Investigating the Intersections of Data and Knowledge Engineering

Jorge Martinez-Gil

Introduction

My research focuses on **Data and Knowledge Engineering**, with a strong particular focus in the intersection between databases and artificial intelligence. Over the last few years, I have performed intensive research on new methods for **Knowledge-Based Systems** that have resulted in publications and software prototypes used in both academia and industry. In the context of this document, I would particularly like to focus my research statement on three fundamental scientific challenges and three applied challenges. In addition, all my research is guided by three general objectives that include accuracy, interpretability, and minimization in terms of energy consumption of all my proposed solutions in line with the priorities stated by the European Green Deal (COM2019-640).

Fundamental Research

By fundamental challenges, I mean the pursuit of pure scientific findings that have no immediate applications on a practical level but necessary to provide the basis for all subsequent applied research. In this context, my preferences are focused on these three challenges:

F1. Fundamental Research on Knowledge Graphs

Knowledge Graphs are a kind of knowledge base used to model information with information gathered from various sources. They can be built automatically and can then be explored to reveal new insights [1]. Currently, Knowledge Graphs offer a wide range of possibilities for research, among which unsupervised entity alignment and automatic knowledge graph completion stand out [2]. Its usefulness in the field of semantic information processing is high because it allows emulating the way in which humans proceed with their decision-making [3]. Currently, there are some challenges to be addressed, especially in methods for the calculation of embeddings that reliably reflect the original graph to proceed with different machine learning tasks with certain guarantees of success [4, 5].

F2. Fundamental Research on Textual Similarity Learning

Textual similarity learning is an exciting area of supervised machine learning in artificial intelligence. It is closely related to regression and classification, but the goal is to learn from a similarity function that measures how similar or related two textual pieces are [6]. Research in this field provides excellent opportunities to advance various scientific and industrial areas, including lots of application domains that are currently overloaded by interoperability problems [7]. The best results are currently obtained through transformers of a neural nature. However, these models are intelligible to human beings. The challenge I am trying to achieve is to improve the interpretability of the resulting models so that people can rely on them [8].

F3. Fundamental Research on Data, Text, and Web Mining

In the context of computer science, mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. There is a multitude of techniques that are almost always dependent on the resources to be mined [9]. Today, there are many opportunities to advance knowledge from databases, text, and even the World Wide Web. Until now, the existing solutions in terms of natural language processing and information retrieval have yielded moderate results. But in recent times very important advances have been made that allow us to go a step further. My challenge consists in the development of new models that allow a more accurate and interpretable mining and that consume much less resources [10].

Applied Research

By applied research challenges, I mean the pursuit of methods and solutions to solve specific and practical issue affecting an individual scenario. To do so, it is usual to rely on results obtained from fundamental research.

A1. Applied Research on Automatic Question Answering

Question answering (QA) is a computer science discipline within the fields of information retrieval and natural language processing (NLP), which is concerned with building systems that automatically answer questions posed by humans. Today's information overload makes these systems the perfect assistant for completing tasks in a wide range of domains. There are two fundamental advances that make it possible to develop models that have never been seen before. On the one hand, models based on Knowledge Graphs that allow a much better understanding of the questions and on the other hand, solutions based on Big Data that allow working with vast amounts of information in a very efficient way.

A2. Applied Research on Big Data

My activity as a lecturer has made me curious about the methods and tools related to big data. While big data focuses on solving problems that, due to the large volume of information they work with, could not be solved by traditional techniques. As the modern information technology enables acquisition and storage of increasingly complex data. My challenge is to develop new software solutions that can integrate a variety of existing methods for dealing with large amounts of information (on the order of hundreds of GBs). For example, Knowledge Graphs of gigantic sizes that cannot be processed in main memory.

A3. Applied Research on Cloud Computing

Cloud computing is the delivery of different services through the Internet, including data storage, servers, databases, networking, and software. Cloud computing is the on-demand availability of computer system resources, especially data storage and computing power, without direct active management by the user. The benefits of cloud computing services include the ability to scale elastically and the possibility to deploy solutions in a very effective and efficient way on a global scale. Again, my activity as a lecturer has stimulated my interest in the development of cloud solutions that can cover many problems related to scalability, security, efficiency, and responsible consumption of resources.

General Objectives

Finally, all my research (both fundamental and applied) is guided by three fundamental objectives: accuracy, interpretability, and sustainable energy consumption. I believe that all three factors are fundamental to building solutions that not only solve problems, but that people can trust and that have a positive impact on society.

G1. Accuracy

By referring to accuracy, I do not make explicit reference to the classical notion of precision since it is usually easy to optimize precision at the cost of low coverage. In the frame of my research, accuracy is understood as high precision with high recall values so that that accuracy would be closer to the so-called f-measure. The truth is that, to date, existing techniques (matching, similarity assessment, link prediction, etc.) have been able to achieve good f-measure values, mainly due to the ability to create an ensemble of methods with high precision together with methods with high coverage or recall [6].

It is expected that research will continue along this line since the problem is not entirely solved yet. Many slight variations of existing solutions are proposed to try to gain a few thousandths more accuracy. But what makes sense to look at the quality of the data on which the models are trained. After all, the results are highly dependent on the dataset with which they are trained, making more sense to develop suitable datasets that fit the users' requirements and cover general and cases that may arise in everyday situations.

G2. Interpretability

Interpretability has gained a lot of specific weight in recent years. This is since during the last decade, there has been a constant race to build models that yield better and better results, but this race has lost sight that in the end, models must be used by people. Models will not be beneficial if the people who use them cannot look at them, study them and understand them. If this is not the case, there will be great distrust and aversion.

While there are methods that have very high levels of interpretability: Consider a regression model (with or without penalty) on a set of basic algorithms. However, empirical evidence shows that easily interpretable models are often quite simplistic, so one must move towards methods capable of complex modeling interactions between algorithms in the ensemble [2]. Even then, these interactions must be easily understandable and reliable.

Further research in this direction is to be needed. Consider, for example, the new techniques based on word embeddings that have extraordinary predictive power but cannot explain to an operator the relationship between input and output parameters. In this case, there is a gap for which researchers have not yet provided a satisfactory answer. To the best of our knowledge, our proposal is the first to pay special attention to the interpretability of the resulting model. My hypothesis that a few hundredths of accuracy do not justify increasingly complicated models to understand is based on the opinion of a multitude of experts who do not find a single answer from a black box model satisfactory and demand to know much more. They demand to be able to understand why the model has arrived at that decision.

G3. Sustainable-energy consumption

Energy saving is one of the major concerns in today's societies. In this context, it is necessary to note that cloud data centers worldwide consumed more than 210 terawatt-hours of electricity last year, which is more than 1% of the world's total electricity consumption. Cloud data centers assume that electricity expenses have become one of their major cost factors. For these reasons, the engineers in charge of maintaining these centers strongly warn that if energy consumption continues to grow, its cost may even exceed the cost associated with the purchase of hardware by a wide margin, not to mention its environmental impact.

Therefore, new solutions should be explored to reduce power consumption. In this context, a new generation of energy-efficient algorithms can operate in cloud computing environments. To date, there have been significant attempts to reduce energy consumption in cloud data centers. The first is based on designing hardware (e.g., processors, memory, and disks) so that running software consumes less power. The second is based on adapting existing software, e.g., using mechanisms for scheduling the execution of instructions to make them more efficient. Symbolic regression-based techniques that do not assume any model or associated structures beforehand can be a good starting point for studying low-carbon footprint solutions.

References

[1] **Jorge Martinez- Gil**, Enrique Alba, José Francisco Aldana Montes: Optimizing Ontology Alignments by Using Genetic Algorithms. **NatuReS** 2008.

[2] **Jorge Martinez-Gil**, Jose F. Aldana-Montes: Evaluation of two heuristic approaches to solve the ontology meta-matching problem. **Knowl. Inf. Syst.** 26(2): 225-247 (2011).

[3] **Jorge Martinez-Gil**, Jose F. Aldana-Montes: An overview of current ontology meta-matching solutions. **Knowl. Eng. Rev.** 27(4): 393-412 (2012).

[4] **Jorge Martinez-Gil**, Jose F. Aldana Montes: KnoE: A Web Mining Tool to Validate Previously Discovered Semantic Correspondences. **J. Comput. Sci. Technol.** 27(6): 1222-1232 (2012)

[5] **Jorge Martinez-Gil**, Ismael Navas-Delgado, Jose F. Aldana Montes: MaF: An Ontology Matching Framework. **J. Univers. Comput. Sci.** 18(2): 194-217 (2012).

[6] **Jorge Martinez- Gil**, José Francisco Aldana Montes: Semantic similarity measurement using historical google search patterns. **Inf. Syst. Frontiers** 15(3): 399-410 (2013).

[7] **Jorge Martinez-Gil**: An Overview of Knowledge Management Techniques for e-Recruitment. **J. Inf. Knowl. Manag.** 13(2) (2014).

[8] **Jorge Martinez- Gil**: An overview of textual semantic similarity measures based on web intelligence. **Artif. Intell. Rev.** 42(4): 935-943 (2014).

[9] **Jorge Martinez- Gil**: Automated knowledge base management: A survey. **Comput. Sci. Rev.** 18: 1-9 (2015).

[10] **Jorge Martinez- Gil**: CoTO: A novel approach for fuzzy aggregation of semantic similarity measures. **Cogn. Syst. Res.** 40: 8-17 (2016).