EPSC Abstracts
Vol. 12, EPSC2018-844, 2018
European Planetary Science Congress 2018

Advanced Techniques for Signal Search and Automatic Classification of Observational Space Data

Al-Ubaidi Tarek (1), Khodachenko Maxim (1), Kern Roman (2), Granitzer Michael (3), Poedts Stefaan (4)
(1) Space Research Institute, Graz, Austria, (2) Know-Center, Graz University of Technology, Austria, (3) University of Passau, Germany, (4) Catholic University of Leuven, Belgium

Abstract

The presentation will outline various approaches in *machine learning* and *content based search* investigated by members of the former *IMPEx-FP7* (http://impex-fp7.oeaw.ac.at/) project consortium, in close cooperation with partners *Know-Center, Graz University of Technology*, and *University of Passau* and discuss some of the numerous possibilities that open up, using these or equivalent techniques in the emerging field of ***e-Science* in conjunction with space science**. In particular, the presentation will focus on applications that allow systems to **automatically classify and pre-select scientific data** and hence speed up scientific workflows significantly by supporting scientists with the cumbersome task of going through vast amounts of data manually, looking for specific patterns, signals and phenomena of interest prior to selecting specific data for closer examination and analysis.

Introduction

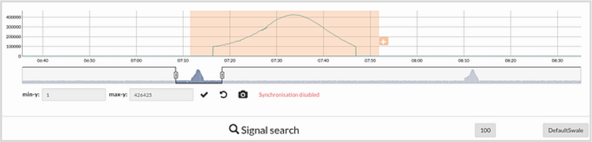
Due to extensive research in the field of information retrieval, search technologies are commonplace today and their algorithmic underpinnings are well understood and proven to scale up to massive amounts of data. While the capability to do complex searches for specific signals and phenomena would come in quite handy when analyzing heterogeneous scientific data, such methods are mostly limited to textual searches and thus will not (directly) apply to cases where the data at hand is time series from sensory data. This is the case for e.g. observational space data or data derived from simulations of various physical processes. In such instances the handling and processing of the data needs to be adapted first and the search paradigm needs to be redefined, since the actual search cannot directly be initiated by key terms entered by the user.

Powerful Tools for Science

A promising solution investigated by the team is a technique known as **query by example** or **content-based search** in the information retrieval community. Data is first transformed into a representation suitable to be managed using an inverted index by adopting and extending *symbolic representation techniques* which are designed to transform continuous data (discrete in the time domain) into a discrete or quantized representation, while keeping the associated information loss minimal (also see *wavelet* analysis). Further approaches from the field of *unsupervised machine learning* can then be applied to obtain temporal patterns of interest and identify a trade-off between frequent and surprising patterns. However, due to the unsupervised nature of the used techniques, the discovered patterns will not be tailored towards specific cases of scientific interest. In order to further limit the identified patterns to a set of candidates that are relevant in the context of specific questions in the realm of space science, techniques from the field of **supervised machine learning** can be established, using a procedure where human annotators provide labelled examples as references. Using these information retrieval and *machine learning* (as well as *deep learning*) methods a system can be built that automatically searches for specific phenomena in large quantities of (observational) data and most importantly also performs automatic classification of scientific data.

Content Based Search and Automatic Classification

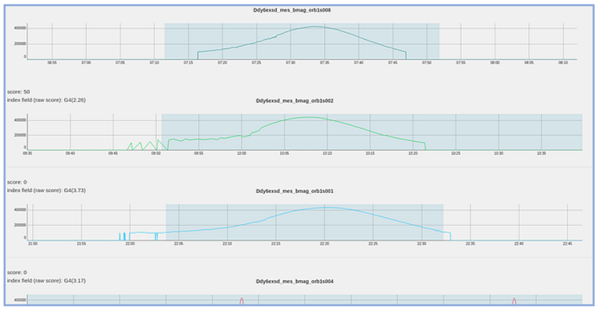
Following a brief depiction of an example workflow for a simple application of the techniques outlined above, where a specific signal of interest is selected and then used as input for a comprehensive *content based search*. As a first step the user selects a portion of the time series data - a graphical interface allowing visual selection (and annotation) of patterns of interest is shown in **Figure 1**.


min-y: 1 max-y: 426425

Synchronisation disabled

Figure 1: The user provides a signal by manually selecting a small portion of data

In this step the user could also provide detailed annotations that specify the signal and allow further enhancing the search capabilities. The system then responds with a ranked list of search results, i.e. signals in the investigated data that resemble the example given above with descending similarity. It should be noted here that the search technology used, scales up to hundreds of Gigabytes of data and beyond. See **Figure 2Error! Reference source not found.** for an example of a possible response generated by the system, given the input selected (**Figure 1**).


score: 50
index field (raw score): G4(2.26)

score: 0
index field (raw score): G4(3.73)

score: 0
index field (raw score): G4(3.17)

Figure 2: The ranked list of signal search results

The first response in particular shows almost identical characteristics as the input signal and is likely to originate from a similar physical environment.

Summary and Conclusions

With new missions leveraging up-to-data capabilities in *telematics* and thus producing ever increasing amounts of observational data, *content based search*, *machine learning* and related technologies can provide a powerful toolset to enhance data analysis and data driven investigations of any kind. Many time consuming tasks can already be sped up and in many instances improved by leveraging current approaches in *machine learning* and *artificial intelligence*. **Now is the time to start building prototype tools** and to carefully analyze scientific workflows in order to gather detailed and relevant requirements for the *e-Science* tools of the future. In this regard, it is crucial that experts in IT and machine learning are closely cooperating with (space) scientists, in order to gain a deep understanding of the problems at hand and to be able to build powerful solutions that will optimally support space science in the 21st century.

Acknowledgements

We would like to thank our colleagues at the ***Catholic University of Leuven***, the ***Skobeltsyn Institute of Nuclear Physics*** and the ***Ministry of Education and Science*** of the Russian Federation (Grant RFMEFI61617X0084) for their support and cooperation and the provision of simulation data as well as observational data that were crucial to build first prototypes and conduct tests of actual *real life* applications of the discussed technologies.