STUDIA INFORMATICA 2014
Volume 35 Number 2 (116)

Mariusz MĄSIOR, Stanisław KACPRZAK
AGH University of Science and Technology, Department of Electronics

DATA ANALYSIS AND MANAGEMENT ENGINE FOR SIGNAL PROCESSING

Summary. This paper presents framework for managing analysis of scientific data. The framework was build on sole purpose of research on signal processing and speech technology but can be successfully adapted to other scientific problems.

Keywords: data management, scientific workflow, signal processing

SYSTEM ANALIZY I ZARZĄDZANIA DANYMI NA POTRZEBY PRZETWARZANIA SYGNAŁÓW

Streszczenie. Artykuł przedstawia środowisko zarządzania analizami danych naukowych. System został stworzony na potrzeby badań nad przetwarzaniem sygnałów i technologią mowy, ale może być z powodzeniem zastosowany w innych problemach naukowych.

Słowa kluczowe: zarządzanie danymi, przetwarzanie sygnałów

1. Introduction

Development of advanced signal processing applications is increasingly data-driven and needs special emphasis on the patterns creation procedures. Image processing and speech technologies demand large amount of data and multiple repeating of training and testing during development. Even very powerful computers spend tens and hundreds of hours to perform single stage of these calculations. Scientific workflows aim to address many of these challenges. They provide formal description of a process for accomplishing a scientific objective, usually expressed in terms of tasks and their dependencies [1]. Small research

teams struggle with complexity of workflow which is large and still growing, as presented in [2].

In our research on phonemic diversity in languages across the world [3,4] we are facing a problem of analysing big amount of audio files [5]. Also, most of our analysis are dependent on previous ones. For that reason we need to have a system that enables easy way to configure workflow of analysis, their inputs and outputs. Also important is to keep results of all analysis up to date. During algorithms development some of the analysis could be changed and system has to rerun computation partially in order to contain only the current data. Fig. 1. shows example of workflow used in beginning stage of our research. In this case root of analysis dependencies tree is always recordings base because all of analyses depend on this data.

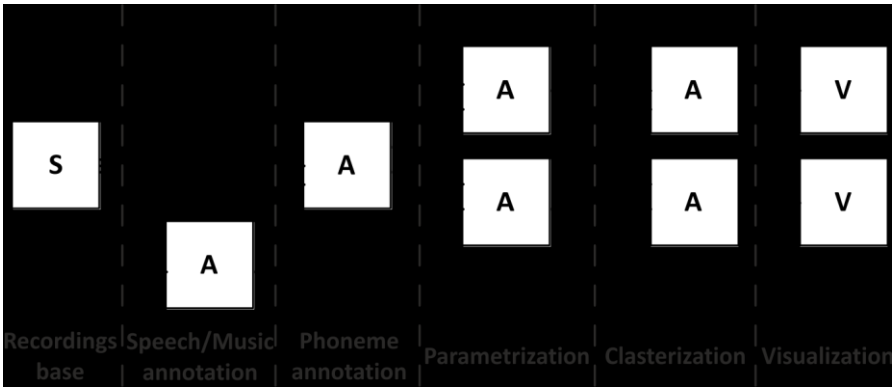

Fig. 1. Example workflow of analyses in multilingual diversity research
Rys. 1. Przykładowy schemat analiz w badaniach nad różnorodnością języków świata

Crucially important in development of advanced signal classification and recognition systems is proper validation of all used procedures and computations. These verifications and validations are applied to every processing block in Fig. 1. and serve as efficiency measurement of every stage of processing.

2. Data Analysis and Relations Mapping Concept

Management of data and scientific processing is a well known problem in many areas of research. There exist at least few advanced scientific workflow systems. Comparison, survey and development directions of most popular tools were presented in [1, 6, 7]. Although these tools are feasible to use in most of researches or can be adopted for it, we believe that a need of different solution in this matter exists.

Common scientific workflow systems are focused on enabling the implementation of advanced processing and visualization of its results. It is a proper approach when processing of scientific data is known and can be entirely implemented. In such cases scientists are

interested in results of data processing and a lot less in processing itself. On the other hand, during development stage of advanced signal processing systems, more important than results of processing is the evaluation of constructed models, algorithms and selection of processing parameters. Procedures of analyses can and will change during research, while set of analysis data is less variant. It implies data-oriented architecture for workflow system and paradigm of optimization data management over processing paths.

The easiest way to meet the defined requirements is by combining an appropriate, dedicated management system with a relational database and calculation components. In our approach all results of analysis are mapped to relational database tables. This allows to describe dependencies in an easy way. It is need just to specify source of data for the analysis and data types that the analysis will produce. All this schema can be converted into tables definition and form of database queries.

The crucial advantage of presented system its functionality that allows reusing of other analysis results rather than recalculating them. Normally, if we want to change something in our workflow (for example parameters of some analysis) using other tools, like Kepler [8], all calculation will be rerun. Our solution doesn't require recomputing of everything. The proper caching of intermediate results and tracking of its timeliness allow to recompute only the analysis with changed input data. This functionality can be achieved in other scientific workflows only by introducing new processing units between all analysis. Although this approach is possible, it is not optimal and the better idea is to build new framework and optimize it for this purpose.

The specificity of data analysed in audio or image processing causes that not all of research data can be directly entered into the database (images, audio or video recordings). A large part of them must be stored in files while the database holds only their descriptions and files location paths.

Presented situation is similar to Object-Relational Mapping (ORM) problem, but solution realization needs different directives. Usage of ORM tools would require prior knowledge of analysis data structure, while this knowledge is gained only during system initialization from analysis configuration files. The mapping is created dynamically, which can be done easily because results of each analysis are usually simple structures with no aggregations. The result of analysis for single portion of data can be merely a row in database table correlated with an adequate analysis.

An attempt of automatic management of data and analysis requires making following assumptions on data schema, possible analyses and its relations:

1. Entire analysis process can be divided into separated, atomic analysis.

2. Each analysis takes as input stored, basic data or results from another analysis (or analyses).
3. Analysis results depend only on its inputs (there is no data not managed by the system).
4. Analysis data can be stored in database table or in a file (or files).

Above mentioned assumptions usually reflect the nature of the typical signal processing algorithms and procedures. More complex processing paths and statistical research methodologies also can be converted into forms corresponding to these assumptions, which is shown in one of following paragraphs.

3. Data Analysis and Management Engine

Fulfilment of the design requirements and the choice of relational database as the primary tool for storage and management of analyses results leads to creation of Data Analysis and Management Engine (codename DAMAGE).

The main concept for the system is the coupling of each analysis with a single table in database. Each of these tables is holding results, specified as outputs of given analysis.

Fig. 2. presents prototype of relational database table correlated with single analysis results. Creation of proper set of indexes on foreign keys in each analysis table allows efficient querying and data management.

Description of entire analysis is done in XML file, prepared by user. This file defines types and schema of analysis results and source of inputs. It also describes also data dependencies, which implies order of previous analyses performance and computational methods of described analysis. Analysis operations can be held by external computational tools (as MATLAB environment) or internal functions (implemented as plugins for workflow system).

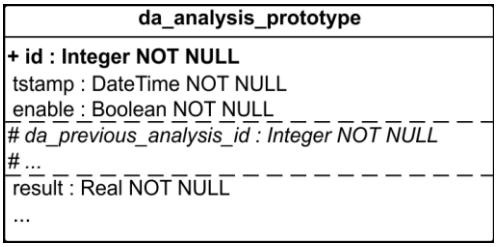

Fig. 2. Prototype of table in database for single analysis
Rys. 2. Szablon tabeli w bazie danych dla pojedynczej analizy

Not all workflow components can be represented completely according to presented model, therefore it requires introduction of special analysis procedures:

- **Source** – Analysis with no input data enclosed in workflow system. Source analysis stores new data from outside the system and provide them for other analyses inside the workflow.
- **Visualization** – Analysis with no table for results. Visualization results are presented outside the system in another form (plots, charts, data summary for download).

Construction of database schema is natural and follows directly the configuration of each analysis in workflow. Such representation of the data can be quite naturally achieved, as is shown in Fig. 3. (DB tables for languages diversity research [4]).

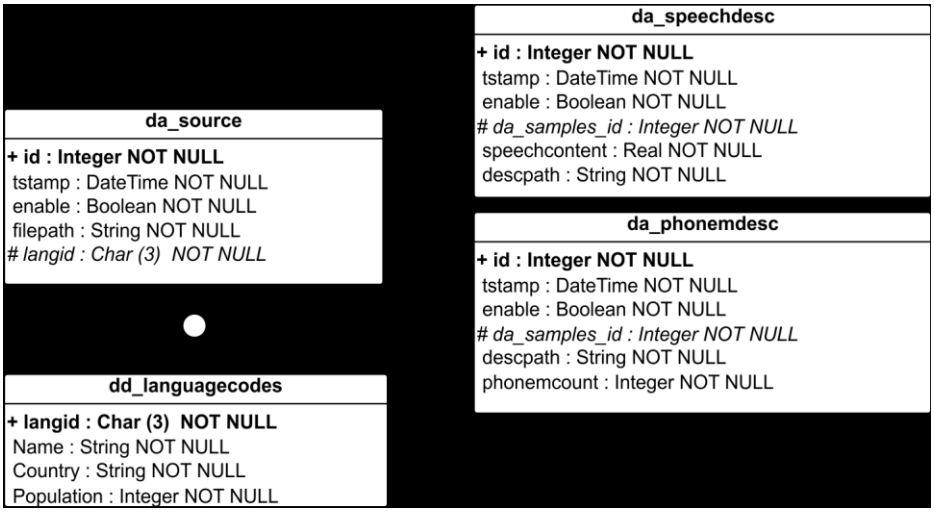

Fig. 3. Entity diagram of initial analyses in multilingual diversity research
Rys. 3. Diagram encji analiz w badaniach nad różnorodnością języków świata

Main task of scientific workflow is to assist in conducting of research calculations, so one of desired, basic functionality has to be capability of performing analysis. The core of management system must be flexible and allow easy attachment of new components and descriptions of analysis mechanisms. This can be achieved by appropriate system structure.

Fig. 4. shows class diagram of the system. Usage of creational design patterns (*Builder* and *Factory*) implements possibility of extending the system and adding new processing tools.

This task requires only an implementation of *AnalysisFactory* and *AnalysisProcessor*. The first one is responsible for initializing external tool and the second for wrapping its execution.

Entire workflow system is designed to work on computational server. Development technology for it was chosen due to the speed of implementation and the availability of additional tools and libraries. Main management core was implemented in C# (Mono framework) under control of Linux. MySQL was used as relational database management system.

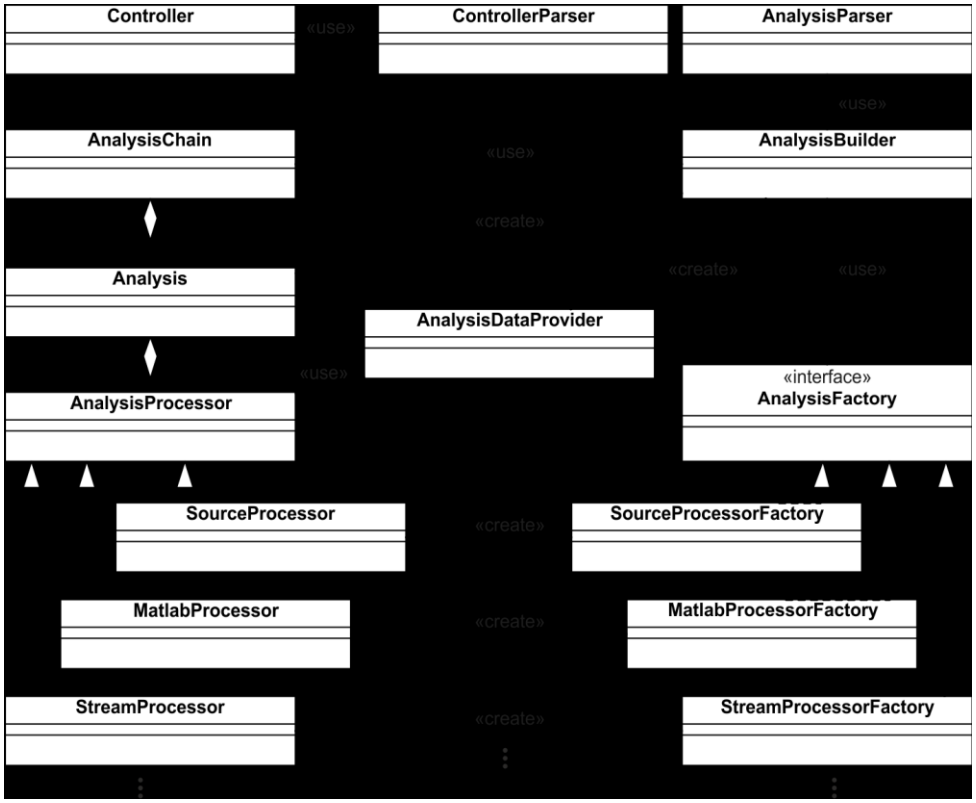

Fig. 4. Simplified class diagram for main core of data management system
Rys. 4. Uproszczony diagram klas jądra systemu zarządzania danymi

Analyses conduction is possible in several ways, at this stage of development. The most useful for signal processing prototyping is MATLAB environment. The workflow system lunches M-file calculation scripts through MATLAB engine, sending and receiving computational data. Also other computing, communication and programming technologies are adapted for cooperation with the system (e.g. standalone application controlled through streams or sockets, native .NET dynamic libraries or R language environment).

4. Complex Data Analysis and Validation Methodologies

Research on advanced machine learning and pattern recognition algorithms that operate on audio, video or images, demands usage of proper testing and validation techniques. Described system allows creation of such techniques and supports statistical mechanisms of performance evaluation.

Fairly frequent and basic scheme for the analysis method selection is performance comparison of an algorithm operating with different parameters or performance comparison of different algorithms. Such action could be arranged through the introduction of additional

analysis that compares different processing paths. Template of such experiment schema is shown in the Fig. 5.

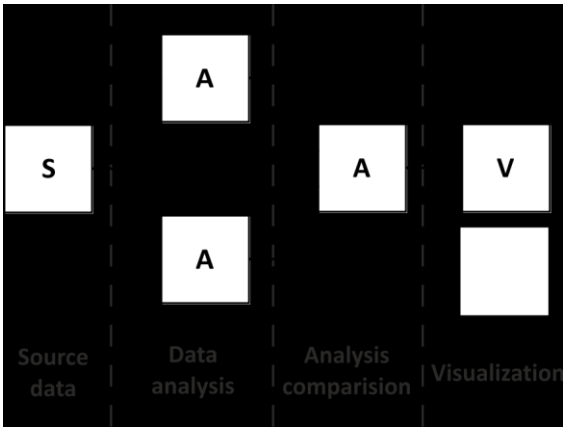

Fig. 5. Analyses relation in simple, algorithms comparison schema
Rys. 5. Relacja analiz dla schematu prostego porównania algorytmów

More advanced methodologies of statistic verification may also be used in the present system. One of the most common and useful for statistical assessing accuracy of processing is a cross-validation procedure [9, 10]. It defines methods of partitioning data into complementary training and testing subsets. Performing the analysis on the first subset and validation on the other allows to estimate the analysis efficiency. The methodology involves multiple rounds of cross-validation, performed on different partitioning for reduction variability of each trial. The final validation results are averaged over the rounds.

Fig. 6. presents schema of general cross-validation technique converted into environment of presented workflow system. Realisation of succeeding cross-validation stages is implemented with additional analyses for performing data partitioning, training and testing.

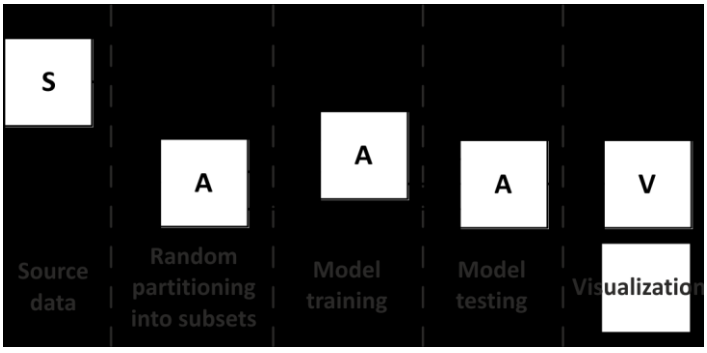

Fig. 6. Analyses relations in general cross-validation schema
Rys. 6. Relacja analiz dla schematu walidacji krzyżowej

5. Conclusions and Future Work

This paper presents basic concepts of Data Analysis and Management Engine (DAMAGE) for automation of scientific workflow. Usage of such system reduces effectively time needed for preparing analyses and may reduce errors caused by usage of wrong or not actual data.

The need to build that system was based on lack of good tools that would be suitable for the needs of signal processing research. We estimate that is actually more efficient to build an entirely new solution instead of putting effort to adapt existing systems.

Future work on presented system will include development and optimization of multi-threading, including dynamic scaling of parallel computation. Another area for increasing efficiency is introduction of functionality for distributed computing. In nowadays pattern recognition technologies (such as automatic speech recognition) proper creation of signal models is extremely time-consuming without multi-cluster computing.

DAMAGE system will also gain new management and administration tools for easy creation of new analyses and real-time analysis performance visualization.

This project has been supported by a grant (decision number DEC-2011/03/B/ST7/00442) from the National Science Centre.

BIBLIOGRAPHY

1. Ludäscher B., Altintas I., Bowers S., Cummings J., Critchlow T., Deelman E., Roure D. D., Freire J., Goble C., Jones M., et al.: Scientific process automation and workflow management. Scientific Data Management: Challenges, Existing Technology, and Deployment. Computational Science Series, 2009, p. 476÷508.
2. Foster I., Geneva C.: Big process for big data. 2013.
3. Ziółko M., Igras M., Kacprzak S.: Phonemes analysis for genealogical tree of world languages. Ways to Protolanguage 3 Conference, Wroclaw 2013.
4. Kacprzak S., Ziółko M., Mąsior M., Igras M., Ruszkiewicz K.: Statisical analysis of phonemic diversity in languages across the world. XIX National Conference Applications of Mathematics in Biology and Medicine, Jastrzebia Góra 2013.
5. Mąsior M., Igras M., Ziółko M., Kacprzak S.: Database of speech recordings for comparative analysis of multi-language phonemes. Studia Informatica, Vol. 34, No. 2B, Wydawnictwo Politechniki Śląskiej, Gliwice 2013, p. 79÷87.

6. Barker A., van Hamert J.: Scientific workflow: A survey and research directions. Parallel Processing and Applied Mathematics Lecture Notes in Computer Science, Vol. 4967, 2008, p. 746÷753.
7. Curcin V., Ghanem M.: Scientific workflow systems – can one size fit all? Biomedical Engineering Conference CIBEC 2008, Cairo International, IEEE 2008, p. 1÷9.
8. Ludäscher B., Altintas I., Berkley C., Higgins D., Jaeger E., Jones M., Lee E. A., Tao J., Zhao Y.: Scientific workflow management and the kepler system: Research articles. Concurr. Comput.: Pract. Exper, 18(10), August 2006, p. 1039÷1065.
9. Geisser S.: Predictive inference: an introduction. Chapman & Hall, New York 1993.
10. Kohavi R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. Morgan Kaufmann, 1995, p. 1137÷1143.

Wpłynęło do Redakcji 22 grudnia 2013 r.

Omówienie

Budowa i rozwój zaawansowanych aplikacji oraz systemów przetwarzania sygnałów wymagają ogromnej ilości danych i położenia szczególnego nacisku na ich właściwe przetwarzanie. Algorytmy przetwarzania obrazów oraz technologii mowy, często oparte na różnych metodach uczenia maszynowego, w trakcie rozwoju są wielokrotnie uruchamiane, trenowane i ewaluowane.

Artykuł przedstawia środowisko do zarządzania analizami danych naukowych. Podobne systemy są tworzone i rozwijane także na potrzeby innych badań ([1, 6, 7, 8]), ale idea ich działania skupia się na zaawansowanych narzędziach budowy różnych ścieżek analiz. W systemach przetwarzania sygnałów, szczególnie w fazie ich rozwoju, same procedury analiz dość często się zmieniają, natomiast zbiory danych zmieniają się rzadziej. Uzasadnionym podejściem jest więc skupienie się na odpowiednim przechowywaniu danych (łatwym i szybkim dostępie do nich) oraz ich zarządzaniu (śledzenie aktualności danych na każdym etapie przetwarzania). Takie podejście umożliwi łatwą rozbudowę tworzonych scenariuszy analiz oraz przyśpieszy ich wykonanie.

Prezentowany system, oparty na tej idei, składa się z kilku komponentów: aplikacji zarządzającej (rys. 4), relacyjnej bazy danych (rys. 2, rys. 3), plików konfiguracyjnych oraz skryptów analiz w różnych środowiskach obliczeniowych (skrypty MATLAB lub języka R, aplikacje i skrypty .NET, Python itp.).

System jest ciągle rozwijany i rozbudowywany o nowe komponenty. Jego ciągłe wykorzystywanie w badaniach nad analizą różnojęzycznych fonemów ([3]) pozwala na dobre określenie wymagań wobec niego i zweryfikowanie jego przydatności.

Addresses

Mariusz MĄSIOR: AGH University of Science and Technology, Department of Electronics, Al. Mickiewicza 30, 30-059 Krakow, Poland, masior@agh.edu.pl.
Stanisław KACPRZAK: AGH University of Science and Technology, Department of Electronics, Al. Mickiewicza 30, 30-059 Krakow, Poland, masior@agh.edu.pl.