Instrumenting Continuous Knowledge Extraction, Sharing, and Benchmarking

Marco Brambilla, Emanuele Della Valle,
Andrea Mauri, and Riccardo Tommasini

Politecnico di Milano, DEIB, Data Science Lab.
Via Ponzio 34/5, I-20133, Milano, Italy
{name.surname}@polimi.it

Abstract. Keeping the pace with the faster and faster evolution of knowledge is becoming a challenge, especially for researchers and knowledge workers. We propose a vision towards a set of (possibly integrated) publicly available tools that can help on this. To this purpose, we expect tools that can improve effectiveness of knowledge extraction, storage, analysis, publishing and experimental benchmarking. This could be extremely beneficial for the entire research community across fields and interests. We describe our vision in this direction and we demonstrate its feasibility with some exemplary tools that we developed and that we shared as public resources to be used by the research community.

1 Introduction

Nanos gigantum humeris insidentes
(Bernard of Chartres, 1115 AD ca.)

Science aims at creating new knowledge upon the existing one, from the observation of physical phenomena, their modeling and empirical validation. This combines the well known motto **"standing on the shoulders of giants"** (attributed to Bernard of Chartres and subsequently rephrased by Isaac Newton) with the need of trying and validating new experiments.

However, knowledge in the world continuously evolves, at a pace that cannot be traced even by large crowdsourced bodies of knowledge such as Wikipedia. A large share of generated data are not currently analysed and consolidated into exploitable information and knowledge [1]. In particular, the process of ontological knowledge discovery tends to focus on the most popular items, those which are mostly quoted or referenced, and is less effective in discovering less popular items, belonging to the so-called long tail , i.e. the portion of the entity's distribution having fewer occurrences [2].

This becomes a challenge for practitioners, enterprises and scholars / researchers, which need to be up to date to innovation and emerging facts. The scientific community also need to make sure there is a structured and formal way to represent, store and access such knowledge, for instance as ontologies or linked data sources.

Our idea is to propose **a vision towards a set of (possibly integrated) publicly available tools that can help scholars keeping the pace with the evolving knowledge**. This implies the capability of integrating informal sources, such as social networks, blogs, and user-generated content in general. One can conjecture that somewhere, within the massive content shared by people online, any low-frequency, emerging concept or fact has left some traces. The challenge is to detect such traces, assess their relevance and trustworthiness, and transform them into formalized knowledge [5].

An appropriate set of tools that can improve effectiveness of knowledge extraction, storage, analysis, publishing and experimental benchmarking could be extremely beneficial for the entire research community across fields and interests.

2 Our Vision towards Continuous Knowledge Extraction and Publishing

We foresee a paradigm where knowledge seeds can be planted, and subsequently grow, finally leading to the generation and collection of new knowledge, as depicted in the exemplary process in Figure 1 [2].

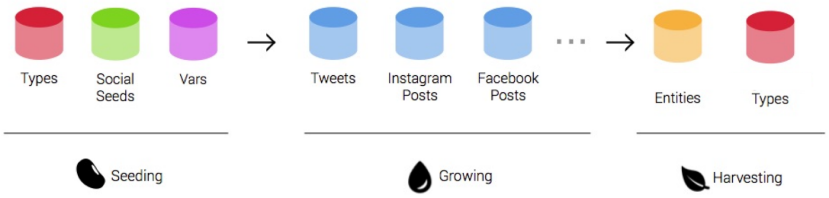

Fig. 1. Exemplary process of knowledge seeding, growing, and harvesting for extracting concepts from social media.

We advocate for a set of tools that, when implemented and integrated, enable the following perspective reality:

- possibility of selecting any kind of source of raw data, independently of their format, type or semantics (spanning quantitative data, textual content, multimedia content), covering both data streams or pull-based data sources;
- possibility of applying different data cleaning and data analysis pipelines to the different sources, in order to increase data quality and abstraction / aggregation;
- possibility of integrating the selected sources;
- possibility of running homogeneous knowledge extraction processes of the integrated sources;

– possibility of publishing the results of the analysis and semantic enrichment as new and further (richer) data sources and streams, in a coherent, standard and semantic way.

This enables generation of new sources which in turn can be used in subsequent knowledge extraction processes of the same kind. The results of this process must be **available at any stage to be shared for building an open, integrated and continuously evolving knowledge** for research, innovation, and dissemination purposes.

3 A Preliminary Feasibility Perspective

Whilst beneficial and powerful, the vision we propose is far from being achieved nowadays. However, we are convinced that the vision is not out of reach in the mid term. To give a hint of this, we report here our experience with the research, design and implementation of a few tools that point in the proposed direction:

1. **Social Knowledge Extractor (SKE) is a publicly available tool for discovering emerging knowledge by extracting it from social content.** Once instrumented by experts through very simple initialization, the tool is capable of finding emerging entities by means of a mixed syntactic-semantic method. The method uses seeds, i.e. prototypes of emerging entities provided by experts, for generating candidates; then, it associates candidates to feature vectors, built by using terms occurring in their social content, and then ranks the candidates by using their distance from the centroid of seeds, returning the top candidates as result. The tool can run continuously or with periodic iterations, using the results as new seeds. Our research on this has been published in [3], a simplified implementation is currently available online for demo purposes at:
`http://datascience.deib.polimi.it/social-knowledge/`,
and the code is available as open-source under an Apache 2.0 license on GitHub at:
`https://github.com/DataSciencePolimi/social-knowledge-extractor`.

2. **TripleWave is a tool for disseminating and exchanging RDF streams on the Web.** At the purpose of processing information streams in real-time and at Web scale, TripleWave integrates nicely with RDF Stream Processing (RSP) and Stream Reasoning (SR) as solutions to combine semantic technologies with stream and event processing techniques. In particular, it integrates with an existing ecosystem of solutions to query, reason and perform real-time processing over heterogeneous and distributed data streams. TripleWave can be fed with existing Web streams (e.g. Twitter and Wikipedia streams) or time-annotated RDF datasets (e.g. the Linked Sensor Data dataset) and it can be invoked through both pull- and push-based mechanisms, thus enabling RSP engines to automatically register and receive data from TripleWave. The tool has been described in [4] and the code

is available as open-source on GitHub at
`https://github.com/streamreasoning/TripleWave/`.

3. **RSPlab enables efficient design and execution of reproducible experiments**, as well as sharing of the results. It integrates two existing RSP benchmarks (LSBench and CityBench) and two RSP engines (C-SPARQL engine and CQELS). It provides a programmatic environment to: deploy in the cloud RDF Streams and RSP engines; interact with them using TripleWave and RSP Services; continuously monitor their performances and collect statistics. RSPlab is released as open-source under an Apache 2.0 license is currently under submission at *ISWC - Resources Track* and is available on GitHub at
`https://github.com/streamreasoning/rsplab`.

4 Conclusions

We believe that knowledge intaking by scholars is going to become more and more time consuming and expensive, due to the amount of knowledge that is being built and shared everyday. We envision a comprehensive approach based on integrated tools that allow data collection, cleaning, integration, analysis and semantic representation that can be run continuously for **keeping the formalized knowledge bases aligned with the evolution of knowledge, with limited cost and high recall on the facts and concepts that emerge or decay**. These tools do not need to be implemented by the same vendor or provider; we instead advocate for opensource publishing of all the implementations, as well as for the definition of an agreed-upon integration platform that allows them all to integrate appropriately.

5 Outlook on Research Resource Sharing

As we envisioned an ecosystem that includes, but is not limited to, modules for extraction, sharing and benchmarking, two research questions require investigation in the immediate future.

First, how can we design and publish new resources for such an ecosystem? Do they exist already? It is important to understand what else is available out there. Researchers commonly support their scientific studies with resources that can benefit the whole community, if released. The release process must comply with a scientific method that ensures repeatability and reproducibility. However, a standard agreed-upon methodology that guide this process does not exists yet.

Second, how should we combine these resources towards shared research workflows? To investigate this research question, we need a platform that enables researchers to deploy their resources and interact with the ecosystem. Therefore, we call for an open discussion about how this integration should be done. References

References

1. Ackoff, R.L.: From data to wisdom. Journal of applied systems analysis 16(1), 3–9 (1989)
2. Brambilla, M., Ceri, S., Daniel, F., Valle, E.D.: On the quest for changing knowledge. In: Proceedings of the Workshop on Data-Driven Innovation on the Web - DDI '16. ACM Press (2016), https://doi.org/10.1145%2F2911187.2914582
3. Brambilla, M., Ceri, S., Valle, E.D., Volonterio, R., Salazar, F.X.A.: Extracting Emerging Knowledge from Social Media. In: Proceedings of the 26th International Conference on World Wide Web - WWW '17. ACM Press (2017), https://doi.org/10.1145%2F3038912.3052697
4. Mauri, A., Calbimonte, J.P., Dell'Aglio, D., Balduini, M., Brambilla, M., Valle, E.D., Aberer, K.: TripleWave: Spreading RDF Streams on the Web. In: Lecture Notes in Computer Science, pp. 140–149. Springer International Publishing (2016), https://doi.org/10.1007%2F978-3-319-46547-0_15
5. Stieglitz, S., Dang-Xuan, L., Bruns, A., Neuberger, C.: Social Media Analytics. Business & Information Systems Engineering 6(2), 89–96 (feb 2014), https://doi.org/10.1007%2Fs12599-014-0315-7