Research and Reviews: Journal of Global Researh in Computer Science

Extended Abstract
Vol. 10, Iss. 1
2019

Data Mining 2016: Scientific discovery by machine intelligence: A new avenue for drug discovery- Carlo A Trugenberger - InfoCodex AG-Semantic Technologies

Carlo A Trugenberger
InfoCodex AG-Semantic Technologies, Switzerland

The majority of massive data is unstructured and of this majority the most important chunk is text. While data mining techniques are well developed and standardized for structured data; numerical data, the realm of unstructured data is still largely unexplored. The general focus lies on information extraction, which attempts to retrieve known information from text. The grail however is knowledge discovery, where machines are expected to unearth entirely new facts and relations that weren't previously known by any human expert. Indeed, understanding the meaning of text is usually considered together of the most characteristics of human intelligence. The ultimate goal of semantic AI is to plan software which will understand the meaning of free text, a minimum of within the practical sense of providing new, actionable information condensed out of a body of documents. As a stepping stone on the road to the present vision i will be able to introduce a completely new approach to drug research, namely that of identifying relevant information by employing a self-organizing semantic engine to text mine large repositories of biomedical research papers, a way pioneered by Merck with the InfoCodex software. I will describe the methodology and a primary successful experiment for the invention of latest biomarkers and phenotypes for diabetes and obesity on the idea of PubMed abstracts, public clinical trials and Merck internal documents. The reported approach shows much promise and has potential to impact fundamentally pharmaceutical research as how to shorten time-to-market of novel drugs, and for early recognition of dead ends. Understanding written language is a key component of human intelligence. Correspondingly, doing something useful with large quantities of text documents that are out of reach for human analysis requires, unavoidably some form of artificial intelligence [5]. This is why handling unstructured data is harder than analyzing their numerical counterpart, for which well-defined and developed mathematical methods are readily available. Indeed, there is as yet no standard approach to text mining, the unstructured counterpart to data mining. There are several approaches to teach a machine to comprehend text [6-8]. The vast bulk of research and applications focuses on natural language processing (NLP) techniques for information extraction (IE). Information extraction aims to identify mentions of named entities (e.g. "genes" in bioscience applications) and relationships between these entities (as in "is a" or "is caused by"). Entities and their relations are often called "triples" and databases of identified triples "triple stores". Such triple stores are the idea of the online 3.0 vision, during which machines are going to be ready to automatically recognize the meaning of online documents and, correspondingly, interact intelligently with human end users. IE techniques are also the main tool used to curate domain-specific terminologies and ontologies extracted from large document corpora. Information extraction, however, is not thought for discovery. By its very design, it is limited to identifying semantic relationships that are explicitly lexicalized in a document: by definition these relations are known to the human expert who formulated them. The "Holy Grail" [9] of the text mining, instead is knowledge discovery from large corpora of text. Here one expects machines to generate novel hypotheses by uncovering previously unnoticed correlations from information distributed over very large pools of documents. These hypotheses must then be tested experimentally. Knowledge discovery is about unearthing implicit information versus the specific relations recovered by information extraction. The present paper is about machine knowledge discovery within the biomedical and pharmacogenomics literature.

This work is partly presented at 3rd International Conference on Big Data Analysis & Data Mining September 26-27, 2016 at London, UK

Biography

Carlo A Trugenberger has earned his PhD in Theoretical Physics in 1988 at the Swiss Federal Institute of Technology, Zürich and his Master's in Economics in 1997 from Bocconi University, Milano. An international academic career in theoretical physics (MIT, Los Alamos National Laboratory, CERN Geneva, Max Planck Institute Münich) led him to the position of Associate Professor of Theoretical Physics at Geneva University. In 2001, he decided to quit academia and to exploit his expertise in Information Theory, Neural Networks and Machine Intelligence to design an innovative semantic technology and co-founded the company InfoCodex AG-Semantic Technologies, Switzerland.

Email: c.trugenberger@infocodex.com

This work is partly presented at 3rd International Conference on Big Data Analysis & Data Mining September 26-27, 2016 at London, UK