IDRA: An ontology driven Cognitive Computing System

Roberto Enea[1]

[1] Guglielmo Marconi University, Via Paolo Emilio 29 00193, Rome, Italy
r.enea@unimarconi.it

Abstract. The problem I intend to address in my PhD research is the leak of quality of the information extracted from Big Data using only a statistical approach. Google Flu is an explanatory example of how the knowledge extraction of Big Data based only on statistical methods could produce low quality results and lead to a misinterpretation of reality. Could it be feasible and effective to design a system that is able to leverage ontologies to balance the limits of statistical methods? Here I present an architecture of an ontology driven Cognitive Computing System that leverages ontologies to filter statistical data.

Keywords: Cognitive Computing, Ontology Extraction, Big Data.

1 Problem statement

The problem I'm addressing in my PhD research is the leak of accuracy that could affect the knowledge extracted from Big Data using only a statistical approach. Several methods and technologies have been developed during the last years in order to extract knowledge from Big Data, mostly based on a statistical approach as the ones that use correlation analysis. One of the paradigmatic examples of application of such methods on Big Data was the Google Flu system [1]. The scope of this system was the forecasting of seasonal flu trends in order to support the CDC (Center of Disease Control and Prevention) in the planning of vaccination campaigns. In the first years of Big Data analysis's spreading, Google Flu has been presented as the best example of valuable knowledge extraction from Big Data. Some years after its release Google Flu started showing all its limitations [2], outputting forecasts that in some cases diverged of 100% from the results gathered by CDC. Those results led Google to put the project in stand-by.

The main idea behind Google Flu was the assumption that users' queries made on Google search engine, that was related to flu, and the real epidemic spreading could be connected by a causation relationship because they were statistically correlated during the testing phase of the system. This is a typical mistake that every researcher could make when he is going to replace a correlation relation with a causation relation. "Cum hoc ergo propter hoc" Latins said. The fact that "correlation proves causation," is considered a logical fallacy that happens when two events occurring together are taken to have established a cause-and-effect relationship. Since correlation analysis is one of the most used methods to extract valuable knowledge from Big Data, the possibility of falling into similar mistakes is around the corner.

2 Relevancy

The relevancy of the problem previously reported is directly proportional to the spreading of the Big Data analysis in Business Intelligence and Decision Making. Nowadays, nearly every aspect of the modern society is impacted by Big Data, involving medical, health care, finance, management and government. The quality of the information extracted is a key element for the future decisions in many fields. Wrong forecasts of Google on Flu Trends are only one example of what a misinterpretation of correlation analysis could cause. The fact that two variables have the same trend during the observation period does not mean that there is a relation between them. For example, Leinweber [3] showed that the S&P 500 stock index was correlated with butter production in Bangladesh, and other strange correlations.

Moreover, as Taleb explains in his recent book [4], when the number of variables grows, the number of fake correlations also grows. This means that as the amount of data grows the quality of the result coming out from statistical methods could not grow but, on the contrary, could decrease. For this reason, finding a way to improve the quality of the result of Big Data analysis will probably be the next main challenge for data scientists. The solution to the problem stated should pass through the use of a mix of technologies. It should involve not only the ones characterized by an inductive reasoning approach (numerical and statistical methods) but also using a deductive approach to verify the quality of the information extracted. The use of knowledge bases (ontologies) and deductive reasoners, that are commonly applied and used in ontology engineering, could be a key element.

3 Related work

Recently there are many papers in scientific literature that are addressing the quality issue of the knowledge extracted from Big Data. Emani et all [5] provide a wide survey of most of the methods and technologies used. It defines and characterizes the concept of Big Data. In addition, a supply chain and technologies for Big Data management are introduced. The Big Data Management process is presented as the set of the following phases:

- **Acquisition**: a Big Data Management System should be able to handle a large amount of data that changes quickly. The infrastructure adopted should ensure efficient memory management and on the other hand a real-time processing of the data.
- **Organization**: BDMS should be able to ingest both unstructured and structured data (text, video, numerical data, etc.)
- **Analysis**: the BDMS has to be able to infer new valuable knowledge from data. In this phase, the system should produce a set of hypotheses, each one tagged with a confidence level.
- **Decision**: the most valuable hypotheses are then the base for the decision making process.

Some of the biggest IT actors are currently trying to implement the process depicted above collecting all the pieces of the "cloud" of Big Data technologies in order to create systems that could be able to support humans in decision making on key fields like health, finance, industry and government, collecting structured and unstructured data mainly from the web and social networks. Such systems have been recently called Cognitive Computing Systems (CCS) because they try to simulate the humans cognitive process. As a matter of fact human approach to the acquisition of new knowledge is composed by the same phases outlined in [5]: acquisition of data, that for humans is the observation of facts, organization that means ascribing a meaning to what is observed (this phase is nothing but the mapping of what we observed with our "knowledge base"), and then the analysis, that means the production of new knowledge inferring the data acquired. A cognitive interpretation of data analysis [6] is becoming popular and is influencing the designing of new data-warehouses.

In this proposal I do not want to present an exhausting state of the art of CCS systems but I'm going to discuss the two systems that are more similar than others to the architecture that I'm going to propose in the next sections. The first example is the IBM Watson. It has been realized inside the DeepQA Project [7] by IBM in order to create an expert system that could be able to succeed in jeopardy game. IBM Watson has a Question Answering interface, chosen to make the human interaction as natural as possible. Behind it there is a software infrastructure realized using the framework UIMA[1] for the elaboration of heterogeneous information sources. IBM Watson needs to ingest a certain amount of data sources like Wikipedia, DBpedia, YAGO2, Wordnet and several other dictionaries and encyclopedias. After the ingestion of those sources it needs to be trained before starting the actual knowledge extraction from Big Data. IBM Watson uses a typical machine learning approach to improve its skills during its activity.

An alternative approach has been introduced by another system that could be also considered a CCS that is the NELL (Never Ending Language Learning) project [8] developed by the machine learning department of Carnegie Mellon University. It differs from Watson because the first one works on specific tasks using a wide and specialized knowledge base, while NELL, starting from a hundred of general categories, encompasses all knowledge available on the web with a minimum human support.

The main aspects of the two reported technologies that I intend to bring into my research project are the domain-driven approach and the strong framework of Watson together with the continuous learning strategy of NELL. The main difference between my approach and the others two is the use of ontologies to guide the knowledge acquisition process.

[1] https://uima.apache.org/

4 Research questions

The first research question I'm going to address, directly springs from the problem statement: could ontology technologies improve the quality of the statistical knowledge extraction from Big Data?

Ontology technologies and statistical ones have a completely different approach to knowledge. The first ones use deductive reasoning to infer new knowledge whereas the second ones use inductive reasoning to achieve the same goal. The first approach produces certain knowledge, the second one returns knowledge characterized by a confidence level. None of them is better than the other and we usually use them both in our cognitive processes. So the previous question becomes: could it be feasible and effective to design an automatic system that is able to leverage both approaches? (hereinafter referred to it as 'RQ1') The CCS model, described in the previous section referring to Watson and NELL, could be the ideal container for this kind of project.

Even though Watson and NELL represent the spearheads of CCS they barely use ontologies and related technologies (reasoning modules in particular): in Watson its use is limited to sources like DBPedia and YAGO2 in order to give to the system basic and general concepts especially in the jeopardy version of the system, while the most part of knowledge is acquired by unstructured text. In NELL the use of ontologies is also very limited. Only few papers have been produced by the NELL research group about this topic [9] [10] [11]. Nevertheless recent studies are demonstrating that ontologies and deductive reasoners could support cognitive process in different areas, from source acquisition through NLP processing [12] to association rules extraction [13] and event correlation [14]. Those are all fields that are traditionally domain of statistical methods. In addition ontologies constitute a privileged recipient of the data extraction process due to their capacity to link and to be linked by external resources.

Moreover the reasoning modules used to infer new knowledge from ontologies are mostly domain-independent. From this consideration springs a sub research question related to RQ1: could it be possible to create a general CCS architecture that can be instantiated by applying a domain ontology, so that the redesign and rework of the CCS architecture could be reduced whenever the application domain is changed? (hereinafter referred to it as 'RQ1.1')

5 Hypotheses

The hypothesis related to RQ1 is strictly connected to the characteristic of Big Data that can be resumed in the five V: Volume, Variety Velocity, Value, Veracity [5]. This means that a system that works on Big Data has at its disposal a continuous large volume of data in different forms and coming from a variety of sources. The direct consequence is that if it detects a fact in the data, it will probably gather hundreds or thousands of versions of that fact (we could consider as a "fact" a statement or a set of statements).

My hypothesis (hereinafter referred to it as 'HYP1') is that if X is a set of versions (intended as contradictory statements that refers to the same subject, uniquely identified, for example "John is male" and "John is female") of the same fact and x_n is the frequency of the n version of the fact, defined as the number of different sources where the fact occurs, the "right" version of the fact will be the one that maximizes x in a reasonable observation interval. In other words, the observation of the fact should converge to the truth or what the most part of the humans considers the truth.

It is important to clarify this hypothesis through an example. Pluto has been recently tagged as a dwarf planet by the International Astronomical Union. Let's consider the two versions of the same fact: (1) *Pluto is a dwarf planet* and (2)*Pluto is a planet*.

A CCS that is going to collect those facts from astronomical sources on the web will probably find that x_1 is bigger than x_2. Of course, it would have found an opposite condition some years ago but anyway it would have been a correct finding because (2) was the correct fact for the most part of the humans. What could happen during the transition? Just after the announcement of the IAU the CCS would have found that x_2 is still bigger than x_1 but increasing the observation window, as soon as the information is going to be fixed in the different sources, x_1 will overcome x_2.

Another hypothesis connected to the previous one is that the convergence velocity of x_n is directly proportional to the number of up to date sources included in the observation set (hereinafter referred to it as 'HYP1.1'). Of course, to make these hypotheses effective, the selection of the sources of the observation set is a key element. I plan to assign this task to the user during the instancing of the CCS (details in section 7) but I don't exclude the future use of provenance algorithms especially to evaluate the confidence level of the sources.

The last hypothesis (HYP2) related to RQ1.1 is that, isolating some components that are domain-dependent like domain ontology, the data sources and the specific algorithms required to elaborate those sources, the rest of the architecture can be domain-independent, so that a CCS can be instantiated providing to it the domain-dependent components without changing the main architecture.

6 Preliminary Result

The architecture I'm going to propose in the next section has been designed in collaboration with the ART[2] research group. One of the main component of the architecture is CODA (Computer-aided ontology development architecture) [15], an already existing triples generator that takes input coming from UIMA annotators and converts them in statements according to a set of PEARL rules. PEARL is a language for projecting (UIMA) annotations over RDF Knowledge Bases [16]. The first improvement that I'm going to introduce in CODA is the managing of uncertain statements, assigning to each generated triple a confidence level, then in the second phase I will add a subsystem that will let the system select the most reliable triples.

[2] Artificial Intelligence at Tor Vergata (art.uniroma2.it)

Moreover my contribution to the novelty of the system is the definition of the general architecture that can be instantiated in different CCSs depending on application's domain(RQ1.1) and the design and development of the sub-modules that will combine certain and uncertain statements (RQ1). Even though the overall of the system is already defined I haven't got preliminary results to show at the moment.

7 Approach

According to the research questions and the hypotheses formulated I propose an architecture for a general CCS that applies a continuous observation of the web, extracting a set of statements that are validated by a domain ontology and a comparative check on different sources. I called this architecture IDRA (Inductive Deductive Reasoning Architecture) because it combines the inductive reasoning commonly used for data extraction and characterized by uncertain information with the deductive reasoning commonly applied to ontologies (RQ1), in which every statement is 100% true or false (Figure 1). IDRA could be considered as a class of architectures instead of a single architecture(RQ1.1 and HYP2). Each instance of the architecture is closely related to the purpose for which the instance was conceived and is determined by the following tuple: **<Sources, Domain Ontology, Evaluation Criteria, Numerical Analysis Tools>.**

We consider the **sources** as a set of fonts on the web where the system can collect data related to the domain it is specialized in. The source is not a predefined and static data corpus but should have the characteristics of Big Data sources (the five V shown above [5]). The **domain ontology** is the ontology that contains all the concepts that are part of the observation field. The **evaluation criteria** are the set of rules translated into statements that allows the system to value the extracted data. The combination of extracted data, domain ontology, and evaluation criteria should allow an ontology reasoner, linked to the resulting ontology, to obtain the desired correct result[3].

As an example, it could be needed to instantiate an architecture that is able to divide a population of individuals into categories in order to achieve a market segmentation and evaluate the best launch strategy for a new product. In this case, the criteria of analysis will consist of the belonging's criteria of individuals to different segments, such as the social class, level of education, geographic area, etc. These criteria will be translated into statements so that the architecture can infer the segmentation from the data. For example, a number of potential job-related statements might be: "software engineers belong to middle class" or "CEOs belong to upper class", etc.

[3] The evaluation criteria have mainly a filtering function so they could be considered optional if the CCS scope is just collecting facts related to a particular domain without any discrimination.

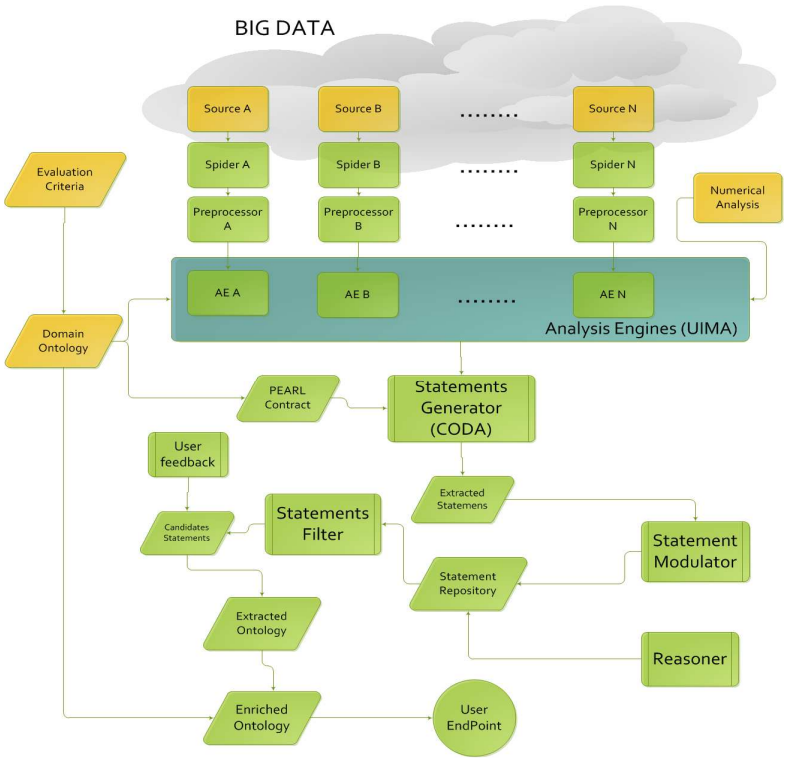

Figure 1: IDRA architecture

Under the name of **Numerical Analysis** are included all those numerical / statistical methods used to support the data extraction and the matching of the latter with the domain ontology. These methods help to define the confidence level of each extracted statement. Each source, as shown in Figure 1, corresponds to a series of three modules that are spider, pre-processor and UIMA's analysis engine. The **Spider** is the module for extracting and collecting raw information directly from the source. It could collect web pages from a set of selected sites, or it could collect posts from Facebook or Twitter. The **Preprocessor** has the purpose of eliminating all the information that is not useful at the next annotation stage. It is of course dependent on the language (s) adopted and the source type. In some cases, the preprocessor might, for example, delete HTML tags from a web page and then stop-words. Always from the preprocessor a stemming operation could be performed for the detection of standardized terms.

The **Analysis Engine** is the source's processing terminal. It represents the process that receives the input data / document and outputs metadata containing annotations. In our case, the main purpose of the AE is mapping data to concepts belonging to the domain ontology. For this purpose, the analysis engine can exploit numerical or statistical methods that, besides providing the mapping between the data and the concepts of the ontology, also return the confidence level. The format of all AE outputs of the system is unique and is represented by the instance of a class implementing the CAS

(Common Analysis System) interface. The CAS is the data model used by UIMA to unify annotations coming out from heterogeneous sources.

The **Statement Generator** is implemented in our architecture by the CODA system. It receives JCAS (Java version of CAS interface) entries from the AE and, in conjunction with the PEARL rule set, where the correspondences between annotations and entities and / or reports of domain ontology are defined, generates a statement list. Each statement is provided with a confidence level depending on:

- The level of trust that the user assigns to the source
- The quality level of the source, computed using methods as in [17] and dependent on the application domain.
- The confidence level of the annotations from which the statement cames out[4].
- The version date of the source document

Extracted statements are processed by the **Statement Modulator**, that takes care of modulation of the confidence level of statements within the statements' repository. It does the following:

- Checks if extracted statements are already in the repository or can be inferred by the statements already present in the repository
- Updates the confidence value if the repository already contains the statement
- If not, adds the statement to the repository

The **Statement Filter** examines the assertions in the repository and extracts some candidates that could be added to the domain ontology. Filtering must surely take place on the confidence interval but also on the number of instances found of the same statement. This is because a confidence value based on a small amount of samples should not be considered reliable. Both parameters must be configurable. The abstract statements will be submitted to the user's judgment for the definitive inclusion in the domain ontology. The process described is cyclic and continuous (as in NELL). According to HYP1 and HYP1.1, after a certain amount of iterations, I expect that the statements tagged with the highest confidence level are the ones commonly considered as true.

8 Evaluation Plan

In order to evaluate and validate the architecture proposed I will design an instance of IDRA for information security purposes. In particular, I want to create a CCS that is able to identify inside a group of persons which are the ones who are most susceptible to social engineering attacks evaluating their activity in social network. The evaluation criteria used will be the ones proposed by [18]. A group of experts in the field, coming from an international security company involved in my research, will

[4] The confidence level in the annotation generally stems from the use of statistical / numerical methodologies. For example, an AE could use matching strings algorithms to create a record that binds an ontology entity with a text element. The confidence level will then be represented by the similarity level calculated by the matching algorithm.

support us on creating a benchmark set and will evaluate the results of IDRA in terms of precision and recall. The precision will measure the ratio between the individuals properly considered as susceptible and all the ones detected by IDRA. The recall will measure the ratio between the individuals properly detected by IDRA and all the ones detected by the experts. Also, the F-Measure will be considered. The result that I expect and that should validate the hypotheses and the approach is the convergence of the profiles defined by the system with the ones defined by the experts as the amount of social networks' data analyzed increases.

9 Reflections

Even if the hypotheses and the approach have some novelty aspects like the use of semantic web technologies to drive the extraction and the selection of facts, the machine learning strategy is similar to the one used by NELL and it achieved significant results even if applied on a general semantic domain. This make me think that a similar strategy could be applied also on specialized domains achieving good results as well. The use of relatively small ontologies since related to specialized fields could also make the elaboration lighter from a computational point of view and at the same time more useful for an industrial use.

Acknowledgements

I want to thank Prof. Pazienza, and Dr Turbati of ART(Artificial Intelligence Research @ Tor Vergata) group for supporting and inspiring my PhD research.

References

[1] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski e L. Brilliant, «Detecting influenza epidemics using search engine query data,» *Nature*, n. 457, pp. 1012-1014, 2009.

[2] D. Butler, «When Google got flu wrong,» *Nature*, pp. 155-156, 2013.

[3] D. J. Leinweber, «Stupid Data Miner Tricks: Overfitting the S&P 500.,» *The Journal of Investing*, n. 16, pp. 15-22, 2007.

[4] N. Taleb, Antifragile: How to Live in a World We Don't Understand, Pengiun Books, 2012.

[5] C. K. Emani, N. Cullot e C. Nicolle, «Understandable Big Data: A survey,» *Computer Science Review*, n. 17, pp. 70-81, 2015.

[6] G. Grolemund e H. Wickham, «A Cognitive Interpretation of Data Analysis,» *International Journal of Statistics*, vol. 82, n. 2, p. 184–204, 2014.

[7] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer e C. Welty, «Building

Watson:An Overview of the DeepQA Project,» *Association for the Advancement of Artificial Intelligence*, pp. 59-79, 2010.

[8] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya e A. Gupta, «Never-Ending Learning,» in *Proceedings of the Conference on Artificial Intelligence (AAAI)*, Austin Texas, USA, 2015.

[9] B. Dalvi, E. Minkov, P. P. Talukdar e W. W. Cohen, «Automatic Gloss Finding for a Knowledge Base using Ontological Constraints,» in *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, Shanghai, China, 2015.

[10] B. Dalvi, W. W. Cohen e J. Callan, «Classifying Entities into an Incomplete Ontology.,» in *AKBC , 2013, 3rd Knowledge Extraction workshop at CIKM 2013*, San Francisco, CA, USA, 2013.

[11] D. Movshovitz-Attias e W. W. Cohen, «Bootstrapping Biomedical Ontologies for Scientific Text using NELL,» in *BioNLP-2012*, Montreal, Quebec, Canada, 2012.

[12] P. Cimiano, C. Unger e J. McCrae, «Ontology-Based Interpretation of Natural Language,» in *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool, 2015, pp. 1-178.

[13] B. Furletti, A. Bellandi, V. Grossi e A. Romei, «Ontology-Driven Association Rule Extraction: A Case Study,» in *International Workshop on Contexts and Ontologies: Representation and Reasoning*, Roskilde, Denmark, 2007.

[14] T. Moser, H. Roth, S. Rozsnyai, R. Mordinyi e S. Biffl, «Semantic Event Correlation Using Ontologies,» in *On the Move to Meaningful Internet Systems: OTM 2009*, Vilamoura, Portugal, 2009.

[15] M. Fiorelli, A. Stellato, M. T. Pazienza e A. Turbati, «CODA: Computer-aided ontology development architecture,» *IBM Journal of Research and Development*, vol. II, n. 58, pp. 1-12, 2014.

[16] M. T. Pazienza, A. Stellato e A. Turbati, «PEARL: ProjEction of Annotations Rule Language, a Language for Projecting (UIMA) Annotations over RDF Knowledge Bases,» in *International Conference on Language Resources and Evaluation (LREC 2012)*, Instanbul, Turkey, 2012.

[17] A. Immonen, P. Pääkkäönen e E. Ovaska, «Evaluating the Quality of Social Media Data in Big Data Architecture,» *IEEE Access*, vol. III, pp. 2028-2043, 2015.

[18] A. Algarni, Y. Xu e T. Chan, «Susceptibility to social engineering in social networking sites: The case of Facebook.,» in *36th International Conference on Information Systems (ICIS 2015)*, Fort Worth, Texas, USA, 2015.