Identifying diachronic topic-based research communities by clustering shared research trajectories

Conference or Workshop Item

Identifying diachronic topic-based research communities by clustering shared research trajectories

Francesco Osborne[1], Giuseppe Scavo[1], Enrico Motta[1]

[1]Knowledge Media Institute, The Open University, MK7 6AA, Milton Keynes, UK
{francesco.osborne, giuseppe.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: 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 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. A similar analysis in the Human-Computer Interaction (HCI) area shows that authors in the HCI community who had a background in User Modeling and Ubiquitous Computing were the ones at the forefront of research on Mobile Devices, once the smartphone became a reality.

TST is integrated within Rexplore [4], a system that combines statistical analysis, semantic technologies and visual analytics to provide support for exploring and making sense of scholarly data. To evaluate our approach we performed an empirical study involving 25 experts from the SW and HCI areas, who were asked to aggregate a set of selected topics to generate the main topic-based communities in their field. The results indicate a high degree of agreement among the experts, confirming that topic-based communities are indeed objective entities that can be recognized by experts. In addition, TST performed at expert level – i.e., its results are statistically consistent with those of the experts.

2 State of the Art

Current approaches to community detection are usually classified according to the strategy they use [1], as either *optimization-based* or *heuristic* methods. The former use either *local search* [4] or *spectral methods* [6], whereas the latter exploit domain-specific assumptions to direct the clustering [7]. Unfortunately these methods tend to rely on topological structures, such as the ones defined by citation or co-authorship networks, and as a result they are not applicable to our scenario, where, as explained in the previous section, we do not have topological structures that completely and correctly define our space. As discussed by Ding et al. [2], it is therefore important to develop novel approaches to community detection, which are able to focus on the relationship between communities and topics and can correctly model their dynamics over time. A first step in this direction is provided by the work of Upham et al. [8], who define an algorithm for identifying topic-based communities which, in addition to the citation graph, also exploits language-level similarities between papers to identify communities. Hence, they are able to group together authors who work on the same topic but are not necessarily related through explicit co-authorship or citation relationships. However, while this approach provides an improvement over purely topological analyses, it seems to us that the focus on publications (rather than authors) and the reliance on language similarities provide too weak a method to detect temporal topic-based communities. In particular, it is not possible in this approach to express explicitly which authors belong to a particular community (or set of communities) at a particular time.

Racherla and Hu [9] identify topic communities by exploiting a topic similarity matrix and assigning a predefined research topic to each document and author. However, this approach is much too limited, as they assume a rigid 1-1 relationship between researchers and topics. In contrast with this work, TST is more flexible and can correctly handle both the situation where a researcher belongs to multiple communities and also that where a community is characterised by a distribution of topics over time.

Semantic technologies have been shown to improve the quality of clusters of different kinds of entities, such as images [10] and tags [11]. Some approaches rely on the detection of latent topics for capturing semantic relationships between keywords, using methods such as *Probabilistic Latent Semantic Indexing* (pLSI) [12] or *Latent Dirichlet Allocation* [13]. For example, the *Author-Conference-Topic* model (ACT) [14] treats authors and venues as probability distributions over topics extracted by means of an unsupervised learning technique. Mei et al. [15] propose a framework to

model topics by regularizing a statistical topic model through a harmonic regularizer, which is based on a graph structure. Differently from these methods, we exploit an automatically generated knowledge base [3] to characterize research topics semantically and we use this as the basis for associating a diachronic semantic topic distribution with each author. The knowledge base is extracted from publication metadata by means of *Klink* [3], an algorithm that combines machine-learning methods and background knowledge to identify research topics and to generate semantic relations between them. Adopting a similar perspective, Erétéo et al. [16] proposed SemTagP, an algorithm which uses existing ontologies to detect communities from the directed typed graph formed by RDF descriptions of social networks and folksonomies. However, their approach is based on label propagation and, in contrast with TST, does not take in account the temporal dimension, which is important for gaining an understanding of community evolution over time and is also being investigated in the emergent field of temporal networks [17].

TST relies on the *Fuzzy C-Means* [18] algorithm, which is a popular unsupervised clustering algorithm that has been applied successfully to a number of real life problems. Clustering techniques (e.g., modularity-based clustering [19] or the k-means algorithm [20]) have also been used by other authors to detect research communities. However, these approaches exploit the similarity between topic vectors associated to publications and, as a result, exhibit limitations when compared to our method. In particular, their topic vectors lack a semantic characterization and, in addition, by focusing on publications rather than authors, they fail to take into account the diachronic dimension. As we will show in Section 4, in contrast with the aforementioned approaches, the use of semantic topics and the adoption a diachronic approach yields a dramatic increase in the quality of the detected communities.

3 Detecting Temporal Topic-Based Communities

We will now discuss the TST approach to identifying clusters of researchers who share common research trajectories – i.e., researchers who appear to work on the same topics at the same time. We refer to these clusters as *temporal topic-based communities* (*TTCs*).

The TST approach for automatically computing TTCs in a given research area, say R, follows three steps:

1. **Semantic topic enrichment**, during which the topic distributions associated with each author are semantically enhanced by taking into account the semantic relationships between research topics.
2. **Topic vector weighing**, during which each component of a topic vector, say T, is given a bonus proportional to the degree of similarity between T and R.
3. **Temporal topic-based clustering**, during which the authors are clustered by means of a Fuzzy C-Means algorithm, using the aforementioned *ATTS* metric.

These steps are discussed in the following sub-sections.

3.1 Semantic topic enrichment

The authors to be clustered are characterized as a collection of topic vectors, one for each year over the examined timeframe, where each value represents the number of publications in a topic during a certain year.

A naive approach here would be to use as topics the keywords associated to the publications. However this method may yield poor results, since, as discussed in [3], the keywords associated to academic publications lack structure and are often noisy. Analogously, the keywords extracted by natural language techniques may also be noisy and may include terms that do not define research areas.

To address this issue we use the Klink algorithm [3], which is able i) to identify keywords that refer to a research area and distinguish them from those which do not and ii) to detect three types of semantic relationships. Specifically, Klink can detect: *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, however T_1 is not a sub-topic of T_2). Hence, the output of an application of Klink to a corpus of publications tagged with keywords is a knowledge base comprising semantic topics structured according to three relations and linked to the relevant publications (and therefore with the relevant authors and organizations).

Taking advantage of this knowledge base, we label with topic T_1 any publication tagged with topic T_2, if T_2 is a sub-topic of T_1 or it is *relatedEquivalent* to T_2. This simple step can yield a dramatic increase in the quality and quantity of data about a certain topic. For example, as a result of applying Klink to a corpus of about 15 million publications in Computer Science, we were able to identify 18 sub-topics of Semantic Web (e.g., "Linked Data", "Semantic Wiki" and "OWL") and 11 *contributesTo* relationships (e.g., "Description Logic"), thus increasing the number of publications in the Semantic Web from 11998 to 20751. In the same way we were able to detect 22 sub-topics of HCI (e.g., "Affective Computing" and "User Interface") and 7 *contributesTo* relationships (e.g., "Task Analysis"), thus increasing the number of publications tagged as "HCI" from 9850 to 93583.

We then build the topic distribution for each author in year *t* as a vector in which each topic is associated with the number of publications in the same year. Finally, for each couple of topics, $<T_1,T_2>$, sharing a *contributesTo*(T_1,T_2) relationship, we assign to T_2 a fraction of the publications in T_1 according to the formula:

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

where $ct(i,T)$ indicates the set of topics associated with the *i-th* publication that is in a *contributesTo* relationship with *T*. $P\left(T \middle| ct(i,T)\right)$ is the probability for a paper with such a set of topics to be also explicitly associated with topic *T* (or with a topic having a *broaderGeneric* or *relatedEquivalent* relationship with *T*) at the time of publication 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*. The exponent φ serves to modulate the *contributesTo* relationship and was empirically set to 0.5. The outputs are semantic topic vectors that include only semantically characterized research areas, whose

associated values are weighed according to the semantic relationships between research areas.

3.2 Topic vector weighing

In most cases it is useful to detect the communities within a certain *main topic* (e.g., Semantic Web), to allow a user to make sense of elements of the research dynamics within the topic. For this reason we take as input only the authors with a significant amount of publications in the main topic. For example in the evaluation we will take in consideration only the authors who have published at least 10 papers in the Semantic Web area in the 2005-2010 interval. Moreover, to highlight the communities strongly related to the main topic, we weigh each topic according to its relationship with the main topic. Given a semantic 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 (empirically set to 2 in the evaluation) that can be tuned to amplify the effect of the weight on the system. Here it is important to emphasise that, as a result of the semantic topic enrichment carried out in the previous step, the co-occurrences used in this formula are actually applied on semantic topics rather than defining standard keyword co-occurrences.

This step can be skipped if the main topic is not defined.

3.3 Temporal topic-based clustering

In the final step of TST, a Fuzzy C-means (FCM) algorithm is applied to the weighted topic vectors to compute a set of fuzzy clusters of authors associated with their distribution of topics over the years. Here, we have adopted a fuzzy clustering technique since most researchers tend to work in more than one community, and a clustering algorithm that forced them to be members of only one would be unfeasible. Moreover, associating authors with a degree of memberships to each community allows for a more granular characterization of their research interests.

FCM is one of the main unsupervised clustering algorithms and has been applied successfully to a number of scenarios. It classifies entities by minimizing the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2 \ , \ 1 \le m < \infty$$

where N is the number of entities (in this case authors), C the number of the chosen centroids, u_{ij} is the degree of membership of x_i in the cluster *j*, *m* is a number ≥ 1, x_i is the *i*th entity, c_j is the centroid of the cluster, and ||*|| is a norm expressing the similarity between any entity and the centroid.

We will not elaborate here on the details of the algorithm since it is well known – see [18] for an in-depth description.

In our case, we need a norm that takes into account the topic vectors over the years. To this end we have introduced a novel similarity measure called *adjusted temporal topic similarity* (*ATTS*).

We first define the *topic similarity* (*TS*) between two authors *A* and *B* in a time interval t_1-t_2 as:

$$TS(A,B,t_1,t_2) = \cos\left(\sum_{i=t_1}^{t_2} \hat{a}_i , \sum_{i=t_1}^{t_2} \hat{b}_i\right)$$

where $\hat{a}_i$ 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.

This metric however does not take into account possible common shifts of interests of the authors. In fact, if author *A* worked on topic T_1 and then shifted to topic T_2, he/she will be considered similar to author *B* who was originally in T_2 and then moved to T_1. To avoid this problem, we need a metric that pays attention to the period of time in which an author addresses a specific topic, rewarding common trajectories. Hence, in order to strengthen the importance of the time factor we compute TS recursively on increasingly shorter time intervals and then average the results. More formally, we define the *temporal topic similarity* TTS between author *A* and author *B* in the interval t_1-t_2 as:

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

The temporal topic similarity covers well the case in which both authors are present in the same time interval. However an author may start publishing after the beginning of the interval or suspend his/her career before the end of it. These cases may be accounted for by introducing a penalty for authors who do not share the entire timeframe. We quantified the penalty *P* as the average TS of *n* authors randomly extracted from the input (*n=500* in the prototype).

Finally, we define the *adjusted temporal topic similarity*, ATTS, as:

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

where I_s is the number of years in which both authors were active, I_{ns} is the remaining number of years, and γ>1 a parameter for weighing their relationship (γ= 2 in the present evaluation). If γ is high, an author active in a good portion of the interval is barely penalized, thus allowing for latecomers to be assigned to the cluster if their topic trajectory is similar enough to the community centroid. Since ATTS is a similarity measure that varies in the interval [0,1], while a FCM needs a distance in the interval [0, ∞], we use as norm the inverse of the ATTS minus 1.

The output of FCM depends on the initial guess on the number of clusters and the candidate centroids. In this scenario there is no absolutely correct initial number of centroids, since even different user experts will suggest a different number of communities. However, we suggest two techniques to select the initial number of centroids. The first, and most conservative one, is to compute the set of clusters for different numbers of centroids and for different random initializations and then select the one with the highest compactness (see Section 5). In this paper we used the PCAES [21] *(Partition Coefficient and Exponential Separation)* as measure for

compactness. This is a cautious approach that will produce very compact communities, but may also miss some of the minor ones. The second approach is the *subtractive clustering method* [22]. This technique estimates the initial centroids by assigning a "potential" to each individual in the dataset, so that an individual with many neighbours will have a high potential. While this approach may build less compact communities, it nevertheless appears to produce results that are very similar to the ones generated by the domain experts (see Section 4).

FCM returns a list of cluster centroids and a partition matrix where each element is associated with its degree of membership to each cluster.

The centroids of the clusters detected by the FCM algorithm are characterized by the topic vectors of the communities for each year in the interval, which can be used to study the community evolution. In fact, by studying the change in the distribution of topics in subsequent years it is possible to detect trends (e.g., a topic is growing considerably and thus may continue to grow in the future) and shifts (e.g., a marginal topic is becoming dominant, such as "Augmented Reality" becoming a more important component of the Virtual Reality community after the introduction of mobile devices). This possibility opens up very interesting scenarios and it is one of the main assets of TST.

By summing the vectors over the years and selecting the topics with the highest values it is possible to label communities according to their most significant topics. For example, a key community, which emerges when analysing the Semantic Web area, has the highest values associated to the topics "Artificial Intelligence", "Knowledge Base" and "Ontology", and therefore we can refer to it as the "AI, KB, Ontology" community.

4 Evaluation

We conducted an empirical study with 25 human experts, 13 from the Semantic Web and 12 from the Human Computer Interaction field. These were chosen among experienced researchers in the two fields. Specifically, we wanted to verify i) if the experts could agree on the main topic-based communities in a field - i.e., if the concept of topic-basic community is clear and well defined enough for human users, and ii) if the proposed method could perform similarly to the experts and thus be considered reliable in detecting this kind of communities.

For setting up the study we first built a dataset covering the SW and HCI areas, by exploiting the Microsoft Academic Search (MAS) API[1], a service that makes it possible to access metadata about authors, publications and keywords. We retrieved authors and papers labelled with HCI or SW or with their first 50 co-occurring topics and we then ran Klink on this dataset to obtain a populated ontology of these two research areas for the semantic topic enrichment phase (see Section 3.1). We then selected as "basic topics" the 35 semantic topics[2] that were most often used as tags for SW or HCI papers in the years 2005-2010 –as a result, some topics that have grown in importance since 2010 may be missing from this sample. We then removed from this

[1] http://academic.research.microsoft.com/

[2] See http://rexplore.kmi.open.ac.uk/data/tce.rtf for a list of the topics used in the experiment.

set highly generic topics (e.g., Artificial Intelligence) to simplify the task for the experts. In fact, these topics tended to be associated with pretty much every single one of the 35 topics used in the experiment and therefore held no discriminatory power. Here, it is important to emphasise that keeping such highly generic topics would have not affected the algorithm, which would have simply assigned them to more than one community with different degrees, while of course it would have complicated significantly the task for the experts.

We used WebSort[3], a card sorting online service, to assist the experts in building the clusters. We allowed for each topic to be associated with only one community at a time, since it would have been cumbersome to ask experts to create overlapping communities or communities characterized by potentially different topics for each year, as our algorithm is able to do. We thus modified the output of the algorithm to follow the same limitations by merging the topic vectors of the different years and assigning each topic only with the community with which it had the highest score. Hence, we gave the experts a collection of "basic topics" related to their field and asked them to aggregate the topics together to shape what they considered to be the main communities in their field. For example an expert in HCI could decide to group together topics such as "Ubiquitous Computing", "Mobile Device" and "Context Aware" and label them as "Mobile Interaction" community.

The SW experts suggested an average of 7.9 ± 2.3 communities, whereas HCI experts suggested an average of 6.7 ± 1.9. We then examined the degree of agreement among experts and with our algorithm. To compute the agreement between two sets of clusters we used the pairwise F-Measure, the harmonic means of the pairwise precision and recall.

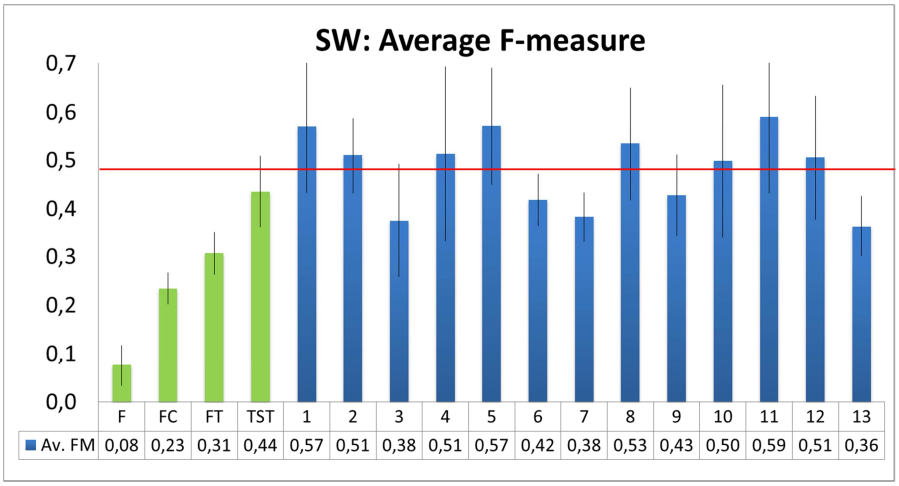
We tested four algorithms on the same dataset: 1) FCM using cosine similarity on regular keywords (labelled *F*), 2) FCM using cosine similarity on semantic topics (*FC*), 3) FCM using ATTS on semantic topics (*FT*) and 4) FCM using ATTS on *weighted semantic topics* (*TST*). We selected as input the set of authors with at least 10 publications about SW/HCI in the 2005-2010 interval. The total amounted to 431 authors for SW, and 458 authors for HCI. The initial centroids were estimated by means of the subtractive clustering method [22].

Figure 2 and Figure 3 show the average degree of agreement of each expert with all the others. For SW, the ANOVA version of the variance test over all experts evidenced statistically significant differences (visible also in the graph), yielding $p=0.02$. Only seven experts exhibited agreement among themselves ($p=0.18$) and they also agreed with the final version of the algorithm, TST ($p=0.12$). Actually there is a fair degree of agreement between the SW experts and our algorithm: the average F-Measure is 0.48±0.04 for the former and 0.44±0.07 for the latter. For HCI, the ANOVA test on experts yielded $p=0.45$. Including as a 'special expert' the final version of our algorithm (TST) yielded $p=0.14$. Since in both cases $p >> 0.05$, we can conclude that there are no statistically significant differences among the experts and between experts and the final version of the algorithm.

The results of the three most basic versions of our algorithm are significantly different, both from the TST version and also from the experts (in all comparisons $p < 0.0001$ with Friedman test for correlated samples). In particular, the version without

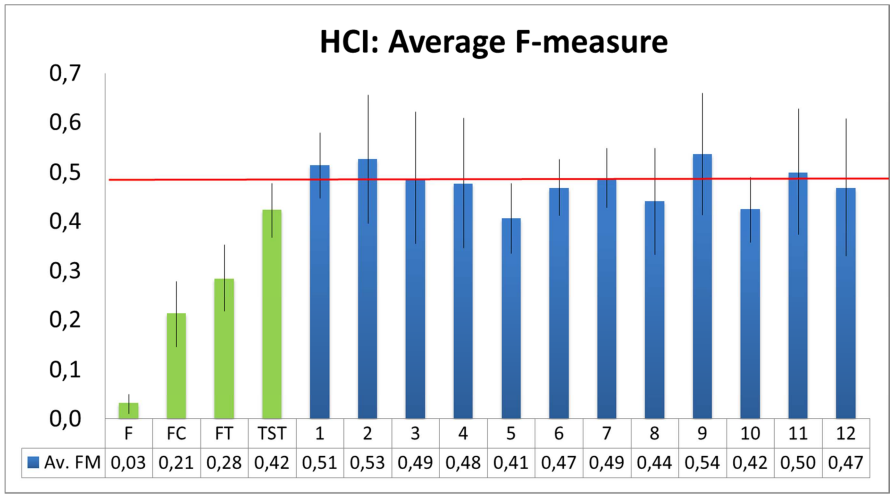
[3] http://uxpunk.com/websort

semantics (F) performed disastrously. The FC and FT version yielded increasingly better results both in SW and HCI, showing how the use of a semantic characterization of topics and the ATTS metric crucially ensures that our method is able to perform consistently with the experts.



	F	FC	FT	TST	1	2	3	4	5	6	7	8	9	10	11	12	13
Av. FM	0,08	0,23	0,31	0,44	0,57	0,51	0,38	0,51	0,57	0,42	0,38	0,53	0,43	0,50	0,59	0,51	0,36

Figure 2. Average F-measure between each expert/algorithm and all the other experts for the SW topic. The red line represents the average F-measure of the experts.



	F	FC	FT	TST	1	2	3	4	5	6	7	8	9	10	11	12
Av. FM	0,03	0,21	0,28	0,42	0,51	0,53	0,49	0,48	0,41	0,47	0,49	0,44	0,54	0,42	0,50	0,47

Figure 3. Average F-measure between each expert/algorithm and all other experts for HCI.

A careful look at the crafted communities evidences that most experts actually agreed on the general picture (the macro-communities) of their field, but sometimes disagreed on how to split some macro-groups, creating sub-groups according to different perspectives. For example in SW the topics “Ontology Engineering” and “Ontology Mapping” are aggregated by some experts within the “Formal Ontology” community, while, according to other experts, they should instead be in two different communities.

Table 1 shows the macro-communities on which most experts agree. We composed it by analysing the labels of the experts and the usual topic components. Thus, for example, an area such as "Ubiquitous Computing/Mobile Device" may either include or not include "Context Aware" according to different experts, but it is usually associated with the same topics and yields similar labels to "Mobile interaction" or "Mobile HCI". SW enjoys 4 size macro-communities on which more than 70% of experts agree, while HCI has 6 of them. Some macro-communities, such as Description Logic in SW and Virtual Reality in HCI, are so well defined that they get almost full agreement. We can say that the skeleton or general frame of the communities appears to be well defined, whereas the details, such as the position of individual fine-grained topics, may vary according to individual experts.

SW Communities	%	HCI Communities	%
Knowledge Base/Des. Logic	100	Virtual Reality	92
Linked Data/Sem. Annotation	100	Information retrieval/WWW	92
Semantic Web Service	77%	Ubiquitous Computing/Mobile Device	83
Ontology Mapping/O. Matching	77%	Interaction Design/Usability Testing	83
Intelligent Agents	69%	Pattern Rec./Gesture Rec./Speech Rec.	75
Ontology Engineering	61%	System Design/Software Engineering	75
WWW/Information Retrieval	61%	AI /Machine Learning/Neural Network	55
Social Semantic Web	46%	Human Robot Inter. /Affective Comp.	42

Table 1. The macro-communities (with more than 40% agreement) in SW and HCI according to the experts.

To study the similarities and differences between the results obtained by our approach and those generated by the experts, we ran TST over an increasing number of clusters (from 4 to 10) to highlight the macro-areas and how they split as the number of clusters grows. Figure 4 and Figure 5 show the result. In most cases the algorithm behaved as a human expert, for example splitting the macro-community "Ontology" in its main sub components as the number of required clusters increased. Our approach found 5 macro communities in both SW and HCI, which can be further split in 10 sub-communities for SW and 9 for HCI.

While here we label each community with the name of the most frequent topics for the sake of simplicity, actually the TTCs are described by a rich distribution of topics over time, which can reveal interesting insights on the dynamics of the research communities. For example, the "Linked Data" community includes a variety of equally represented topics up to 2007, such as "Query Language", "Semantic Annotation" and "Information Retrieval", while from 2008 we see the strong onset of the actual "Linked Data" topic. This reflects an interesting dynamics, where the different research areas that were addressing alternative challenges associated with research on Semantic Web eventually converged on "Linked Data" once a number of underlying technologies became sufficiently mature. In the same way, by analysing the topic distribution of the "Virtual Reality" community, we can see the onset of topics such as "Mobile Device" and "Augmented Reality" after 2007, which help to analyse the impact of the introduction of smartphones (the first iPhone was realized in 2007) and anticipate the vast amount of work that will be done on these topics in the following years.

All macro-communities in Table 1 are detected by our algorithm, except for "Social Semantic Web" for SW and the "AI-Reasoning" for HCI. "Social Semantic Web" is usually composed by topics such as "Social Networks" and "Semantic Wiki". The experts found it natural to aggregate these research areas into one category that today is becoming more and more important. The algorithm did not, because according to the dataset this area did not have enough authors and publications to be considered as a main community during the time frame in question. In sum, this was an unfortunate consequence of not being able to run the experiment on the most recent data (the MAS API did not provide us with much data after 2010).

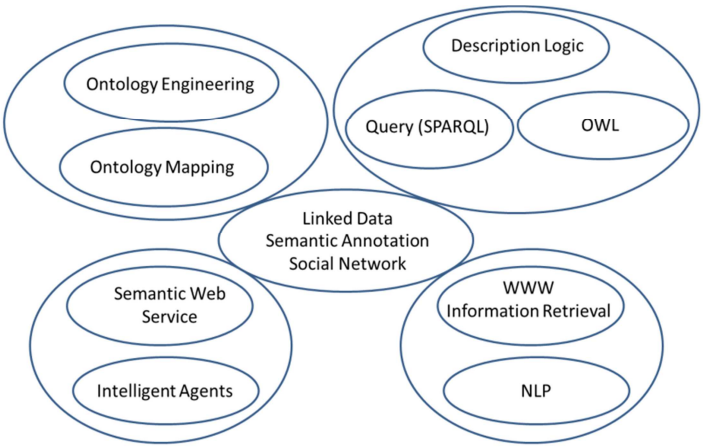


Figure 4. The SW main communities and how they are split in sub-communities by our algorithm. To increase the readability of the image, only the most important topics are shown.

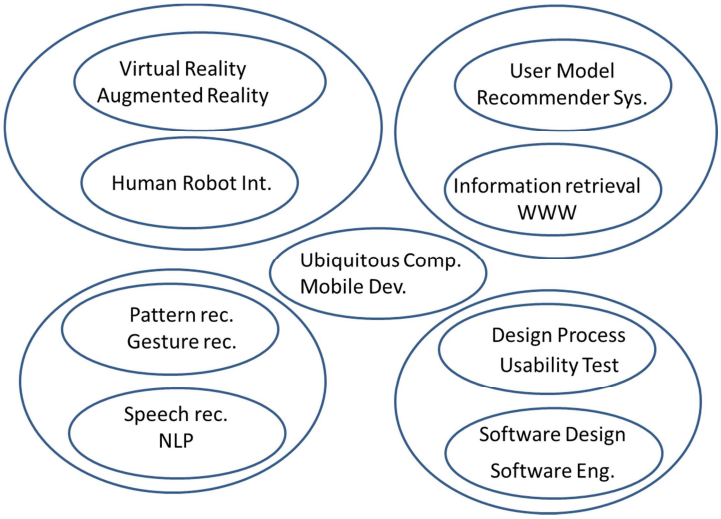


Figure 5. The main communities in HCI and how they are split in sub-communities by our algorithm. To increase the readability of the image, only the most important topics are shown.

The "AI-Reasoning" macro-community is a particularly interesting case, since it is an abstract category where different human experts placed AI techniques, such as

Machine Learning, Neural Networks, User Model and Data Mining. To a human in fact it makes sense to have this kind of abstract categorization of techniques that can be applied in different fields. The algorithm instead is designed to assign each one of these topics to the communities who mostly use them. For example, Machine Learning was associated in most years with the Pattern Recognition and the Information Retrieval/World-Wide-Web communities; Data Mining and Mobile Device with IR /WWW and User Model mostly with Recommender Systems.

In conclusion, human experts are able to create abstract categories, when appropriate, while TST cannot do this (unless an abstract category emerges from the clustering process). TST detects categories on the basis of the trends and practical use in a research area. We believe that these two perspectives are actually complementary: we need the abstract classification provided by experts in order to identify groups of generically applicable techniques/tools relevant to different communities, but we also need the data-driven perspective, to understand by which communities and in which context these are used.

5 Evaluation of Cluster Compactness

In this section we briefly present an evaluation of the *compactness* within each community cluster. We do so by using a standard validity index for fuzzy clustering, *PCAES* [21]. PCAES varies between $-n$ and n, where n is the number of clusters. A large PCAES value means that each cluster is compact and well separated from the others. We ran the different versions of the algorithm to find n communities under SW/HCI, with $4<n<10$. Figure 6 shows the average *PCAES* for SW and HCI over 20 runs: the best performance for all three techniques is reached for HCI in correspondence of n=4, whereas for SW the best overall performance corresponds to the use of TST with n=5. FC and FT obtain the best result with n=4. These values are slightly inferior (but still within two standard deviations) to the values of 7.9 ± 2.3 for SW and 6.7 ± 1.9 for HCI indicated by the experts, possibly because they tend to favor a more articulate classification, even at the cost of some less well-defined communities.

We have thus chosen n=4 and n=5 as the number of clusters on which to run a statistical evaluation of the performance of the three techniques, and in particular of TST relative to the other two, based on the Wilcoxon non-parametric test for correlated pairs. In the SW case, for n=5 we obtain p=0.005 for both TST vs. FT and for TST vs. FC. For n=4, the difference gets less marked, with p reaching the threshold of 0.05 in both comparisons. FT and FC have essentially similar behaviours for both values of n (p=0.35). In the HCI case, using n=4 (best value for all three techniques), the comparison of TST relative to FT and to FC evidences in both cases statistically significant differences, respectively with p= 0.01 and 0.005. Using n=5, TST still dominates over FC (p=0.02) but no longer over FT (p=0.23).

This confirms that TST is able to produce significantly more compact clusters, in particular when using the optimal value for n, mainly due to the use of topic vector weighing (see Section 3.3). We obtained similar results by selecting the maximum *PCAES* over 20 runs. In the SW case, given 4 clusters we obtained PCAES=2.09 for TST, 1.42 for FT, and 1.17 for FC (with 5 clusters the values were 2.89, 1.19 and

1.16). In HCI, given 4 clusters, we obtained *PCAES*=0.79 for TST, -0.32 for FT, and -0.28 for FC.

Interestingly, the cluster sets in SW seems to be more compact than the HCI ones. The results seem to contradict the human experts, who actually showed a higher degree of agreement when composing HCI communities. However, what is considered the best clustering according to these metrics is not always perceived as such by human experts. The reasons why HCI clusters have a lower PCEAS may in fact simply lie in the fact that HCI authors tend to address more heterogeneous themes and work across different communities. On the contrary a number of people working in the Semantic Web tend to publish most of their work within a particular community. We thus may need novel evaluation metrics to be able to take in account the peculiarities associated with different topic-based research communities.

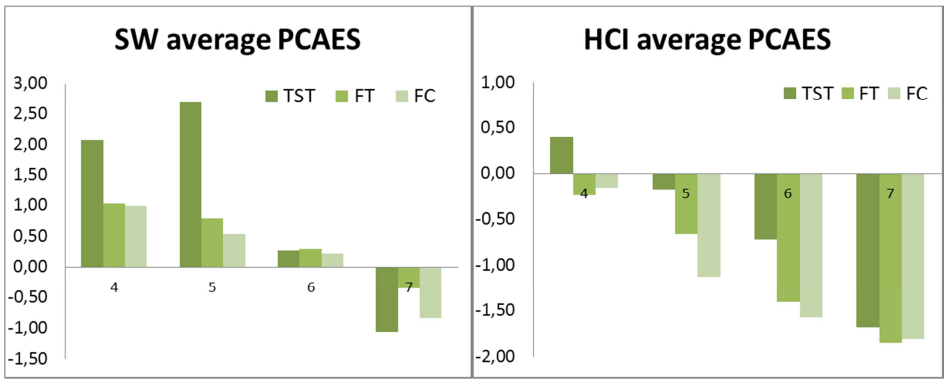


Figure 6. Average PCAES for Semantic Web and HCI.

6 Conclusions

In this paper we have presented TST, a novel approach to automatically detect diachronic topic-based communities –i.e., communities of researchers who work on semantically related topics at the same time.

The user study presented in this paper shows that our approach yields results that are statistically consistent with those obtained from domain experts. The study also shows that the adoption of i) a semantic characterization of the research topics (see Section 3.1), ii) the topic vector weighing (see Section 3.2) and iii) the ATTS metric (see Section 3.3) dramatically increases the quality of the detected communities. Moreover, according to the PCAES index, the use of topic vector weighing also increases significantly the degree of compactness of the detected communities.

Our approach opens up many interesting directions of work. Currently we are working on a novel method to automatically detect different kinds of patterns in the research flow, such as the merging/splitting of different communities or the occurrence of topic shifts within a community. In addition, we also plan to build on this approach to develop effective methods to measure the impact of specific events on the research environment, such as the introduction of a new technology or the award of a new grant. Such functionality is of particular importance to research managers and funding bodies, who need better tools to measure the impact of policy

decisions. Finally, we plan to work on a predictive technique, aimed at forecasting the behaviour that a community is likely to exhibit in the short and medium term.

References

1. 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)
2. Ding, Y. 2011. Community detection: topological vs. topical. Journal of Infometrics, 5(4).
3. Osborne, F., Motta, E.: Mining Semantic Relations between Research Areas. 11th International Semantic Web Conference (ISWC 2012). Boston, MA.
4. Osborne, F., Motta, E., Mulholland, P.: Exploring Scholarly Data with Rexplore. In Proceedings of the 12th International Semantic Web Conference. (2013)
5. Smyth Guimera, R., Amaral, L. A. N.: Functional cartography of complex metabolic networks. Nature, 433(7028), 895-900. (2005)
6. Smyth, S., White, S. A spectral clustering approach to finding communities in graphs. 5th SIAM International Conference on Data Mining (pp. 76-84). (2005)
7. 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)
8. Upham, S. P., Rosenkopf, L., Ungar, L. H.: Innovating knowledge communities. Scientometrics, 83(2), 525-554. (2010)
9. Racherla, P., Hu, C.: A social network perspective of tourism research collaborations. Annals of Tourism Research, 37(4), 1012-1034. (2010)
10. Wang, S., Jing, F., He, J., Du, Q., Zhang, L.: Igroup: presenting web image search results in semantic clusters. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 587-596). ACM. (2007)
11. Schrammel, J., Leitner, M., Tscheligi, M.: Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2037-2040). ACM. (2009)
12. Hofmann, T.: Probabilistic latent semantic indexing. 22nd ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 50–57), Berkeley, CA. (1999)
13. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1033. (2003)
14. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: extraction and mining of academic social networks. 14th Int. Conference on Knowledge Discovery and Data Mining, pp. 990-998. (2008)
15. Mei, Q., Cai, D., Zhang, D., Zhai, C.: Topic modeling with network regularization. 17th international conference on World Wide Web (pp. 101-110). ACM. (2008)
16. Erétéo, G., Gandon, F., & Buffa, M.: Semtagp: semantic community detection in folksonomies. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on (Vol. 1, pp. 324-331). IEEE. (2011)
17. Holme, P., & Saramäki, J.: Temporal networks. Physics reports, 519(3), 97-125. (2012).
18. Bezdek, J. C., Ehrlich, R., Full, W. 1984. FCM: The fuzzy c-means clustering algorithm. Computers and Geosciences, 10(2), 191-203. (2012)
19. Van Eck, N. J., Waltman, L.: Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523-538. (2010)
20. Yan, E., Ding, Y., Jacob, E.: Overlaying communities and topics. Scientometrics, 90(2), 499-513. (2012)
21. Wu, K. L., Yang, M. S.: A cluster validity index for fuzzy clustering. Pattern Recognition Letters, 26(9), 1275-1291. (2005)
22. Chiu, S. L.: Fuzzy model identification based on cluster estimation. Journal of intelligent and Fuzzy systems, 2(3), 267-278. (1994)