IEEE Second International Conference on Data Stream Mining & Processing
August 21-25, 2018, Lviv, Ukraine

CoLiTec Software for the Astronomical Data Sets Processing

Sergii Khlamov
Laboratory of astrometry
Institute of Astronomy,
Kharkiv National University
Kharkiv, Ukraine
sergii.khlamov@gmail.com

Vadym Savanevych
Western Radio Technical Surveillance Center
State Space Agency of Ukraine
Mukachevo, Ukraine
vadym@savanevych.com

Olexander Briukhovetskyi
Western Radio Technical Surveillance Center
State Space Agency of Ukraine
Mukachevo, Ukraine
izumsasha@gmail.com

Artem Pohorelov
Computer engineering and management
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
artempogorelov@gmail.com

Vladimir Vlasenko
Western Radio Technical Surveillance Center
State Space Agency of Ukraine
Mukachevo, Ukraine
vlasenko.vp@gmail.com

Eugen Dikov
Scientific Research Design and Technology Institute of Micrographs
Kharkiv, Ukraine
endikov@gmail.com

***Abstract*—Nowadays, quick technological progress provokes creation of a big amount of the information that can be fed in different forms. There are a lot of different fields of science that use high dimensional data sets to analyze them in their researching. So, we need the data pre-processing methods and data reduction models to simplify input data sets by reducing unnecessary information. The paper deals with an approach of CoLiTec (Collection Light Technology) software to process in automated mode the different types of astronomical information which is fed online in the form of data sets or streams. Also the benefits of an usage of the OnLine Data Analysis System (OLDAS) was described. OLDAS helps with solving of the following Data Mining problems, such as clustering, classification and identification.**

Keywords—data, set, stream, series of frames, processing, CoLiTec, OLDAS.

I. Introduction

The 21st century is closely connected with a huge revolution that characterized by tremendous technological progress. This progress provokes appearing of a big amount of different data sets, streams, but data growth is ahead of computing abilities of the existed machines. That's why it is very important to optimize data stream processing by using only necessary input data to allow improving the computing abilities of machines. So, data mining has become under the interest that has attracted a huge number of researches and experimentation to improve efficiency and productivity in different fields of interest.

What is the Data Mining? Data mining can be defined as a process during which previously unknown, nontrivial and hidden information will be collected in a larger dataset [1]. The data mining goal is extraction of the potentially useful information from the given large input online stream or data sets using necessary relationships, associations or patterns within the data, and transformation received data into subsets with understandable structure. These formed subsets will be used for the effective analysis and using in the future.

II. Problem Statement

The new networks of automated ground- and space-based observation systems and new surveys projects lead to a fast growing of the astronomical data sets. These sets can be fed in different forms, e.g. files stream, video stream, physical data saved on the different servers. For example, the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) [2] currently contains two telescopes with 1.8-m aperture that located at Haleakala in Hawaii (Fig. 1).

Fig. 1. Panoramic Survey Telescope and Rapid Response System

Each telescope has a field of view that equals to 3-degree. Both of them are equipped with the largest CCD-camera, which records about 1.400 millions of pixels per image. Each image requires about 2 gigabytes of storage and exposure times will be up to one minute. Also the time for computer processing (saving image to the storage) is equal to one minute or even more. Since images are taken on a continuous basis as an online data stream, more than 10 Terabytes of data are obtained every night.

Also the Large Synoptic Survey Telescope (LSST) [3] currently is under construction. It is a wide-field survey with reflecting telescope, which has a primary mirror with diameter 8.4 meters. The design of a telescope includes three mirrors. Both of them have very wide field of view that equals to 3.5-degree. Telescope uses a CCD-camera with resolution of 3.2-gigapixel (Fig. 2).

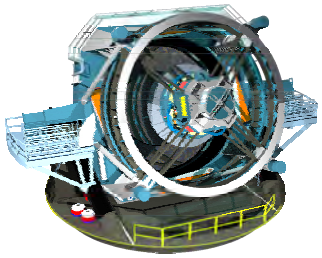

Fig. 2. Large Synoptic Survey Telescope (LSST)

LSST will take images of the full sky every few nights. There will be about 200 thousands of uncompressed images per year that equals to 1.28 petabytes. The managing and effective data mining and processing of the received from telescope data sets will be a very difficult part of the LSST project. Approximate requirements for the servers are about one hundred teraflops of power and about 15 petabytes of storage.

Such a big amount of the observational data requires a big volume of the hard disks or clusters. All these observational data can't be collected on a single universal local server. So, the disturbed systems as well as clouds technologies should be used for this goal, but how to process the big and high dimensional data streams in an efficient way?

III. Task solution

The problems with classification and identification the data sets are related to Data Mining problems. They can be resolved in semi-automatic or automatic mode. Before the applying of data mining algorithms, intelligent methods to formed useful data, the pre-processing of it is recommended or even required. Because data mining can accidentally causes the misunderstanding when producing the significant results that cannot be used in the future to predict behavior and cannot be reproduced. Often, such results can appear during investigating of too many hypotheses and performing not properly testing of statistical hypotheses.

That's why the correct data selection is a necessary pre-processing step in the analyzing high-dimensional data sets or streams. It is considered to be a complex and time-consuming problem. So, the main goal of optimization of the data stream processing is an enhance the accuracy of classification and reducing the execution time.

During pre-processing step the anomaly detection and data cleaning are performed. Anomaly detection allows the identification of the records with an unusual data. Some of them except data errors can be interesting for the further investigation. Data cleaning removes noisy, irrelevant, redundant or missing data.

After pre-processing step the remaining useful information in data set will be categorized into clusters, based on the specific attributes. Clustering therefore breaks down data sets into subsets, whereby the different elements are assigned to the appropriate groups while the similar data are grouped together. Then created subsets will be classified by applying known structure to the new data.

Approach of CoLiTec (Collection Light Technology) software (http://neoastrosoft.com) [4-6] can resolve data mining problems by using of the OnLine Data Analysis System (OLDAS) that covers different scientific and technological fields.

The main purpose of it is to bring together astronomical observation results of ground- and space-based observation systems, provide astronomers with the full-scope instruments for accessing and analyzing of the collected data.

IV. CoLiTec software

Software for data sets processing in automated mode is necessary for the most effective astronomical observations. This data set can be a survey given as series of frames. Modern astronomical systems and telescopes, e.g. Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) [2] and Large Synoptic Survey Telescope (LSST) [3], allow taking a lot of frames of considerable sky area in one night.

This is a big amount of the raw data sets that should be processed. The main goal of this can be achieved using the Data Mining approach.

This approach is provided by CoLiTec software [5, 6] that allows processing of the input data sets or streams in real time. The visual confirmation of results after processing is also available.

With help of the CoLiTec software you can process the observation data that is continuously formed during observation (online stream). The processing pipeline includes corrupted data rejection, brightness equalization of frames, astrometry, photometry, detection of the moving objects and others.

The CoLiTec software guarantees not only a high efficiency of data sets processing but also a high accuracy of the data measurements [4, 5, 7]. We provided the comparison of statistical characteristics of positional CCD-measurements between CoLiTec and Astrometrica [8] software with the same set of CCD-frames. This comparison demonstrated that the limits for reliable positional CCD-measurements with CoLiTec software are wider than those with Astrometrica one for the area of extremely low signal-to-noise ratio [9].

The On-line Data Analysis System (OLDAS) is a very important part of the CoLiTec software. Using OLDAS you can process the data sets and streams as soon as they are successfully saved on the storage or uploaded to the server. This approach allows speeding up of the processing with preventing the collision. Also it provides the immediate notification about the emerging issues for user.

Also OLDAS provides ability to process the Big Data in real-time. For example, the data set that includes frames can be used for real-time photometry. The result of processing can be represented as light curves of the investigated variable stars. These light curves will be created and visualized on our server.

Some another CoLiTec software features are the following: intraframe processing (estimation of the objects position, astrometry and photometry reduction), interframe processing (detection of the moving objects and trajectories) and confirmation of the most interesting objects at the night of their preliminary discovery.

According to Data Mining approach CoLiTec software performs the following.

A. *Pre-processing*

During pre-processing step CoLiTec software in OLDAS mode starts processing with the input data set or stream as soon as they successfully received. These raw data will be moderated before using in computing process. At this stage unsupported and corrupted frames will be rejected. Only useful information from data set will be used in computing process.

B. *Clustering*

The remaining useful information in data set will be categorized into clusters with help of specified attributes. CoLiTec software uses the following attributes: equatorial coordinates, telescope, filter type, object under investigating and others. According to these attributes the appropriate input data from set will be separated into subsets with similar data.

C. *Classification*

Created after clustering subsets of data will be classified by applying known astronomical structure of the raw data that specified in Flexible Image Transport System (FITS) standard by NASA [10]. FITS is the most commonly used digital file format in astronomy. It is designed specifically for scientific data and includes as well as various astrometric, photometric or calibration information as the image metadata. After input data set classification the FITS files are sent to the processing pipeline.

D. *Identification*

While processing pipeline starts receiving of the classified FITS files it identify types of them. For example, is this FITS file a raw light frame or maybe it is service master-frame that is used in frame's calibration (e.g. bias, dark, darkflat, flat). If this is a raw light frame the processing pipeline starts computing process.

E. *Processing*

Computing process consists of two stages: intraframe and interframe processing. Intraframe processing is designed to estimate the position of all objects (stars, galaxies, asteroids, comets) in the frame at current moments. Also calibration, background alignment and brightness equalization are performed at this stage (Fig. 3). In OLDAS mode the brightness equalization process has the following workflow:

- online searching for the frames in specified directories;
- searching for the required additional frames (bias, dark, darkflat, flat) if they are not specified;
- creating of the appropriate master-frames;
- applying the inverse median filter.

Interframe processing is used to detect and estimate moving objects trajectories. The core of CoLiTec software based on the preliminary objects detection by the accumulation of statistics that performed by multi-valued transformation of the objects coordinates that corresponds to Hough space-time transformation.

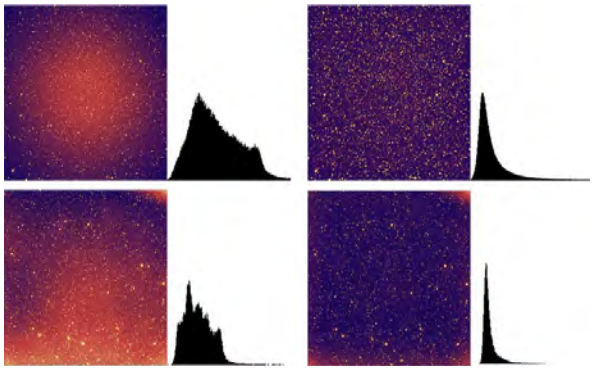

Fig. 3. Brightness equalization with histograms (before and after).

CoLiTec has abilities for detecting very slow, very fast objects and objects with near-zero apparent motion [5-7]. Range of velocities for detection is from 0.8 to 40 pixels per frame. For example, the fasters NEO is K12C29D (40 pixels per frame) or the object with near-zero apparent motion is ISON C/2012 S1 comet (0.8 pixels per frame) [11].

Also CoLiTec software has the following features: automated detection of the faint moving objects (SNR > 2.5); working with the huge field of view (< 10 degrees); automated calibration and brightness equalization; automated astrometric and photometric reduction; deciding system of the processing results allows adapting the user settings and inform user about correct results at the each stage of processing; automated objects rejection with bad or unclear observations; multi-threaded processing support; multi-cores systems support with the ability to manage individual treatment processes.

CoLiTec software has the system for monitoring processing messages with a detailed logging of handling process and tracking system of all running modules allows correct managing and terminating processes at any stage without data losses.

Also CoLiTec software includes pipeline for digital video processing. It is presented in form of the flexible platform for receiving and processing video in any resolution. Also the pipeline allows an easy integration of the different modules for improving the image quality and detection of the moving objects.

F. *Summarization*

After pipeline processing CoLiTec provides the different forms of data set representation, including results visualization and generation of report to different services.

CoLiTec software equipped with the modern viewer of processing results LookSky with a user-friendly GUI (Fig. 4).

LookSky viewer can be run without the main program. It can be used for independent review of the processing results by CoLiTec during data processing of the main program. With help of blinking method human can’t carefully analyze all interesting objects in the frame.

The particularly serious difficulty for this is the frames analyzing from the wide-field telescopes with a huge aperture. Because about several tens of asteroids with slight shine can be present simultaneously in telescope's field of

view. That's why automatic series of frames processing is necessary to the modern astronomer.

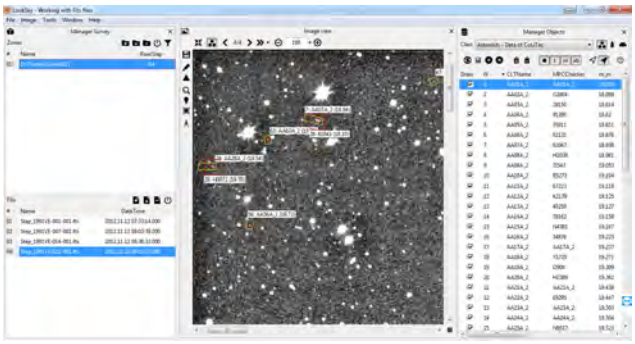

Fig. 4. LookSky viewer of processing results.

V. Conclusions

The quick technological progress, new surveys projects and networks of automated ground- and space-based observation systems lead to the fast growing of astronomical data sets that can be provided in the different form and volume. A huge amount of the raw data sets are needed to be processed, but trend of the data growth is ahead of computing abilities of the existed machines. So, we suggest using Data Mining approach for all goals that connected to the processing of a big amount of data. As described in the paper the Data Mining approach is very useful for the optimization of data stream processing. It allows using only necessary input data to improve the computing abilities of machines. The good example of using the data mining principles and stages of processing (anomaly detection, clustering, classification, identification, and summarization) is the CoLiTec software [4-6].

With help of the CoLiTec software more than 1,500 asteroids were preliminary discovered, including 5 NEO, 21 Trojan asteroids of Jupiter and 1 Centaur. CoLiTec software was used in about 700,000 observations, during which four comets (C/2010 X1 (Elenin), P/2011 NO1 (Elenin), C/2012 S1 (ISON) [11], P/2013 V3 (Nevski)) were discovered.

Acknowledgment

The authors thank observatories and institutes that have implemented CoLiTec software for observations. We express our gratitude to Mr. W. Thuillot, coordinator of the Gaia-FUN-SSO network [12], for the approval of CoLiTec as well-adapted software for the frames processing for all Gaia-FUN-SSO members (https://gaiafunsso.imcce.fr).

CoLiTec software is available at the following website: http://neoastrosoft.com.

References

[1] D. Peralta, S. del Rio, S. Ramirez-Gallego, I. Triguero, J. Benitez, and F. Herrera, "Evolutionary feature selection for big data classification: A map reduce approach," Mathematical Problems in Engineering, vol. 2015, Article ID 246139, pages 11, 2015.

[2] L. Denneau, R. Jedicke, Tommy Grav, Mikael Granvik, Jeremy Kubica, and Andrea Milani, "The Pan-STARRS Moving Object Processing System," Publications of the Astronomical Society of the Pacific, vol. 125, pp, 357-395, 2013.

[3] M. T. Tuell, H. M. Martin, J. H. Burge, W. J. Gressler, and C. Zhao, "Optical testing of the LSST combined primary/tertiary mirror," Proc. SPIE 7739, Modern Technologies in Space- and Ground-based Telescopes and Instrumentation, 77392V, 23 July 2010.

[4] V. E. Savanevych, O. B. Briukhovetskyi, N. S. Sokovikova, M. M. Bezkrovny, I. B. Vavilova, Yu. M. Ivashchenko, L. V. Elenin, S. V. Khlamov, Ia. S. Movsesian, A. M. Dashkova, and A. V. Pogorelov, "A new method based on the subpixel Gaussian model for accurate estimation of asteroid coordinates," Monthly Notices of the Royal Astronomical Society, vol. 451 (3), pp. 3287–3298, 2015.

[5] V. E. Savanevych, S. V. Khlamov, I. B. Vavilova, A. B. Briukhovetskyi, A. V. Pohorelov, D. E. Mkrtichian, V. I. Kudak, L. K. Pakuliak, E. N. Dikov, R. G. Melnik, V. P. Vlasenko, and D. E. Reichart, "A method of immediate detection of objects with a near-zero apparent motion in series of CCD-frames," A & A, Worldwide astronomical and astrophysical research, vol. 609, A54, pages 11, 2018.

[6] S. Khlamov, V. Savanevych, O. Briukhovetskyi, and A. Pohorelov, "CoLiTec software – detection of the near-zero apparent motion," Proceedings of the International Astronomical Union: Cambridge University Press, vol. 12(S325), pp. 349–352, 2017.

[7] S. V. Khlamov, V. E. Savanevych, O. B. Briukhovetskyi, and S. S. Oryshych, "Development of computational method for detection of the object's near–zero apparent motion on the series of CCD–frames," Eastern-European Journal of Enterprise Technologies, vol. 2, iss. 9 (80), pp. 41-48, 2016.

[8] H. Raab, "Astrometrica: Astrometric data reduction of CCD images" 2012, Astrophysics Source Code Library, record ascl:1203.012.

[9] V. E. Savanevych, A. B. Briukhovetskyi, Yu. N. Ivashchenko, I. B. Vavilova, M. M. Bezkrovniy, E. N. Dikov, V. P. Vlasenko, N. S. Sokovikova, Ia. S. Movsesian, N. Yu. Dikhtyar, L. V. Elenin, A. V. Pohorelov, and S. V. Khlamov, "Comparative analysis of the positional accuracy of CCD measurements of small bodies in the solar system software CoLiTec and Astrometrica," Kinematics and Physics of Celestial Bodies, vol. 31 (6), pp. 302–313, 2015.

[10] D. C. Wells, E. W. Greisen, and R. H. Harten, "FITS - a Flexible Image Transport System", A & AS, vol. 44, p. 363, 1981.

[11] Minor Planet Center, COMET C/2012 S1 (ISON). Available at: http://www.minorplanetcenter.org/mpec/K12/K12S63.html.

[12] W. Thuillot, B. Carry, J. Berthier, P. David, D. Hestroffer, and P. Rocher, "Gaia-FUN-SSO: a network for ground-based follow-up observations of Solar System Objects," Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, SF2A-2014, pp. 445-449, , 2014.