, Distributed Computing, Sensor Networks	Web Design, Web Engineering, Cosmic Ray, E-Commerce
HCI, Design Process, Learning Process, Web Accessibility	Ubiquitous Computing, Communication Systems
E-Commerce, Information Technology	Web Interface, Eclipsing Binaries, Variable Stars

Table 3. Left Panel: main communities of WWW in 2000-2010. Right Panel: Topics associated with the main WWW communities in 2000-2010 according to RCMB and FC.

6 Conclusions

In this paper we have presented the Research Communities Map Builder (RCMB), an algorithm which builds maps of research communities, describing for each year in a given interval the main communities within a given topic, the topic shifts and a number of other key events. The resulting map of research communities can be used in a variety of ways to study the evolution of the communities within a given topic, their interactions, and how they are influenced by each other or by external events, such as a technological breakthrough.

For the future, we plan to build on this work and develop methods able to forecast topic shifts and key events, e.g., to estimate the probability that a new topic will emerge in a certain community or that two communities will merge in the coming years. We also intend to develop new knowledge-based techniques, able to provide comprehensive explanations for the identified dynamics.

References

1. Yan, E., Ding, Y., Milojević, S., Sugimoto, C.R.: Topics in dynamic research communities: An exploratory study for the field of information retrieval. Journal of Informetrics, 6(1), 140-153. (2012)
2. Van Eck, N. J. Waltman, L.: Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523-538. (2010)
3. Upham, S. P., Rosenkopf, L., Ungar, L. H.: Innovating knowledge communities. Scientometrics, 83(2), 525-554. (2010)
4. Yan, E., Ding, Y., Jacob, E.: Overlaying communities and topics. Scientometrics, 90(2), 499-513. (2012)
5. Zhao, Z., Feng, S., Wang, Q., Huang, J. Z., Williams, G. J., Fan, J.: Topic oriented community detection through social objects and link analysis in social networks. Knowledge-Based Systems, 26, 164-17. (2012)
6. Osborne, F., Motta, E.: Mining Semantic Relations between Research Areas. In Proceeding of the 11th International Semantic Web Conference (ISWC 2012). Boston, MA. (2012)
7. Ding, Y.: Community detection: topological vs. topical. Journal of Informetrics, 5(4), 498-514. (2011)
8. Upham, S. P., Small, H.: Emerging research fronts in science and technology: patterns of new knowledge development. Scientometrics, 83(1), 15-38. (2010)
9. Osborne, F., Scavo, G., Motta, E.: Identifying diachronic topic-based research communities by clustering shared research trajectories. In ESWC 2014. Crete, Greece. (2014)
10. Osborne, F., Motta, E., Mulholland, P.: Exploring Scholarly Data with Rexplore. In Proceedings of the 12th International Semantic Web Conference. (2013)
11. Flake, G. W., Lawrence, S., Giles, C. L., Coetzee, F. M.: Self-organization and identification of web communities. Computer, 35(3), 66-70. Chicago. (2002)
12. Smyth Guimera, R. Amaral, L. A. N.: Functional cartography of complex metabolic networks. Nature, 433(7028), 895-900. (2005)
13. Racherla, P. Hu, C.: A social network perspective of tourism research collaborations. Annals of Tourism Research, 37(4), 1012-1034. (2010)
14. Hofmann, T.: Probabilistic latent semantic indexing. In the 22nd Conference on Research and Development in Information Retrieval (pp. 50–57). Berkeley, CA. (1999)
15. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1033. (2003)
16. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: extraction and mining of academic social networks. In Proceeding of KDD 2008, pp. 990-998. (2008)
17. Bezdek, J. C., Ehrlich, R., Full, W.: FCM: The fuzzy c-means clustering algorithm. Computers and Geosciences, 10(2), 191-203. (1984)
18. Olsson, D. M., Nelson, L. S.: The Nelder-Mead simplex procedure for function minimization. Technometrics, 17(1), 45-51. (1975)
19. Sethi, I. K., Patel, N. V.: Statistical approach to scene change detection. In Symposium on Electronic Imaging: Science & Technology. SPIE. (1995)
20. Chiu, Stephen L.: Fuzzy model identification based on cluster estimation. Journal of intelligent and Fuzzy systems 2.3 (1994): 267-278. (1994)
21. T. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American, 284(5), 28-37. (2001)
22. Hendler, J.: Where are all the Intelligent Agents? A Letter from the Editor in Intelligent Systems IEEE, May/June 2007. (2007)
23. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In The semantic web (pp. 722-735). Springer. (2007)
24. Pérez, J., Arenas, M., Gutierrez, C.: Semantics and Complexity of SPARQL. In Proceedings of ISWC 2006 (pp. 30-43). Springer Berlin Heidelberg. (2006)
25. Wu, K. L. and Yang, M. S.: A cluster validity index for fuzzy clustering. Pattern Recognition Letters, 26(9), 1275-1291. (2005)