Semantic Culturomics (vision paper)

Fabian M. Suchanek
University of Paris-Saclay
Paris, France
fabian@suchanek.name

Nicoleta Preda*
University of Versailles
Versailles, France
nicoleta@preda.fr

ABSTRACT

Newspapers are testimonials of history. The same is increasingly true of social media such as online forums, online communities, and blogs. By looking at the sequence of articles over time, one can discover the birth and the development of trends that marked society and history – a field known as "Culturomics". But Culturomics has so far been limited to statistics on keywords. In this vision paper, we argue that the advent of large knowledge bases (such as YAGO [37], NELL [5], DBpedia [3], and Freebase) will revolutionize the field. If their knowledge is combined with the news articles, it can breathe life into what is otherwise just a sequence of words for a machine. This will allow discovering trends in history and culture, explaining them through explicit logical rules, and making predictions about the events of the future. We predict that this could open up a new field of research, "Semantic Culturomics", in which no longer human text helps machines build up knowledge bases, but knowledge bases help humans understand their society.

1. INTRODUCTION

Newspapers are testimonials of history. Day by day, news articles record the events of the moment – for months, years, and decades. The same is true for books, and increasingly also for social media such as online forums, online communities, and blogs. By looking at the sequence of articles over time, one can discover the trends, events, and patterns that mark society and history. This includes, e.g., the emancipation of women, the globalization of markets, or the fact that upheavals often lead to elections or civil wars. Several projects have taken to mining these trends. The Culturomics project [27], e.g., mined trends from the Google Book Corpus. We can also use the textual sources to extrapolate these trends to the future. Twitter data has been used to make predictions about election results, book sales, or consumer behavior. However, all of these analyses were mostly restricted to the appearance of keywords. The analysis of the role of genders in [27], for instance, was limited to comparing the frequency of the word "man" to the frequency of the word "woman" over time. It could not find out which men and women were actually gaining importance, or in which professions. This brings us to the general problem of such previous analyses: They are mostly limited to counting the occurrences of words. So far, no automated approach can actually bring deep insight into the meaning of news articles over time. Meaning, however, is the key for understanding roles of politicians, interactions between people, or reasons for conflict. For example, if a sentence reads "Lydia Taft cast her vote in 1756", then this sentence gets its historic value only if we know that Lydia Taft was a woman, and that she lived in the United States, and that women's suffrage was not established there until 1920. All of this information is lost if we count just words.

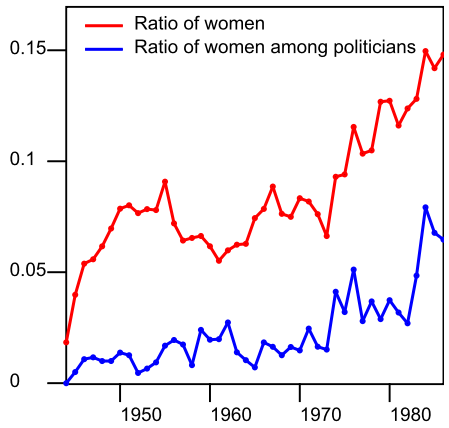

Figure 1: The gender gap, mined from Le Monde and YAGO

We believe that this barrier will soon be broken, because we now have large commonsense knowledge bases (KBs) at our disposal: YAGO [37], NELL [5], TextRunner [4], DBpedia [3], and Freebase (`http://freebase.com`). These KBs contain knowledge about millions of people, places, organizations, and events. The creation of these KBs is an ongoing endeavor. However, with this vision paper, we take a step ahead of these current issues in research, and look at what can already be achieved with these KBs: If their knowledge is combined with the news articles, it can breathe life into what is otherwise just a sequence of words for a machine. Some news organisations[1] participate already in the

*This work was supported in part by KISS, a research project funded by French ANR Call INS 2011

Proceedings of the VLDB Endowment, Vol. 7, No. 12

[1] `http://developer.nytimes.com/docs/semantic_api`

effort of annotating textual data with entities from KBs. Once people, events, and locations have been identified in the text, they can be unfolded with the knowledge from the KB. With this combination, we can identify not just the word "woman", but actually mentions of people of whom the KB knows that they are female. Figure 1 shows a proof of concept that we conducted on YAGO and the French newspaper Le Monde [15]. It looks at all occurrences of people in the articles, and plots the proportion of women both in general and among politicians. Women are mentioned more frequently over time, but the ratio is smaller among politicians. Such a detailed analysis is possible only through the combination of textual data and semantic knowledge.

2. SEMANTIC CULTUROMICS

> Semantic Culturomics is the large-scale analysis of text documents with the help of knowledge bases, with the goal of *discovering*, *explaining*, and *predicting* the trends and events in history and society.

Semantic Culturomics could for example answer questions such as: "In which countries are foreign products most prevalent?" (where the prevalence can be mined from the news, and the producer of a product, as well as its nationality, comes from the KB), "How long do celebrities usually take to marry?" (where co-occurrences of the celebrities can be found in blogs, and the date of marriage and profession comes from the KB), "What are the factors that lead to an armed conflict?" (where events come from newspapers, and economic and geographical background information comes from the KB), "Which species are likely to migrate due to global warming?" (where current sightings and environmental conditions come from textual sources, and biological information comes from the KB). None of these queries can be answered using only word-based analysis. The explanations that Semantic Culturomics aims at could take the form of logical rules such as "A politician who was involved in a scandal often resigns in the near future". Such rules can *explain* particular past events by pointing to a general pattern of history with past instances. They can also be used to make predictions, and to deliver an explication as to *why* a certain prediction is made.

Semantic Culturomics would turn around a long-standing paradigm: Up to now, all information extraction projects strive to distill computer-understandable knowledge from the textual data of the Web. Seen this way, human-produced text helps computers structure and understand this world. Semantic Culturomics would put that paradigm upside down: It is no longer human text that helps computers build up knowledge, but computer knowledge that helps us understand human text – and with it human history and society.

3. STATE OF THE ART

Digital Humanities and Culturomics. The Digital Humanities make historical data digitally accessible in order to compare texts, visualize historic connections, and trace the spread of new concepts. The seminal paper in this area, [27], introduced the concept of "Culturomics" as the study of cultural trends through the quantitative analysis of digitized texts. This work was the first large-scale study of culture through digitized texts. Yet, as explained above, it remains bound to the words of the text. The work has since been advanced [20, 1], but still remains confined to counting occurrences and co-occurrences of words. Closer to our vision, the GDELT project [21] annotates news articles with entities and event types for deeper analysis. The focus is on the visualisation of trends. In contrast, Semantic Culturomics aims also at providing *explanations* for events, which become possible by the background knowledge from the KB.

Event prediction. A recent work [33] mined the New York Times corpus to predict future events. This work was the first that aimed at predicting (rather than modeling) events. Of particular interest is the ability to bound the time point of the predicted events. The authors make use of key phrases in the text, as well as semantic knowledge to some degree. A recent follow-up work [34] extended the analysis to Web queries. Another approach modeled causality of events by using background data from the Linked Open Data cloud [32]. These works were the first to address the prediction of events at large scale. [32] goes a long way towards the identification of events and causality. In a similar vein, Recorded Future[2], a company, has specialised in the detection and the prediction of events with the help of a KB [36]. However, these works built classifiers for predictions rather than explicit patterns in the form of logical rules that we aim at. Furthermore, Semantic Culturomics would model the interplay between text and semantic knowledge in a principled way, and thus unify the prediction of future events with the modeling of past trends.

Predictive analytics. Businesses and government agencies alike analyze data in order to predict people's behavior[3]. There is a business-oriented conference[4] dedicated to these projects. Therefore, we believe that this endeavor should preferably be studied also in a public, academic, space. Furthermore, predictive analytics is mostly centered on a specific task in a specific domain. A model that can predict sales of a certain product cannot be used to predict social unrest in unstable countries. Semantic Culturomics, in contrast, aims at a broader modeling of the combination of textual sources and knowledge bases.

Social Media Analysis. Recently, researchers have increasingly focused on social media to predict social trends and social movements. They have used Twitter data and blogs to predict crowd phenomena, including illnesses [18], box office sales, the stock market, consumer demand, book sales, consumer behavior, and public unrest (see, e.g., [16] and references therein). Other Web data has been used to predict the popularity of a news article [13] or to analyze elections [39]. These works have demonstrated the value of Twitter for event prediction. However, they always target a particular phenomenon. We believe that what is needed is a systematic and holistic study of textual data for both explanation of the past and prediction of the future.

Machine Reading. Several projects have looked into mining the Web at large scale for facts [5, 4, 28, 38]. Recent work has mined the usual order of events from a corpus [40], the precedence relationships between facts in a KB [41], and implicit correlations in a KB [19]. Several of these methods can be of use for Semantic Culturomics. However, they can

[2] http://www.recordedfuture.com
[3] http://www.forbes.com/sites/gregpetro/2013/06/13/what-retail-is-learning-from-the-nsa/
[4] http://www.predictiveanalyticsworld.com/

only be an ingredient to the project, because Semantic Culturomics aims at mining explicit logical rules, together with a temporal dimension, from text and KBs.

Enabling Technologies. Our vision of Semantic Culturomics can build on techniques from entity recognition, event detection, rule mining, and information extraction. We detail next how these techniques would have to be advanced.

4. CHALLENGES

Mining text in combination with knowledge bases is no easy endeavor. The key challenges would be as follows:

Modeling hybrid data. KBs contain knowledge about entities, facts, and sometimes logical axioms. Text, on the other hand, determines the importance of an entity, the co-occurrence of entities, the location of entities in time, the type of events in which an entity is involved, the topic of an entity, and the actions of entities. Thus, Semantic Culturomics has to operate on a hybrid space of textual data and semantic knowledge. KB information is usually represented in RDF. RDF, however, cannot model time, let alone textual content. Other approaches can represent hybrid data, but do not allow KB-style reasoning [22, 12, 44, 6].

Semantic Culturomics calls for a new data model, which can represent entities and their mentions, textual patterns between entities, the dimension of time, and out-of-KB entities. In analogy to an OLAP data cube, this data model could be called a "Semantic Cube". It should support truly hybrid query operations such as: do a phrase matching to find all text parts that contain names of entities with a certain property; choose one out of several disambiguations for a mention; given a logical rule, remove all facts that match the antecedent, and replace them by the succedent; dice the cube so that the text contains all paraphrases of a relation name. The goal is to develop a query language that subsumes all types of analyses that can be of interest on hybrid data of text and semantic KBs in general.

Identify events and entities. Given, for example, a history of the newspaper articles of a certain region, we want to be able to predict the crime rate, voting patterns, or the rise of a certain person to political prominence. In order to mine trends from a given text corpus, we have to develop methods that can load the textual data (jointly with the structured knowledge) into a Semantic Cube. This requires first and foremost the identification of entities and events in the textual corpora.

There is a large body of prior work on information extraction, and on event mining in news articles [7, 24, 43]. However, most of this work is non-ontological: It is not designed to connect the events to types of events and to entities of the KB. Several works have addressed the problem of mapping entity mentions to known entities in the KB (e.g., [14, 26]). However, these works can deal only with entities that are known to the KB. The challenge remains to handle new entities with their different names. For example, if Lady Gaga is not in the KB and is mentioned in the text, we want to create a new entity Lady Gaga. However, if we later find Stefani Germanotta, in the text, then we do not want to introduce a new entity, but rather record this mention as an occurrence of Lady Gaga with a different name.

Empower rule mining. The goal of Semantic Culturomics is not only to mine trends, but also to explain them. These explanations will take the form of logical rules, weighted with confidence and support measures. Rule mining, or inductive logic programming, has been studied in a variety of contexts [11, 23, 29, 10, 8, 31, 17]. Yet, for Semantic Culturomics we envision rules that cannot be mined with current approaches.

We would like to mine *numerical rules* such as "Mathematicians publish their most remarkable works before their 36th anniversary", or "The spread between the imports and the exports of a country correlates with its current account deficit". Previous work on numeric rule mining [31, 25] was restricted to learning intervals for numeric variables. Other approaches can learn a function [17, 9], but have been tested only on comparatively small KBs (less than 1000 entities) – far short of the millions of entities that we aim at.

We also aim to mine *temporal rules* such as "An election is followed by the inauguration of a president". These should also predict the time of validity of literals. First work in this direction [30] has been tried on just toy examples.

Another challenge is to mine of rules with *existential variables*, such as "People usually have a male father and a female mother". Such rules have to allow several literals in the succedent, meaning that Horn rule mining approaches and concept learning approaches become inapplicable. Statistical schema induction [42] can provide inspiration, but has not addressed existential rule learning in general.

We would also need *rules with negation*, such as "People marry only if they are yet not married". Such rules have been studied [31], but not under the Open World Assumption. In this setting, learning rules with negation risks learning the patterns of incompleteness in the KB rather than negative correlations in reality. Furthermore, there exist many more statements outside the KB than inside in the KB, meaning that we risk mining a large number of irrelevant negative statements.

Finally, we want to mine rules that take into account the *textual features* that the hybrid space brings. These are features such as the importance of an entity or the textual context in which an entity (or a pair of entities) appears. [35] mines rules on textual phrases, but does not take into account logical constraints from the KB. If we succeed in mining rules that take into account textual features, the reward will be highly attractive: Finally, we will be able to explain *why* a certain event happened – by giving patterns that have led to this type of events in the past.

Privacy. Predicting missing facts means also that some facts will no longer be private. For instance, consider a rule that can predict the salary of a person given the diploma, the personal address, and the employment sector. Smart social applications could warn the user when she discloses information that, together with already disclosed information, allows predicting private data. The intuition is that automatic rule mining could reveal surprising rules that humans may not directly see or may ignore, as shown in [2].

5. CONCLUSION

In this vision paper, we have outlined the idea of *Semantic Culturomics*, a paradigm that uses semantic knowledge bases in order to give meaning to textual corpora such as news and social media. This idea is not without challenges, because it requires the link between textual corpora and semantic knowledge, as well as the ability to mine a hybrid data model for trends and logical rules. If Semantic Culturomics succeeds, however, it would add an interesting twist to the digital humanities: semantics. Semantics turns the

texts into rich and deep sources of knowledge, exposing nuances that today's analyses are still blind to. This would be of great use not just for historians and linguists, but also for journalists, sociologists, public opinion analysts, and political scientists. They could, e.g., search for mentions of politicians with certain properties, for links between businessmen and judges, or for trends in society and culture, conditioned by age of the participants, geographic location, or socio-economic indicators of the country. Semantic Culturomics would bring a paradigm shift, in which no longer human text is at the service of knowledge bases, but knowledge bases are at the service of human understanding.

6. REFERENCES

[1] O. Ali, I. N. Flaounas, T. D. Bie, N. Mosdell, J. Lewis, and N. Cristianini. Automating news content analysis: An application to gender bias and readability. In *WAPA*, 2010.

[2] N. Anciaux, B. Nguyen, and M. Vazirgiannis. Limiting data collection in application forms: A real-case application of a founding privacy principle. In *PST*, 2012.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. DBpedia: A Nucleus for a Web of Open Data. In *ISWC*, 2007.

[4] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open Information Extraction from the Web. In *IJCAI*, 2007.

[5] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010.

[6] D. Colazzo, F. Goasdoué, I. Manolescu, and A. Roatis. Rdf analytics: Lenses over semantic graphs. In *WWW*, 2014.

[7] A. Das Sarma, A. Jain, and C. Yu. Dynamic relationship and event discovery. In *WSDM*, 2011.

[8] L. Dehaspe and H. Toironen. Discovery of relational association rules. In *Relational Data Mining*. 2000.

[9] N. Fanizzi, C. d'Amato, and F. Esposito. Towards numeric prediction on owl knowledge bases through terminological regression trees. In *ICSC*, 2012.

[10] L. Galárraga, C. Teflioudi, K. Hose, and F. M. Suchanek. Amie: association rule mining under incomplete evidence in ontological knowledge bases. In *WWW*, 2013.

[11] B. Goethals and J. Van den Bussche. Relational association rules: getting warmer. In *Pattern Detection and Discovery*. 2002.

[12] J. Han. Mining heterogeneous information networks by exploring the power of links. In *ALT*, 2009.

[13] E. Hensinger, I. Flaounas, and N. Cristianini. Modelling and predicting news popularity. *Pattern Anal. Appl.*, 16(4), 2013.

[14] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *EMNLP*, 2011.

[15] T. Huet, J. Biega, and F. M. Suchanek. Mining history with le monde. In *AKBC*, 2013.

[16] N. Kallus. Predicting crowd behavior with big public data. In *WWW*, 2014.

[17] A. Karalič and I. Bratko. First order regression. *Machine Learning*, 26(2-3), 1997.

[18] V. Lampos and N. Cristianini. Nowcasting events from the social web with statistical learning. *ACM Trans. Intell. Syst. Technol.*, 3(4), Sept. 2012.

[19] N. Lao, T. Mitchell, and W. W. Cohen. Random walk inference and learning in a large scale knowledge base. In *EMNLP*, 2011.

[20] K. Leetaru. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 16(9), 2011.

[21] K. Leetaru and P. Schrodt. Gdelt: Global data on events, language, and tone, 1979-2012. In *International Studies Association Annual Conference*, 2013.

[22] C. X. Lin, B. Ding, J. Han, F. Zhu, and B. Zhao. Text cube: Computing ir measures for multidimensional text database analysis. In *ICDM*, 2008.

[23] F. A. Lisi. Building rules on top of ontologies for the semantic web with inductive logic programming. *Theory and Practice of Logic Programming*, 8(3), 2008.

[24] W. Lu and D. Roth. Automatic event extraction with structured preference modeling. In *ACL*, 2012.

[25] A. Melo, M. Theobald, and J. Voelker. Correlation-based refinement of rules with numerical attributes. In *FLAIRS*, 2014.

[26] P. N. Mendes, M. Jakob, A. Garcia-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In *ICSS*, 2011.

[27] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Holberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 2011.

[28] N. Nakashole, M. Theobald, and G. Weikum. Scalable knowledge harvesting with high precision and high recall. In *WSDM*, 2011.

[29] V. Nebot and R. Berlanga. Finding association rules in semantic web data. *Knowledge-Based Systems*, 25(1), 2012.

[30] M. C. Nicoletti, F. O. S. de Sá Lisboa, and E. R. H. Jr. Automatic learning of temporal relations under the closed world assumption. *Fundam. Inform.*, 124(1-2), 2013.

[31] J. R. Quinlan. Learning logical definitions from relations. *Machine learning*, 5(3), 1990.

[32] K. Radinsky, S. Davidovich, and S. Markovitch. Learning to predict from textual data. *J. Artif. Intell. Res.*, 45, 2012.

[33] K. Radinsky and E. Horvitz. Mining the web to predict future events. In *WSDM*, 2013.

[34] K. Radinsky, K. M. Svore, S. T. Dumais, M. Shokouhi, J. Teevan, A. Bocharov, and E. Horvitz. Behavioral dynamics on the web: Learning, modeling, and prediction. *ACM Trans. Inf. Syst.*, 31(3), 2013.

[35] S. Schoenmackers, O. Etzioni, D. S. Weld, and J. Davis. Learning first-order horn clauses from web text. In *EMNLP*, 2010.

[36] Staffan Truvé. Big Data For the Future: Unlocking the Predictive Power of the Web. Technical report, Recorded Future, 2011.

[37] F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In *WWW*, 2007.

[38] F. M. Suchanek, M. Sozio, and G. Weikum. Sofie: a self-organizing framework for information extraction. In *WWW*, 2009.

[39] S. Sudhahar, T. Lansdall-Welfare, I. N. Flaounas, and N. Cristianini. Electionwatch: Detecting patterns in news coverage of us elections. In *EACL*, 2012.

[40] P. P. Talukdar, D. T. Wijaya, and T. M. Mitchell. Acquiring temporal constraints between relations. In *CIKM*, 2012.

[41] P. P. Talukdar, D. T. Wijaya, and T. M. Mitchell. Coupled temporal scoping of relational facts. In *WSDM*, 2012.

[42] J. Völker and M. Niepert. Statistical schema induction. In *ESWC*, 2011.

[43] D. Wang, T. Li, and M. Ogihara. Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs. In *AAAI*, 2012.

[44] P. Zhao, X. Li, D. Xin, and J. Han. Graph cube: on warehousing and olap multidimensional networks. In *SIGMOD*, 2011.