Forecasting Technology Migrations by means of the Technology-Topic Framework

Francesco Osborne, Andrea Mannocci, Enrico Motta

Knowledge Media Institute, The Open University, MK7 6AA, Milton Keynes, UK
{francesco.osborne,andrea.mannocci,enrico.motta}@open.ac.uk

Abstract. Technologies such as algorithms, applications and formats usually originate in the context of a specific research area and then spread to several other fields, sometimes with transformative effects. However, this can be a slow and inefficient process, since it not easy for researchers to be aware of all interesting approaches produced by unfamiliar research communities. We address this issue by introducing the Technology-Topic Framework, a novel approach which uses a semantically enhanced technology-topic model and machine learning to forecast the propagation of technologies across research areas. The aim is to foster the knowledge flow by suggesting to scholars technologies that may become relevant to their research field. The system was evaluated on a manually curated set of 1,118 technologies in Semantic Web and Artificial Intelligence and the results of the evaluation confirmed the validity of our approach.

Keywords: Scholarly Data, Semantic Web, Technology Propagation, Technology Spreading, Bibliographic Data, Scholarly Ontologies.

1 Introduction

Researchers constantly reuse ideas, methods and materials from different research areas and need to be aware of the most recent results which are potentially relevant to their work. For example, Semantic Web technologies were first created by research communities in areas such as Artificial Intelligence, Knowledge Base Systems, Formal Ontology and others. Subsequently these technologies contributed to a variety of other research areas, e.g., Information Retrieval, Human Computer Interaction, Biology, and others. However, given the steady increase of the rate of production of scientific knowledge, it is becoming increasingly harder for researchers to track all potentially relevant results produced by all potentially relevant research communities.

We address this issue by introducing the Technology-Topic Framework (TTF), a novel approach which uses a semantically enhanced technology-topic model and machine learning to forecast the propagation of technologies to research areas. TTF characterises the evolution of technologies as a set of matrices representing the number of documents associated with a research topic during a year and applies machine learning on these data to forecast the research field that will likely adopt a technology in the following years. The aim is to foster the knowledge flow by suggesting to scholars technologies that may be relevant to their research field.

2 Technology-Topic Framework

The Technology-Topic Framework takes as input three knowledge bases: i) a dataset of research papers, described by means of their titles, abstracts, and keywords; ii) an ontology of research areas, describing topics and their relationships, and iii) a list of input technologies, associated to the relevant publications in the research paper dataset.

In the study presented in this paper, we used as dataset a dump of the Scopus database in the 1990-2013 period, containing about 16 million papers in the field of Computer Science. As a reference topic ontology, we adopted the Computer Science Ontology (CSO), created to represent topics in the Rexplore system [1] and currently trialled by Springer Nature to classify proceedings in the field of Computer Science [2], such as the well-known LNCS series. CSO was created by applying the Klink-2 algorithm [3] on the Rexplore dataset, which consists of about 16 million publications, mainly in the field of Computer Science. It includes about 17k topics linked by 70k semantic relationships. Finally, the list of technologies comprises a manually curated dataset of 1,118 technologies in Semantic Web and Artificial Intelligence. We first selected an initial set of about 2,000 technologies by running TechMiner [4] on a set of 3,000 papers in Semantic Web. We then manually cleaned and enriched the resulting dataset by discarding wrong entities that were not explicitly described as technologies in research papers and by adding 500 other technologies extracted from Wikipedia pages listing Artificial Intelligence and Machine Learning algorithms and methods.

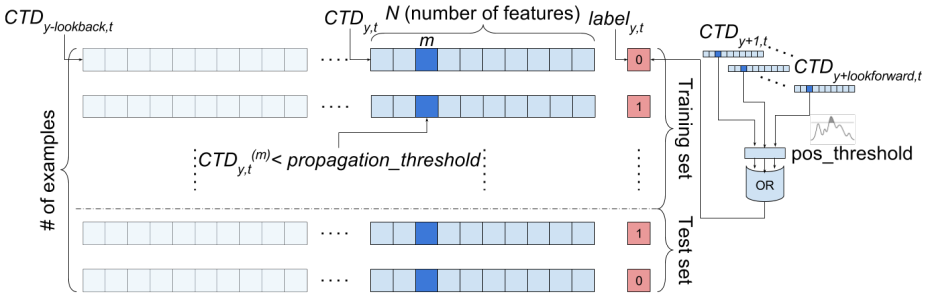

Figure 1. Construction of examples for topic *m*.

TTF builds for each year a matrix that characterises technologies in terms of their number of publications in different research topics. To this end, it exploits the topic ontology, associating to each paper i) all the topics in CSO whose label is found in the title, the abstract or the keyword set, as well as ii) all *skos:broaderGeneric* and iii) all the *relatedEquivalent* areas of the topics in the initial set. Then, for each technology, it counts the number of papers for each topic in each year. The result is a sequence of matrices, one matrix for each year, in which rows represent technologies, columns represent topics, and cells contain the number of publications of a technology for a given topic in a given year.

The forecasting of technology propagation is treated as M separate classification problems, one for each topic of interest. For the m^{th} topic, the sequence of technology-topic matrices is processed to extract *examples* to be fed to the machine learning models. For each topic m, we select as examples only the ones in which the technology t is associated in year y with fewer than $propagation\_threshold$ publications (2 by default) in m. Each example is characterized by the cumulative topic distribution (*CTD* in the

figure) for the year y together with the *CTD*s of the last $lookback$ years (2 by default). The example is labelled as positive if the technology will become associated with at least $pos\_threshold$ publications for topic m (5 by default) over a span of $lookforward$ years (5 by default).

3 Evaluation

We evaluated TTF on 1,118 technologies and 173 topics in the field of Computer Science during the 1990-2013 period[1]. We selected as *training set* examples in the 1990-2004 period and as *test set* examples in the 2005-2008 period. We chose these intervals as they allowed us to label the examples in the test set using a window of five years (2009-2013). We considered only examples about technologies which existed for no more than 5 years and we simulated a realistic situation by assuming 2005 as current year and not using any information successive to that year to label the examples in the training set.

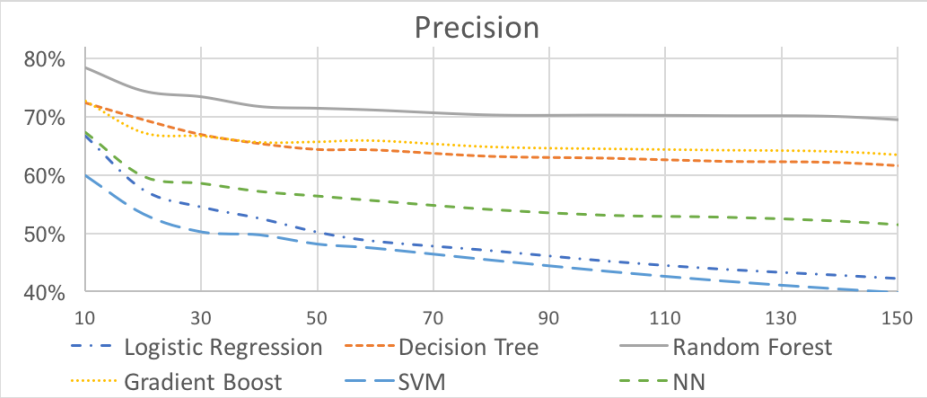

Figure 2. Average precision of the six machine learning approaches on the first *n* topics.

We selected the 173 topics which were associated with at least 30 positive examples in both the training and the test sets in the period under analysis and trained a classifier for each of them. Each topic classifier was trained on average on 5,136 ± 240 examples (for a total of 888,633 examples) and was evaluated on 679 ± 90 examples (for a total of 117,516 examples). We tested six machine learning algorithms: Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, Feed Forward Neural Network, and Gradient Boosting. The tuning of hyper-parameters used for each model was performed by a twofold cross-validation over the training set.

Figure 2 and Figure 3 show respectively the precision and recall obtained by the six algorithms on the first *n* topics, ordered by the number of positive labels in the test set. Random Forest yielded the best result in terms of precision. For the first 20 topics, its precision was over 74.4%, significantly higher ($p<0.0001$) than the value of 69.4% obtained with Decision Tree and 67.2% with Gradient Boosting. Also, considering the first 100 topics, Random Forest scored best, with 70.2% versus 62.9% of Decision Tree

[1] The evaluation materials, the background knowledge, and the code are available at http://rexplore.kmi.open.ac.uk/TTF

and 64.4% of Gradient Boosting ($p<0.0001$). Conversely, Gradient Boosting performed best in terms of recall. For the first 20 topics, it scored 47.2%, significantly higher than the value of 44.7% for Random Forest ($p=0.038$) and the value of 42.5% for Decision Tree ($p<0.0001$). For the first 100 topics, the Gradient Boosting recall was 35.1%, again significantly higher ($p<0.0001$) than 32% for Random Forest and 31.5% for Decision Tree.

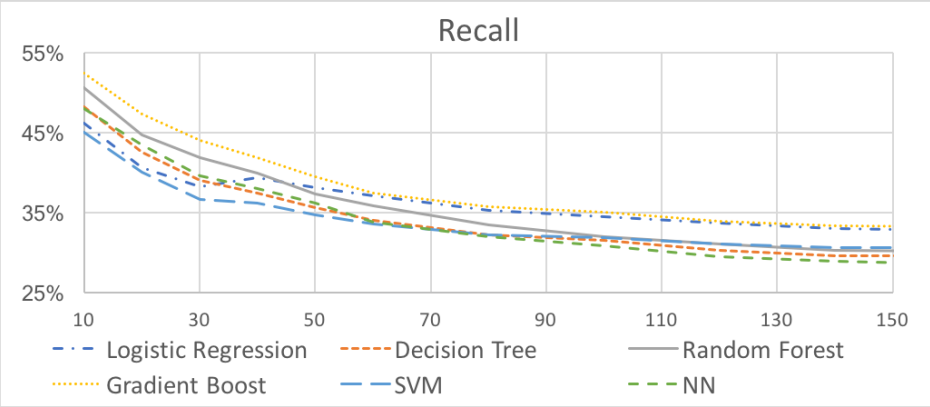

Figure 3. Average recall of the six machine learning approaches on the first *n* topics.

4 Conclusions

The evaluation confirms that TTF is able to learn from historical spreading patterns and forecast technology propagation with good precision. For example, TTF was able to forecast the propagation of Semantic Web formats (e.g., OWL, SKOS, SWRL) to several research areas, such as Bioinformatics, Social Networks, e-Learning, and so on.

As next step, we plan to enrich the forecasting model by considering text generated features and possibly deriving additional features from external knowledge bases and social media. We also intend to include in the analysis a wider set of fields, including Biology, Social Science and Engineering. Finally, we plan to create a web application for suggesting to researchers technologies which may contribute to their field.

References

1. Osborne, F., Motta, E., Mulholland, P.: Exploring scholarly data with Rexplore. In The Semantic Web–ISWC 2013 (pp. 460-477). Springer Berlin Heidelberg. (2013)
2. Osborne, F., Salatino, A., Birukou, A. and Motta, E.: Automatic classification of springer nature proceedings with smart topic miner. In International Semantic Web Conference (pp. 383-399). Springer International Publishing. (2016)
3. Osborne, F. and Motta, E.: Klink-2: integrating multiple web sources to generate semantic topic networks. In International Semantic Web Conference (pp. 408-424). Springer International Publishing. (2015)
4. Osborne, F., de Ribaupierre, H. and Motta, E.: TechMiner: Extracting Technologies from Academic Publications. In Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings 20 (pp. 463-479). Springer International Publishing. (2016)