Sinteza
Međunarodna naučna konferencija Univerziteta Singidunum

Impact of Internet on Business Activities in Serbia and Worldwide
Uticaj Interneta na poslovanje u Srbiji i svetu

DOI: 10.15308/SINTEZA-2014-88-93

INTELLIGNECE AND SECURITY DATABASE (ISD) VS. 'HARD CORE' SCIENCES DATABASES (HSCD): CHALLENGES AND OPPORTUNITIES

Anđelka Kovačević[1], Milan Dimitrijević[2], Luka Popović[3]
[1]Faculty of Mathematics, Department of astronomy, University of Belgrade
[2]IHIS, Serbia
[3]Astronomical Observatory, Belgrade, Serbia

Abstract:
Intelligence and security databases (ISD) are facing the specific challenges and opportunities of information overload and the ultimate need for advanced intelligence analyses and investigations. We summarize analogies with hard core science databases (HCSD) challenges(the case of synoptic astronomy) and point out HCSD experiences which could be applied on problems facing ISD.

Key words:
databases.

INTRODUCTION

We are in an avalanche of data today. In many fields, data is collecting at never known before rates. Conclusions and decisions that previously were done through guesswork, or on thorough models, now is made using the data itself. Such Big Data analysis now governs almost eveiy area of our modern society, including telecommunication networks, retail, manufacturing, financial services, life sciences, homeland intelligence and defense systems and physical sciences.

Using sensors, experiments, long term monitoring campaigns and computer simulation, Hard Core Sciences (HCS) data is growing in volume and complexity at an enourmous rate. Today, the cost of producing the data is very high:satellites, large telescopes, particle accelerator, genome sequencing and supercomputing centers are just of some examples of information generators that cost billions.

The difficulties facing the intelligence and security (IS) data systems are of the same kind as it was mentioned above in the case of HCS. As the sensors used in the various IS surveillance missions improve, the data volumes are increasing with a projection that sensor data volume could potentially increase to the level of Yottabytes (10^{24} Bytes) beyond 2015. At present, IS survey campaigns such as the Global Hawk [1] system, are potent of producing 10's to 100's of Terabytes [2] over a period of hours. In contrast, the capability of transporting or storing this data is not keeping pace with projected growth of data.

In Table 1 is given some impression of the relative size of the data sets being considered in IS. If we assume that the earth has surface area of $5*10^{14}$ m^2 and that we could allocating 1byte/m^2 with resolution of 1 m^2/sec, it could be seen from Table 1 expected amount of collected data after certain time. So, IS data volumes are in many cases comparable to those encountered in other data intensive enterprises, particularly in HCS.

TABLE 1 ILLUSTRATION OF SCALES OF IS DATA

time of data collecting	collected information
1hour	1.8 Exabytes
30 days	1.3 Zettabytes
365 days	16 Zettabytes
36500 days	1.6 Yottabytes

Both, IS and HCS data becomes useful information only by means of using the scientific methods, i.e., validation of models and induction of rules from observations. Up to now, database technology has evolved mostly targeting financial applications, where correctness and completeness are imperative, with large a priori knowledge to prepare the system for fast response. However, for intelligence and scientific databases, none of this is possible anymore. Beside the enormous amount of data to be processed, the users do not always know exactly what they are looking for and they not always care for a complete answer; of the greatest importance is searching for interesting patterns. The Volume of the data, as a major challenge, is the one that is most easily recognized.

However, there are other challenges, as it is in Variety and Velocity [3]. By Variety, it is usually meant heterogeneity of data types, representation, and semantic interpretation. By Velocity, it is meant both the rate at which data arrive and the time in which it must be acted upon. The last is very important in the cases such as hurricanes, tsunami, earthquakes, near earth asteroids flybys as well as crime and terrorist acts. Beside this three major V's we could found more V's challenges such as:Veracity (verifying inference-based models from comprehensive data collections) , Variability, Venue, Vocabulary, etc.

In HCS, the problems start already during data acquisition, e.g. when the data tsunami requires us to make decisions, currently in an ad hoc manner, in the sense what data to save and what to discard, and how to store what we keep reliably with the right metadata. Similarly with counter terrorism, large quantities of data are likely to be gathered and analysed to support the process of tracking and predicting terrorist activity. Even data of no apparent importance could be of greater significance at a later stage of investigation. As a result the variety of information is much greater than for conventional police investigations and it may not be possible in advance to predict all the kinds of information that one needs to hold.

Beside this, the value of data enormously rises when it can be linked with other data, thus data integration is a major creator of value. Since most data is directly generated in digital format today, we have the opportunity and the challenge both to influence the creation to later linkage and to automatically link previously created-historical data. Here, we review the overall challenges facing the ISD as regards large data, but with the objective of putting these in the context of similar challenges facing other large enterprises, such as HCSD. It is useful to examine one of these HCSD: the newly coming field of synoptic astronomy as it is an example in which the response to large data volumes is connected to the most important scientific goals.

DATA CHALLENGES

The case of synoptic astronomy

Astronomy is a "Big Data" science flourishing from the synergy of computer science and applied mathematics-particularly statistics. Fundamentally, modern astronomy has been established on the digital pictures of the Kosmos. Each pixel can have between ten and a few thousand attributes. At the image resolution of modern astronomical instrumentation, the entire sky requires a peta-scale database regime. Specifically, astronomy is coping to answer questions about our Universe by combining the observations with innovation in imaging analysis, non-parametric statistics, inference through machine learning, and high dimensional hypothesis testing and regression statistics.

There are some issues driving the current data challenges in astronomy. We are in a vastly different data regime in astronomy than we used to be even ten or fifteen years ago. Over the past three decades, we have been engineered telescopes that are 30 times larger, with detectors which are 3,000 times more powerful in terms of pixels. The rise in sensitivity of these detectors (for an example the multi- gigapixel cameras which is in line with similar improvements in homeland defense and intelligence systems) is a consequence of Moore's Law[4]—they can collect up to a hundred times more data than was possible even just a few years ago. This exponential increase induces that the collective data of astronomy doubles every year, and that can be very tough to capture and analyze.

The principal driving forces of change of astronomy were large digital sky surveys, most notably the Sloan Digital Sky Survey (SDSS; [5]), but also DPOSS [6], 2MASS [7],etc, which have been static surveys of the sky. A new wave is coming with the birth of synoptic sky surveys covering large areas of the sky in repetition (an example is CRTS; http:// crts.caltech.edu [8] and planned facilities for this decade and beyond, such as the Large Synoptic Survey Telescope (LSST; [9]) and the Square Kilometer Array (SKA [10]). So astronomy is transformed in short period of time from a panoramic digital sky-photography to a panoramic digital Universe-cinematography. To have a more quantitative impression on the data rates of synoptic astronomy, a useful is to compare it to the Large Hadron Collider (LHC) at CERN. At zenith of its capacity (all four experiments are running simultaneously), LHC generates ~1.8 GB/s and requires the largest distributed computing network in the world to handle its output. Since the network can transfer data at ~1 GB/s, we could use it as a fiducial value, denoted as 1 LHC.

TABLE 2 SURVEY DATA RATES IN TERMS OF THE DATA RATE OF THE LARGE HADRON COLLIDER (1 LHC = 1 GB/S)

Survey	Wavelength	Operation start	Data rate [LHC]
ASKAP	radio	2012	2.8
GAIA	optical	2013	0.005
LOFAR	radio	2013	50-200
LSST	optical	2018	0.7
SKA	radio	2020	~30000

As can be seen, from Table 2, the real challenge is with the new kind of radio surveys. Also, associating and relating these data to themselves and to other data will increase their volume and complexity.

These data-extreme surveys define the computational frontiers of astronomy in the next decades. Extrapolating current disk space growth rates to 2030 the entire LSST catalog (~200 TB) can be confined onto a single disk with plenty of space for associated data. However, conventional relational database technology will almost certainly not scale comparably. Relational database management system (RDBMS) [11] do not perform well beyond ~100 TB in size and so alternate solutions, such as the NoSQL class of distributed storage technologies for structured data, will be necessary for any of the larger surveys. In the context of the CAP (Brewer's) theorem [12], NoSQL stores often compromise consistency in favor of availability and partition tolerance.

NoSQL is fundamentally about simple key-value or document-style schema (collected key-value pairs in a "document" model) as a direct alternative to the explicit schema in classical RDBMSs. It allows the informatics engineer to treat things asymmetrically, whereas traditional approaches have enforced rigid uniformity across the data model. The reason this is so interesting is because it provides a different way to deal with changes in data model, and for larger data sets it makes interesting opportunities to manage volumes and performance.

There are also databases that combine two or more of following properties: document-oriented databases, key/value stores, graph databases, column-oriented databases, in- memory databases, and other database types. An attempt to better match needs of managing scientific data is SciDB [13], which is a column-oriented entity (rather than row-oriented like a RDBMS) that uses arrays as first-class objects rather than tables but is still ACID.

The Variety of data in astronomy is important challenge, because it allows us to discriminate subtle new classes of objects (e.g. Class Discovery). Class Discovery assumes a distinctive attribute separation and discrimination of classes. The separation of classes improves when the"correct" criterion are chosen for investigation. To a computer scientist, "clustering" is a discovety process (see [14], [15]) that groups objects and their similarity is maximized within the group, while the similarity between objects in different groups is minimized ([16], [17]). In astronomy, we often group objects into "populations" with distinct properties. There is a large overlap between astronomical population and informatics"cluster."There are many different types of astronomical populations. If the properties discriminating the populations are spatial (positions on the sky or in space), the populations identified may be real physical "clusters" of objects. For example, it has been well known that galaxies tend to form "galaxy clusters." Spatial clusters are among very common populations in astronomy. However, populations with similar physical and image parameters may exist both within the spatial clusters and independent of them. Similarly, IS domain tends to have different type of clustering in their cyberspace. Building clustering algorithms for astronomical data poses a many challenges due to both the characteristics of the data discussed as well as the types of the desired clusters. The clusters may be of variable sizes and densities, and of arbitrary shapes. For spatial clustering, algorithms have mostly a dynamic-modeling approach to measure the similarity between two clusters. Two clusters are merged in the case when the discrepancy of parameter values between the clusters is comparable to the internal scatter of the parameter values within each cluster. An example of such a clustering algorithm is Chameleon [18]. Another approach to the problem of identifying astronomical populations is unsupervised clustering: for example, the expectation maximization (EM) algorithm with mixture models to detect groups of interest, making descriptive summaries, and building density estimates for large data sets. Moreover, it would be of great opportunity using genetic algorithms to devise improved detection and supervised classification methods. This would be especially interesting in the case of interaction between the image (pixel) and catalog (attribute) domains. Clustering techniques could be used to detect rare, or in certain way unusual objects, e.g., as outliers in the parameter space, to be selected for further investigation. It is also possible to use semi-autonomous artificial intelligence (AI) or software agents to explore the large data parameter spaces and report on the occurrences of unusual instances or classes of objects. All mentioned is also importnat for IS data exploration.

The exploration of observable parameter spaces (OPS), created by combining of large sky surveys over a range of wavelengths will be one of main scientific purposes of astronomy. A complete observable parameters space axes include the object coordinates (positioning domain), velocities or redshifts, fluxes at a range of wavelength (spectrophotometric domain), and the measured time (MJD, UTC, time domain) etc (see [19]). We will notice that these domains are exactly domains of IS survey too. Astronomy and IS surveys cover some solid angle, over some wavelength range and with some dynamical range of fluxes (see Fig. 1). So, it is not the physical realm we want to study, actually it is parameter space embedded in cyber space. In the cyberspace, the data can be viewed as a set of n points or vectors in an m- dimensional parameter space. The magnitude of range of n could be many millions or billions and m could be within the range of few tens to hundreds and thousands. As we already mentioned possibilities of clustering algorithms, the data may be clustered in k statistically distinct classes, which could be modeled. This is a computationally high non trivial problem. However, not all parameters may be equally important, and lowering dimensionality of set of parameters would be an important task.

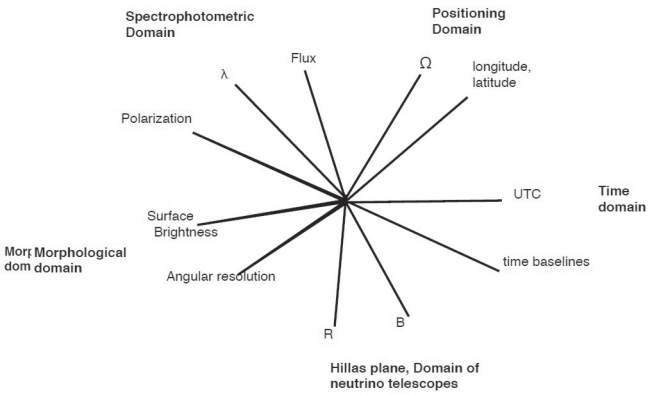

Figure 1. A schematic illustration of the observable parameter space of astronomy and IS.All axes are measured independently and they ar mutually orthogonal. Note that here is depicted non electromagnetic channels such as neutrinos, which parameter space is Hillas plane (RxB, where R is size of acceleration region and B is typical magnetic field of the source). Other non- electromagnetic channels, such as gravity waves parameter space, are not given.

Techniques such as automated pattern recognition and classification tools could be used to discover sources with

particular image morphology, or to employ AI techniques to search through panoramic images (from multiple wavelengths) for unusual image patterns. The typical search of such databases is not possible with traditional indexing technique (established with relational databases) since it is impossible to build an index on all possible linear combinations of attributes. However, what is possible is using the data as geometric objects (points) in the k dimensional space. So data can be quantazied into containers (collecting objects of similar properties)

Beside this, both astronomy and IS needs common reference frame for the sky and earth respectively which can be used by different astronomical/IS databases, making easy cross- referencing among the catalogues. This problem in astronomy was adressed in the study [20], which suggested hierarchical subdivisions (starting from octahedron base set,each spherical triangle can be recursively divided into 4 sub-triangles of approximately equal areas. Each subarea can be divided into additional sub-areas, etc.). Using such 3D Cartesian representation of angular coordinates simplifies finding objects within certain spherical distance-just by testing linear combinations of 3D coordinates. So tessellation and Cartesian coordinates merge into efficient storage. The SDSS implemented this algorithm. In such a way, the scientific databases will be the "virtual sky" that astronomers will study and mine.

It is well known, that new scientific understanding will flow from the discovered knowledge, which is derived from the avalanche of information content, which is extracted from the massive data collections. Knowledge discovery in databases (KDD ,[13]) is the process of extracting useful knowledge from data. In the process of data mining, the application of specific algorithms to discover rare or previously unknown types of object or phenomenon, is a particular step. KDD is inherently interactive and iterative. Common KDD functions are classification, cluster analysis, and regression. There are several astronomy-specific data mining projects such as: AstroWeka (http:// astroweka.sourceforge.net/), Grist (Grid Data Mining for Astronomy; http://grist.caltech.edu), the Laboratory for Cosmological Data Mining http:// lcdm.astro.illinois.edu), the LSST Data Mining Research study group ([21]), the Transient Classification Project at Berkeley [22], and Palomar Transient Factory (http:// www.ptf.caltech.edu/ iptf).

There are various algorithms developed and used as data mining tools in astronomy: Bayesian Analysis for example to separate galaxies from stars among the many thousands of objects detected in large images; Decision Trees - used in the identification of cosmic ray pollution in astronomical CCD images; Neural Networks - more recently applied in the classification of different galaxy types within large databases of galaxy data ; Support Vector Machines (SVM) - used in the determination of the photometric redshift estimate for distant galaxies or for forecasting solar flares and solar wind-induced geostorms. Also, some other other methods have been applied to astronomical data mining including principal component analysis (PCA), kernel regression, random forests, and various nearest- neighbor methods

The case of Intelligence and Security (IS)

There is agreement in opinions that the threats we could expect have expanded beyond the typical military or counter- intelligence threats of the past, especially those of the Cold War. This enlarged range of threat falls into a major category and two sub-categories. The major category can be marked as 'non-conventional' threats, those that do not fall into the state- on-state category. They include environmental threats, threats of pandemic disease, terrorism and transnational crime.This broad class of non-conventional threat can be divided between those threats of a human agency (terrorism, crime, people smuggling and trafficking) and those of a non-human agency (climate change and other types of environmental threat, natural disasters, pandemic disease). These two sub-categories are, however, closely linked, as demonstrated by [23] .

Law enforcement work has always handled large amounts of information in the form of textual data: for example case notes and reports. Dealing with unexpected kinds of data and investigating potential connections between disparate facts or elements are characteristics that are not well supported by current technology. As we already emphasized, information technology (IT) has been supported the storage, acquisition and analysis of the data through the use of record-based structures, often in relational database systems [11].

Such design is inappropriate when it is not possible to know all the types of information that may need to be stored. Also, particularly important in the context of intelligence analysis, is that record-based databases tend to obscure possible connections among the entities and facts. The record formats used by any usual application could be both very different and complex and it is difficult to develop general- purpose software that can explore the relations that is implicit in the data.There are other issues that remain unresolved in the storage and exploitation of intelligence: how to represent levels of certainty (which as also problem in astronomy) and the fact that much of the intelligence gathered may be contradictory or of doubtful origin. Another major issue, which was already mentioned in the case of astronomy, is the collation step in intelligence cycle and combination of data from a variety of different database designs.

It is very important to note that relational database systems are the best solution for many types of problems, especially when data is highly structured and volume is less than 10 terabytes. However, a new class of problem is emerging when dealing with large amount of data of volumes greater than 10 terabytes. Although relational database architectures are capable of running in a Data Cloud, many current such database systems fail in the Data Cloud in following manners. First, beside limitations in volume, such database requires highly specialized components to fulfill all the time small amount of increasing in scale. Second, and critically important to Intelligence Analysis, is the object-relational impedance mismatch that occurs as complex data is normalized into a relational table format.

There are investigation of exploit cloud computing for astronomy, which conclusions could be applied to IS due

to similarity of data. A team of astronomers [24], demonstrated the calculation of an large set of periodograms of light curves obtained by the Kepler mission, as an example of how the Amazon cloud can be used to generate a new science product. Although the costs presented in their study were low, these costs can grow significantly as the number of curves grows, or as the set of search parameters are enlarged. They concluded that commercial clouds may not be best solution for large- scale computations, due to applications that are best suited for commercial clouds are those that are processing- and memory- intensive. On the other hand commercial clouds applications that are I/O-intensive, which are the most suitable for astronomy and IS where is often involved processing large quantities of image data, they are uneconomical to run because of the high cost of data transfer and storage. They require high-throughput networks and parallel file systems to achieve best performance.

Despite the high costs of using clouds, the virtualization technologies used in commercial clouds could be more efficient for astronomy and IS when they are used within a data center. There is now a movement towards providing academic clouds, such as those being built by FutureGrid (http://futuregrid.org/) or the US National Energy Research Scientific Computing Center (NERSC) (http: //www.nersc.gov/ nusers/systems/magellan/) that will build virtual environment capabilities to the scientific community. Also, the CADC (Canadian Astronomy Data Center) is adjusting its entire operation to an academic cloud called CANFAR (Canadian Advanced Network for Astronomical Research), "an operational system for the delivery, processing, storage, analysis, and distribution of very large astronomical datasets. The goal of CANFAR is to support large Canadian astronomy projects." To our knowledge, this is pioneering astronomy archive that has migrated to cloud technologies. It can be considered as a protomodel of the archive of the future, and consequently its performance should be monitored by large the community of potential users. Also, we believe that this solutions could be applied on IS databases.

ANALOGIES BETWEEN IS AND ASTRONOMICAL DATABASES

Comparing astronomical with IS database demands, there exist tremendous analogies between two disciplines which we summarized in Table 3.

We know that pipeline image processing of the data streams in astronomy and IS databases will be possible using parallel processors. More interesting challenges are presented by the archiving and mining tasks. Storage technology is rapidly evolving, so that keeping all the data online will almost certainly be possible. What is more important, we need now to discover ways to search for correlations in the resulting massive database. While the required data hardware and software for the key science programs present challenges, assuring opportunity for unanticipated science using such huge databases presents an even grater challenge. Designing optimal data handling and search routines will be an exciting demand. Crafting the software pipeline and developing efficient database management tools and the algorithms for data mining will present more of a challenge than the pre- processing computational capacity. The demands of this post- processing will be hardware and software intensive. The effort invested in software, data system design, tools for visualizing and analyzing data, and the science data analysis, may be comparable to that spent on instruments.

TABLE 3 ANALOGIES BETWEEN ASTRONOMICAL AND IS DATABASES CHALLENGES

DATABASES		*astronomy*	*IS*
Challenges	*Domain*	OPS: -time -positioning (geographical and celestial) -spectrophotometry	OPS: -time -positioning (geographical) -spectrophotometry
	Variety data	finding: -classes -clusters	finding: -terrorist networks -terrorist classes -terrorist clusters
	Imaging	to discover sources: with particular image morphology multiwavelength images with unusual patterns	to discover sources: with particular image morphology multiwavelength images with unusual patterns
	Hardware	-most of computation to be done locally to data disks -bandwidth to disks -new technologies of storage:holographic	most of computation to be done locally to data disks -bandwidth to disks -new technologies of storage:holographic
	Software	data mining~statistics expressed as algorithms scalability with the number of data vectors and number of dimensions	data mining~statistics expressed as algorithms scalability with the number of data vectors and number of dimensions
	Language	XML, KML	Global Justice XML
Methodlogy		Knowledge Discovery from Databases	Knowledge Discovery from Databases
Contri-butions	*Scientific*	Discovery Informatics X-informatics	Discovery Informatics X-informatics criminology, terrorism research

The enormity of these astronomy and IS data sets creates statistical challenges beyond the computational. For example, many statistical techniques used by astronomers today have been optimized to deal with the small size of existing data sets. In order to fully understand and characterize the data, higher-order correlations are often necessary [25]. Unfortunately, directly computing the n point correlation function (npcf) is extremely computationally expensive. A direct computation of the npcf will require enumerating all possible n-tuples of data points. Since there are $O(N^n)$ n-tuples for N data points, this is prohibitively expensive for even modest-sized data sets and low- orders of correlation. So, algorithms which scale much worse than linearly will be unacceptable computationally. At the same time, the main source of errors will be various systematic effects.We should think to develop approximate statistical techniques, where the approximation is within some boundaries, which algorithm has a non-polynomial scaling.Furthermore, regardless the size

of the data set, there will always be features and scales for which the estimation error will be important, so the need for developing statistically efficient estimation schemes and methods for assessing estimation error will always remain. Combining statistical efficiency with computational efficiency will be a constant challenge, since the more statistically accurate estimation methods will often be the most computationally intensive. Up to now, in astronomy potentially devastating near-Earth objects go undetected. However new techniques of extracting relevant image parameters can be used on the petascale imaging data to automatically find such objects. Similar image-mining techniques can be very relevant in IS investigation as well. Finally, data visualization will present an impressive challenge. Efficient methods for statistical visualization and sampling of large databases are required. User- reconfigurable trees of image feature catalogs driving multi-dimensional displays could help, but the opportunities here are largely unexplored.

CONCLUSION

We analyzed challenges and opportunities of new generation astronomical databases and IS databases. In the sense of data characteristics, they both face the information avalnche and information overload problem. In the sense of technology development, they both are searching for new paths, methodologies, and innovative use of existing techniques. In terms of scientific contributions, they both may add new insights and knowledge to various academic disciplines.

Having on mind the unique challenges (and associated opportunities) of information overload and the pressing need for advanced criminal and intelligence analyses and investigations, we believe that the Knowledge Discovery from Databases (KDD) methodology [17], which has achieved significant success in other information-intensive, knowledge- critical domains including business, engineering, biology, astronomy, physics and medicine, could be critical in addressing the challenges and problems facing IS databases.

Aknowledgement

This paper is within the project 176002 of Ministry of Education, Science and Technological Development of Republic of Serbia.

REFERENCES

[1] The Official Web Site of US Air Force:http://archive.is/2012121203343 5/http://www af.mil/information/factsheets/factsheet.asp?fsID=13225
[2] M. Duchaineau, "Sensor based video processing", Presentation to JASON, June 23, 2008.
[1] Y. Genovese and S. Prentice, "Pattern-Based Strategy: Getting Value from Big Data", Gartner. Special Report, June 2011.
[3] J. Gray and A. Szalay, "2020 computing: science in an exponential world" , Nature, vol. 440, 413-414, 2006.
[4] D.G. York et al. (the SDSS team), "The Sloan Digital Sky Survey: technical summary", Astron. J. , vol. 120, 1579 -1587, 2000.
[5] S. G. Djorgovski, R. Gal, S. Odewahn, R. de Carvalho, R. Brunner, G. Longo, and R. Scaramella, "The Palomar Digital Sky Survey (DPOSS)", In: Wide Field Surveys in Cosmology, eds. S. Colombi, Y. Mellier, and B. Raban, Gif sur Yvette: Editions Frontieres, pp. 89-99, 1998.
[6] M. Skrutskie, R. M. Cutri, R. Stiening, M. D. Weinberg, S. Schneider, at al. (the 2MASS team), "The Two Micron All Sky Survey (2MASS)" Astron. J, vol 131, pp.1163-1183, 2006.
[7] A. J. Drake, S. G. Djorgovski, A. Mahabal, E. Beshore, S. Larson, et al.:"First results from the Catalina Real-Time Transient Survey.". Astrophys. J. vol. 696, pp.870-884, 2009
[8] Large Synoptic Sky Survey. http://lsst.org
[9] Square Kilometer Array. http://skatelescope.org
[10] E. F. Codd, "A Relational Model of Data for Large Shared Data Banks", Communication of the ACM, vol. 13, no. 6, pp.377-387, 1970.
[11] N. Lynch and S. Gilbert, "Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services", ACM SIGACT News, vol. 33, issue 2, pp. 51-59, 2002.
[12] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, " Knowledge discovery and data mining: towards a unifying framework", Proc. International Conf. Knowl. Disc. Data Mining 2, pp. 82-88, Portland, 1996.
[13] M. Stonebraker, R. Agrawal, U. Dayal, E. J. Neuhold, A. Reuter, "DBMS Research at a Crossroads: The Vienna Update," Proc. of the 19th VLDB Conference, pp.688-692, 1993.
[14] M. S. Chen, J. Han and P. S. Yu, "Data Mining: An Overview from a Database Perspective", IEEE Transactions on Knowledge and Data Engineering, 8^6), pp. 866-883, 1996.
[15] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data", Prentice Hall, 1988.
[16] L. Kaufman, and P. J. Rousseeuw, " Finding Groups in Data: an Introduction to Cluster Analysis", John Wiley & Sons, 1990.
[17] G. Karypis, R. Aggarwal, V. Kumer and S. Shekhar, "Multilevel Hypergraph Partitioning: Applications in VLSI Domain", IEEE Transactions of VLSI Systems, 7(1), pp. 69-79, 1999.
[18] S. G. Djorgovski, A. A. Mahabal, A. J. Drake, M.J. Graham, and C. Donalek,"Sky surveys", in Planets, Stars, and Stellar Systems-Volume 2: Astronomical Techniques, Sofftware, and Data, eds. T. Oswalt and H. Bond,Springer Netherlands, 2013, pp.223-281
[19] A. S. Szalay, J. Gray, P. Kunszt, A. Thakar, and D. Slutz, "Large Databases in Astronomy," Mining the Sky, Proceedings of MPA/ESO/ MPE workshop, Springer, 2001, pp. 99-118
[20] K. D. Borne, M. A. Strauss and J. A. Tyson, "Data Mining Research with the LSST,"American Astronomical Society, AAS Meeting 211, Bulletin of the American Astronomical Society, vol. 39, pp.983-983, 2007.
[21] J. S. Bloom, D. L. Starr, N. R. Butler, P. Nugent, M. Rischard, D. Eads, and D. Poznanski, "Towards a Real-time Transient Classification Engine," Astronomiische Nachichten, vol. 329, pp.284-287, 2008.
[22] T. Homer-Dixon and J. Blitt, "Introduction: A Theoretical Overview", in Ecoviolence: Links Among Environment, Population and Security, eds. T. Homer-Dixon and J. Blitt, Rowman and Littlefield Publishers Inc, 1998, pp.1-17
[23] G. B. Berriman, E. Deelman, G. Juve, M. Regelson, P. Plavchan, "The Application of Cloud Computing to Astronomy: A Study of Cost and Performance"in Proceedings of the 2010 e-Science in Astronomy Conference, Brisbane, Australia, pp.1-7, 2010.
[24] S. White, " The hierarchy of correlation functions and its relation to other measures of galaxy clustering", Monthly Notices of the Royal Astronomical Society, vol. 186, pp. 145-154, 1979.