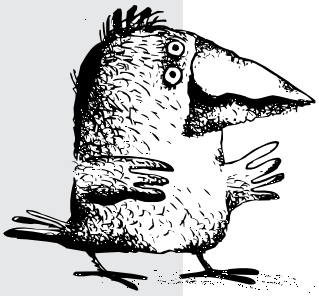
HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

School of Business Informatics

Proceedings of the Russian-French Workshop in Big Data and Applications



October 12–13, 2017
Moscow

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

School of Business Informatics

Proceedings of the Russian-French Workshop in Big Data and Applications

October 12–13, 2017
Moscow

Higher School of Economics Publishing House
Moscow 2018

УДК 004.6
ББК 32.97
Р93

Р93 **Proceedings** of the Russian-French Workshop in Big Data and Applications. October 12–13, 2017, Moscow [Electronic resource] / Executive editor V. V. Kornilov ; National Research University Higher School of Economics. — Electronic text data (8.0 MB). — Moscow: HSE Publishing House, 2018. — ISBN 978-5-7598-1808-3.

УДК 004.6
ББК 32.97

Published by Higher School of Economic Publishing House
<http://id.hse.ru>

ISBN 978-5-7598-1808-3

Content

GEOMARKETING MIX OPTIMIZATION USING A FUZZY SPATIAL AND MULTISCALE SEGMENTATION

Jérôme Baray (Prof.)
IRG, University of Paris-East (UPEC), France

Martine Pelé (Prof.)
LARGEPA, University of Paris 2 Pantheon Assas, France

Abstract. *This article introduces a new method for optimising the marketing-mix of a product or service by taking into account supply and demand features including their geographical location. Introducing the concept of geomarketing mix, the method using a factorial analysis and a fuzzy clustering enables to automatically detect business and strategic opportunities.*

Keywords: *spatial segmentation, geomarketing mix, retail, clustering.*

Introduction

Taking into account both supply and demand, the four leverage parameters of "geomarketing mix[1]" (price, product, promotion, place or distribution incorporating the notion of space) have been variously studied. But it is retail locations optimization and distribution intrinsically including high spatial components which have interested the most, managers, geographers, logisticians, mathematicians, economists, computer scientists and operational researchers. Promotion, especially outdoor advertising, flyers and letterbox advertising have been addressed with similar methods to those used in facility location optimization.

The study of consumer demand involving numerous qualitative and quantitative hierarchical components changing in time and space is difficult to implement. It is therefore necessary to define first the relevant geographical region and the category of products or services to handle. Spatial data mining knowledge discovery is the process of

[1] Geomarketing mix — according to our approach, marketing mix taking into account the spatial parameter. Its optimization consists in specifying the 4*P* values everywhere in the investigated geographic area.

analyzing large data set of different sources and to extract useful information. This technique splits in predictive and descriptive spatial data mining and shows a promising future in the field of optimization for management problems.

A global optimization of the marketing mix approach means, at first, to get a picture of the market concerning the type of product or service offered, taking into account the varying spatial and temporal parameters often neglected in traditional marketing approach. The analysis step requires an accurate description or a detailed segmentation of demand, supply and environment in the study area e.g. trading area and competition analysis. Given the strengths and weaknesses of the company, a good marketing mix forms a coherent set of decisions on price, product, communication and distribution of the product or brand to achieve the firm's objectives. This step called positioning reflects a strategic choice of the company that agrees to waive certain customers and certain market segments in order to make it more attractive offer in other parts of the market. The difficulty stems from the fact that all information about demand and supply is not freely available and a lot of expenditures, time and efforts are needed to gather a minimum of essential data to make an accurate analysis and then possibly take strategic marketing decisions.

This article presents a method to both spatially segment a market and optimize the marketing mix components. In the first part, we will present the latest methods used in geographical segmentation and their limits, which is traditionally a premise to a good geomarketing mix choice. Given a certain range of products, the main strategic lever in marketing decisions is price [Gauri, Sudhir, Talukdar, 2008]. Indeed, according to classical methods, this price for each product is usually determined in different areas following a geographic segmentation of demand and local analysis of environment and competition. Then, after reviewing general marketing mix optimization methods associated with a geographical approach, global spatial advanced pricing models will be considered as well as their limits. In a third part, we introduce the new geomarketing mix optimization system which comprises a pre-treatment of available data linked to supply or/and demand in the studied geographical region, a multiscale visualization, a cluster detection and a market positioning according to a specific firm's strategy. An example of implementation will show how price and other marketing-mix variables can be easily determined according to areas presenting homogeneous features in terms of supply.

Spatial segmentation methods used in marketing studies

The segmentation of demand in homogeneous consumers' areas derived from the commonly named geodemographic segmentation is based on the facts that spatially closed people tend to have similar social features. And reversely, shoppers in a neighborhood have similar preferences and needs [Mitchell, 1983; Hinesty, 2012]. Geographic segmentation describes where people are and demographic segmentation specifies who people are [Kirdar, 1997]. Geo-demographic segmentation doesn't classify only according to geographics and demographics but sometimes also includes corporate socio-economic data [Ryan, 1991] and lifestyles [Powers, 1990].

On the other hand, in marketing studies, the segmentation of supply enables to outline areas with analogous product features in terms of price ranges, quality, product specificities, distribution means. The easiest way of segmentation simply uses basically national (countries), regional, environmental (i.e. warm versus cold climates) or density (i.e. urban, suburban, or rural) classifications. But more elaborated ways of geographical classifications have emerged. According to geo-segmentation comparisons, no particular algorithm shows any superiority and no proof can give a clue of its accuracy [Grekousis, Hatzichristos, 2012]. Nevertheless, data input quality will condition the quality and precision of the output data and of the spatial clustering. Using huge, artificial and illogical input territories to aggregate in homogeneous spatial segments will lead to a poor and unreliable classification results. Spatial clustering is an important issue in geomarketing and generally speaking in geodemographics as many strategic problematic are depending on an optimized geographic segmentation i.e.:

- the allocation of concession areas to outlets in the automobile, banking, real estate or banking sectors;
- the identification of nearly homogenous areas to perform adapted marketing policies;
- a pre-processing of data based on consumers' demand to define areas' centroïds as demand nodes in a location models [Baray, 2003].

Before proposing a new method, we first outline practical methods in use. We then proceed to describe the current state of the art by presenting, first, the work of various authors and point out the limits of such methods in processing geomarketing data.

State of the art in geographical clustering methods

Among the most famous ways of classifying geographical territories are the *k*-means used in most commercial software, and for bigger database, artificial neural networks, genetic algorithms, and fuzzy logic algorithms. But generally speaking, clustering methods can be divided into partitioning methods, hierarchical methods, density based methods and grid based methods [Benassi, Bocci, Petrucci, 2011].

Among partitioning and unsupervised hard clustering, the *k*-means [MacQueen, 1967] consists in choosing *k*-points or initial centroïds into the geographical space including the objects to clusters or classes. These points represent initial group centroids. Each object is assigned to its closest centroïd according a distance metric. Then the position of each centroïd in every cluster is recalculated. The assignment of objects and recalculation of centroïd locations is repeated until no centroïd longer moves. The *k*-medoïds method can use other metric distances than Euclidean for example the Manhattan or the Minkowski distance and tries to minimize the sum of distances between the points and their closed center [Kaufman, Rousseeuw, 1987]. The representative center of each cluster is an actual point. The method is analog to the *p*-median used to optimize outlets and other facilities locations. It is more robust to outlier values than the *k*-means. These hard clustering methods based on classical set theory allow each point to belong only to one group. In fuzzy clustering, every point has a degree of belonging to clusters. The fuzzy *c*-means, widely used in geo-demographic studies, is identical to the *k*-means but with a probability for each point to be part of any of the *c*-segments. This method has been for example applied to delineate urban housing submarket [Hwang, Thill, 2007] with a hedonic analysis showing that houses prices are linked to socioeconomic characteristics of residents, structural characteristics of housing units, and location characteristics of neighborhood. The other method currently applied to geo-demographic problems is the Gustafson-Kessel algorithm [Grekousis, Hatzichristos, 2012] which extends the fuzzy *c*-means using adaptive distance norm, to detect clusters of various geometrical shapes. Whereas the *k*-means and the fuzzy *c*-means takes the hypothesis that clusters have spherical shapes, the Gustafson-Kessel system can detect elliptical regions. The method has been performed in geomarketing research to spatially segment the metropolitan area of Athens taking into account 130 demographic, lifestyle, and economy variables [Ibid.]. Ge-

netic clustering improves the *k*-medoïds algorithm by providing better initial representative centers [Brunsdon, 2006; Fernández et al., 2005].

Whereas partitioning methods gives a single result according to a pre-defined number of classes, hierarchical methods are creating a tree or dendogram with a hierarchy of clusters (agglomerative clustering) either by grouping clusters little by little into larger ones or by dividing bigger clusters in smaller ones (divisive clustering). Hierarchical clustering can't reach usually the performance of *k*-means concerning its objective function but is sometimes used to build an initial set of centroïds later improved by the *k*-means or fuzzy clustering. Hierarchical methods are rarely used in geodemographics.

Recent density clustering methods have been developed to determine other cluster shapes than spherical. Thus, spaces of high densities separated by boundaries of low densities are detected. The method usually determines the density of each point within a radius and states that this region is dense when it counts a preset number of points in the surroundings. Thus, the quality of this clustering depends on the good choice of the 2 parameters (the radius and the minimum number of points). The first density based method was the DBSCAN algorithm (Ester et al., 1996) later improved by the OPTICS algorithm [Ankerst et al., 1999] building first a reachability-plot to facilitate the 2 parameters choice.

Grid based clustering models the multidimensional space into a finite number of cells and a grid structure. The processing time is then improved depending on the number of points and variables. The grid structure clustering uses either statistical analysis of the information stored in cells i.e. STING detecting high density regions [Wang et al., 1997], a density based method i.e. CLIQUE examining the densities in sub-dimensional spaces [Agrawal et al., 2005] or a wavelet transform i.e. WaveCluster [Sheikholeslami, 2000]. In the field of geomarketing data, we can also mention the use of morphological functions to delineate trading areas from an ultra fine resolution or pixels grid [Baray, Cliquet, 2007] or of convolution functions after a filtering or smooth processing [Ibid., 2004]. In continuation of this research, the grid and morphological method have been used to detect urban areas measuring the road densities in each grid cell [Yuan et al., 2012]. More and more public national and international offices provide or use demographic with a grid format i.e. INSEE (France), Denmark Statistic, OECD & European Commission. Ready to use geographical segmentations based on sociostyles and usually built with *k*-means or fuzzy clusterings are proposed worldwide

by public or private organisms like Claritas Prizm (US), Tapestry (US), CAMEO (UK), ACORN (UK) and MOSAIC (UK).

Limits of existing methods

The *k*-means is the most classical method used and has some drawbacks. The unstable segmentation result depends on the value of *k* as well and of the initial values of the k initial centroïds choosen usually at random. Unable to detect concave shapes and clusters of extreme sizes, the algorithm also needs an important processing time as well as the *k*-medoïds or partitioning around medoïds (PAM). Density based methods are not efficient for a high number of variables and can hardly be used for geo-data mining. They also rely on the initial choice of the 2 parameters of radius and minimum neighbourhood points. Contiguity models used in the work of Hofstede, Wedel and Steenkamp [2002] are discrete: consumers are located in territories (regions, Länder, ...) that are aggregated according to their similarity toward perceptual attributes (quality feeling, distance, ...). Besides some inaccuracies regarding the allocation of consumers (territory centroids, homes, shops) and the type of distance used (euclidean, road or time distances) as well as a weak sample (only a couple of consumers by region), this method does not allow to overcome the artificial boundaries of predetermined areas. The accuracy of the spatial segmentation then is very low (regional boundaries) and needs a large computation time.

Another limit of these approaches is that variables have various formats i.e. age, earnings, education levels and usually pre-processing of data only includes mean centering of variables which does not reduce multicollinearity [Gatignon, Vosgerau, 2005]. Little research has been done on data formats suitable for spatial segmentation. One can mention the work of Ding and He [2004] about principal component analysis for unsupervised reduction of dimensions before a *k*-means clustering in the case of medical data and gene types related to human lymphoma. It is clear that the PCA treatment makes it easier to detect patterns in an optimized hyperspace of a reduced variables number.

A problem specific to geomarketing and geodemographic data is that areas of customer profiles are usually blurred: spatial segments do not usually have accurate boundaries and regions rather interpenetrate and show mixed consumer profiles on shared strips of border lands. The same statement can be made for product or services features related to

offer and a concurrence geomarketing analysis. Classical fuzzy clustering gives a probability of each point to belong to a cluster but doesn't clearly specify the cluster cores with highly defined profiles and on the other side the mixed-profile border geographical areas. In traditional clustering, elements should be mutually exclusive and collectively exhaustive; a point or an area can only belong to one cluster set and every zone must be taken into account. Moreover, spatial clustering can lead to results where contiguity of areas belonging to a same cluster is not guaranteed. A cluster can sometimes gather remote or isolated points or areas. The spatial segmentation quality and logic is visually appreciated through a mapping created with an geographical information system. The clustering algorithm used in geodemographics only range in descriptive methods but marketing managers need automatic systems offering expert recommendations to ease decision making. The present challenge in spatial segmentation is at last to integrate large and possibly disparate data sources and to minimize the computational time especially in case of real-time clustering.

Methods used in one-step geomarketing mix optimization

Location and price problems should be decided as far as possible as a whole when a company wishes to establish on a geographic market where competitors are already present and active. The main final objective of a company is indeed to maximize its revenue. Competitive location models including pricing issue are rare. Plastria [2009] reported these unusual systems and proposed a location-allocation model including the price parameter. Hotelling [1929] was the first author to describe a linear competitive market with the equilibrium conditions in the principle of minimum differentiation. His model was later criticized by D'aspremont, Gabszewicz and Thisse [1979] and a large number of models based on spatial economy and industrial organization have emerged in the field of geography and operational research [Serra, Revelle, 1995]. Game theory was then used by taking into account competition in a location and distributed quantities problem [Hakimi, 1986; Wendell, McKelvey, 1981]. By extending the Hotelling model, it was shown that the existence of equilibrium depends on the distribution of demand in the geographical area when both the price and location are decision variables [Eiselt, 1993].

In practice, there are two main approaches when a company chooses its retail locations and price. Some authors take the case of a simultaneous choice of prices and locations. On the other hand, other models adopt the Hotelling formulation with the concept of Nash equilibrium: location and price are then determined in two successive steps with the goal of maximizing profit. The choice of location precedes indeed most often policy decisions price. This approach is called perfect Nash equilibrium in subgame. Companies reflect on this principle of the selected locations to determine their level of prices. One could argue that the reverse is also true and that some companies choose their locations in areas corresponding to their brand and therefore a political price determined in advance e.g. the establishment of luxury products shops in gentrified neighborhoods.

A new model of location — general allowance, the PMaxCap was then designed [Serra, Revelle, 1999]. This model is an extension of the model of maximum coverage with pricing variable in a competitive environment. Consumers are represented by points in *i*. Assume that all overhead costs of each POS are equivalent. The assumption of this model is that consumer *i* will patronize an outlet, if the total cost including the cost of travel and the cost of the product is the lowest in the market. It should be noted that this model does not take into account the elasticity of demand. The PMaxCap is to date the only attempt to build a practical model to optimize the price, with the spatial variable. But this basic theoretical model can hardly be implemented on practical cases as it does not take into account competition and only includes distance as the optimization parameter ignoring the other marketing mix components.

An approach based on the automatic detection of strategic opportunities

The data to be taken into consideration in order to optimise store location and the selling price as a whole are numerous and of various types i.e. costs of production, price of other items in the range, strategy regarding the desired pricing policy, prices set by competitors, elasticity of the demand in terms of price, gap between the actual price and the price perceived by consumers, forecasting data on future demand. On the other hand, although price cannot be set without considering the other marketing mix elements and positioning in terms of the competitive advantages of the company, data on geographical supply, demand

and environment is only partially available. Against this background, we propose a new method derived from using the latest data mining tools. As illustration will show how optimization is possible where the only available data is the spatial breakdown of supply and its features.

A spatial model for setting optimal marketing parameters

To be optimal, the selling price and other parameters like the stores' location or product features must take account of the competition and assume a suitable sales strategy. Within this framework, the new method with the four following stages will detect strategic price ranges after building what could be named a *spatialized factorial map*. The *first stage* of this method is to collect all possible data about the company's products and its competitors. This quantitative and qualitative data might concern the various physical characteristics of the products as well as marketing attributes such as price or distribution modes. It should also be included the spatial locations of the products or its stores e.g. latitude and longitude. The *second stage* consists of obtaining new independent variables through a factorial analysis (ACP or AFC) to obtain a spatial representation of the products on these factorial axes: the geographical coordinates will be considered at the same level than the other quantitative variables. Thus, the factorial hyperspace will tend to reveal regions corresponding to closed products in terms of geographical proximity but also of product features. In traditional studies, factorial analysis is performed only on the product specifications before implementing a spatial analysis or mapping the results. The demand with all consumers' features could be processed the same way: the factorial analysis would then represent extended trading areas in a multidimensional space gathering closed consumers in terms of vicinity and demographic characteristics. To be clear, this new and original factorial chart is a deformed geographical map as it includes latitude and longitude mixed to the offer variables in linear combinations to build the factorial axis. In the *third stage*, factorial analysis high densities areas are outlined: different techniques can be used such as hierarchical classification, clustering tools or morphological analysis. The purpose of this step is to determine the offer in each identified cluster in characteristics of location and marketing mix parameters. Morphological analysis is a mathematical technique used in the automatic detection of contours and the analysis of shapes. It has been successfully performed in the precise delimitation of trading areas

[Baray, Cliquet, 2007]. Adapted to large data amounts, it includes the two basic transformations, dilation and erosion. The *final stage* consists in taking account of a particular firm's strategy i.e. concurrence predation or avoidance. Competition avoidance means to rather target niche markets and in this case, it corresponds on the factorial/spatial chart to the lacunar and intercluster zones that are underexploited and show a smaller representation concerning offer. These zones revealed by morphological analysis can also be described in terms of the ideal product characteristics, such as price which gives the company its competitive advantage. If available, spatial demand level data can be plot on the same chart as illustrative variables and a same process applied to detect high demand level zones: good business opportunities will then meet low supply densities zones with high demand densities zones if a niche market strategy is decided.

An empirical study on the french second-hand car market

Let us consider a database describing second-hand cars for sale on the web in France by dealers (first stage of the method — 2114 vehicles for sale from the year 2008 on the website <www.autoscout24.fr>). The available variables for each car are: brand and model of the vehicle, number of kilometres on milestone, horsepower, month of first registration, price, and dealer geographical location (the zip code has been geocoded to pick up the geographical coordinates of the relevant cities in grs80 format, the traditional format of GPS systems).

A principal component analysis (PCA) of the continuous variables logically shows a close correlation between price and power: the negative values of the first factorial axis correspond to high horsepower and high prices (second stage of the method). The uniqueness of this PCA is to have integrated spatial coordinates X and Y to form geographical clusters defining a product that is close in terms of product characteristics. A hierarchical classification in ascending order, taking the six first PCA factors as input variables, shows that the type of vehicles sold is quite different from one region to another. Three main regions can be identified nationally: the North/North West area in which vehicles often have low horsepower, low price and are foreign made (Nissan, Fiat, Lancia); the South-West/Centre area comprising more French (Renault, Citroën) and Spanish (Seat) produced cars, and the East area with higher horsepower and higher priced German or Peugeot cars. Figure 1 shows

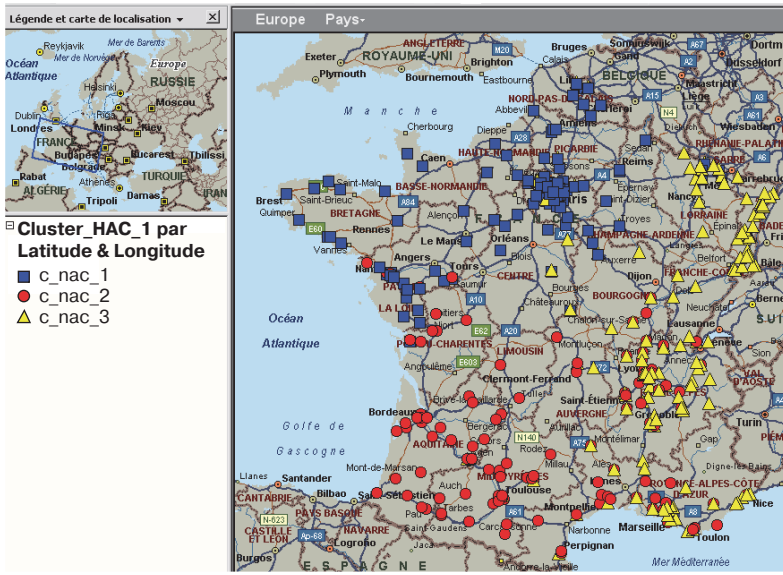


Fig. 1. Representation of the three clusters that characterise the types of vehicles sold

the spatial clusters with fuzzy boundaries. On the deformed geographical factorial map, longitude would correspond rather to the negative values of first factorial axis and latitude to the negative values of the second factorial axis.

Let us examine the vehicles' characteristics distribution in the factorial space and on its main representation with factorial axis 1 and 2. One notices lacunae (areas on the factorial chart where the density and offer is low or void). By using this method, we can therefore detect those lacunae which could represent an opportunity in sales terms, and identify their specific characteristics of price, vehicle features, etc. A morphological transformation (dilation) was applied to the factorial representation of the individuals following the factorial axes 1 and 2. The effect of this transformation is to reveal by a smoothing and juncture process of the closest points, the areas appearing in black where supply is dense (third stage of the method). The negative picture of this representation, consequently, shows the zones where supply is weak. These zones have been identified, automatically, numbered (18 zones), and located by the

factorial coordinates of their centroïds corresponding to specific cars' features. Zone or lacuna 17 shows, for example, that there is a notable lack of supply for vehicles more than 4 months old, with approximately 72 CV steam horses (fiscal power), priced at around 10599 euros. A seller might therefore operate a differentiation strategy by positioning itself on that particular segment. If one takes the view that factors such as the horsepower or age of the vehicles do not amount to a competitive lever within the company's control, all that needs to be done, therefore, is to neutralise this parameter in the factorial analysis and to consider it as an illustrative factor. By contrast, one could consider other factors easier to handle for the firm. Let us imagine that the store has not yet been created: it is then possible to integrate the geographical coordinates X and Y in the PCA as we did before, and the detection of gaps in supply will give indications where the outlet should be located, as well as the characteristics of the vehicles that should be sold in areas that are free from competition.

This analysis examines only product supply, but it would also be a simple matter to integrate the spatial demand level. All that is required is to add additional constraints to the present target zones in the factorial space corresponding to the product characteristics desired by the potential customers. For example, a market survey might show that consumers have a preference for more economical cars of less than 100 steam horses: the zones to target would therefore be limited to zones 5 and 17 with prices of around 16,774 and 10,599 € respectively. This method to optimize price, location and other marketing mix parameters in the geographical space has a number of noteworthy advantages:

- it takes into account all the measurable characteristics of the competing products and can be applied to an existing product or service, or to a product or service still to be launched. This method will advocate all features to build a strategic and global marketing plan, including the spatial price variations and the different locations to target in case of a retail network;
- the 4P can thus be described in their entirety using this method. Each zone in the factorial space corresponds to different 4P characteristics provided that this marketing data on spatial supply and demand are available and included in the factorial analysis;
- the selection of price range/characteristic zones for products and for the competition can be ongoing and dynamic: the databases must in that case be regularly updated;

- the prices and characteristics of products can be planned over time with a defined marketing strategy: in this case a path or route must be traced in the factorial space in order to plan for intended price changes;
- specific product ranges can be chosen: each individual product will correspond to a given zone in the factorial space;
- the method can incorporate and process very large databases taking account of supply and demand in space-time (morphological analysis is rapidly processed). In this last case, a filter (constraints) limits the acceptable zones in the factorial space;
- the method offers a geographical fuzzy clustering based on either homogenous groups of customers, or on regions of similar competitive products.

Limitations and prospects

The method presented makes possible, by exact positioning, to give a product or service a determined place in relation to competition through data mining techniques. While hierarchical ordering techniques can only produce groups of individuals, morphological analysis when applied to a factorial map allows both zones occupied by competing products and empty zones to be identified. Used as part of the optimization based on competitive advantage, a spatial positioning strategy thus offers the possibility of targeting zones where the competitive pressure is weak, and shows the marketing parameters to be taken into account, such as price, product characteristics and the best retail areas to target. The example given above shows a factorial map limited to two axes, but the analysis could equally well be done in a space with n factorial axes, and, likewise, it could incorporate qualitative variables if a factorial correspondence analysis (FCA) were used instead of the PCA.

This method lends itself both to niche market strategies and, at the other end of the scale, more aggressive predatory pricing strategies that target zones in the factorial space already occupied by concurrence. The question is how competition would react when confronted with this new positioning. It would be advisable, therefore, to refer to game theory and the laws of probability to study which positioning would be the safest and the least risky over time, and this could be the subject of a future research.

Conclusion

The method set out in this article spatially optimizes the price of a product or service as well as other interdependent elements in the marketing mix by taking account of supply and demand as well as a defence or an aggressive marketing strategy. It opens up a new way of exploiting large amounts of data to define the exact frontiers between certain ranges of products that either exist already or have yet to be designed. The development of the Internet makes it easier to gather abundant data on supply or demand in order to provide regular or almost instantaneous recommendations and marketing mix adaptation. The application fields encompass mass retail, producing companies and the services sector (tourism, real estate agencies, retail banking, insurance and so on). From a different angle, the same procedure can be used for monitoring information and for the continuing analysis of competitors' strategies by following the paths of their 4*P* choices within the factorial and geographical space. Further research could lead not only to improvements in the method by incorporating complex strategies, but also a better understanding of the interactions between supply and demand.

References

Agrawal R., Gehrke J., Gunopulos D., Raghavan P. Automatic Subspace Clustering of High Dimensional Data. Data Mining and Knowledge Discovery. Springer Netherlands, 2005. No. 11 (1). P. 5–33.

Ankerst M., Breunig M. M., Kriegel H.-P., Sander J. OPTICS: Ordering Points To Identify the Clustering Structure. ACM SIGMOD International Conference on Management of Data. ACM Press., 1999. P. 49–60.

Baray J., Cliquet G. Delineating Store Trade Areas through Morphological Analysis // European Journal of Operational Research. 2007. No. 182 (2). P. 886–898.

Baray J., Cliquet G. Delineating Store Trade Area through Filtering and Convolution // International Journal of Quantitative Management. 2004. No. 10 (1). P. 1–14.

Baray J. Optimisation de la localisation commerciale : une application du traitement du signal et du modèle p-médian // Recherche et Applications en Marketing. 2003. No. 18 (3). P. 31–44.

Benassi F., Bocci C., Petrucci, A. Spatial Datamining for Clustering: From the Literature Review to an Application Using Red Cap. Working

Paper. Dipartimento Di Statistica. Italy: Università degli Studi di Firenze, 2011.

Brunsdon C. A Cluster Based Approach to the Zoning Problem Using and Extended Genetic Algorithm. Proceedings of the GIS Research UK 14th Annual Conference. UK: The University of Nottingham, 2006. April 5–7.

D'aspremont C., Gabszewicz J.J., Thisse J.F. On Hotelling's Stability in Competition // Econometrica. 1979. No. 47. P. 1145–1150.

Ding C., He X. K-means Clustering via Principal Component Analysis. Proceeding. ICML 04 Proceedings of the twenty-first international conference on Machine learning. N.Y., 2004.

Eiselt H.A., Laporte G. The Existence of Equilibria in the 3-facility Hotelling Model in a Tree // Transportation Science. 1993. No. 27 (1). P. 39–43.

Fernández V., García Martínez R., González R., Rodriguez L. Genetic Algorithms Applied to Clustering. Buenos Aires: School of Engineering; University of Buenos Aires, 2005.

Gatignon H., Vosgerau J. Moderating Effects: The Myth of Mean Centering. Working Paper. Insead. 2005.

Gauri D.K., Sudhir K., Talukdar D. The Temporal and Spatial Dimensions of Price Search: Insights from Matching Household Survey and Purchase Data // Journal of Marketing Research. 2008. No. 45 (2). P. 226–240.

Grekousis G., Hatzichristos T. Fuzzy Clustering Analysis in Geomarketing Research // Environment and Planning B: Planning and Design. 2013. No. 40 (1). P. 95–116.

Grekousis G., Hatzichristos T. Comparison of Two Fuzzy Algorithms in Geodemographic Segmentation Analysis: The Fuzzy C-Means and Gustafson–Kessel Methods // Applied Geography. 2012. No. 34. P. 125–136.

Hakimi S. P-median Theorems for Competitive Locations // Annals of Operations Research. 1986. No. 6 (4). P. 75–98.

Hinesty M.D. An Alternative Segmentation Approach: Anticipating Consumer Preferences Based on Early Cultural Experiences. DBA Thesis // Grenoble Ecole de Management. 2012.

Hofstede F.T., Wedel M., Steenkamp J-B.E.M. Identifying Spatial Segments in International Markets // Marketing Science. 2002. No. 21 (2). P. 160–177.

Hotelling H. Stability in Competition // Economic Journal. 1929. No. 39. P. 41–57.

Hwang S., Thill J.C. Using Fuzzy Clustering Methods for Delineating Urban Housing Submarkets. Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems. N.Y., 2007.

Kaufman L., Rousseeuw P.J. Clustering by Means of Medoids / Statistical Data Analysis Based on the L1-Norm and Related Methods / ed. by Y. Dodge. North-Holland, 1987. P. 405–416.

Kirdar I.Ü. Tourism Market Segmentation for National Tourism Organisations and Its Practical Importance for National Tourism Offices Abroad. University of Surrey. PhD in Tourism Marketing. School of Management Studies for the Service Sector. 1997.

Labbé M., Hakimi S. Market and Location Equilibrium for Two Competitors // Operations Research. 1991. Vol. 39 (5). P. 749–756.

MacQueen J.B. Some Methods for Classification and Analysis of Multivariate Observation. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967. No. 1. P. 281–297.

Mitchell A. The Nine American Life Styles. N.Y.: Warner, 1983.

Plastria F. Vanhaverbeke L. Maximal Covering Location Problem with Price Decision for Revenue Maximization in a Competitive Environment // OR Spectrum. 2009. No. 31 (3). P. 555–571.

Powers T. Marketing Hospitality. N.Y.: John Willey & Sons, 1990. P. 47–83.

Ryan C. Recreational Tourism: A Social Science Perspective. L.: Routledge, 1991. P. 167–204.

Sander J., Ester M., Kriegel H.-P., Xu X. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications // Data Mining and Knowledge Discovery. Berlin: Springer-Verlag, 1998. No. 2 (2). P. 169–194.

Serra D., Revelle C. Competitive Location in Discrete Space. Facility location // Springer. 1995. P. 337–356.

Serra D., Revelle C. Competitive Location and Pricing on Networks // Geographical Analysis. 1999. No. 31. P. 109–129.

Sheikholeslami G., Chatterjee S., Zhang A. Wavecluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases // The International Journal on Very Large Data Bases. 2000. No. 8 (3–4). P. 289–304.

Wang W., Yang J., Muntz R. STING: A Statistical Information Grid Approach to Spatial Data Mining. Proceedings of the 23rd International Conference on Very Large Data Bases. 1997. P. 186–195.

Wendell R., McKelvey R. New Perspectives in Competitive Location Theory // European Journal of Operations Research. 1981. No. 6 (2). P. 174–182.

Yuan N.J., Zheng Y. Segmentation of Urban Areas Using Road Networks. Microsoft Research Technical Report MSR-TR-2012-65. 2012.

BIG DATA PASSIVE ACOUSTIC FOR BAIKAL LAKE (ОЗЕРО БАЙКАЛ) SOUNDSCAPE & ECOSYSTEM OBSERVATORY

H. Glotin (Prof.)
M. Poupard (PhD st.)
R. Marxer (Assic. Prof.)
M. Ferrari (PhD st.)
J. Ricard (IGR)
V. Roger (PhD st.)
J. Patris (Assic. Prof.)
F. Malige (Dr.)
P. Giraudet (Assic. Prof.)
J.-M. Prevot (IGR)

EADM group, UMR CNRS LSIS — Université de Toulon;
Aix-Marseille Université, France

M. Komarov

National Research University Higher School of Economics,
Moscow, Russia

Abstract. *Lake Baikal, the world's deepest freshwater lake, $^1/_5$ of the world's fresh water, is the "Russia's Galápagos": much of the fauna here is found nowhere else on Earth. The Baikal lake ecosystem is mainly based on three endemic species: a Baikal seal (Pusa sibirica), a fish (Comephorus baikalensis and Comephorus dybowskii) and a copepod (Epischura baikalensis). These species are isolated and vulnerable or endangered, mostly due to noise, chemical pollution, global warming. The Baikal seal is on of the main superpredators of the lake and plays a crucial role in the balance of the ecosystem. Like any mammal, seal need to browse and adapt to its environment, and exhibit a large diversity of movement and diving times. Environmental factors, such as anthropogenic sounds and prey availability, play a significant role in such behaviour. One of their main threats is water pollution, and anthropogenic noise is a threat gaining importance as the human activity increases in the area.*

B2O is proposed to be the first long term bioacoustic observatory of the Baikal lake. The project will bring knowledge on seal presence, seasonality, habitat use, and foraging behaviour and success. The project will deliver a continuous passive acoustic monitoring system, seal tracking data, knowledge about seal swimming behaviour in response to changes of their ecosystem.

B2O focuses on passive acoustics big data monitoring to study multiple aspects of the Baikal lake system, such as animals vocal/biosonar behaviour, velocity

and trajectory, abundance data, and vessel attendance and movements. Three or two dimensional tracking of the fauna could be produced from multiple channel hydrophone recordings. These tracks will provide insight about seal movements and behaviour according to the context. The revealed tracks, headings and speed, will aid to understand the population and predation of the seal. Furthermore, our passive acoustic tracking system captures social behaviour cues, providing information about groups presence that can be valuable in planning protection of the superpredator of the Baikal lake, thus its ecological balance.

The analysis of seals vocal cues may also contribute to the knowledge of population structure and trend, by the use of size/age/sex frequency or temporal structures correlates of individual vocalisations. In addition, the study of their interaction with surface ships may reveal how the animals react to surface traffic and modify behaviours that are primarily addressed to maintaining their energy intake.

Anthropophony may dramatically reduce the acoustic communication and sensing space of animals, reducing their ability to acoustically detect incoming ships. Recent studies show that 50% of the aquatic noise is generated by 15% of ships and that in some coastal and other high-traffic areas, ship noise has reached levels that degrade habitat for endangered species. An objective of **B2O** *is to assess these noise levels and correlate them with the behaviour of the animals. These results could directly feed into the development of new approaches to the management of the lake's environment.*

Moreover, the passive acoustic monitoring of the seal population could indirectly assess the impact of global warming on the behaviour and adaptation of the species. Water and environment characteristics could be jointly analysed with the obtained results to identify any effects of context change.

Overall B2O will provide a new view of the Baikal lake, by characterizing and revealing its underwater and airborne soundscape. B2O will produce new knowledge on the lake's habitat and data that will be of great use to scientific and economic actors of the region and around the world.

The main objectives of the B2O project are:

***1. Characterise the underwater soundscape of the Baikal lake: Acoustic detection of different species composing the food chain and background noise (ice, boat)** (2 years)*

1.1. *Installation of robust acoustic observatory and luxmeter (first weeks)*
1.2. *Recording long time series of Baikal soundscapes (seasonal variability)*
1.3. *Describe the different sounds and their correlation with luminance*
1.4. *Identify the sound sources: big data (8 TB/year) indexing*
1.5. *Investigate properties of sound propagation across seasons and sources*

***2. Determine the acoustic repertory of seal** (18 months)*

2.1. *Determine underwater vocalisations of seal (calls, burst, growl) / big data classification*
2.2. *Ethoacoustics of seal emission within pollution, vessels, season / big data clustering*
2.3. *Classify seal airborne/underwater vocalisations, pup/mum interaction*

3. Determine seal position and trajectories: big data tracking *(6 months)*

3.1. *Seal 3D tracking*

3.2. *Matching field model*

3.3. *Monitoring variability of acoustic propagation in Baikal with seals signal*

Keywords: *Big data, Soundscape analysis, Ethoacoustics, Baikal Seal.*

Baikal lake and main elements of its trophic chain

Lake Baikal, the world's oldest and deepest freshwater lake, is situated in Russia in north of the Mongolian border. Some refer to Baikal as the country's "sacred sea". To scientists, it's "Russia's Galápagos": much of the fauna here is found nowhere else on Earth. Lake Baikal is the largest freshwater lake in the world, containing 22–23% of the world's fresh water [Schwarzenbach et al., 2003]. At 636 km long by 79 km wide, Lake Baikal has the largest surface area of any freshwater lake in Asia, and is the deepest lake in the world at 1642 m.

Concerning the water temperature, 3 parameters may change it: location, depth of the lake and seasons. The entire lake surface is covered by ice during 4–5 months, from early January to May–June. On average, the ice thickness is 0.5 m to 1.4 m. Its surface temperature in August is between 10 and 12°C. Lake water is very clear (visibility = 40 m) due to high numbers of planktonic animals eating organic matters.

The pelagic food chain is simple, composed by phytoplankton (*Aulacoseira baicalensis*), mesozooplankton (*Epichura baicalensis*), macrozooplankton amphipod (*Macrohectopus*), fish (*Coregonus autumnalis migratorius* and four species of cottoid fishes), and finally Baikal seal (*Phoca sibirica*) [Yoshii et al., 1999]. More than half the species found in Lake Baikal are unique to this place (endemic species).

The diversity of Amphipoda in Baikal is really important (300 species with **95% of endemism**). These animals can measure until 90 mm. Actually, we don't know the acoustic emissions of theses species. The Baikal lake is a threatened ecosystem, principally because of global warming, epidemics (1988 a disease wipe out between 5000 and 10,000 seals), increasing of tourism or chemical pollution.

There are also some fisheries in the lake: 50 registered commercial fishing enterprises are enumerate in Baikal. Recreational fishing and ice-fishing are unrestricted in the lake [Brunello et al., 2006]. Different

fisheries management is now implemented in Baikal for maintain the regulation of fishing [Matveyev, Samusenok, 2015]. They study the impact of fishing on species but not the impact of noise of fishing.

ISU (Irkutsk State University) try to develop a composite index reflecting the health status of aquatic ecosystems in Baikal. *B2O* could be efficient to improve this index. Increasing vessel traffic, fishing boat, excursion boat on the lake, could changes the phytoplankton composition and upset all this vulnerable and unique aquatic food chain. *B2O* acoustic monitoring will enable to know more about theses species, and anticipate their damage, that could also directly results from noise pollution.

1. The Baikal seal (*Pusa sibirica*)

The Baikal seal (Pusa sibirica), is an endemic species to Lake Baikal. The Baikal seal is one of the smallest true seals and the only pinniped species that lives solely in freshwater [Randall et al., 2002]. Concerning its history, scientifics are not still sure of this provenance. But they suggere that they descent from the ringed seal Phoca hispida and that the species may have been isolated geographically for about 500,000 years (<http://www.pinnipeds.org>). Seal population is estimated at approximately 100,000 animals [Kutyrev et al., 2008; < http://www.iucnredlist.org>]. Adult seal measures 1.4 m in length with a body mass from 63 to 70 kg (Seal Conservation Society, 2007). The animal weight can vary during the year (lowest weight in summer) [Pastukhov, 2007]. There are a little sexual dimorphism (males are larger than female). So it's possible than acoustics emissions would be different between males and females. This hypothesis is also applicable for pup (up to 3.5 kg, 70 cm [Ibid., 1971]. For example, Thomas et al. [1982] have described the repertory of vocalization of Weddell seal (Leptonychotes weddelli). They found 34 different calls (based on differences in frequency range, duration, repetition rate, number per series, presence or absence of harmonics, auxiliary sound usage, and contour). They have stated a distinct sex-related differences in vocalization usage. Seals males can produced more vocalizations than females.

Concerning the distribution of seal, they are present in all lake, but it depend of the season and environmental factors. Globally seasonal movement of seal is driven by ice more than food [Ivanov, 1938]. So in october, seals move towards bays, lagoons and river deltas, mostly along

the eastern shore where ice forms and expands out into the lake. After this (January to May) they spread throughout the lake in deep water. Adults spends time in North of lake whereas immatures lives in south. Pups born in winter when baikal lake is frozen.

Since 2008, the Baikal seal has been classed as a Least Concern species on the IUCN Red List (IUCN Red List of Threatened Species). Principally due to chemical threats. In fact, introduce pollutants in ecosystem can have potential to impair fertility, diseases and immunocompetence [De Swart et al., 1995; Kannan et al., 2000]. But the most serious future threat to the survival of the seal may be global noise pollution (anthropic activities), in addition to the climate warming, which has the potential to affect a closed cold-water ecosystem such as that of Lake Baikal (Burkanov, 2016).

Fig. 1. The Baikal seal (*Pusa sibirica*)

Concerning reproduction, they are polygamous and territorial. Males mark the female's with a strong odor which can be smelled by another male if he approached. But it can be possible that males emit reproduction sound, to call females.

The main food source of seal is the golomyanka (cottoid oilfish, endemic to baikal lake). Particularly in winter and spring, 90% of its food is composed of *Golomyankas* [Pastukhov, 2007].

Diving patterns are not well documented but the mean of juveniles are between 10 and 50 m depth, and for adult a maximum of 300 m [Kutyrev et al., 2006]. During the day, dives are deepest, due to rise of their prey during the night. Times of depth are between 2 and 10 min, but the longest recorded dives exceeded 40 min during winter [Stewart et al., 1996].

Concerning acoustic communication of this species, we know anything. Based on some preliminary analyses we conducted on basic recordings available online [<http://www.bbc.co.uk/nature/life/Baikal_Seal>], our team supposes and will use these facts:

- seal can emit some burst for echolocation (Fig. 2, 3), at low frequency, for long range detections of about 300 m. This burst might be detectable at a range of 2 km;
- some growls (Fig. 4) and vocalizations that might includes individual signatures, at least the size the animal (shift of the central frequency peak, inter-pulse interval variation);
- different calls are emitted between males and females;
- the previous hypothesis shall be applicable for pup — mother calls airborne and underwater;
- it can be possible that males emit reproduction sound, to call females.

Some pre-analysis of sea calls are given below. Acoustic monitoring of baikal seal is a central key to know the ecology, movement and communication of this animal. If we have access to these informations, it would be possible to elaborate program to protect them at short and long term.

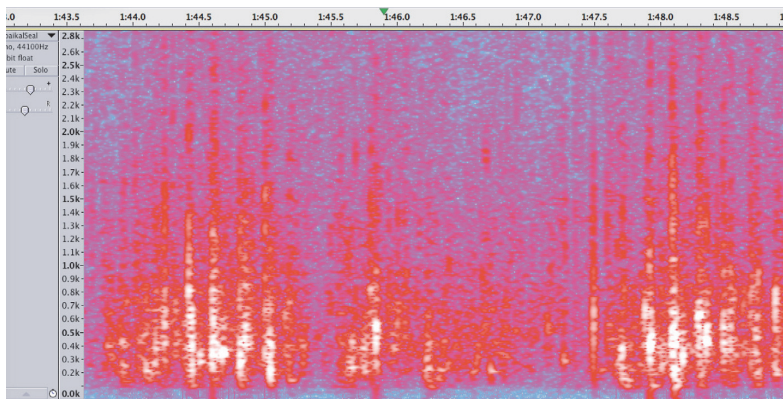


Fig. 2. Spectrogram (abscissa: seconde; ordinata: Hz) of a burst of a Baikal seal [BBC]

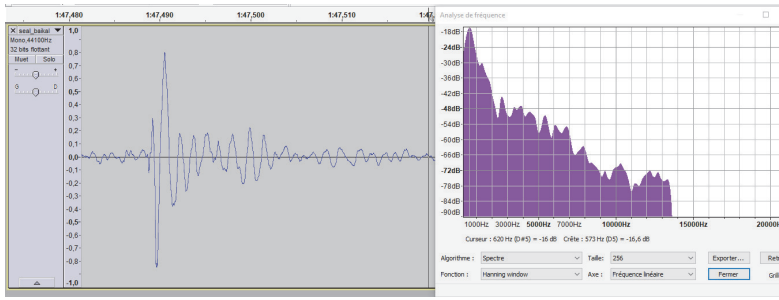


Fig. 3. Representation of a Burst of Baikal seal and its spectrum associate [<http://www.bbc.co.uk/nature/life/Baikal_Seal>]. Duration of the burst — 20 ms. A frequency peak is visible at 2500 Hz. The burst is composed of two main pulsations that could depend of the size of the animal

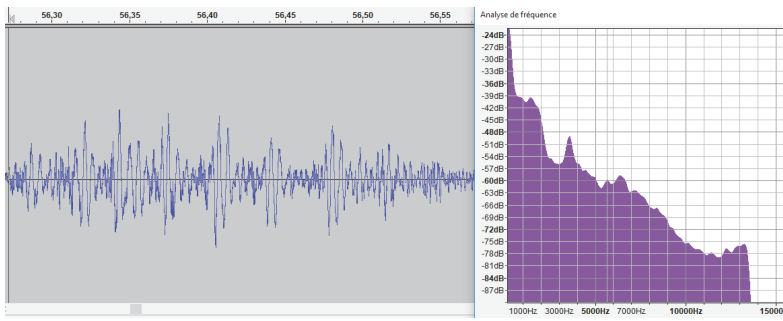


Fig. 4. Representation of growls of Baikal seal and its spectrum associate [BBC]. Duration of the growls 200 ms (ten times the duration of a burst Fig. 3). The frequency pick is around 3500 Hz

2. The main food of the Baikal seal

2.1. Endemic *Comephorus*

Baikal oilfish or golomyankas are pelagic and also endemic to Lake Baikal, is only genus in the family *Comephoridae*. Two species of golomyankas are present: *C. baikalensis* (21 cm) and *C. dybowskii* (16 cm), they have a translucent bodies [Froese et al., 2012]. They are composed by high lipid content, it for this reason that they are main food of seals. Golomyankas are present throughout the entire water column (from surface

until 1 km of deep), but they primarily occur deeper than 100 m [Ibid., 2016]. They move close to the surface (10–25 m) during the night to feed Burkanov, 2016]. Golomyanka is the most populous fish in Baikal with a biomass of 150,000 tons (70% fishes are *golomyankas*) [Froese et al., 2016]. We can see small group of fishes near the lake bottom (20 individuals). Concerning food, they feed the planktonic *Epischura baikalensis*, the amphipod crustacean *Macrohectopus branickii* and larvae of sculpins [Miyasaka et al., 2006]. ISU, particularly the Department of Vertebrate Zoology and Ecology already study the ecology of a number of endemic Baikal *Cottoidei* fish. We could correlate these abundance informations with the seal detections.

Fig. 5. Endemic *Comephorus*

We don't know if this species can produce sound. But they are so numerous and they represent the largest concentration of fish in lake. At least their presence will produce specific sound diffraction of seals calls, or noise from the boat, or variations in time and frequency of the wave pattern of the calls of the seal.

2.2. Endemic *Epischura baikalensis*

Pelagic zooplankton of Lake Baikal is composed of a small number of species. *Epischura baikalensis* are the dominant pelagic zooplankton species (85% of total biomass) [Penkova, 1997; Afanasyeva, 1998]. This third species is also endemic to Lake Baikal. *Epischura baikalensis* is a species of copepod in family *Temoridae*, and measure up to 2 mm. They migrate annually through the water column, because they need low

temperatures (5–10°C, dies above temperatures of 15°C). Biomass and abundance of this species vary seasonally and annually, with maximum abundance observed at the end of August and beginning of September. They play the main role in filtering and purifying Baikal's water [Reid, 1996].

Fig. 6. Endemic *Epischura baikalensis*

Again the acoustic quality of the water of the lake shall depend of the concentration and position of the plankton. We shall observe based on the quantity of measured light incoming in the lake, variation of the background noise, due to the variation of the concentration of the plankton in space and time, as we do observe in the ocean soundscape.

So, the aquatic food chain of Lake Baikal is based on theses four species. For this reason, the food web is very vulnerable, and if one element of the chain coming to disappear, all of lake ecosystem will be destroyed. Acoustic monitoring could give informations on the state of health and variability of this food chain.

Setting a permanent Passive Acoustic Observatory

The knowledge from this trophic chain is limited due to its seasonality and it is exclusively validated with visual observations. Observations are only performed during daytime, and they are too sparse to provide reliable conclusions. It would be interesting to add animal's velocity/trajectory, abundance data, from *B2O* systems. Assessing the

effects and the propagation of underwater noise on seals also requires large databases that characterise animals' behaviour, movements and associated noise levels.

We would particularly be interested in the deepest area of the lake (−1600 m, middle of the lake, Fig. 7) in order to detect acoustic seal emissions during deep dives. The red plan in Fig. 7 represents the possible monitored area using our three passive acoustic stations.

First, we wish to place one station in the middle of the west shore of the lake (East station, Fig. 7), at −10m, with 40 m of cable, simply connected to our aerial system with a battery car. The acoustic recorder has already been tested (it is produced by University of Toulon for this research purpose by SMIoT of [<http://smiot.univ-tln.fr/>]). This station would be in stereo (two hydrophones) in order to process source separation and estimation of the azimuthal movement of the seal.

Then we would complete these observations with another fixed stereo station on the West shore, at the other extremity of the transect in order to get a maximal view of this area of interest.

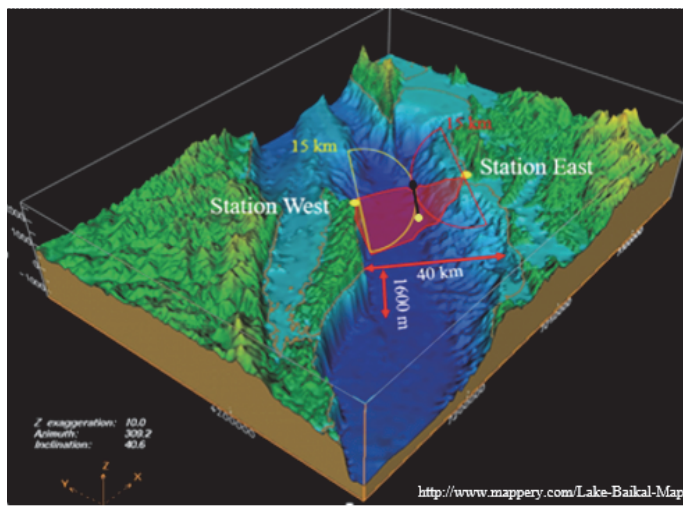


Fig. 7. *B2O* expected location. Yellow dots: positions on the shore of the West and East stations. Each station is a stereo recorder with hydrophones at −10m to avoid waves troubles. The station in the middle (black dash) will be deployed by cable at −40m, in summer from a boat at day time, and in winter from ice platform (one week 24/7)

Another hydrophone station could complete the observatory, at winter time at a depth of –40 m, simply on the Baikal ice in the middle of the transect. At summer time, only several days of recording at the same place would be possible using a small boat and hydrophone and recorder than at winter.

We would produce 3*D* seal tracks. Our autonomous passive acoustic observatories are designed to produce tracks on large area, possibly of hundreds of seals each day. The revealed tracks and headings, speed of the seal will help to describe their behaviour in interactions with their preys and anthropophony. As the system works on diving animals, unseen at the surface, it provides a completely novel information. We will also track surfacing social groups if they are socializing and producing social sounds. In this case we manage to protect whole groups rather than individuals. This research is primordial in trying to understand the threats such as noise pollution.

This new information on how animals move in the 3*D* space to exploit habitat resources and to develop their social behaviours will provide completely new insights into their behaviour, and provide valuable information for their management. The detection of seals also provides information on the population structure and trend based on the size/age/sex of the detected animals.

The acoustic sensing approach will also permit the development of new approaches to lake environment management. Recent studies show that 50% of the ocean noise is generated by 15% of ships [Veirs, 2017] and that in some coastal and other high-traffic areas, ship noise has reached levels that degrade habitat for endangered species marine wildlife [Hatch 2012; Erbe 2014; Viola 2017]. According to Veirs [2016], ship noise can also have an impact on marine mammals' ability to echolocate their preys. Because the work of Veirs demonstrates that high frequency components of ship noise interfere with echolocation, we may infer that it can also interfere with environment sensing and thus with ship detection.

Besides the primary tasks of the project, *B*2*O* will provide a new view of the Baikal soundscape, an area that is of primary importance for the whole region and for its endemic fauna. The analysis of the acoustic space will give a long term view of the combination of biological sources and of the anthropogenic noise sources, including far sources, such as pile driving activities taking place in a range of hundreds of km [Sciacca, 2016].

Methods

Objective 1: Characterise the underwater soundscape of Lake Baikal

1.1. Installation of robust acoustic & underwater/ice lux-meter stations

Our project is based on advanced scientific instrumentation to observe long term bioacoustic and light series. This instrumentation has been elaborated by our team at Toulon in the Information Numeric Prevention interdisciplinary Group and JASON project 2015–2016. It is now produced and sold at industrial quality for advanced research programs of environmental survey by the university Toulon technological platform SMIoT [<http://smiot.univ-tln.fr>].

The recording card is the JASON high velocity audio card (Fig. 8), used to monitor up to 5 channels at 2 mHz sampling rate each. It is equipped by Cetacean research hydrophones and a simple battery, pro-

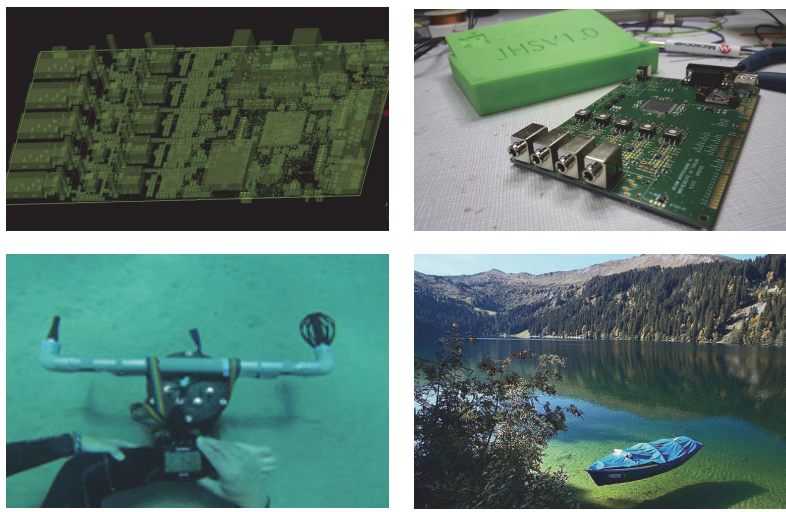


Fig. 8. Top: The JASON HYPERSOUND DAQ version 4 x 2 mHz x 16 bits x x low power = High Velocity + Long autonomy. Bottom: Example of our autonomous stereo acoustic station, here placed at –10 m on Med. coast, for a full week, equipped of JASON recorder. It is also developed with 4 channels (called Blue 4*D*). Bottom right: Example of shore of the Baikal where will be placed the Blue 4*D*

grammed to record N minutes each M minutes ($M = 0$ for continuous recordings), 24/7, during 2 or 3 weeks depending of the battery type and the storage (usually a simple HDD 2 To).

The second system is a luxmeter (Fig. 9), called qualilife-Light and developed by SMIoT to record variation of the light in the –40 m the Baikal, joint to the central station. The activity of seals and other organisms shall depend on this parameter, and we will then correlate bioacoustic activities with it.

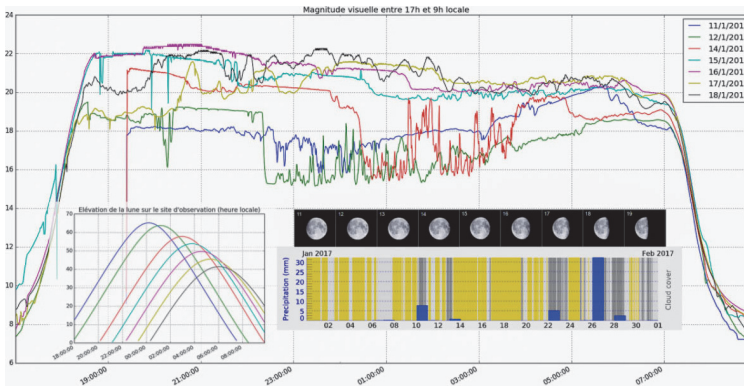


Fig. 9. Qualilife-Light: result of high dynamic monitoring of nocturnal moon light during one week, with different weather conditions

1.2. Recordings long time series of lake soundscapes

Our approach is based on long term bioacoustic monitoring correlated with light and boat traffic. Detection of the seals would be optimized for RT operation (as in Glotin et al. Boussole / DECAN / VAMOS Pelagos report, 2017). Three dimensional or two dimensional tracking of the ecosystem will be processed on the whole corpus when at least 4 hydrophones (quadratic solver [Glotin, 2009]) are recording.

So for a long period (one year) we could describe the different sounds emitted from the ecosystem, correlation with luminance.Then, we will identify the sound sources: big data (8 TB/year) indexing (6 months). Finally, we will investigate properties of sound propagation across seasons and sources (6 months).

Objective 2: Determine the acoustic repertory of seal

Once material was installed, and long records were effectuated, it will be possible to identified and classified sounds that we interest.

2.1. Determine underwater emissions of seal: big data classification

The first part of acoustic study of baikal seal is to classify underwater emissions of animals (calls, burst, growl), it's the **big data classification on the To recorded by *B2O*.**

Some studies has already done different repertories of seals emissions [Mirhaj et al., 2004]. They found different types of calls (Fig. 10).

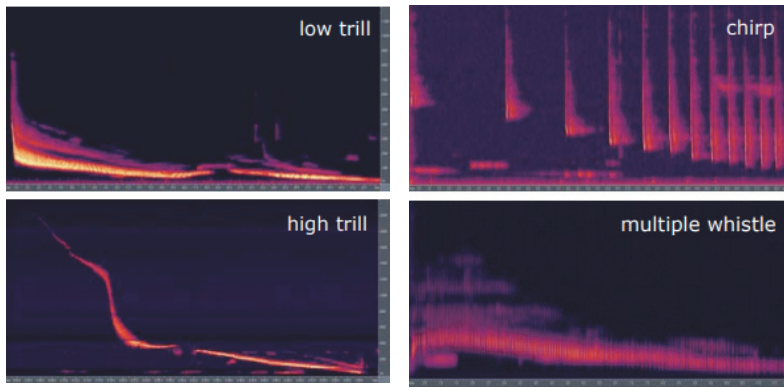


Fig. 10. Spectrograms of selected Weddell seal calls [Mirhaj et al., 2004]

Our objective is to built a repertory of baikal seal calls, and to compared theses calls with other species (Fig. 10). Abgrall et al. [2003] have proved regional difference of vocalizations for Weddell seal (*Leptonychotes weddellii*). They found differences in the mesogeographic range and temporal range. But seals were separate from over 600 km. Baikal lake measure 636 km, so it is possible that there would be differences in vocalizations patterns through the lake. So this repertories could depend on geographic positions.

Once these repertories will be established, showing the different vocalization, it will be possible to correlate some of these emissions with external data such as pollution, seasons or boat traffic as presented in next section.

2.2. Ethoacoustics: variations of seal emission according to external data

Poupard et al 2017 shown the feasibility to cluster acoustic patterns of dolphins/whales and the consistency of this pattern linked to their behavior (Fig. 11 and 12). In this figure each point represent a whistle/vocalization. Distance between points is really important: things being close together we can say that they are similar, and things being far apart we can confirm they are different.

On the Fig. 11 we can see two distinct clusters (0 and 1). Cluster 0 is composed only of Anthropogenic Pressure (AP). So there are variations of whistles structure, when behaviors changed, particularly with AP condition.

On Fig. 12 we can see the same representation but with vocalization in presence of different food stimuli in water column. This research

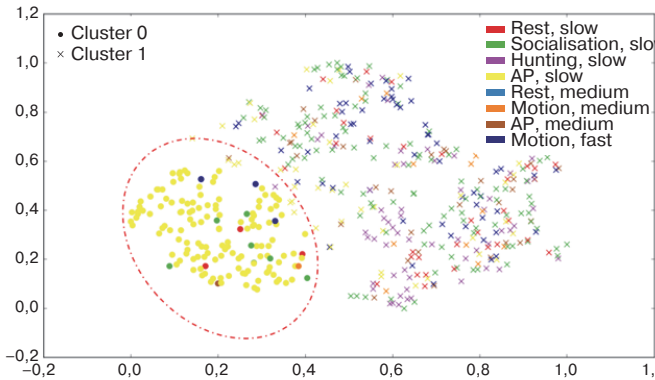


Fig. 11. Visualizing dolphin Sa whistles in 2-dimensions with *t*-SNE as a function of velocity and behaviors (with 4 important features), according to BNP clustering, *AP = Anthropogenic Pressure

use DMS (Dimethyl sulfide) and krill extracts as chemical stimuli. Attractiveness of these compounds have been compared to control solvent: Clay and CTL (control) respectively. The hypothesis is that, in presence of stimuli, whales have not the same communication compared to a neutral environment. We can see 2 different clusters, the first one is composed by vocalizations made in presence of Krill and DMS (chemical stimuli). So, acoustics parameters are differents when preys are present in field.

Our research should contribute to the evolution needed for the setting up of acoustic survey systems. These methods are therefore an excellent way for studying the dynamics of marine mammals populations, and to analyze impacts of humans on marine life. In fact, this technique is used for Mysticete (baleen whales) and odontocetes. We will apply our model to monitor the behavior, ecology and trajectories of Seal.

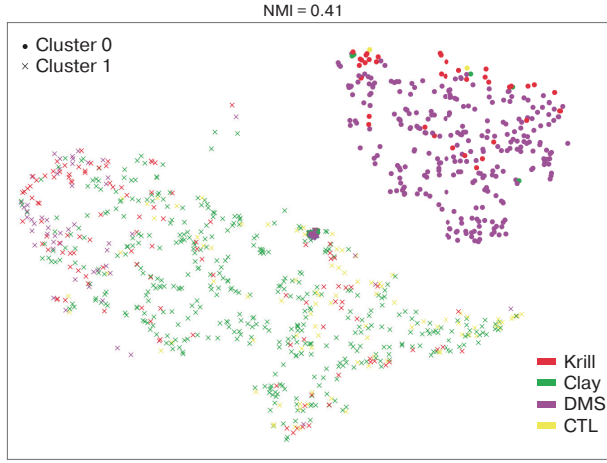


Fig. 12. Visualizing humpback whale vocalizations in 2-dimensions with *t*-SNE as a function of different stimuli (with all features), according to BNP clustering (2 clusters)

2.3. Classify seal vocalizations in airborne environment

All pinnipeds can communicate vocally and visually, underwater or in airborne environment. In fact, there might be different airborne vocalizations. Stirling and Warneke [1971] have studied the structure and number of airborne vocalizations of the Australian fur seals. Four types of calls were identified:

pup attraction call given by adult females;

- female-attraction call given by pups;
- high-intensity guttural threat given by adult males;
- repetitive barking call given by adult males.

They found also differences between two species. Somes studies proved individual signatures for pups vocalizations from grey seal (*Halichoerus grypus*) [McCulloch et al., 1999].

It will be interesting to study seal vocalizations in airborne environment particularly after births (end of winter, on the ice). During this period animals are more present in North of lake because the longer winter keeps the ice frozen longer (preferable for pupping) [Pastukhov, 2007]. So, we could install two stations in stereo (two hydrophones), with a camera. It will be possible to make correlation vocalizations and behaviours on the ice between mums and pups.

Objective 3: Determine seal positions/trajectories by big data

The third objective is to determine seal positions, in order to have new knowledge of the seal underwater behaviour in the lake. We could also explore correlation between seals positions with other characteristics of the lake such as temperatures, pollutions indices or preys quantities. The lake is a close environment so it is perfect setting for such study, given the possibility to install multiple passive acoustic systems with range overlap.

3.1. Recordings and 3*D* monitor presence of Seal

Seal are obviously present all the year. The 2 shore recording stations, and the central station (Fig. 7) will allow large scale and long term passive acoustics monitoring of the seals, and possibly their prey. *Copepods* would also be monitored by passive acoustics at a row scale (to be confirmed).

The two recording stations (East and West) fixed on the shore will be, in the final system, each equipped of 4 hydrophones (4*D*) to solve the localisation of the source. The central station in summer time will be stereo to allow simple source separation. At winter time, we propose a quadri-hydrophone central station, fixed to the ice platform to get a high definition view of the movements of the fauna directly observed from the middle of the lake.

The complete large scale *B*2*O* observatory, composed of the two observatories on West and East stations, plus the middle observatory on the ice or on a boat, offers a large and continuous observation of fauna in a volume of nearly 40 km × 4 km × 1.6 km (the last dimension is depth).

The computed tracks will give insights into the seal behaviour according to their context including vessel noise/AIS traffic.

DYNI has the know how to track online by passive acoustics on stereo station (Fig. 11, 12) and in 3*D* [Glotin et al., 2009; 2014], with already proof of concept on astrophysic small array of hydrophones [Benard, 2011], as shown in Fig. 13.

Fig. 13. BOMBYX with stereo antenna pointed to South to observe the megafauna in front of Toulon (Med. Sea) (credit Dyni)

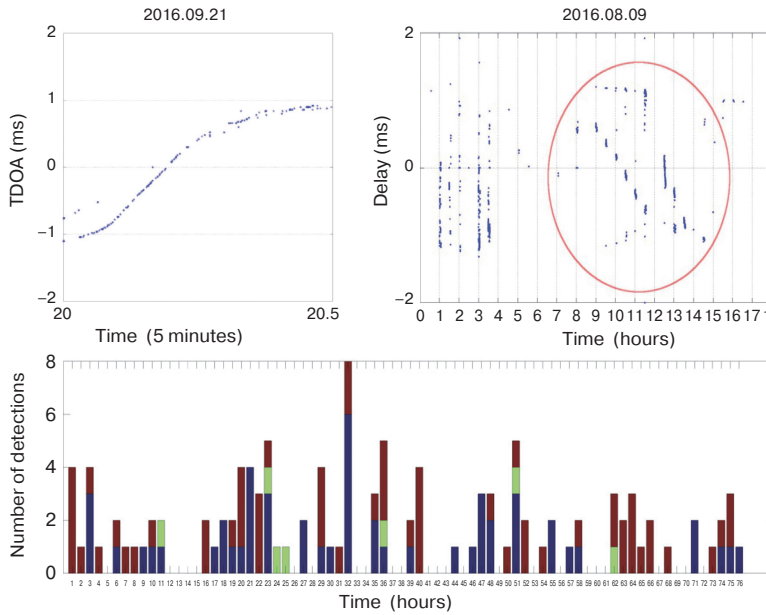


Fig. 14. Example of monitoring of Pm versus time from stereo BOMBYX. Top left: Time Delay of Arrival showing acoustic detections of Pm going from East to West in 5 mn nearby BOMBYX the 21/09/2016. Top right: Same, but long range Pm detections (inside circle), possibly 15 km, going from West to East in 7 hours the 6/08/2016. Bottom: Total Pm countings and directions in the 0–15 km range of BOMBYX (Red: from East to West; Blue: inverse; Green: unknown, on 76 days of summer 2016 [Glotin et al., 2016])

For the full 3*D* monitoring, we will need to complete the *B*2*O* array up to 4 hydrophones with JASON SMIoT [< http://smiot.univ-tln.fr>] high velocity recorder (up to 5 × 500 mHz Sampling Rate) to estimate the range, depth and azimuth of the seals swimming in a range of 1 to 3 km to a station.

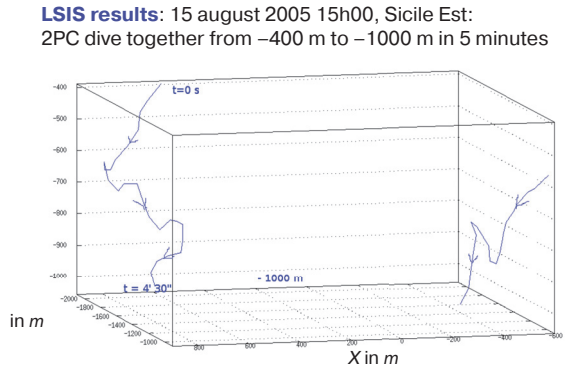


Fig. 15. 3*D* Passive acoustic tracking of 2 Sperm whales diving together, computed from the 2 m aperture 4 hydrophones array of NEMO astrophysical obs. (placed 2 km away from the whales) having similar geometry to MEUST (credit DYNI [Benard et al., 2011])

Moreover, we propose an airborne acoustic station in order to analyse the acoustic communication between pups and mums. This station will be in stereo, to allow separation of each source (males, females, pups).

A budget will have to be developed for a specific implementation of the proposal, however the LSIS DYNI team has availability for some of the equipment required by the project and has vast experience of more than 10 years in developing, installing and maintaining, and processing High Performance Computing (Deep learning) on passive acoustic monitoring observatories.

References

Abgrall P., Terhune J.M., Burton H.R. Variation of Weddell Seal (Leptonychotes weddellii) Underwater Vocalizations over Mesogeographic Ranges // Aquatic Mammals. 2003. No. 29 (2). P. 268–277.

Afanasyeva E.L. Life Cycle of Epischura Baicalensis Sars (Copepoda, Calanoida) in Lake Baikal // Journal of Marine Systems. 1998. No. 15 (1). P. 351–357.

BBC. Nature Wildlife. Life Animals Mammals Baikal Seal. <http://www.bbc.co.uk/nature/life/Baikal_Seal>.

Bénard F., Glotin H., Giraudet P. Whale 3*D* Monitoring Using Astrophysic NEMO ONDE Two Meters Wide Platform with State Optimal Filtering by Rao-Blackwell Monte Carlo data association // Journal of Applied Acoustics. 2011. Vol. 71 (2010). P. 994–999.

Brunello A.J., Molotov V.C., Dugherkhuu B. et al. Lake Baikal. Experience and Lessons Learned. South Lake Tahoe: Tahoe-Baikal Institute, 2006.

Chen F., Shapiro G.I., Bennett K.A. et al. Shipping Noise in a Dynamic Sea: A Case Study of Grey Seals in the Celtic Sea // Marine Pollution Bulletin. 2017. No. 114 (1). P. 372–383.

Cunningham K.A., Reichmuth C. et al. High-Frequency Hearing in Seals and Sea Lions // Hearing Research. 2016. No. 331. P. 83–91.

De Swart R.L., Harder T.C., Ross P.S. et al. Morbiliviruses and Morbilivirus Diseases of Marine Mammals. Infectious Agents and Diseases 4. 1995. P. 125–130.

Erbe C., Williams R., Sandilands D., Ashe E. Identifying Modeled Ship Noise Hotspots for Marine Mammals of Canadas Pacific Region. PLoS ONE 9, e89820. 2014.

Froese R., Pauly D. (eds). Species of Comephorus // FishBase. 2012.

Froese R., Pauly D. (eds). Comephorus Dybowskii // FishBase. 2016.

Glotin H., Giraudet P., Ricard J. et al. Visées aériennes de Mammifères marins jointes aux obs. acoustiques sous marines de Bombyx et Antares sur Physeter m. Pelagos, 2017.

Glotin H., Bartcus M., Roger V. et al. Scaled Unsupervised Arctic Bioacoustics: Joint Narwhal Call and Click Train Monitoring. LSIS RR, 2015. <http://sabiod.org>.

Glotin H., Mishchenko A., Giraudet P. Joint Dopler & Time-Delay of Arrival for Efficient Passive Acoustic Tracking — Application to Bioacoustics. Brevet INPI, 2014.

Glotin H., Giraudet P., Bénard-Caudal F. Multiple Sources Real-time Tracking Using Transitivity Constraints. Patent 2007/06162 FR, EU, USA 8638641, Australia 2008327744, New Zealand 606802. 2009–2012.

Hatch L.T., Clark Ch.W. et al. Quantifying Loss of Acoustic Communication Space for Right Whales in and around a U.S. National. Marine

Sanctuary // Conservation Biology. Wiley-Blackwell, 2012. No. 26. P. 983–994.

Ivanov T.M. Biology and Harvest of Baikal Seal // Izv. Biol.-geograf. NII pri Vost.-Sib. gos. universitete. Irkutsk, 1938. Vol. YIII. No. 1–2. P. 5–119.

Kannan K., Blankenship A.L., Jones P. Toxicity Reference Values for the Toxic Effects of Polychlorinated Biphenyls to Aquatic Mammals // Human & Ecological Risk Assess. 2000. No. 6. P. 181–201.

Kastak D., Schusterman R.J. Low-frequency Amphibious Hearing in Pinnipeds: Methods, Measurements, Noise, and Ecology // Journal of the Acoustical Society of America. 1998. No. 103 (4). P. 2216–2228.

Kutyrev I.A., Lamazhapova G.P., Erofeeva L.M., Zhamsaranova S.D. Comparison of Microanatomical and Cytological Characteristics of Mesenteric Lymphatic Node of Adolescents and Adult Individuals of Baikal Seal (Pusa sibirica) // Zoologichesky. 2006.

Kutyrev I.A., Pronin N.M., Imikhelova L.S. et al. Baikal Nerpa. The Passport and Bibliography. Ulan-Ude: Russia HURNAL, 2008. Vol. 85 (7). P. 886–892.

Matveyev A.N., Samusenok V.P. The Fishes and Fishery in Lake Baikal // Aquatic Ecosystem Health & Management. 2015. No. 18 (2). P. 134–148.

McCulloch S., Pomeroy P.P., Slater P.J. Individually Distinctive Pup Vocalizations Fail to Prevent Allo-suckling in Grey Seals // Canadian Journal of Zoology. 1999. No. 77 (5). P. 716–723.

Mirhaj M., Plötz J., Bornemann H. et al. Underwater Calls of Weddell Seals in the Weddell Sea. SCAR Open Science Conf. Antarctica & the Southern Ocean in the Global System. Antarctica, 2004. P. 25–31.

Miyasaka H., Dzyuba Y.V., Genkai-Kato M. et al. Feeding Ecology of Two Planktonic Sculpins, Comephorus Baicalensis & Comephorus Dybowskii, in Lake Baikal // Ichthyological Research. 2006. No. 53 (4). P. 419–422.

Pastukhov V.D. Ecological Characteristic of Baikal Seal and Features of Its Sustainable Use: Avtoref. ... diss. kand. nauk. Irkutsk, 1971.

Pastukhov V.D. The Face of Baikal — Nerpa. 2007. September 27. <www.bww.irk.ru>.

Patris J., Asch M., Glotin H. et al. High Performance Modeling of the Propagation of Biological Sounds in the Ocean: New Numerical Tools in the Study of Cetaceans / Racket in the Oceans: Why Underwater Noise Matters, How to Manage It? École Polytechnique, 2017.

Patris J., Malige, F., Komatitsch D. et al. Naval Acoustics, A New Tool to Model Underwater Propagation of Whales Songs. 2017.

Penkova O.G. Zooplankton in the Ecosystem of Lake Baikal. Probl. Conservation of Biol. Diversity of South. Sib. Mezhregion. Scientific-practical. Conf. Kemerovo, 1997. May 19–22.

Poupard M., Glotin H. Automatic Ethoacoustic Data Mining: Applications on Spotted Dolphin Whistles and Humpback Whale Vocalizations. DYNI LSIS Research Report, 2017.

Randall R.R., Brent S.S. et al. National Audubon Society Guide to the Marine Mammals of the World. N.Y.: Alfred A. Knopf Publishing, 2002.

Reid J.W. Epischura baikalensis. The IUCN Red List of Threatened Species. 1996.

Sciacca V., Viola S., Pavan G. et al. Shipping Noise and Seismic Airgun Surveys in the Ionian Sea: Potential Impact on Mediterranean Fin Whale. Proceedings of Meetings on Acoustics, 27 040010. 2016.

Schwarzenbach R.P., Philip M.G., Dieter M.I. Environmental Organic Chemistry. 2nd ed. Wiley Interscience, 2003. P. 1052.

Shapiro G., Chen F., Thain R. The Effect of Ocean Fronts on Acoustic Wave Propagation in the Celtic Sea // Journal of Marine Systems. 2014. No. 139. P. 217–226.

Stewart B.S., Petrov E.A., Baranov E.A. et al. Seasonal Movements & Dive Patterns of Juvenile Baikal Seals, Phoca Sibirica // Marine Mammal Science. 1996. No. 12 (4). P. 528–542.

Stirling I., Warneke R.M. Implications of a Comparison of the Airborne Vocalizations & Some Aspects of the Behaviour of the Two Australian Fur Seals, Arctocephalus Spp., on the Evolution & Present Taxonomy of the Genus // Australian Journal of Zoology. 1971. No. 19 (3). P. 227–241.

Thomas J., Pastukhov V., Elsner R., Petrov E. Phoca Sibirica // Mammalian Species. 1982. No. 188. P. 1–6.

Thomas J.A., Kuechle V.B. Quantitative Analysis of Weddell Seal Underwater Vocalizations at McMurdo Sound, Antarctica // The Journal of the Acoustical Society of America. 1982. No. 72 (6). P. 1730–1738.

Van Opzeeland I., Kindermann L. et al. Insights into the Acoustic Behaviour of Polar Pinnipeds Current Knowledge & Emerging Techniques of Study // Animal Behaviour: New Research. N.Y., 2008.

Veirs S., Veirs V., Wood J.D. Ship Noise Extends to Frequencies Used for Echolocation by Endangered Killer Whales // PeerJ. 2016. No. 4. e1657.

Veirs S., Veirs V., Williams R. et al. A Key to Quieter Seas: Half of Ship Noise Comes from 15% of the Fleet. 2017.

Viola S., Pavan G. et al. Continuous Monitoring of Noise Levels in the Gulf of Catania-Western Ionian Sea. Study of Correlation with Ship Traffic // Marine Pollution Bulletin. 2017.

Yoshii K., Melnik N.G., Timoshkin O. et al. Stable Isotope Analyses of the Pelagic Food Web in Lake Baikal // Limnology and Oceanography. 1999. No. 44 (3). P. 502–511.

<http://www.iucnredlist.org>.

<http://www.wikiwand.com/en/Lake_Baikal>.

<http://smiot.univ-tln.fr>.

SPECIFIC FEATURES OF BIG DATA PROCESSING AND THE CONCEPT OF INFORMATION

Peter Golubtsov
Lomonosov Moscow State University, National Research University Higher School of Economics, Moscow, Russia

Abstract. *The Data in "big data" sets, as a rule, have a huge volume, are distributed among numerous sites and are constantly replenished. As a result even a simplest analysis of big data faces serious difficulties. To apply traditional processing all the relevant data has to be collected in one place and arranged in the form of convenient structures. Only then the corresponding algorithm processes these structures and produces the result of analysis. In the case of big data, it can be just impossible to collect all the relevant data on one computer, and even impractical, since one computer would not be able to process them in a reasonable time. An appropriate data analysis algorithm should, working in parallel on many computers, extract from each set of raw data some intermediate compact "information", gradually combine and update it, and finally, use the accumulated information to produce the result. Upon arrival of new pieces of data, it should be able to add them to the accumulated information and eventually renew the result. We will discuss specific features of such well-arranged intermediate form of information, reveal its natural algebraic properties, and present several examples. We will also see that in many important data processing problems the appropriate information space may become equipped with an ordering which reflects the "quality" of the information. It turns out that such an intermediate form of information representation in some sense reflects the very essence of the information contained in the data. This leads us to a completely new, 'practical' approach to the notion of information.*

Keywords: *distributed data collection and processing, big data systems, parallel processing, information representation, canonical information, linear estimation, algebra of information, quality of information, information space.*

Introduction

Recently there has been a sharp surge in research related to big data. Indeed, it was found that large amounts of data may contain unexpected valuable information. Many interesting examples can be found in [Mayer-Schönberger, Kenneth, 2013]. We can say that in big data problems, as a rule, we are talking about extracting certain hidden information and presenting it in a form suitable for interpreting or making decisions. Such processes usually involve several stages in which information is extract-

ed from the original data, transformed, transmitted, accumulated and, eventually, converted to a form convenient for interpretation.

The use of the term "information" has recently increased significantly, especially in the context of data analysis. Usually it is understood too broadly and informally. However, in the author's opinion, such an increased frequency of use of this term indicates an increasing need for a more accurate and formal understanding of the phenomenon of information. Can the area of big data bring us closer to this understanding?

Studies related to big data systems are aimed at the problems of processing large amounts of distributed data and have, as a rule, a practical and technical orientation. At the same time, most of research on information theory is carried out in the context of the probability theory and mathematical statistics and is of predominantly theoretical interest.

Perhaps the most applied part of information theory, originating in Shannon's works [Shannon, Weaver, 1949], is related to the transmission of messages in the presence of interference. It is not so much about the "meaning" or quality of information, but about its quantity. A special place in mathematical statistics is occupied by Fisher's information, described by matrices [Barra, 1971; Borovkov, 1998]. It provides a more detailed reflection of the concept of information and, in particular, has an important additive structure in which the union of independent statistics corresponds to the sum of their information matrices. Despite numerous studies on information theory, the problem of formalizing the concept of information, reflecting precisely the meaning of the information contained in the data, is still far from a satisfactory solution. In this connection, we mention the papers [Golubtsov, 1991; 1999; 2002], in which instead of defining the information contained in the data, the informativeness of the sources of data is investigated. Within such approach, the algebraic structure of information sources and the partial order that allows comparing their informativeness naturally arise.

At the moment, the areas of interest of big data and different approaches to the notion of information are poorly connected. However, as noted above, the problem of big data requires a more precise, formal description of the very concept of information and information processes. This is necessary for constructing effective tools for converting information, based on mathematical (for example, algebraic) properties of information. In this regard big data problems might soon become the main driver and beneficiary of the general information theory. In this paper, we will try to show how some formalization of the concept of

information and its algebraic properties can follow simply from the consideration of the problem in the context of big data.

What distinguishes the problems of "big data" among data analysis problems? Big data, usually, have a huge volume, are distributed among numerous sites and are constantly replenished. As a result, even the simplest analysis of big data faces serious difficulties. Indeed, the traditional approaches to information processing assume that the data intended for processing is collected in one place, organized in the form of convenient structures (for example, matrices), and only then the appropriate algorithm processes these structures and produces the result of the analysis. In the case of big data, it is impossible to collect all the data needed for a research project on a single computer. Moreover, it would be impractical, because one computer would not be able to process them in a reasonable time. As a result, there emerges a need to transform existing algorithms, leading to their "parallelization", or even to develop new approaches to data processing, which, by the very formulation of the problem, could process separate data fragments independently and in parallel. The corresponding data analysis algorithm must, working on many computers in parallel, extract from each set of source data some intermediate compact "information", gradually combine and update it and, finally, use the accumulated information to produce the result. Upon the arrival of new pieces of data, it should be able to add them to the accumulated information and, eventually, update the result.

We will discuss the features of such a well-organized intermediate form of information, reveal its natural algebraic properties, and present several examples. We will also see that in certain data processing problems the appropriate information space may become equipped with an ordering which reflects the "quality" of the information. It turns out that such an intermediate form of information representation in some sense reflects the very essence of the information contained in the data. This leads us to a completely new, 'practical' approach to the notion of information.

Traditional data processing in big data context

Let us focus on the following features of information processing problems in big data systems:

a) typically, such problems deal with huge amounts of data;

b) usually such data is not collected in one place, but distributed over numerous, possibly remote computers;

c) new data is constantly emerging and has to be promptly included in the processing.

Traditional processing methods usually do not take into account this specificity and require a serious revision if they need to be applied to big data problems.

Let us briefly consider a standard approach to data processing (in an extremely simplified form). Problems of this kind include estimation, decision-making, learning, classification, etc. Usually, in problems with a small fixed data set, the processing consists in applying some transformation (algorithm, method)which represents the processing $\boldsymbol{P}$, to a data set and obtaining the result of processing (for example, an estimate of some unknown value) (Fig. 1).

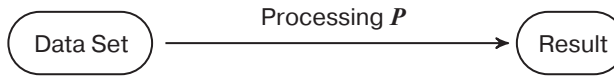


Fig. 1. Standard approach to data processing

It is critical here that all the data is collected in one place, presented in the form of suitable structures, say, matrices, and is ready for the processing transformation to be applied to them. If the data is distributed in many different locations, to apply the processing they must first be collected in one place and organized in the form of suitable structures (Fig. 2). Double dashed arrows hereinafter denote the transfer of data in the original form.

The disadvantages of this approach to the processing of distributed data are quite obvious. Transferring large amounts of raw data would cre-

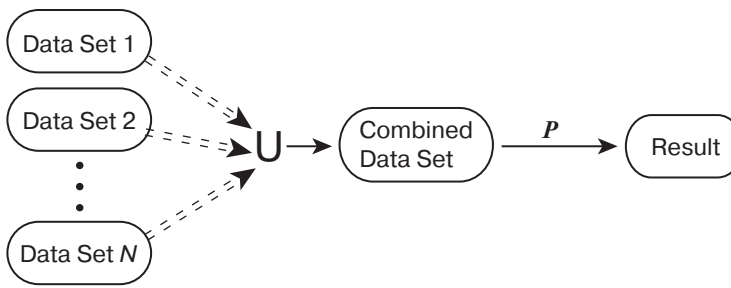


Fig. 2. Standard processing scheme for distributed data

ate excessive traffic. Keeping the combined data set in one place would require huge amounts of memory. Processing all data on one computer would require excessive computational and time resources. As new data become available, the combined data set would grow and, as a consequence, require ever-increasing (potentially infinite) storage resources. Besides, the processing algorithm would need to be reapplied to the constantly increasing amount of accumulated data.

Extracting intermediate information as a critical step in the processing

Consider the following modification of the processing pattern, which allows to overcome the drawbacks outlined above. Suppose that the complete processing algorithm $\boldsymbol{P}$ admits a factorization into two phases $\boldsymbol{P} = \boldsymbol{P}_2 \circ \boldsymbol{P}_1$ (Fig. 3), where $\boldsymbol{P}_1$ extracts some intermediate information from the original data and $\boldsymbol{P}_2$ computes the result based on the extracted intermediate information.

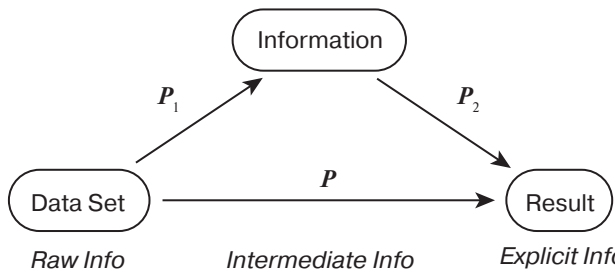


Fig. 3. Splitting the processing into two phases

The choice of an appropriate intermediate form of information representation is determined by the considered data processing problem. We will call such form **canonical information**.

In a certain sense, the nodes of the diagram in Fig. 3 reflect the presentation of information in different forms:

a) Data Set — information in **raw** (original) form;

b) Result — information in an **explicit** (convenient for interpretation) form;

c) Information — information in an intermediate (convenient for processing) **canonical** form.

Such a form of information representation must be complete, that is, contain all the information necessary for the calculation of the result (this is reflected by the commutativity of the diagram in Fig. 3) and compact, that is, to have the smallest possible size, ideally not depending on the amount of represented data. Below we will discuss in more detail the desirable properties of canonical information.

Now, if the complete processing can be split into two phases, indicated above, the processing scheme for distributed data, shown on Fig. 2 can be modified to the form shown on Fig. 4.

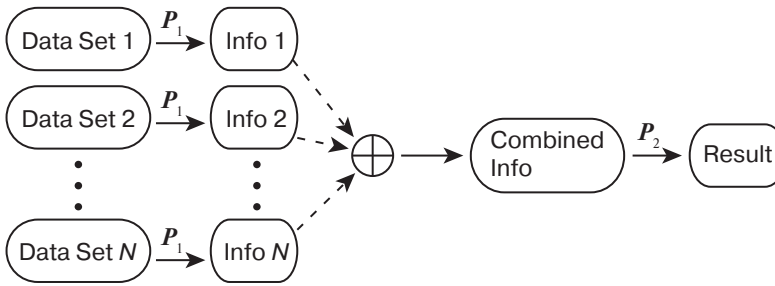


Fig. 4. Modified scheme for processing distributed data. Dashed arrows indicate transfer of compact canonical information

Such a scheme overcomes all the drawbacks of the standard distributed data processing scheme noted above. Only compact fragments of the selected intermediate information are transmitted (shown by dashed lines). Storing all the combined information would require small amounts of memory, possibly the same as the volumes required for storing separate parts of the intermediate information. Pieces of the intermediate information are extracted in parallel from separate data sets via phase $\boldsymbol{P}_1$. If the main part of the processing is concentrated in the first phase, the second phase $\boldsymbol{P}_2$, consisting in constructing the result from compact accumulated information, would not require serious computing and time resources. As new data become available, it would only be necessary to extract intermediate information from them and "add" it to the accumulated information. In this case, the processing algorithm would need to be reapplied to compact information of fixed volume.

Let us note that in the above arguments we assume the existence of a composition operation (addition) of individual fragments of canonical information. Moreover, we assume that the union of the two data sets is

represented by the composition of the corresponding fragments of canonical information (Fig. 5).

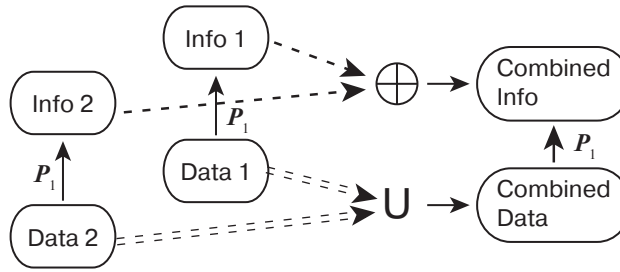


Fig. 5. Relation between the composition of fragments of canonical information and the union of the original data sets

This can be written as $\boldsymbol{P}_1(D_1) \oplus \boldsymbol{P}_1(D_2) = \boldsymbol{P}_1(D_1 \cup D_2)$, where $D_1 \cup D_2$ is the union of two data sets.

Finally, note that the distributed data processing scheme presented in Fig. 4. perfectly "fits" the architecture of distributed data storage and analysis systems, such as, for example, Hadoop [White, 2015].

Example 1. Computing mean value

Consider the following extremely simple problem as our first illustrative example. While the problem itself is quite trivial, we will assume that the volumes of data sets and the number of such sets are extremely large.

Suppose that the data set $D = (x_1, x_2, \ldots, x_n)$ is a sequence of n real numbers and the processing goal is to compute their mean value $X = \frac{1}{n}\sum_{i=1}^{n} x_i$:

$$(x_1, \ldots, x_n) \xrightarrow{\boldsymbol{P}} X = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Fig. 6. Standard approach for processing the single data set

If the original data is contained in N sets $\left(x_1, \ldots, x_{n_1}\right)$, ..., $\left(z_1, \ldots, z_{n_N}\right)$, located on different computers, then to process them using

this algorithm, one would have to collect them in one place and apply the transformation $\boldsymbol{P}$ (Fig. 7).

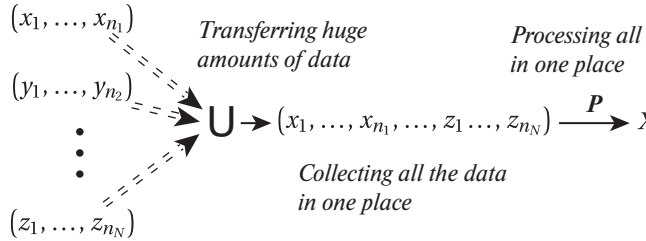


Fig. 7. Combining the raw data for processing

Such a scheme would require to transfer large amounts of raw data, store and process the complete huge set $\left(x_1, \ldots, x_{n_1}, \ldots, z_1, \ldots, z_{n_N}\right)$ on one computer. Upon arrival of a new dataset, one would have to appended it to the already existing combined set and recalculate the result X.

However, it is quite obvious that the computation of X can be split in two steps. Let $S = \sum_{i=1}^{n} x_i$. Then $X = \frac{S}{n}$. Thus, all the information sufficient for calculating X can be represented by a couple (n, S) and the whole processing P can be divided into two stages $\boldsymbol{P} = \boldsymbol{P}_2 \circ \boldsymbol{P}_1$ (Fig. 8).

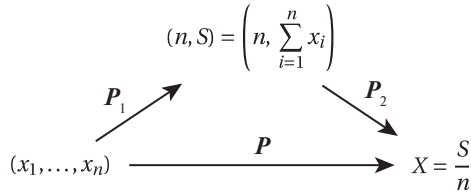


Fig. 8. Factorization of the processing into two stages by extracting the canonical

In the problem under consideration the couple (n, S) is a convenient intermediate form of presenting information about the initial data — the canonical information. Note that it is specified by two numbers, regardless of the amount of data it represents.

As a result of introducing canonical information and factorizing the algorithm $\boldsymbol{P}$ into two phases, the distributed data processing scheme presented in Fig. 7 can be transformed into the form shown on Fig. 8. From each separate fragment of the data, the canonical information

(n_j, S_j) is extracted, which is subsequently combined and used to compute the result.

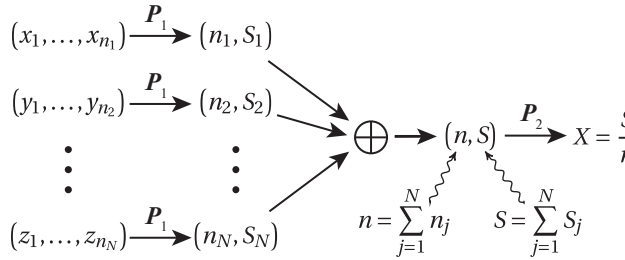


Fig. 9. Revised distributed data processing scheme

Let us emphasize the main features of such a revised scheme. The extraction of canonical information (n_j, S_j) from the data set (transformation $\boldsymbol{P}_1$) can be carried out "in situ" in parallel and independently. As a result, the distribution of input data contributes to improving the efficiency of processing due to parallelization. Only compact fragments of the selected canonical information of the same volume (2 numbers) are transmitted, which does not depend on the volume of the initial data set. The addition of the pieces of the canonical information is maximally simplified and is determined by the componentwise addition of the couples (n_j, S_j):

$$(n_1, S_1) \oplus (n_2, S_2) = (n_1 + n_2, S_1 + S_2).$$

Storing all combined canonical information also requires the same small amount of memory (2 numbers). Since the main part of the processing is concentrated in the first phase, the second phase $\boldsymbol{P}_2$, i.e. constructing the result from the compact accumulated information (n, S), does not depend on the amount of the initial data and does not require significant computing and time resources. As new data become available, it will only be necessary to transform it to canonical form, "add" it to the accumulated information, and reapply the transformation $\boldsymbol{P}_2$ to the updated compact information.

The couples of the form (n, S) can be considered as elements of a set endowed with an additional structure — the canonical **information space** $\mathfrak{I}_1$. In this example, $\mathfrak{I}_1 = \mathbb{N}\times\mathbb{R}$, where $\mathbb{N} = \{0, 1, \ldots\}$ is the set of natural numbers and $\mathbb{R}$ is the set of reals. Moreover, the space

$\mathfrak{I}_1$ is equipped with a composition operation $\oplus$, defined, componentwise.

Example 2. Adding sample variance to the processing goal

Now let us slightly modify our previous example by modifying the processing goal. Suppose that the data set is the same as before: $D = (x_1, x_2, \ldots, x_n)$, but in addition to the sample mean $X = \frac{1}{n}\sum_{i=1}^{n} x_i$ we need to compute the sample variance $V = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - X)^2$ as well.

$$(x_1, x_2, \ldots, x_n) \xrightarrow{\boldsymbol{P}} (X, V) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i,\ \ \frac{1}{n-1}\sum_{i=1}^{n}(x_i - X)^2\right).$$

Fig. 10. Standard processing approach for the modified goal

Since computing V involves X and all the original x_i it might seem that we would need to keep all the original data to perform such computation. However, since

$$\sum_{i=1}^{n}(x_i - X)^2 = \sum_{i=1}^{n} {x_i}^2 - X\sum_{i=1}^{n} x_i - \left(\sum_{i=1}^{n} x_i\right)X + nX^2 = T - \frac{1}{n}S^2,$$

where $T = \sum_{i=1}^{n} {x_i}^2$, we can just modify our previous form of canonical information by adding to it T. As a result, $V = \frac{1}{n(n-1)}(nT - S^2)$ and the appropriate factorization of the processing $\boldsymbol{P}$ has the form shown on Fig. 11.

As a result, we arrive at a new form of canonical information, suitable for the modified processing problem. Now canonical information is represented by a triple (n, S, T) and the new information space $\mathfrak{I}_2 = \mathbb{N}\times\mathbb{R}\times\mathbb{R}_+$, where $\mathbb{R}$ is the set of nonnegative reals. Again, the space $\mathfrak{I}_2$ is equipped with a componentwise composition operation $\oplus$:

$$(n_1, S_1, T_1)\oplus(n_2, S_2, T_2) = (n_1 + n_2, S_1 + S_2, T_1 + T_2),$$

which corresponds to the combination of two datasets, i.e. $\boldsymbol{P}_1(x_1, \ldots, x_{n_1}) \oplus \boldsymbol{P}_1(y_1, \ldots, y_{n_2}) = \boldsymbol{P}_1(x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2})$. Any sin-

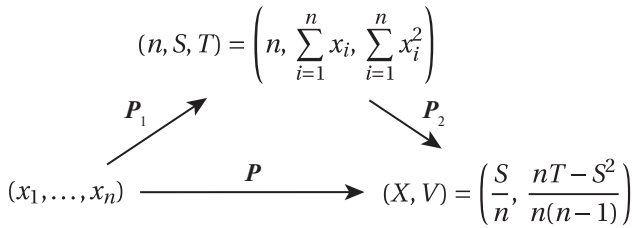


Fig. 11. Factorization of the processing into two stages for the modified goal

gle observation x can be added to the collected canonical information "on the fly" via the following updating operation:

$$(n, S, T) \oplus x = (n + 1, S + x, T + x^2).$$

It is easy to see that all the nice features of the revised two-stage processing scheme mentioned above are valid for this modified problem as well.

Moreover, these two examples show that a more elaborate goal may require a more elaborate information space. More precisely, $\mathfrak{I}_1$ can be considered as a subspace of $\mathfrak{I}_2$. This illustrates that certain relations between processing goals (for the same types of data) should lead to the hierarchy of information spaces.

Example 3. Linear estimation problem

As our next example we will consider a widely used linear estimation process. Consider a scheme of linear measurement of the form [Pyt'ev, 1982; 1989]:

$$y = Ax + v,$$

where $x \in \mathcal{D} = \mathbb{R}^m$ is the unknown vector of the m-dimensional Euclidean space — the measurement object, $y \in \mathcal{R} = \mathbb{R}^k$ is the measurement result, $A: \mathcal{D} \to \mathcal{R}$ is the linear mapping, represented by an $k \times m$-matrix, describing the distortions of the measuring system, The vector $v \in \mathcal{R}$ is the random noise vector with zero mean $\mathsf{E}v = 0$ and the given covariance matrix $S = \mathrm{cov}(v)$, which can be considered as a linear operator $S: \mathcal{R} \to \mathcal{R}$ — the covariance operator of the random vector $\in \mathcal{R}$.

It is easy to see that the covariance matrix of the random vector is symmetric and non-negative definite. We will consider only the measurements in which the matrix S is positive definite, $S > 0$, and hence is invertible. In essence, this means that the noise ν is "possible in all directions", that is, there is no proper subspace $\tilde{\mathcal{R}} \subset \mathcal{R}$ such that $\nu \in \tilde{\mathcal{R}}$ with probability one.

Thus, the measurement data is represented by the triple (y, A, S) and includes the measurement result y and the measurement model described by the pair (A, S).

The problem of linear estimation of an unknown vector x consists in constructing a linear mapping $R: \mathcal{R} \to \mathcal{D}$ such that the estimate $\hat{x} = Ry$ is maximally close to x. A detailed and general consideration can be found in [Pyt'ev, 1982; 1989]. Formally, let us consider the error in estimating $\mathsf{E}\|Ry-x\|^2$

$$\mathsf{E}\|Ry - x\|^2 = \mathsf{E}\|R(Ax + \nu) - x\|^2 = \|(RA - I)x\|^2 +$$
$$+2\mathsf{E}\langle (RA - I)x, R\nu\rangle + \mathsf{E}\|R\nu\|^2 = \|(RA - I)x\|^2 + \mathrm{tr}RSR^*.$$

Since there is an unknown vector x in the expression for $\mathsf{E}\|Ry-x\|^2$, we define the estimation error provided by the operator R as

$$H(R) = \sup_{x\in\mathcal{D}} \mathsf{E}\|Ry-x\|^2$$

It is easy to see that if $RA \neq I$ then $\|(RA-I)x\|^2$ can take arbitrarily large values and, consequently,

$$H(R) = \begin{cases} +\infty, & \text{if } RA \neq I, \\ \mathrm{tr}RSR^*, & \text{if } RA = I. \end{cases}$$

Thus, the linear mapping R provides a finite estimation error $H(R)$ if and only if $RA = I$. It is easy to see that the last equation is equivalent to the requirement that the estimate $\hat{x} = Ry$ is unbiased, i.e. $\mathsf{E}Ry = x$. Thus, the problem of linear estimation can be regarded as the problem of conditional minimization:

$$\min_{R:\mathcal{R}\to\mathcal{D}} \{\mathrm{tr}RSR^* \mid RA = I\}.$$

It has a solution if and only if $A^*S^{-1}A: \mathcal{D} \to \mathcal{D}$ is nonsingular. In this case the optimal estimate, known as the best linear unbiased esti-

mate (BLUE), and the corresponding estimation error are given by the expressions:

$$\hat{x} = Ry = \left(A^* S^{-1} A\right)^{-1} A^* S^{-1} y,$$
$$\mathsf{E}\|\hat{x}-x\|^2 = \operatorname{tr}\left(A^* S^{-1} A\right)^{-1}.$$

Thus, the processing procedure $\boldsymbol{P}$ consists in converting the original data, represented by (y, A, S) into the processing result: the optimal estimate $\hat{x}$ of the vector x:

$$(y, A, S) \xrightarrow{\boldsymbol{P}} \hat{x} = \left(A^* S^{-1} A\right)^{-1} A^* S^{-1} y.$$

Note, that the mapping $\boldsymbol{P}$ is not defined everywhere, but only if the operator $A^* S^{-1} A$ is invertible.

Linear estimation in the case of many independent measurements

Now let there be many independent measurements of the same unknown vector $x \in \mathcal{D}$:

$$y_i = A_i x + v_i, \qquad i = 1, \ldots, n,$$

where $y_i \in \mathcal{R}_i$ are measurement results, $A_i : \mathcal{D} \to \mathcal{R}_i$ are linear mappings, and $v_i \in \mathcal{R}_i$ are independent random vectors with zero means $\mathsf{E} v_i = 0$ and covariance operators $S_i : \mathcal{R}_i \to \mathcal{R}_i$. In general, the measurement spaces $\mathcal{R}_i = \mathbb{R}^{k_i}$ can be different.

To process n such measurements, one would have to collect all the pieces of data (y_i, A_i, S_i) in one place, reorganize them in the form of block matrices, possibly very large dimensions, and apply the transformation $\boldsymbol{P}$ to the combined data (Fig. 12).

Here

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathcal{R}, \quad A = \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{pmatrix} : \mathcal{D} \to \mathcal{R}, \quad S = \begin{pmatrix} S_1 & 0 & \cdots & 0 \\ 0 & S_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & S_1 \end{pmatrix} : \mathcal{R} \to \mathcal{R},$$

$$\mathcal{R} = \mathcal{R}_1 \times \mathcal{R}_2 \times \cdots \times \mathcal{R}_n, \qquad \dim \mathcal{R} = \sum_{i=1}^{n} \dim \mathcal{R}_i = \sum_{i=1}^{n} k_i.$$

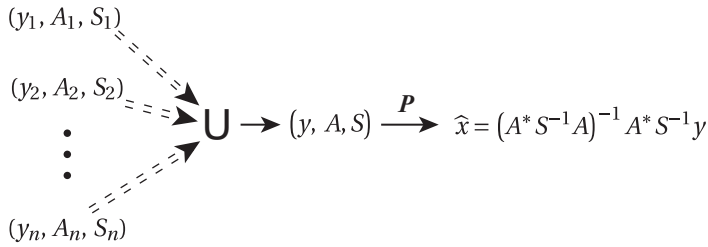


Fig. 12. The standard scheme of linear estimation for a large number of measurements

For a large number of measurements, the dimension of the combined data can become extremely large, which would make this approach unfeasible. Besides, the addition of new data would lead to the increase in the dimensions of the merged data, which in turn would require increasing resources for their storage and processing (application of the transformation $\boldsymbol{P}$).

Parallelizing processing by extracting the intermediate information

Let us show that the data processing in the linear estimation problem can be divided into two phases $\boldsymbol{P} = \boldsymbol{P}_2 \circ \boldsymbol{P}_1$, where the first phase $\boldsymbol{P}_1$ extracts some compact intermediate information from the initial data, and the second $\boldsymbol{P}_2$ calculates the estimation result based on this intermediate information. Moreover, our goal will be to find such a factorization that the application of the transformation $\boldsymbol{P}_1$ to the combined data set can be replaced by the parallel application of $\boldsymbol{P}_1$ to individual data and the subsequent "addition" of the extracted information fragments.

As we have just seen, the vector y and matrices A, and S describing the combined data can become extremely large, which would make application of the transformation $\boldsymbol{P}$ impossible. However, it can be shown that the main parts of the expression $\hat{x} = \left(A^* S^{-1} A\right)^{-1} A^* S^{-1} y$ can be decomposed into pieces:

$$A^* S^{-1} y = A_1^* S_1^{-1} y_1 + \cdots + A_n^* S_n^{-1} y_n,$$
$$A^* S^{-1} A = A_1^* S_1^{-1} A_1 + \cdots + A_n^* S_n^{-1} A_n.$$

This implies that all the information needed for further processing related to the i-th piece of data (y_i, A_i, S_i) can be represented by a pair (v_i, T_i), where

$$v_i = A_i^* S_i^{-1} y_i \in \mathcal{D}, \qquad T_i = A_i^* S_i^{-1} A_i : \mathcal{D} \to \mathcal{D},$$

and T_i is a non-negative definite operator. Obviously, the pair (v, T) in which $v = v_1 + \cdots + v_n$ and $T = T_1 + \cdots + T_n$ will correspond to the combined data (y, A, S).

We will call the pair $(v, T) = \left(A^* S^{-1} y,\ A^* S^{-1} A\right)$ the **canonical information** for the data (y, A, S), and the set $\mathfrak{I}$ of all such pairs canonical **information space** for the problem of linear estimation of a vector from the space $\mathcal{D}$. It can be shown that $\mathfrak{I}$ consist of all the pairs (v, T) in which $v \in \operatorname{im} T$. Thus,

$$\mathfrak{I} = \{(v, T) \mid T \in \mathbb{S}_{\mathcal{D}}^+, v \in \operatorname{im} T\},$$

where $\mathbb{S}_{\mathcal{D}}^+$ is the set of nonnegative definite operators on $\mathcal{D}$ — a convex cone in the linear space $\mathbb{S}_{\mathcal{D}}$ of selfadjoint operators on the space $\mathcal{D}$. If $\dim \mathcal{D} = m$, then $\dim \mathbb{S}_{\mathcal{D}} = \frac{m(m+1)}{2}$. Thus, $\mathfrak{I} \subset \mathcal{D} \times \mathbb{S}_{\mathcal{D}}^+$ is a convex cone in the $\frac{m(m+3)}{2}$-dimensional vector space $\mathcal{D} \times \mathbb{S}_{\mathcal{D}}$. It implies, in particular, that any element of the information space $\mathfrak{I}$ can be represented by $\frac{m(m+3)}{2}$ numbers.

Obviously, the process of linear estimation can be divided into two phases $\boldsymbol{P} = \boldsymbol{P}_2 \circ \boldsymbol{P}_1$, where the first phase $\boldsymbol{P}_1$ consists in constructing the canonical information:

$$(v, T) = \boldsymbol{P}_1(y, A, S) = \left(A^* S^{-1} y, A^* S^{-1} A\right),$$

and the second phase $\boldsymbol{P}_2$ calculates the estimation result based on this information (Fig. 13):

$$\hat{x} = \boldsymbol{P}_2(v, T) = T^{-1} v.$$

As was shown above, the combination of the initial data (y_1, A_1, S_1) and (y_2, A_2, S_2) can be represented by the composition of the corresponding pieces of canonical information (v_1, T_1) and (v_2, T_2), defined as

$$(v_1, T_1) \oplus (v_2, T_2) = (v_1 + v_2, T_1 + T_2).$$

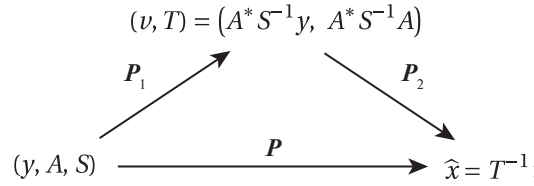


Fig. 13. Splitting data processing into two phases

This can be written as $\boldsymbol{P}_1(y_1, A_1, S_1) \otimes \boldsymbol{P}_1(y_2, A_2, S_2) = \boldsymbol{P}_1((y_1, A_1, S_1) \cup (y_2, A_2, S_2))$, where $(y_1, A_1, S_1) \cup (y_2, A_2, S_2)$ is combining two data sets into one (Fig. 14).

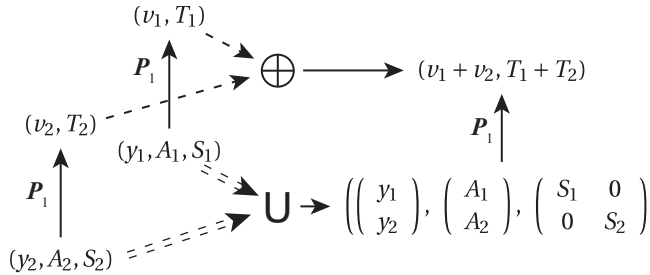


Fig. 14. The correspondence between the composition of fragments of canonical information and combining sets of input data

As a result of the introduction of canonical information and the factorization of algorithm P into two phases, the data processing scheme presented in Fig. 12 can be transformed as follows (Fig. 15). From each individual fragment (y_1, A_1, S_1) of the data, the canonical information (v_1, T_1) is extracted, which is subsequently combined and used to calculate the estimation result.

Let us outline the main features of such a modified scheme. The amount of memory required to store information in the canonical form does not depend on the volume of the represented original data and is $\frac{m(m+3)}{2}$ real numbers (m-dimensional vector and symmetric $m \times m$ matrix). Computing the canonical information (v_j, T_j) from the i-th set of data (transformation $\boldsymbol{P}_1$) can be performed on the computers, where the data is located, in parallel and independently. Only compact fragments of the canonical information of the same volume are transferred. The addi-

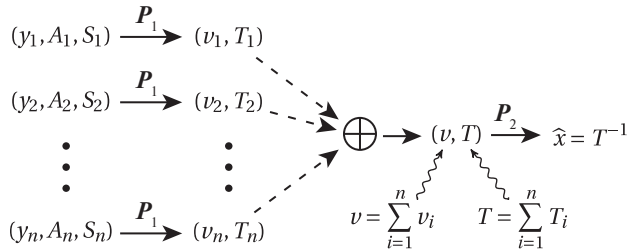


Fig. 15. Modified scheme for processing distributed data

tion of the parts of canonical information is maximally simplified and is determined by the componentwise addition of the pairs (v_i, T_i). Resource requirements for the second phase $\boldsymbol{P}_2$, consisting in constructing the result from the compact accumulated information (v, T), is determined only by the dimension m of the space of unknown x and does not depend on the volume of the original data. As a new data becomes available, it would only be necessary to extract from it the canonical information and "add" it to the accumulated information. In this case, the final processing $\boldsymbol{P}_2$ would have to be reapplied to the compact information of a fixed volume.

As a result, the distribution of the initial data contributes to an increase in processing efficiency due to the natural parallelization of the algorithm.

Quality of information

In the problem of linear estimation, considered above, our goal was to construct an estimate $\hat{x}$, that is, $\boldsymbol{P}(y, A, S) = \hat{x}$. The corresponding estimation error is $\mathsf{E}\|\hat{x}-x\|^2 = \mathrm{tr}Q$, where $Q = \left(A^* S^{-1} A\right)^{-1} = T^{-1}$ represents the covariance matrix of $\hat{x}$, i.e., $Q = \mathrm{cov}(\hat{x})$. Matrix Q also allows to determine the errors in estimating the individual components of the vector x since $\mathsf{E}\left(\hat{x}_j - x_j\right)^2 = \mathrm{var}(\hat{x}_j) = Q_{jj}$.

Moreover, the smaller the covariance matrix, the less the estimation error: that is, if $Q \le \tilde{Q}$, then $Q_{jj} \le \tilde{Q}_{jj}$ and $\mathrm{tr}Q \le \mathrm{tr}\tilde{Q}$. It means that if Q and $\tilde{Q}$ are covariance matrices for two estimates of x, then the estimate with the smaller covariance matrix would provide better precision in all respects. Here we define the partial order on the set of symmetric matrices of the same dimension as follows:

$$Q \geq \tilde{Q} \Leftrightarrow Q - \tilde{Q} \geq 0,$$

that is, Q is greater or equal than $\tilde{Q}$ if $Q - \tilde{Q}$ is nonnegative definite.

We will say that the information (v, T) is not worse (not less accurate) than $(\tilde{v}, \tilde{T})$ and write $(v, T) \succcurlyeq (\tilde{v}, \tilde{T})$ if $T \geq \tilde{T}$. If $(v, T) \succcurlyeq (\tilde{v}, \tilde{T})$ and $(\tilde{v}, \tilde{T}) \succcurlyeq (v, T)$, then we say that (v, T) and $(\tilde{v}, \tilde{T})$ have the same accuracy and denote this $(v, T) \approx (\tilde{v}, \tilde{T})$. Obviously, this is equivalent to the condition $T = \tilde{T}$. It is easy to see that more accurate information provides more accurate estimation. Indeed, let $T \geq \tilde{T}$ and the couples (v, T) and $(\tilde{v}, \tilde{T})$ allow to construct the corresponding estimates, that is, T and $\tilde{T}$ are invertible. According to [Pyt'ev, 1982], this implies that $T^{-1} \leq \tilde{T}^{-1}$ and hence $Q \leq \tilde{Q}$, where Q and $\tilde{Q}$ are the covariance matrices of the corresponding estimates.

Note that the above accuracy ordering emerges on the information space quite naturally. It turns out that the above concept of the accuracy of information leads to the same ordering on the set of models (A, S) of linear measurement as the concept of the quality of measurement models in [Ibid., 1989; 1982] or informativeness of information transformers in [Golubtsov, 1992], while the definitions of the corresponding partial orders in these two approaches are quite different and more complex.

Properties of canonical information

Let us summarize the properties of the canonical information spaces $\mathfrak{I}$ defined above. These properties not only represent an independent interest, but can serve as an example of the general properties of information spaces that arise in the tasks of processing large volumes of distributed data.

Existence for any source dataset. Any source dataset must allow the presentation of information in a canonical form. Note that the calculation of the final result may not be possible for some data. Strictly speaking, the transformation $\boldsymbol{P}$ is not everywhere defined. At the same time, we require $\boldsymbol{P}_1$ to be defined everywhere.

As we have seen, the information contained in the data (y, A, S) may not allow to construct the result of the estimation. Namely, if $A^*S^{-1}A$ is singular, then the estimate of the unknown vector can not be produced. Nevertheless, the canonical information (v, T) can be constructed for any initial data. Note that even the complete absence of measurements

(carrying zero information) can be represented in the canonical form. Formally, any measurement (y, A, S) in which $A = 0: \mathcal{D} \to \mathcal{R}$ is a zero mapping, does not carry any information about the vector being measured. Any such measurement corresponds to the canonical information $\mathbf{0} = (0,0)$, i.e. $v = 0 \in \mathcal{D}$ and $T = 0: \mathcal{D} \to \mathcal{D}$.

Completeness (or **sufficiency**). The canonical form retains all the information contained in the original data, namely, it leads to the same result as the original data from which it was derived. Formally, this means that $\boldsymbol{P}(\mathrm{D}) = \boldsymbol{P}_2(\boldsymbol{P}_1(\mathrm{D}))$ for any data D from the domain of definition of $\boldsymbol{P}$. This property resembles the concept of sufficiency in mathematical statistics.

Composition operation $\oplus$. On the canonical information space $\mathfrak{I}$, a composition operation $\oplus$ is defined that represents the combination of the corresponding fragments of data. Moreover, $(\mathfrak{I}, \oplus, \mathbf{0})$ is a commutative monoid, that is, the following properties hold for any $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathfrak{I}$:

a) $\boldsymbol{a} \oplus \boldsymbol{b} = \boldsymbol{b} \oplus \boldsymbol{a}$;

b) $(\boldsymbol{a} \oplus \boldsymbol{b}) \oplus \boldsymbol{c} = \boldsymbol{a} \oplus (\boldsymbol{b} \oplus \boldsymbol{c})$;

c) $\boldsymbol{a} \oplus \mathbf{0} = \boldsymbol{a}$.

In addition, the monoid $(\mathfrak{I}, \oplus, \mathbf{0})$ also has the cancellation property:

d) $\boldsymbol{a} \oplus \boldsymbol{b} = \boldsymbol{a} \oplus \boldsymbol{c} \Rightarrow \boldsymbol{b} = \boldsymbol{c}$,

but does not have invertible elements other than $\mathbf{0}$, i.e. there is no "negative" information.

Preorder $\succcurlyeq$, reflecting accuracy of information. This relation naturally appears in the linear estimation problem and is intrinsically related to the estimation precision.

On the canonical information space $\mathfrak{I}$ a relation $\succcurlyeq$ is defined that has the following properties:

a) $\boldsymbol{a} \succcurlyeq \boldsymbol{a}$ (reflexivity);

b) $\boldsymbol{a} \succcurlyeq \boldsymbol{b}$ & $\boldsymbol{b} \succcurlyeq \boldsymbol{c} \Rightarrow \boldsymbol{a} \succcurlyeq \boldsymbol{c}$ (transitivity).

Besides, the algebraic structure of the information space is consistent with the order structure:

c) $\boldsymbol{a} \succcurlyeq \mathbf{0}$ — any information is more accurate than the lack of information;

d) $\boldsymbol{a} \oplus \boldsymbol{b} \succcurlyeq \boldsymbol{a}, \boldsymbol{b}$ — the composition of two fragments of information is more precise than each of them individually;

e) $\boldsymbol{a} \succcurlyeq \boldsymbol{b}$ & $\boldsymbol{c} \succcurlyeq \boldsymbol{e} \Rightarrow \boldsymbol{a} \oplus \boldsymbol{c} \succcurlyeq \boldsymbol{b} \oplus \boldsymbol{e}$ — the more accurate the fragments of information, the more accurate is the result of the composition.

Uniqueness of data representation in the canonical form. For any original data there should be a unique representation in the information space $\mathfrak{I}$, consistent with the operation of the composition. In particular, since the estimation result does not depend on the order of the data in the source set, the canonical information should not depend on the order of the data.

Finally, we will mention two "practical" properties of this form of representing intermediate information. They are more of a technical nature, related to the implementation of the corresponding algorithms.

Compactness. The information presented in the canonical form should occupy a small (preferably minimal) volume, if possible, independent of the amount of data presented. In the estimation example the canonical form occupies a fixed volume $\frac{m(m+3)}{2}$ numbers.

Efficiency. The presentation of the intermediate information in canonical form should ensure the efficient implementation of all stages of data processing. Specifically, in the estimation example:

1. Extracting canonical information from the original data requires several matrix multiplications for the matrices determined by the individual data fragments. Besides, extraction of canonical information from individual fragments can be performed in parallel.

2. The combination and accumulation of canonical information reduces to the addition of vectors and matrices of fixed dimension and requires insignificant computational resources.

3. The calculation of the result on the basis of the accumulated canonical information requires solving a system of linear equations of fixed size $m \times m$ (or inversion of the corresponding matrix). Even with the constant arrival of new data, updating of the estimate can be carried out only from time to time.

Conclusion

Let us emphasize that a purely technical attempt to "parallelize the algorithm" actually leads to the construction of a special type of information representation that has natural algebraic properties. In a sense, this representation reflects the very essence of the information contained in the data. We can say that the very need to effectively manipulate huge distributed data sets puts forward new requirements for the conceptualization and formalization of the notion of information.

In the examples considered above, the choice of the canonical form of information is quite obvious. In general, the choice of compact intermediate information may not be obvious or even impossible. In this regard, it seems important to identify a class of problems in which it is possible to extract fairly compact intermediate information and find effective methods for constructing suitable information spaces.

In many practical problems the processing $\boldsymbol{P}$, transforming the initial data into the final result of processing, has a specific "origin", namely, it optimizes the quality of the result (e.g., estimation precision). Due to the optimization statement of the original problem, the concept of the quality of the solution (the accuracy of the estimate) induces an ordering on the information space that reflects the "quality" of the information. As shown in [Golubtsov, 1999; 2002; 1992; 1998; 2004], such natural ordering and algebraic structure always arise when investigating the informativeness of various classes of information sources, including, for example, multi-valued [Golubtsov, Filatova, 1992] and fuzzy [Golubtsov, 1994]. It can be expected that such an ordering, consistent with the algebraic structure of the information space, will always appear in the context of problems of optimal decision-making in distributed systems.

In this paper, we tried to minimize formalism in order to focus on the informative side of the problem. We outlined the basic requirements for well-organized intermediate information. This, in turn, leads to the question of choosing in some sense the optimal or ideal form of intermediate information. Such problems would require further formalization and research.

References

Barra J.-R. Notions foundamentales de statistique mathematique. Paris: Dunod, 1971.

Borovkov A.A. Mathematical Statistics. Gordon and Breach, 1998.

Golubtsov P.V. Axiomatic Description of Categories of Information Transformers // Probl. Peredachi Inf. 1998. Vol. 35. No. 3. P. 60–80 (engl. transl.: *Golubtsov P.V.* Problems Inform // Transmission. 1999. Vol. 35. No. 3. P. 259–274).

Golubtsov P.V. Informativity in the Category of Linear Measurement Systems // Probl. Peredachi Inf. 1992. Vol. 28. No. 2. P. 30–46 (engl. transl.: *Golubtsov P.V.* Problems Inform // Transmission. 1992. Vol. 28. No. 2. P. 125–140).

Golubtsov P.V. Informativity in the Category of Multivalued Information Transformers // Probl. Peredachi Inf. 1998. Vol. 34. No. 3. P. 60–80 (engl. transl.: *Golubtsov P.V.* Problems Inform // Transmission 1999. Vol. 34. No. 3. P. 259–276).

Golubtsov P.V. Information Transformers: Category-Theoretical Structure, Informativeness, Decision-Making Problems // Hadronic Journal Supplement. 2004. Vol. 19. No. 4. P. 375–424.

Golubtsov P.V. Measurement Systems: Algebraic Properties and Informativity // Pattern Recognition and Image Analysis. 1991. Vol. 1. No. 1. P. 77–86.

Golubtsov P.V. Monoidal Kleisli Category as a Background for Information Transformers Theory // Information Processes. 2002. Vol. 2. No. 1. P. 62–84.

Golubtsov P.V. Theory of Fuzzy Sets as a Theory of Uncertainty and Decision-Making Problems in Fuzzy Experiments // Probl. Peredachi Inf. 1994. Vol. 30. No. 3. P. 47–67 (engl. transl.: *Golubtsov P.V.* Problems Inform // Transmission. 1995. Vol. 30. No. 3. P. 232–250).

Golubtsov P.V., Filatova S.A. Multivalued Measurement-Computer Systems // Mathematical Modeling. 1992. Vol. 4. No. 7. P. 71–86.

Mayer-Schönberger V., Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. N.Y.: Houghton Mifflin Harcourt, 2013.

Pyt'ev Yu.P. Mathematical Methods of Experiment Interpretation. M.: HSE, 1989.

Pyt'ev Yu.P. Pseudoinverse Operators. Properties and Applications. Mat. Sb. 1982. No. 118 (160). P. 19–49 (engl. transl.: *Pyt'ev Yu.P.* Math. USSR Sb. 1983. No. 46).

Pyt'ev Yu.P. Reduction Problems in Experimental Investigations. Mat. Sb. 1982. No. 120 (162). P. 19–49 (engl. transl.: *Pyt'ev Yu.P.* Math. USSR Sb. 1984. No. 48).

Shannon C.E., Weaver W. The Mathematical Theory of Communication. University of Illinois Press, 1949.

White T. Hadoop: The Definitive Guide. O'Reilly, 2015.

FORSMEDIA — SOCIAL NETWORK ANALYTICS FOR CUSTOMER PROFILE

Olga Gorchinskaya
FORS Group, Moscow, Russia

Abstract. *There is no doubt that the information about clients is very important for any companies. Nowadays, social networks are a valuable source of information about people. We present ForSMedia — an innovative software product for the analysis of social networks customer profiles.*

Keywords: *ForSMedia, social networks, customer profiles, data analysis, big data technologies.*

ForSMedia is an advanced platform for social media data analysis for customer profiles enrichment. The important differentiator of the solution is its ability to process very large number of customer profiles automatically to find new information about customer interests, hobby, favorite music etc.

The motivation of the project was rather obvious. The information about clients is very important for companies in all industries. The more we know about people we are dealing with the more success we can achieve. Traditionally main sources of knowledge about clients were internal systems of organization. Now when social networks (SN) are becoming more and more popular a lot of interesting data can be discovered there. New big data technologies allow extract this data and convert it to information valuable for business.

What SN can tell us about people? First of all it is attributes explicitly indicated by SN users in their profile. But not only this — important facts are contained in SN implicitly. Reading posts, descriptions and other texts we can discover interesting facts about hobbies, opinions, favorite movies and so on. It is very important to get not only explicit but also implicit information

Taking into account this motivation, we created ForSMedia with the following functionality:

- Social network data acquisition;
- Customer identification in social networks;
- Linguistic processing and text enrichment from posts, subscription groups, comments;

- Revealing new customer attributes by data mining technics;
- Merging social networks user data into unified customer profile;
- Analysis and Discovery.

Let's look at each function in more details.

Collecting data from SN. We are collecting information from Social Networks, pre-processing it and store in our Hadoop cluster. It allows us to find information about clients within reasonable time. ForSMedia provides several methods to collect and monitor SN data. The first method is to use public API provided by any popular Social Network. It is very easy to integrate API into any application and this is the advantage of the method. But the quality of API can be very low. Some of SN API are very open and allow to get a lot of different data but some of them gives almost nothing. It is always possible to parse web pages, extract data and then monitor it using crawlers. This method is very labor & time consuming, but here we doesn't depend on SN owner. And the third method is to buy data from a company which collect and update all data. The main advantage — high speed of data acquisition but it may expensive. The most reasonable way to collect and monitor data is to combine all three methods depending on the requirements of a particular project.

Customer Identification in SN. Customer identification in SN means the detection of all user profiles in SN that correspond the customer data provided for identification. The initial data for identification may be first name, last name, date of birth, city. Usually it is enough to find profiles but of course additional information like company, address, phone number can narrow the search. It is important to take into account that SN data are incomplete, not standard and sometime invalid. So short name or nick name can be indicated instead of full name, and some users of SN specify only day & month of birth without the year. It means that the user profile can correspondent to a given customer only with some confidence. ForSMedia supports confidence identification level based on our original algorithms of data cleansing & normalization and score calculation.

Getting more information about SN users. After detection of user profiles which correspondent to each customer a number of attributes in the user profile are available. ForSMedia provide standardization & cleansing of names, cities, addresses and so on. It is important to mention that valuable information about SN user is contained implicitly in posts, comments, description of subscription group and other texts.

ForSMedia uses special linguistics & statistics technics to extract interests, favorite movies, etc. from these texts.

The results of text processing can be not accurate due to the ambiguity of natural language and statistical assumptions that are not always true in real life. In ForSMedia approach is the joint usage of both technics than improve the quality and accuracy of results. But in all cases a level of confidence is provided for any fact extracted from text.

Creating unified customer profile. There may be several SN users that correspond to the same customer with high confidence or identification. In this case these users should be merged into unified customer profile. It can be done in different ways depending on the requirements and tasks. For example constructing the list of interests we can select only interests presented in every matched profile or it is possible to unite all interests of all matched users into one extended list of interests. In fact, the merging algorithm can be more complicated than simple unions and intersections. We can use also confident rates resulted from linguistic processing.

Analysis of data is based on Data Discovery technology that allows using intuitive search and discovery in contrast with traditional Business Intelligence. As a tool we are using Oracle Bid Data Discovery.

ForSMedia is based on Open source software. All data are stored in Hadoop cluster with the usage of Hbase and Hive and processing is implemented with R, Python and fact extraction software from Russian linguistic company RCO. ForSMedia can be installed on any hardware platform that meets the requirements. We have also the version of ForSMedia that is certified on a special software-hardware platform — Oracle Big Data Appliance and Oracle Exalytic to deliver speed, reliability and scaleability. The first machine is Hadoop Machine and the second one is Oracle Analytics Machine.

DISTRIBUTED COLLECTION, PROCESSING AND ANALYSIS OF SENSOR DATA

Vladimir V. Korkhov

St. Petersburg State University, St. Petersburg, Russia

Abstract. *This paper presents an overview of approaches to the tasks of collecting, processing and analyzing data from sensors in various practical scenarios: collecting and processing data in unreliable low-performance computing infrastructures based on system-on-chip devices; detecting anomalies in sea vessel telemetry data; controlling smart-home lighting based on sensor data analysis, machine learning and prediction of light usage scenarios.*

Keywords: *distributed data collection and processing, sensor data analysis, system-on-chip devices, sea vessel telemetry data, smart homes, machine learning, Internet of Things.*

1. Introduction

The tasks of collecting, processing and analyzing data from sensors are becoming increasingly important in the development of modern systems for monitoring, analyzing and predicting the operation of technical systems. The constantly increasing complexity of the information environment leads to the need to collect and analyze more and more diverse data, which becomes possible with the development of the hardware and software base for the implementation of data processing systems.

In this paper we outline several data collection, processing and analysis scenarios that are investigated and implemented at the Department of Computer Modelling and Multiprocessor Systems of St. Petersburg State University (SpbSU). There are a number of motivating reasons:

- sensor data generated in the field might require immediate preprocessing: *need for cheap light-weight* infrastructure *e.g. based on SoC (system-on-chip) devices*;
- handling groups of sensors and data processing on unreliable and low-performance SoC devices requires workload management software: *light-weight fault-tolerant scheduler for distributed data processing*;
- detailed sensor data on object state, conditions and behaviour can be used to detect anomalies and predict characteristics: *intelligent* ***analysis*** *of telemetry data e.g. sea vessel motion, conditions and traffic data*;

• building models of object usage scenarios based on historical sensor data to control devices: e.g. *smart lighting control based on luminance and motion sensors and prediction models.*

We consider several practical scenarios aligned with main directions of research and development inspired by these motivating points:

• using low-cost unreliable resource-constrained hardware for data collection and processing: data collection from distributed sources (sensors); data pre-processing in the vicinity of the source;

• detection of anomalies in sea vessel movement using telemetry data: processing large amounts of data using distributed computing systems; intelligent data analysis, detection of dependencies, anomalies, prediction of characteristics and properties of the system;

• using IoT-related technologies for smart lighting, water consumption control: use of sensory data for equipment management, organization of direct interaction of devices without intermediary links.

One of the latest trends in the management of distributed devices using data from sensors installed on them is the Internet of Things (IoT) technology [Chaouchi, 2010]. Significant part of existing approaches and solutions for organizing distributed computing systems can be adapted and applied in the context of IoT. In particular, solutions are needed to organize distributed data collection and processing based on microcomputers (single-board microcomputer devices, e.g. Raspberry Pi and Intel Edison) using scheduling, workload distribution and load balancing systems [Gankevich et al., 2016*b*]. One example of such hardware and software systems is the system for data collection and processing developed in St. Petersburg State University with dynamic load distribution on unreliable and low-power resources [Korkhov et al., 2017]. The purpose of this system is to organize data processing in real time and conduct general computing alongside high-tech equipment to minimize the transfer of sensor data to remote servers. The result of the development was the software and hardware system for organizing distributed processing of data on low-power and unreliable resources (Raspberry Pi and Intel Edison single-board microcomputers), connected by wireless technologies, supporting dynamic load balancing, fault tolerance (including failover to the master node), the elasticity of the computer system. The system is integrated with the popular package for distributed data processing Apache Spark [Zaharia et al., 2010]: the developed system can be used both in conjunction with Apache Spark, and separately. A study of the performance of the

developed system showed that on devices of the microcomputer class the created system exceeds the standard version of Apache Spark by several times.

Another area of work with sensor data is the intellectual data analysis to identify dependencies, anomalies and predict future properties and system characteristics. One of the examples of areas requiring such analysis is the analysis of data on the movement of sea vessels, detection of abnormal situations, analysis of the actions of the captain and the crew of the vessel. The application of big data technologies, machine learning methods, allows to identify unusual situations based on ship telemetry data and position data of other vessels to predict dangerous approaches, prevent and avoid them [Lind et al., 2016].

Finally, we consider scenarios of applying IoT technologies to use data from sensors for controlling technical systems without direct human intervention. One of the examples is the smart lighting control system based on the collection and processing of data from motion and light sensors installed in corridors and rooms, machine learning to discover patterns of room usage and subsequent automatic control of room lighting to save electricity while keeping high level of comfort for room users.

2. Using unreliable resource-constrained hardware for data collection and processing

The problem of building distributed computing infrastructures for data collection and processing has been around for many years. One of the well-known technologies for building large-scale computing infrastructures is grid computing. It provides means to connect heterogeneous, dynamic resources into a single metacomputer. However, being focused on high-performance computing systems, grid technologies do not suit well other classes of basic hardware. One of such examples are low-performance, low-cost unreliable microcomputers similar to Raspberry Pi or Intel Edison, sometimes also called System-on-Chip (SoC) devices. To be able to execute distributed applications over a set of such devices extensive fault-tolerance support is needed along with low resource usage profile of the middleware.

We have developed and presented an approach to orchestrate distributed computing and data processing on microcomputers with help of custom scheduler focused on fault tolerance and dynamic re-

scheduling of computational kernels that represent the application. This scheduler, the latest open generation of which is named Bscheduler [Bscheduler], provides its own low-level API to create and manage computational kernels. The scheduler is built on the ideas and approaches presented in [Gankevich et al., 2016*a*; *b*]. In addition, we considered possibilities to integrate the scheduler into Apache Spark [Zaharia et al., 2010] data processing framework instead of the default scheduler used by Spark. This opened possibilities to use a wide range of existing Spark-based programs on the underlying microcomputer infrastructure controlled by the scheduler. The project aimed to solve the following main tasks:

- Develop automatic failover and high-availability mechanisms for computer system.
- Develop automatic elasticity mechanism for computer system.
- Enable adjusting application algorithm precision taking into account current number of healthy cluster nodes.
- Adjust load distribution taking into account actual and heterogeneous monitoring data from cluster nodes.
- Adjust micro-kernel execution order to minimize peak memory footprint of cluster nodes.

The task of data processing on resource-constrained and unreliable hardware emerges within the framework of sensor real-time near-field data processing. The implementation of the system, allowing to carry out the processing in the field, will allow one to quickly respond to sudden changes in sensor readings and reduce the time of decision-making. The implementation of general-purpose computations in such a system allows one to use the same hardware and software system for diverse high-tech equipment. Figure 1 presents evaluation results of

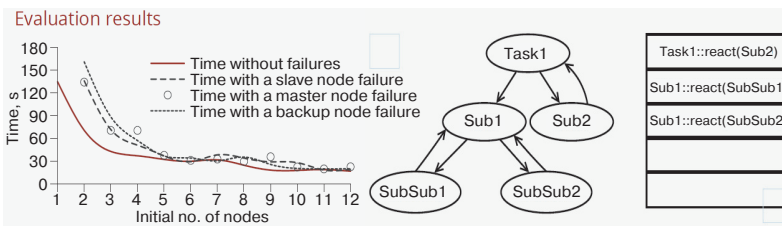


Fig. 1. Scheduler evaluation results and inter-task transactional communication scheme

the scheduler performance with various types of failures (worker node failure, master node failure) and depicts the scheme of computational kernels hierarchy.

The following was achieved as the final outcome of the project [Korkhov et al., 2017]:

- Fault-tolerant scheduler implemented running standalone or with Apache Spark (with Spark Streaming supported).
- Master-node fault tolerance is supported.
- Dynamic resource discovery, composition and re-configuration of distributed cluster.
- Optimized for running on unreliable and resource-constrained microcomputer hardware.
- Running in heterogeneous and dynamic hardware and networking environment.
- Integrated microcomputer and cluster monitoring API.

The task of data processing on low-power and unreliable computers arises within the task of processing sensor data in real time and in the field. The implementation of the system, which allows processing in the field, gives an opportunity to respond quickly to abrupt changes in the sensor readings and to shorten the decision-making time. The implementation of general purpose computing in such a system allows using the same hardware and software system for heterogeneous high-tech equipment.

3. Detection of anomalies in sea vessel movement using telemetry data

Analysis of sea vessel telemetry data is often needed to determine characteristics and behavior of the vessel both during the trip and after the trip is finished. One of the important tasks is to detect anomalies in ship movement and operation to analyze correctness of crew actions, detect abnormal situations caused by other ships approach and discover unusual combinations of ship telemetry data which includes: latitude, longitude, vessel speed and course over ground and over water, rate of turn, depth, wind speed and angle, values of roll, pitch, sway, surge, heave, yaw, fuel consumption. Currently many machine learning methods are developed for anomaly detection and they can be applied to this task with additional modifications and tuning to particular scenario of sea navigation. Figure 2 shows sample anomalies on roll/pitch distribution.

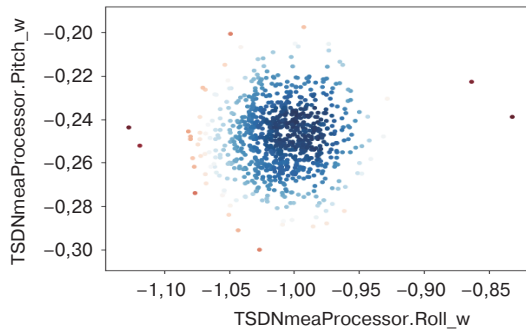


Fig. 2. Roll/Pitch anomalies

Classes of anomalies of interest that can be detected and used to signal about increased risk of collisions, unexpected events, non-optimal operation mode etc.:

- Maneuvering anomalies
  – Abrupt (rough) maneuvers; turns/speed distribution.
- Prerequisites for collisions
  – Depth/speed; maneuvers with course and speed, taking into account CPA (Closest Point of Approach);
  – TCPA (Time to CPA); fuzzy logic for maneuvers with course and speed.
- Stability
  – Detection of 1% of the most noticeable rolls; roll/pitch distribution.
- Marginal conditions
  – Drift angle/speed distribution.
- Manipulating controls
  – Shaft RPM control; Rudder control.

Collected telemetry data together with weather forecast can be used to predict different parameters of ship behavior, in particular fuel consumption. Various methods of prediction and their combinations can be applied, in particular linear regression, gradient boost and random forest methods. Figure 3 illustrates obtained results for fuel consumption prediction (linear regression + gradient boost). Python and Apache Spark are used for the implementation.

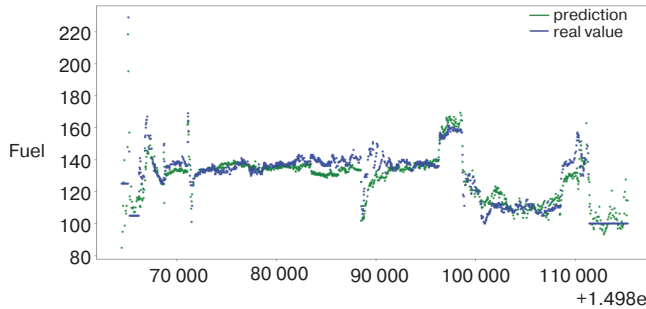


Fig. 3. Fuel consumption prediction

4. Internet of Things and Cyber-Physical Systems

Using the Internet of Things technologies enables automatic and consistent use of sensor data to control technical systems without direct human intervention [DEUS ME6]. One of the examples of projects in this field that SPbSU is working on is the implementation of a smart lighting system based on the collection and processing of data from motion and luminance sensors installed in corridors and rooms. Based on the statistics on the collected data, the model of using the premises is constructed using the methods of machine learning. This model is further used to automatically control indoor lighting, which ultimately will significantly improve economic efficiency through the rational use of electricity with the provision of a high level of light comfort for people using the premises.

Our projects using IoT/CPS-related technologies for smart lighting, water control, navigation:

- Smart lighting control: pilot project on cloud-based smart-home lighting control based on sensor data analysis, machine learning and prediction of light usage scenarios.
- Water sensors: using autonomous GSM data transmitters, which periodically transmit measurements of water consumption to the data collector and analysis servers; build an hourly estimation of current urban water pipes load and produce their usage forecasts.
- Indoors location and navigation with beacons.

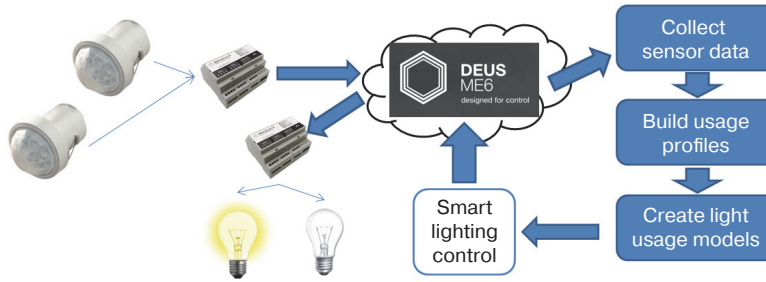


Fig. 4. Schematic view of data collection from light sensors, analysis and processing with help of DEUS platform

Cloud-based intelligent lighting control system (Fig. 4):

• Secure platform that collects and consolidates data on lighting and energy consumption from various sites to enable energy-efficient lighting control (currently based on DEUS ME6 solution [DEUS ME6]).

• Distributed sensors and lighting devices interconnect and exchange data on light use.

• Data analysis algorithms: electricity consumption for lighting in dynamics; monitor current status of lighting devices.

• Machine learning: train light control algorithms, discover light usage patterns, predict light use and automatically control light behavior based on this knowledge.

• Cloud service: create models of rooms by uploading room plans, make the arrangement of devices, combine the devices in groups and describe initial rules of light functioning.

Smart infrastructure for water consumption control:

• Based on distributed sensors, water meters, GSM communication between devices, data collection and analysis to fully automate water consumption technical billing.

• Big data analysis allows users to build an hourly estimation of current urban water pipes load and produce their usage forecasts.

• System of customer profiles is used to estimate sectoral water consumption.

• Profiles allow users to accurately estimate and predict water consumption of similar objects in one class, even with incomplete data.

• The system of such devices is self-sufficient and constantly expanding. Currently, there are about 13,000 of such devices that cover half of the city, where about 2 million people live.

Indoors location and navigation with Bluetooth Low Energy (BLE) Eddystone beacons:

- Using beacons to navigate patients in hospitals.
- Eddystone is an open, scalable BLE beacon format that allows developers to create contextually aware experiences on both Android and iOS devices; supports multiple types of broadcast signals, or in BLE terms, "frames".

5. Conclusion

There is a variety of existing sensor data in many fields that needs to be collected, processed and analyzed. In this paper we presented a short overview of possible problems and approaches with prototype implementations, starting from computing infrastructures for data collection and processing up to smart systems based on data analysis and machine learning methods for automatic control of technical systems, devices and processes.

The research was partially supported by Russian Foundation for Basic Research (project No. 16-07-01111).

References

Bscheduler. <https://igankevich.github.io/bscheduler> (access: 10.12.2017).

Chaouchi H. The Internet of Things. L.: Wiley-ISTE, 2010.

DEUS ME6: Cloud lighting management system. <https://me6cloud.com/> (access: 10.12.2017).

Gankevich I., Tipikin Y., Korkhov V., Gaiduchok V. Factory: Non-stop Batch Jobs without Checkpointing. International Conference on High Performance Computing and Simulation, HPCS. 2016*a*. No. 7568441. P. 979–984.

Gankevich I., Tipikin Y., Korkhov V. et al. Factory: Master Node High-Availability for Big Data Applications and Beyond / by ed. O. Gervasi et al. ICCSA 2016*b*. Pt II: Lecture Notes in Computer Science., 2016. Vol. 9787. P. 379–389.

Korkhov V., Gankevich I., Iakushkin O. et al. Distributed Data Processing on Microcomputers with Ascheduler and Apache Spark. Lecture Notes // Computer Science 2017. Vol. 10408. P. 387–398.

Lind M., Hägg M., Siwe U., Haraldson S. Sea Traffic Management — Beneficial for all Maritime Stakeholders // Transportation Research Procedia. 2016. No. 14. P. 183–192.

Zaharia M., Chowdhury M., Franklin M.J. et al. Spark: Cluster Computing with Working Sets. HotCloud 2010. June 2010.

BIG DATA: AN INFORMATION SYSTEMS APPROACH

Jacky Akoka
Faten Atigui
Isabelle Comyn-Wattiau
Fayçal Hamdi
Elena Kornyshova
Nadira Lammari
Elisabeth Métais
Cédric du Mouza
Nicolas Prat
Samira Si Saïd-Cherfi
CNAM, Paris, France

Abstract. *Several challenges and issues characterize the research on Big Data including in social media data, mobile data, web data, and network data. While much has been written in terms of technologies and algorithms, much less has been written about methodologies and frameworks that could enable researchers and practitioners to more efficiently tackle the Big Data challenges and issues they face. The aim of the ISID Group at CNAM CEDRIC is to propose methodologies, approaches, and frameworks in order to cope with these challenges. This paper presents issues of the Big Data research from the Information System point of view and describes different research topics of the ISID Group.*

Keywords: *Big Data, Information System, reverse engineering, process methodology, security, quality, data warehouse, Social Web.*

Several challenges and issues characterize the research on Big Data. Let us mention some of these issues:

1. There is clearly a need to define a Big Data process methodology.

2. Big Data has challenging security and privacy problems requiring new approaches. Let us remind that privacy and security are the most important Big Data issues including conceptual significance. Big Data anonymization and privacy can be combined in an integrated framework.

3. Given the range of Big Data applications, there is a need to develop methodologies to tackle the data quality issues. It is obvious that bad data quality in the context of Big Data can lead companies to disas-

trous situations. Big Data quality needs to be conceptualized in a framework.

4. Approaches for data integration issues that arise in many real-life settings are needed.

5. Issues of Big Data information provenance require new approaches.

6. Since Big Data is unstructured, traditional analysis methods are insufficient to analyze huge volume of data.

7. Representation of unstructured data coming from different sources require the use of specific methods.

8. Theoretical foundations for Big Data especially based on an ontology is missing.

9. An ontology-based approach to Big Data analytics remains to be developed. It can be used to address the semantic challenges presented by big and unstructured data sets.

10. Framework for Big Data-driven risk management is yet not available.

11. A conceptual model for Big Data is not well developed.

12. Big Data warehouse conceptual model, based on MDA, has been proposed but not fully tested and validated.

Reverse engineering and Integration of Big Data

Chikofsky and Cross [1990] defined reverse engineering to be "analyzing a subject system to identify its current components and their dependencies, and to extract and create system abstractions and design information." Existing reverse engineering methods and tools focus on extracting the structure of a legacy system with the goal to reengineer or to reuse it. In the context of Big Data, incorporating data from various sources (internal and external) in a data warehouse or in a Big Data warehouse requires a data model. However the variety of data, its volume, and its velocity create problems difficult to solve. Conceptual modeling of Big Data appears to be one of the challenges facing researchers. Various conceptual modeling techniques, such as ontologies, semantic, RDF schemas, SPARQL Language, etc., are not well suited for Big Data conceptualization. We propose a new approach based on reverse engineering and schema integration techniques.

Processes and Methods in Big Data

Situational Method Engineering (SME) offers a wide range of approaches allowing to adapt methods and processes to a given situation: method family configuration [Kornyshova, 2011], method extension [Deneckère, 2001], contextual reconstitution from method components [Ralyté, 2001] etc. Information Systems engineering methods should be revised to fit the context of Big Data and to be able to deal with scalability, velocity, variety, variability and other Big Data issues. Our approach to manage this problem is to use the SME techniques to adapt different processes and methods in the context of Big Data in the following manner. Methods or processes conceived for the same purpose are analyzed to identify the common and variable components, then a meta-method (or a meta-process) is constructed from the initial ones, finally a set of guidelines is suggested to define a method or to execute a process adapted to a given case. These contextual method definition and process execution are done using decision-based guidance.

Data and Data Interlinking Quality

Several Data and Linked Open Data issues are actually investigated within the ISID team. The identification of quality defects related to data and data interlinking quality. Several recent contributions from literature pointed out the lack of quality [Halpin, 2010; Zaveri, 2015]. The challenge is to investigate several interlinked data sources from several domains and try to qualify the underlying quality defects. Once quality criteria identified, it is necessary to associate to each criterion a set of assessment methods and algorithms. The detection of quality defects raises the problem of correction that is not sufficiently addressed in literature where quality stands more for evaluation than correction. The proposed solutions should be implemented through a prototype or a platform to support Data Interlinking evaluation and improvement. Finally, the problem of scalability should be addressed as Open Data is also Big Data and developed solutions should be scalable.

Data Security & Privacy

Following a security incident, a security analysis is often required. It focuses first on the traces of computer systems (logs) in order to re-

construct what has happened and deduce the attacker's mode of operation. Due to the explosion of connected objects and the proliferation of online and surfing activities carried out on social networks, Internet induces the appearance of numerous and various traces. To cope with this enormous amount of data, security analysts have equipped themselves with computer tools known as SIEM (Security Information and Event Management). The latter, given the difficulty of automating the complex process of human reasoning, generate confusing results and multitude of false positives and false negatives. For instance, the HuMa project will capitalize on the complex reasoning of security experts in order to introduce guidance in SIEM tools and to permit to security analysts to react on the fly.

Data Warehouse Modeling

We have worked on data warehouse and dimensional modeling for many years. At the era of Big Data warehousing, we aim to define logical models for each noSQL database family (column-based, graph, key-value, and document) and mapping rules between UML/logical and physical noSQL levels. Until now, the noSQL community does not refer to logical or conceptual models. We claim that adding such models will considerably facilitate the handling of Big Data warehouses.

Social Web and Recommendations

Micro-blogging platforms such as Twitter, Pinterest, Instagram, Weibo or Tumblr all share this mechanism of selecting interesting people to follow and being followed by other users which is now well established in the Internet culture. But with this success, microblogging platforms began to be very crowded and users started having issues to keep up with all the content available. We investigate specific collaborative filtering methods to recommend items of interest, based on both content and topology of the underlying social network, which can scale up to very large datasets [Constantin16]. In this context we have also investigated edge-partitionning solution to allow intensive graph-computation on very large graphs [Li, 2016]. Finally we have proposed in [Pozo et al., 2016*a*] a distributive collaborative filtering system based on the semantic knowledge of the domain in order to help the user in the e-commerce area. The main issue with this approach concerns

cold-start. In [Pozo et al., 2016*b*] we have used bloom filters and parallelization to deal with scalability.

References

Chikofsky E., Cross J. Reverse Engineering and Design Recovery: A Taxonomy // IEEE Software. 1990. No. 7 (1). January. P. 13–17.

Constantin C., Dahimene R., Grossetti Q., du Mouza C. Finding Users of Interest in Micro-blogging Systems. EDBT. 2016. P. 5–16.

Deneckere R. Approche d'extension de méthodes fondée sur l'utilisation de composants génériques. PhD thesis. University of Paris 1 Panthéon-Sorbonne, 2001.

Halpin H., Hayes P.J., McCusker J.P. et al. When Owl: Sameas Isn't the Same. An Analysis of Identity in Linked Data // The Semantic Web–ISWC 2010. 2010.

Kornyshova E., Deneckère R., Rolland C. Method Families Concept: Application to Decision-Making Methods. EMMSAD. L., 2011.

Li Y., Constantin C., du Mouza C. A Block-Based Edge Partitioning for Random Walks Algorithms over Large Social Graphs // WISE. 2016. Vol. 2.

Pozo M., Chiky R., Metais E. Enhancing Collaborative Filtering by Using Implicit Relations in Data. LNCS Transactions on Computational Collective Intelligence (TCCI). 2016*a*. Vol. 9655.

Pozo M., Chiky R., Meziane F., Metais E. An Item/User Representation for Recommender Systems based on Bloom Filters. RCIS. 2016*b*.

Ralyté J., Rolland C. An Assembly Process Model for Method Engineering // Proceedings of CAISE 2001. Berlin: Springer, 2001.

Zaveri A., Rula A., Maurino A. et al. Quality Assessment for Linked Data: A Survey // Semantic Web. 2015. No. 7 (1).

BIG DATA AND THE NEXT INFORMATION REVOLUTION

Mikhail Lugachev (Prof., Dr. Sc.)
IBS, Moscow, Russia

Abstract. *The concept of permanent revolution was formulated in the XIX century by Karl Marx and became a subject of constant debate in humanities circle. In contrast-scientific and technological revolutions are natural components at all steps of human development. Their permanence is commonly recognized imperative, followed by numerous confirmations with a convincing inevitability. Information and industrial revolutions taking place now in the world are such evidences. Experts declare today the fourth industrial revolution. Peter Drucker farely predicted the fourth information revolution. It is interesting that the most important trait of both revolutions is the artificial intelligence which functions in the sphere of Big Data and Internet of Things. The application field (not the only) is the economy-its structure and content. Experts state the emergence of information capitalism and the information economy — innovations obtaining special and revolutional traits.The article is devoted to analysis of main components of the innovations and offers the ways how they should be reflected in the curriculum for modern economists and managers.*

Keywords: *information revolution, industrial revolution, information capitalism, Big Data, curriculum.*

It is believed that the modern world has gone through several stages of development of capitalism. The natural evolutionary way trade capitalism appeared to ensure that the successful functioning of financial capitalism emerged. Industrial capitalism came into the world economy through the struggle and the resistance of the working masses thanks to the industrial revolution. Today, researchers are talking about the information revolution, as a fait accompli and we consider a necessary consequence of this phenomenon: information capitalism.

At the end of the last century, Peter Drucker in his article "The next information revolution" [Drucker, 1998] stated: "So far, for 50 years, the information revolution has centered on data — their collection, storage, transmission, analysis, and presentation. It has centered on the 'T' in IT. The next information revolution asks, What is the MEANING of information, and what is its PURPOSE? And this is leading rapidly to redefining the tasks to be done with the help of information, and with it, to redefining the institutions that do these tasks." After almost two decades, we see how visionary was this prediction. IT giant Google has

become a world leader in the use of market information processes and its experience already provides a basis for generalizations. Chief Economist at Google H. Varian in a series of recent publications [Varian, 2010; 2014] analyzes business activities of their company, and notes that from the outset, with the advent of the computer as a means of implementing the transaction — fundamentally changed not only the accumulation of data and the use of processes, but also own decision ecosystem.

Decision ecosystem is largely determined by the provision of information processes, representing the information lifecycle — from data collection, storage and processing them — pending a decision on the basis of information obtained from the data collected. Today, in this environment are actively developing Big Data.

With the advent of Big Data decision-making environment changes reached fundamentals of the global economy — the processes of capital accumulation. Data become an asset. Among large IT data are converted into an instrument of profit directly from the information processes. This business is completely of IT origin: automatically search, data collection and storage, processing, search for users — potential customers, delivering them with information, account activity and payment from customers. Big Data demonstrate the basic property of IT instruments — adding value: data on the Internet are not worth anything, but after going through the algorithms for Big Data acquired consumer properties information may interesting to advertisers. For this feature, advertisers pay generously. Shoshana Zuboff [2015]: "I explore the proposition that 'Big Data' is above all the foundational component in a deeply intentional and highly consequential new logic of accumulation that I call **surveillance capitalism**. This new form of information capitalism aims to predict and modify human behavior as a means to produce revenue and market control."

Internet users — citizens of the virtual state, which is ruled by the information capitalism — are totally rightless: they are not asked when one picks up their data and — without any consent they collect information in the form of advertising. How dumb slaves they dutifully perform their duties: they are clicking on the interesting icon in the social networks and marketplaces. There is nothing forcing them to act in this manner — except formed need in comfortable collection the information you need to know. So they hope to reduce the level of uncertainty of the existence. But freedom of the uncertainty is not freedom at all, or freedom of choice in the surveillance capitalism. Beneficiaries of in-

voluntary information servitude do not care about the creators of their capital: they do not care who they are, what are the circumstances of their existence. Their actions are directed primarily to users attempting to tie them stronger to the network by offering new search features or unusual sources of information forcing them to increase the number of clicks, and so to increase their capital.

However, Google deliberately ignores the existence of the consent of the monitored transactions for the discreet monitoring: a large part of companies and citizens "lend" their data unknowingly. For others — the trials are planned in the company's part of the overall business process and the possible costs of the super-lawyers obviously overlap obtained profits.

This manner of doing business (popular and for other mega-integrators) helped to accelerate some world economic processes and the entire economic picture of the world has changed dramatically over the past decade. According visualcapitalist.com — among the 5 largest companies in the world by capitalization was only one in 2006 related to IT — Microsoft. Other places confidently held companies representing traditional business: oil and the TOTAL EXXON, a diversified corporation — General Electric and financial giant — Citibank.

In 2016 on the first place was the APPLE, on the second — Alphabet, parent, GOOGLE, on the third — Microsoft, on the fourth — the world's leading Internet-trade — Amazon. On the fifth broke Facebook company. No one of leaders does have any relation to the physical product that provides only communication services on the Internet.

Peter Drucker has not found at the end of the twentieth century sufficient grounds for declaration of the fourth information revolution. According to him at the time it was not recorded such changes, which could be compared in scale to have taken place during the previous one-third of the revolutions — the appearance of book printing. Then, thanks to the new possibilities of copying texts and images the world has received new learning opportunities — were secular books and any universities, surface maps of Earth pushed travel horizons and new lands were opened. It shook the religious world order. Nothing comparable with that global impact in the world community the actual information technology has demonstrated yet — according to P. Drucker. It seems that today, 20 years later, he would have changed his mind. The fourth information revolution with the advent of Big Data in the real economy — was accomplished. There was also formed global information

economy — all the properties and effects of which we have yet to learn. Researchers believe that we are on the way to an information civilization.

Whether the information capitalism becomes the dominant logic of accumulation in our time? What alternative path to the future can be associated with these competing forms? We are at the beginning of a new story that will lead us to new answers.

Modern education must respond to the major challenges of contemporary development of the information economy. The questions discussed in this article are waiting for answers in the educational process of economists. It seems that this situation should be of interest to specialists departments of economics and political economy. Indeed, history is repeating itself. Including economic. As for the industrial revolution came into the world of industrial capitalism, information capitalism has entered into the world economy after the accomplished information revolution. A new economic phenomenon requires careful study by professionals and beginners economists. There are many new and immediately you can see that from the old habitual knowledge may be used for the study of this economic innovation. In fact — the real unlimited resources and zero marginal costs give rise to a new economy. In some sectors of the information economy is dominated by the unusual composition of capital: Intangible Assets (Data), the productive forces — also virtual — are software agents. As in the information capitalism will shape political life when fundamental workers are real and virtual robots, and the relationship between employers and employees takes place in the absence of the social contract. The behavior of voters — people are fully controlled, and their decisions may be subject to a targeted correction. What kind of society will form the information capitalism?

In any case, there is a wide field for research and development of new curricula, relevant major challenges of the real world economy.

References

Drucker P.F. The Next Information Revolution // Forbes ASAP. 1998. August 24.

Varian H.R. Computer Mediated Transactions // American Economic Review. 2010. No. 100 (2).

Varian H.R. Beyond Big Data // Business Economics. 2014. No. 49 (1).

Zuboff S. Big Other: Surveillance Capitalism and the Prospects of an Information Civilization // Journal of Information Technology. 2015. No. 30.

RULE-BASED CLASSIFICATION APPROACH: CLOSED ITEMSETS VS RANDOM FORESTS

Tatiana Makhalova
Sergei O. Kuznetsov
National Research University Higher School of Economics,
Moscow, Russia

Abstract. *Random Forests are proved to be a good model for classification problem. Nowadays they are widely used as a part of deep architectures. The quality of the Random Forests are highly depend on such parameters as the size of bootstrap samples, the minimal size of leafs or the depth of trees as well as the size of subsets of features used to grow a tree. In this work we study the question of computing a set of rules in a deterministic way in order to build an ensemble of classifiers (i.e. closed itemsets) having accuracy comparable to Random Forests.*

Keywords: *Random Forests, Rule-Based Classification, Big Data, Deep Neural networks.*

1. Introduction

Random Forests is one of the best models both for classification and regression tasks. They were proposed in 2001 [Breiman, 2001], they rapidly spread to a variety of the different domains and demonstrate high performance [Marchese Robinson et al., 2017; Huo, Shi, Chang, 2016; Youssef et al., 2016], especially in combination with boosting techniques [Freund, Schapire, 1995].

Recently, in the "Big Data" decade, Deep Networks are be coming more and more popular, they can successfully solve a broad spectrum of different tasks. Nevertheless, Random Forests (RFs) is still a very useful model, even in "Big Data"-applications. One of the most promising approaches is to incorporate this model into a deep network. In [Bengio et al., 2009] it was proposed to use trees on the top layers of a neural network, while the lower layers of the network are used to perform data regularization (in an unsupervised manner).

In a similar way, concept lattices are proposed as underlying architecture of neural networks [Kuznetsov, Makhazhanov, Ushakov, 2017]. The distributed algorithms for dealing with RFs as well as neural-tree RFs have been proposed in [Yildiz, Alpaydin, 2013]. Deep Neural Decision Forests [Kontschieder et al., 2015] combine the ideas of deep

convolutional networks and RFs to create a model based on stochastic, differentiable decision trees. In a similar way lazy classification using concept-based classifiers can help to avoid a biased classification of a single decision tree [Kuznetsov, 2013*a*; *b*].

In this paper we study the problem of concise representation of Classification RFs in order to boost the classification stage. In Section 2 the basic approaches to building ensembles of trees are described and the issue is related to the similar-tree generation is discussed. Section 3 introduces a theoretic model called Formal Concepts that allows overcoming the mentioned difficulty. In Section 4 we intro duce a new "Concept-based" classifier (CBC), which is based on formal concepts and represents all similar rules as one classifier. In the second part of this section we also describe the results of the comparative analysis of CBCs and RFs.

2. Ensembles of trees. The constructing approaches

The general approach to building an ensemble of trees is to create several trees trained on different subsamples of data. For example, bagging implies computing trees on randomly selected subsets of objects. Statistical properties of the resulting model depends on the size of subsampled sets. Random Forest combines trees trained on the whole set of objects which are described by a random subset of attributes. Combination of the two approaches is also applied in practice. In RFs of this kind the more shallow trees in an ensemble, the more similar classification rules comprise the ensemble. To illustrate this problem let us consider a small example.

Example. The input data is given by Table 1. The rows of the table correspond to objects and the columns correspond to attributes. A unit table entry means the respective object has the respective attribute. The last column stays for the target attribute.

Based on the given data we compute a random forest, the trees in the forest are only 4 types given on Fig. 1. Moreover, all the classification rules extracted from trees describe exactly the same subset of object, i.e. they are equivalent in terms of classification.

The reasonable questions that follows directly from the example above is "Whether such kind of repetition has sense and can we get rid of exhaustive computations?" It is clear that both training and classification stages involve excessive computations. Further, we tackle this issue by means of Formal Concept Analysis (FCA).

Table 1

A formal context with 6 attribute and a target

Name	4 legs	Hair	Yellow-brown	Grown	Black-brown	Target: animal
Sphynx cat	1	1	0	1	1	1
Dog	1	1	0	1	1	1
Cat	1	1	1	1	0	1
Leopard	1	1	1	1	0	1
Fur coat	0	1	1	0	0	0
Chair	1	0	1	0	0	0
Sunflower	0	0	1	1	0	0
Balloon	0	0	0	1	1	0

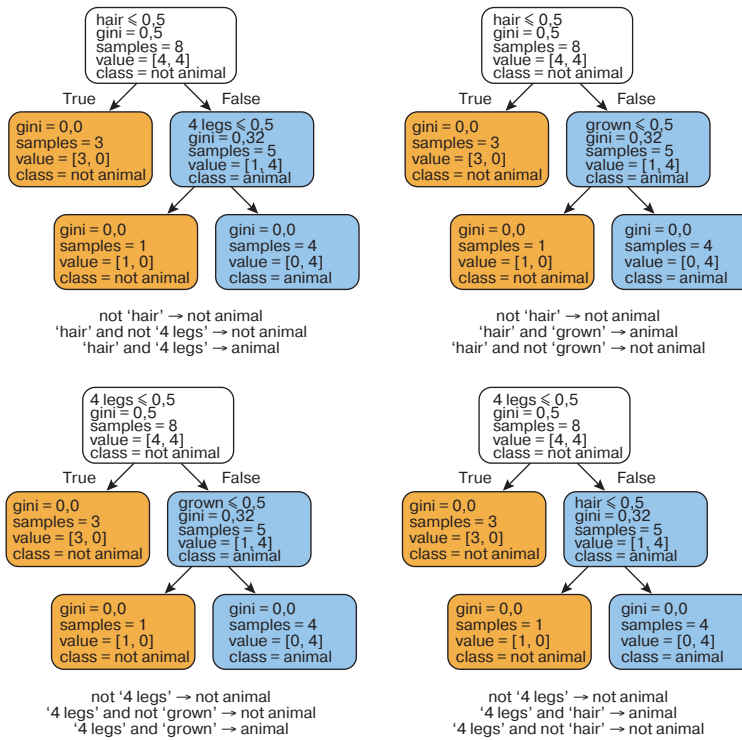


Fig. 1. Different trees of a random forest computed on data given in Table 1

3. Formal Concept Analysis

Here we give the main definitions of FCA [Ganter, Wille, 1999]. A formal context is a triple (G, M, I), where G is called a set of objects, M is called a set of attributes and $I \subseteq G \times M$ is a relation called incidence relation, i.e. $(g, m) \in I$ if object g has attribute m. The derivation operators $(\cdot)'$ are defined for $A \subseteq G$ and $B \subseteq M$ as follows:

$$A' = \{m \in M \mid \forall g \in A : gIm\}$$
$$B' = \{g \in G \mid \forall m \in B : gIm\}$$

A' is the set of attributes common to all objects from A and B' is the set of objects sharing all attributes from B. The double application of $(\cdot)'$ is a closure operator, i.e. $(\cdot)''$ is extensive, idempotent and monotone. Sets $A \subseteq G$, $B \subseteq M$, such that $A = A'$ and $B = B''$ are said to be closed.

A (formal) concept is a pair (A, B), where $A \subseteq G$, $B \subseteq M$ and $A' = B$, $B' = A$. A is called the (formal) extent and B is called the (formal) intent of the concept (A, B). A partial order $\leqslant$ is defined on the set of concepts as follows: $(A, B) \leqslant (C, D)$ if $A \subseteq C$ $(D \subseteq B)$, a pair (A, B) is a subconcept of (C, D), while (C, D) is a superconcept of (A, B). The partially ordered set of concepts of a context forms a lattice, i.e. each pair of concepts has supremum and infimum wrt. $\leqslant$.

Example. Let us consider formal concepts from the running example (see Table 1). The formal concepts are maximal subtables filled with ones. As can be seen from Fig. 2, concept "({cat, dog, leopard, Sphynx cat}, {4 legs, tail, grown})" represents all the possible accurate trees in the Random Forest (see Fig. 1). Thus, classification with one classifier can be much faster than classification with a bunch of equivalent trees.

4. Concept-Based Classifiers

In this section we introduce a classification model that is based on formal concepts. We also describe the results of a comparative study showing that CBCs have accuracy comparable with that of RFs.

4.1. Classification Model (2)

Let us consider a formal context $K_{\pm} = (G, \{M, w\}, I)$, where $G = G_+ \cup G_- \cup G_\tau$ is a set of objects comprised of positive examples G_+, negative examples G_- and undetermined examples G_τ which need to be

classified. The objects are described by attributes from set M and a target attribute w is defined only for G_+ and G_-, i.e. $gIw = '+'$ for $g \in G_+$ and $gIw = '-'$ for $g \in G_-$.

A set of classifiers S is computed for context $K = (G_+ \cup G_-, M, I)$ incrementally, from the most general concepts to more specific, where the latter describe smaller subsets of objects.

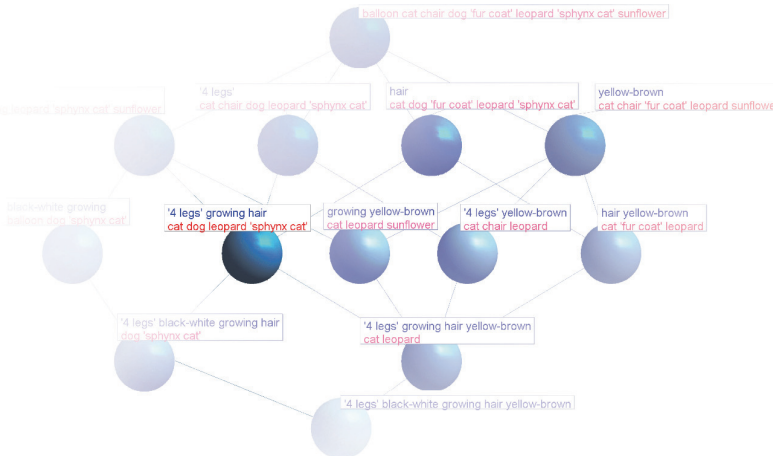


Fig. 2. A concept lattice computed on a context from Table 1

For each concept $C = (A, B)$ the rate of positive/negative examples is computed and the class of majority of examples is assigned to the concepts, i.e.

$$class(C) = \arg\max_{s \in \{+,-\}} \left(\frac{|A_s|}{|A|} \right).$$

To classify an object $g \in G_\tau$ we consider its attributes g' and compute the number of positive and negative concepts those intents are included in g', i.e. $n_s = |\{B \subseteq g' \mid (A,B) \in S, class((A,B)) = s\}|$. The class of majority of the included concepts is assigned to g:

$$class(g) = \arg\max_{s \in \{+,-\}} \left(\frac{n_s}{|\{C \in S \mid class(C) = s\}|} \right).$$

4.2. Experiments

To study the performance of the proposed model we randomly generated 4 bunches of datasets (see details in [Guyon, 2003]). These bunches differ from each other's by the rate of informative/random attributes. All of the datasets consist of 100 object and 20 attributes with the following quantities of informative/random attributes: 20/0, 15/5, 10/10, 5/15. Each dataset is randomly split into training set of 75 objects and test set of 25 objects.

We measure the accuracy of models depending on their depth, i.e. the amount of objects that are classified by leaves of the trees/the most specific concepts.

As we see from Fig. 3, CBCs performs not worse than RFs. More than that, RFs tend to overfitting as the depth of a model increases.

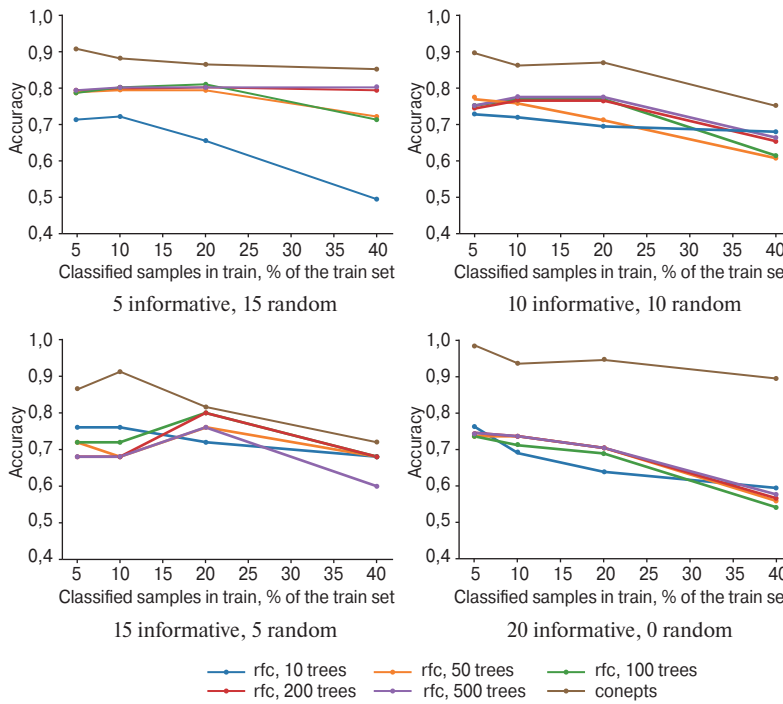


Fig. 3. The accuracy of concept-based classifiers and random forest depending on the depth of a model

Conclusion

In this paper we address the problem of computing concise Random Forests. We have proposed a new classification model based on formal concepts and study its accuracy. The experiments performed on synthetic data show that accuracy of the newly-proposed model is at least not worse than the accuracy of Random Forests.

One of the directions of future work is to generalize the proposed approach to classification of non-binary data, which can be done by using the extension of FCA to arbitrary partially ordered descriptions called Pattern Structures [Ganter, Kuznetsov, 2001].

References

Bengio Y. et al. Learning Deep Architectures for AI. Foundations and Trends // Machine Learning. 2009. No. 2 (1). P. 1–127.

Breiman L. Random Forests // Machine learning. 2001. No. 45 (1). P. 5–32.

Freund Y., Schapire R.E. A Desicion-theoretic Generalization of Online Learning and an Application to Boosting // European Conference on Computational Learning Theory. Springer, 1995. P. 23–37.

Ganter B., Kuznetsov S. Pattern Structures and Their Pr jections / Conceptual Structures. Broadening the Base. 2001. P. 129–142.

Ganter B., Wille R. Formal Concept Analysis Mathematical Foundations. 1999.

Guyon I. Design of Experiments of the Nips 2003 Variable Selection Benchmark. 2003.

Huo, J., Shi T., Chang J. Comparison of Random Forest and svm for Electrical Short-term Load Forecast with Different Data Sources. Software Engine Ering and Service Science (ICSESS), 7th IEEE International Conference. 2016. P. 129–142.

Kontschieder P., Fiterau M., Criminisi A., Rota Bulo S. Deep Neural Decision Forests. Proceedings of the IEEE International Conference on Computer Vision. 2015. P. 1467–1475.

Kuznetsov S.O. Fitting Pattern Structures to Knowledge Discovery in Big Data / ed. by P. Cellier, F. Distel, B. Ganter. Pro c. 11th International Conference on Formal Concept Analysis (ICFCA 2013), Lecture Notes in Artificial Intelligence. Springer, 2013. Vol. 7880. P. 254–266.

Kuznetsov S.O. Fitting Pattern Structures to Knowledge Discovery in Big Data / ed. by P. Maji, A. Ghosh, M.N. Murty et al. Pro c. 5th International Conference Pattern Recognition and Machine Intelligence (PReMI'2013), Lecture Notes in Computer Science. Springer, 2013. Vol. 8251. P. 30–41.

Kuznetsov S.O., Makhazhanov N., Ushakov M. On Neural Network Architecture Based on Concept Lattices // ISMIS. 2017. P. 653–663.

Marchese Robinson R.L., Palczewska A., Palczewski J., Kidley N. Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets // Journal of Chemical Information and Modeling. 2017. Vol. 57 (8). P. 1773–1792.

Yildiz O.T., Alpaydin E. Regularizing Soft Decision Trees // Information Sciences and Systems. Springer, 2013. P. 15–21.

Youssef A.M., Pourghasemi H.R., Pourtaghi Z.S., Al-Katheeri M.M. Landslide Susceptibility Mapping Using Random Forest, Boosted Regression Tree, Classification and Regression Tree, and General Linear Models and Comparison of Their Performance // Landslides. 2016. No. 13 (5). P. 839–856.

OPEN DATA IN THE ERA OF DIGITAL ECONOMY

Mikhail Parfentev

Department for IT and data processing, Analytical Center
for the Government of the Russian Federation, Moscow, Russia

Abstract. *Open data is data that anyone can access, use or share. According to the McKinsey's report open data — public information and shared data from private sources — can help create $3 trillion a year of value in the seven areas of the global economy. Another survey of economic potential conducted by the National Research University Higher School of Economics (NRU HSE) shows that open data usage could save about 50 billion roubles in the sphere of transport of Moscow.*

Nowadays, we do not underestimate the role of such kind of data. This working paper is devoted to the open data analysis in Russia and key figures of the Open data portal of the Russian Federation in the era of digital economy.

Keywords: *open data, open government, open data portal, digital economy.*

Body text

A few years ago it was a very debatable topic on the role of open data and its interconnections with big data. However, in Russia by the term "open data" we assume, first of all, "open government data" that is why it is possible to observe open data through Venn diagram as a node with logical relations between government data, open government concept and open data (Fig. 1).

To illustrate the role of open data let's have a look at the public value in money terms. The diagram represents the potential of open data in seven areas from education up to consumer finance (Fig. 2).

According to the McKinsey's report open data — public information and shared data from private sources — can help create $3 trillion a year of value in the following seven areas of the global economy.

Another survey of economic potential conducted by NRU HSE that open data usage could save about 50 billion roubles in the sphere of transport of Moscow (NRU HSE, 2015). So, we do not underestimate the role of such kind of data.

Speaking about open data in Russia it is necessary to analyse key figures of the Open data portal of the Russian Federation (Table 1).

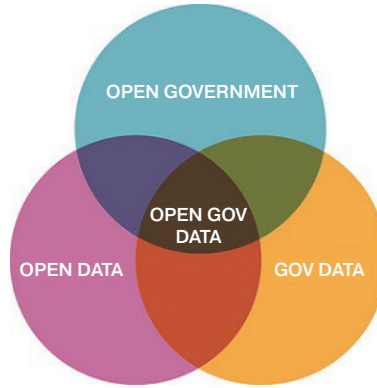


Fig. 1. Open data Venn diagram

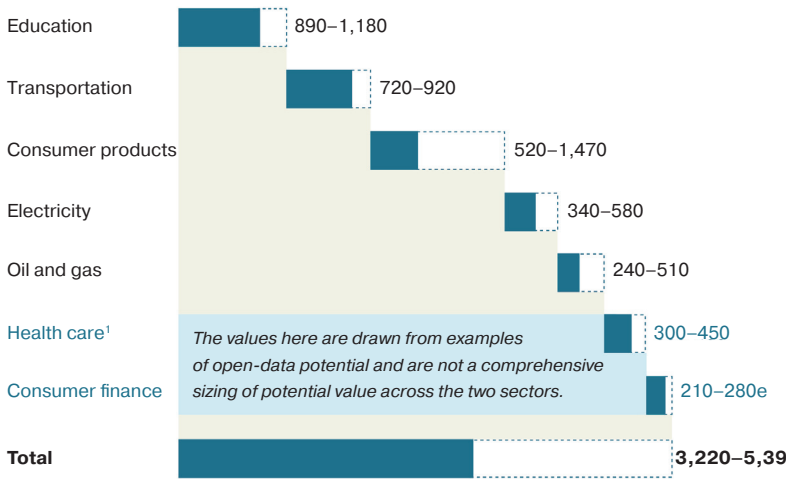


Fig. 2. Potential of open data [Manyika et al., 2013]

Today we have about 5 thousand registered users on the open data portal of Russia from more than 8 hundred organizations of different governmental levels. About twenty five percent of registered users are open data publishers and we have more than 14 thousand datasets. The diagram represents TOP-7 topics of datasets except the Government topic, which is marked by the majority of 57% of all datasets (Fig. 4).

Table 1

Open data portal of Russia stats

	2016				2017		
	Q1	Q2	Q3	Q4	Q1	Q2	Q3
Number of datasets	7 128	8 372	10 402	12 584	13 155	13 966	14 252
Number of organizations	485	549	635	700	773	828	837
Number of publishers	1 824	2 073	2 748	3 244	3 513	3 792	1 572
Number of datasets per organization	14	15	16	17	17	16	17
Number of views	1 119	1 378	1 567	1 956	2 246	2 599	2 512
Average number of downloads per dataset	4	4	4	4	5	6	6
Number of downloads	33 060	41 232	47 106	61 691	71 162	90 549	97 250

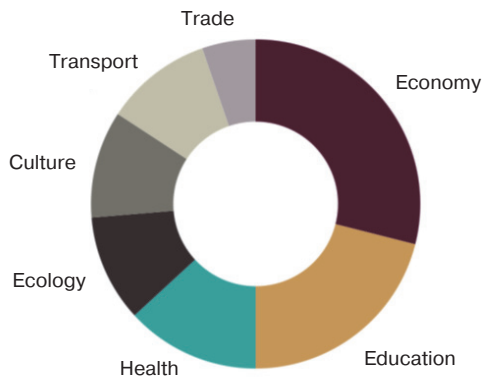


Fig. 4. Open data datasets

Nowadays, there are more than two thousands of governmental information systems in Russia. Some of information systems consists data, which could be published in the format of open data and make an additional value (Table 2).

Table 2

Open data and governmental information systems

Supervisor	Description	Open data datasets
Ministry of Economic Development	Government websites' rankings	2 datasets
The Federal Treasury	Public procurement	3132 datasets
Ministry of finance The Federal Treasury	Budget data	68 datasets
Federal State Statistics service	Official statistics	5503 indicators
The Federal Treasury	State and municipal institutions	16 datasets
General Prosecutor's Office	Governmental inspections	48 datasets

Each time before publishing we need to identify the type of data: is it closed, shared or open; what are potential risks of publishing; how to do the publishing process much easier? It means that publisher should define the format, the update frequency and the place where they will publish: on the website of the public body, on the local portal or on the federal portal. All open data publishers should follow Methodological guidance and technical requirements called the Russian open data standard.

Table 3

Diversity of open data datasets

Dataset	Number of records	Update frequency	In bulk
Unified timetable of sports events	46 890	N/a	16,36 MB
Unified Register of small and medium-sized businesses	> 6 433 xml's	Monthly	2,92 GB in archive
State Register of Medicinal Products	> 100 000	Quarterly	22 MB
Jobs of the All-Russian base of job vacancies in Russia	344 951 vacancies 1 136 482 jobs	Daily	> 800 MB
Executive proceedings against legal persons	> 3 millions	Daily	> 2,3 GB

Current version of Russian open data standard 3.0 allows publishing a huge variety of open data (Table 3).

Open data positively impacts on the prospects for attracting investment. Potential investors may be interested in the data such as statistics, census, labor skills, tariffs and regulations, land relations or national information infrastructure. Such kind of open data could be involved to maintain investment in new originating projects. It allows to decrease potential costs connected with special characteristics of the national government especially for foreign investors concerned on risks.

Nowadays the monitoring and implementation of the world's open data programs are represented by two major initiatives: "The barometer of public data" (Open Data Barometer), published by the Foundation for the development of the World Wide Web and the Open Data index, developed by the Open Knowledge Foundation. Both reports are updating on the annual basis. It is fruitful to take a higher position in both rankings for countries seeking to attract foreign investment.

The importance of data for investment decisions can be exemplified by the practices of the Millennium Challenge Corporation. The Corporation uses a set of indicators in order to estimate the eligibility criteria of countries to receive funding. In spite of the fact that these indicators are designed by third parties (including — the World Bank), they reflect aspects associated with the economy, health, education and environment. The Corporation uses these indicators for the national "scorecard" analyzed by the foreign investors making their own decisions.

Investors involve state governmental public data to the data-driven decision-making process. The national open data program is considered as an indicator of relative openness of the state, and it is an important factor for investors.

Conclusion

In conclusion, it is necessary to highlight the fact that data management question was incorporated into two of five developmental tracks of the Digital economy program in Russia: infrastructure and regulation. Thus, open data as well as data as a whole still plays an important role in the era of digital economy.

References

All-Russia contest “Open data in Russian Federation”. <http://opendatacontest.ru>.

Artamonov R.E., Datiev S.B., Zhulin A.B. Assessment of the Socio-economic Impact of Open Publication of Data on the Example of Moscow’s Public Transport Data. M.: HSE, 2015.

Manyika J., Chui M., Groves P. et al. Open Data: Unlocking Innovation and Performance with Liquid Information. McKinsey Global Institute; McKinsey Center for Government; McKinsey Business Technology Office, 2013.

Methodological Guidance and Technical Requirements. < http://data.gov.ru/metodicheskie-rekomendacii-po-publikacii-otkrytyh-dannyh-versiya-30>.

HIGH-PERFORMANCE SOLUTIONS BASED ON DESKTOP-GRIDS AND COMBINED INFRASTRUCTURES

Mikhail Posypkin

Federal Research Center "Computer Science and Control"
of the Russian Academy of Sciences, National Research University
Higher School of Economics, Moscow, Russia

Abstract. *The desktop grids (volunteer grids) can collect huge amount of cheap computational resources harnessing idle cycles of personal computers. Desktop grids can be used together with traditional sources of computing power such as service grids, supercomputers and clouds. In this paper we outline some of relevant approaches.*

Keywords: *desktop grids, volunteer computing, BOINC, combined distributed infrastructures.*

Desktop grids (DGs) is a relatively new technology for assembling resources of PCs from over the world for solving hard computational problems. Volunteers donate the idle resources of their personal computers ("clients") by connecting them to a project server that manages the computational process.

The primer tool for desktop grid computing is BOINC (Berkeley Open Infrastructure for Network Computing) [Anderson, 2004]. BOINC is a software system that makes it easy for scientists to create and operate public-resource computing projects. It supports diverse applications, including those with large storage or communication requirements. PC owners can participate in multiple BOINC projects, and can specify how their resources are allocated among these projects.

Potentially DGs can collect a huge computational power however its efficient utilization faces lots of significant challenges, e.g. heterogeneity and unreliability of computational resources, limited network bandwidth, limited connectivity among nodes. Overcoming these issues have led to a noticeable progress in DG technologies resulted in several new technologies, namely:

- virtualization to cope with binary incompatibility of client PCs;
- building combined distributed infrastructures based on DGs, service grids and clouds.

Virtualization copes with binary compatibility, licensing and safety issues of desktop grid applications. The primer tool for desktop grid computing BOINC supports "VM apps" — applications that run in VirtualBox virtual machines. This provides several benefits:

- You don't need to build application versions for different platforms. You can develop your app in your environment of choice (say, Debian Linux), and then bundle the resulting executable with a virtual machine image containing an appropriate runtime environment. The application can then be run on all platforms (Windows, Mac OS X, all versions of Linux) with no additional work on your part.
- Virtual machines provide the strongest available security sandbox: a VM app cannot access or modify the host system. This makes it feasible to deploy untrusted applications.
- VM apps don't need to have their own checkpoint/restart mechanism — BOINC provides one.

We should also notice that volunteer computing is not a replacement for traditional approaches such supercomputers, service grids or clouds. It is just another approach to increasing computational resources. In the past a lot of efforts have been made to combine difference sources of computing power. The 3G-Bridge [Kacsuk, Farkas, Fedak, 2008] combines the advantages of the service and desktop grids concepts. A building block of this infrastructure is bridging between the different grid types. 3G-Bridge provides the special BOINC client application so it can represent itself as a very powerful machine towards the BOINC server. It can start a wrapper application specified in a configuration file that can be used to handle the work unit in basically any way based on its description file produced by the modified BOINC client.

The CluBORun tool [Afanasiev et al., 2015] is the technology for integrating idle computing cluster resources into volunteer computing projects. The main principles of this technology are the following: only standard cluster user credentials and only idle computing cluster resources (i.e. the resources that are not employed by other cluster users) are used. The CluBORun tool was successfully applied to boost the performance of volunteer computing projects SAT@home and OPTIMA@home.

References

Anderson D.P. Boinc: A System for Public-resource Computing and Storage. Grid Computing. Proceedings. 5th ACM International Workshop. IEEE, 2004. November. P. 4–10.

Kacsuk P., Farkas Z., Fedak G. Towards Making BOINC and EGEE interoperable. eScience. 4th International Conference. IEEE, 2008. December. P. 478–484.

Afanasiev A.P., Bychkov I.V., Manzyuk M.O. et al. Technology for Integrating Idle Computing Cluster Resources into Volunteer Computing Projects.5th International Workshop on Computer Science and Engineering. M., 2015. P. 109–114.

TEMPORAL DATA MINING

Vera Shalaeva
Laboratoire d'nformatique de Grenoble (AMA group),
Université Grenoble Alpes, Grenoble, France

Abstract. *Nowadays almost each object in the world has sensors and able to emit enormous amount of temporal data. With growing quantity of data, the needs to efficiently process and analyze time series also increased. There are a lot of applications where it's necessary to mine temporal data such that genomic analysis, information retrieval, finance, energy data analytics, airplane tracking and so forth. There are machine learning algorithms that were modified to deal with temporal data. However, there is few general purpose tools to deal with Big temporal data both the machine learning experts and non-experts can use.*

Keywords: *temporal big data, time series, classification.*

Building a complete software to deal with time series is the relevant industrial and scientific goal of IKATS (Innovative ToolKit for Analysing Time Series) project[1]. In the scope of this project the aim is to set tools for preprocessing, classification and clustering of temporal data. To complete the development of an end-user oriented software, one requires interactive tools for visualization of data, results and workflows.

To achieve all these goals, the architecture of the project was defined as three primary blocks: pre-processing, machine learning, visualization.

Pre-processing step includes the extraction the data from database with data cleaning, dimension reduction and so forth.

Machine learning block of the project has to be able to accomplish the clustering, classification, pattern search tasks on time series.

And the last block is responsible for visualization. All results and models generated by the machine learning tools have to be graphically represented. The visual representation must be interpretable and reveal the link between dataset and the models for the targeted users that are either data scientists, engineers or domain experts.

[1] IKATS (Innovative ToolKit for Analysing Time Series) is research and development project. The consortium is composed of LIG (Laboratoire d'Informatique de Grenoble), CS (Communication & Systèmes) — Designer, integrator and operator of information systems, Airbus — Leading aircraft manufacturer and EDF (Électricité de France) — electricity generator.

One important task in temporal data mining and for IKATS project is time series classification. All methods in temporal classification can be divided in the following groups:

- feature based classification, where first some features are extracted from time series on which conventional classification methods can be applied next;
- distance based classification. These methods measure and use pairwise similarity distances between all input time series. The most of the popular method in this group is 1-NN method with DTW distance;
- model based classification, where assumed the classes of time series are generated under some model. During the training, the parameters of this model are learned. On the classification step, model assigns a probability to each class and the time series is associated with the class having the highest likelihood. Neural network algorithm are related to this category.

In the context of temporal classification task, we are working with the algorithm Classification Trees for Time Series [Douzal-Chouakria, Amblard, 2012]. This method modifies conventional decision tree algorithm which split the dataset at each node by using features of data. Instead of feature extraction from temporal dataset, we use distances between time series. At each node of a tree the algorithm searches for the best split pair of series according to an evaluation criterion such as Gini impurity index. Each sub-node of tree node comprises the time series that more similar with one time-series from the split pair than with another. The algorithm allows also to explore different distance functions at each node and to find the most split significant time interval by dichotomy search. The split process continues while sub-node is not represented by one pure class called a leave. In the classification step, time series traverses the tree and class of achieved leave is assigned.

Besides the accuracy of classification, the advantage and interest of this method is its high level of interpretability. During the visualization step it's possible to have clear representation of learned model. However, to be able to include this method in the project tool we face the challenge of scalability. The current version of the algorithm has the high complexity $O\left(\log_{\frac{1}{\alpha}}(T)KN^3\right)$, where N is number of time series, K — number of explored distances, T — the time series length and α is the cover rate for dichotomous search of the most significant time interval. Therefore, in our PhD work, we focus on different approaches to decrease the algorithm complexity without losing in classification accuracy.

References

Camerra A., Palpanas T., Shieh J., Keogh E. iSAX 2.0: Indexing and Mining One Billion Time Series. 10th International Conference on Data Mining. IEEE, 2010.

Chaudhari P., Rana D.P., Rana R.G., Mistry M.N.J. et al. Discretization of Temporal Data: A Survey // International Journal of Computer Science and Information Security (IJCSIS). 2014.

Douzal-Chouakria A., Amblard C. Classification Trees for Time Series // Pattern Recognition Journal. 2012.

Esling P., Agon C. Time-series Data Mining // ACM Computing Surveys (CSUR). 2012.

Laboratoire d'Informatique de Grenoble. <https://www.liglab.fr/>.

The AMA team. <http://ama.liglab.fr/>.

Université Grenoble Alpes. <http://www.univ-grenoble-alpes.fr/>.

Xing Z., Pei J., Keog E. A Brief Survey on Sequence Classification // ACM SIGKDD Explorations Newsletter. 2010.

IMPLEMENTATION OF DATA PROCESSING CENTER FOR SPACE VLBI PROJECTS

M.V. Shatskaya
A.A. Abramov
N.A. Fedorov
S.F. Likhachev
S.I. Seliverstov
D.A. Sichev

Astro Space Center, P.N. Lebedev Physical Institute, RAS,
Moscow, Russia

E.A. Isaev

National Research University Higher School of Economics,
Pushchino Radio Astronomy Observatory of Astro Space Center,
P.N. Lebedev Physical Institute, RAS, Moscow, Russia

Abstract. *IT support of two VLBI projects are described: currently running Radioastron project and future Millimetron project.*

Keywords: *Radioastron, Spektr-R, Millimetron, Data Processing Center, data center, radiointerferometry, VLBI.*

Radioastron — is the international project led by the Astro Space Center of Lebedev Physical Institute, Moscow, Russian Federation. 10 m Space Radio Telescope is the main payload of Spektr-R spacecraft. The project goal is to create together with a ground based radio telescopes the huge Ground to Space interferometer with a baseline up to 350 km. After successful launch on 18 July, 2011 the Radioastron missions started systematic investigations of the Universe at broad radio frequencies range.

Millimetron — observatory ("Spektr-M" project) is a 10-meter Space Telescope aimed at solving a wide range of astronomical problems in the wavelength range from far-infrared to millimeter. "Millimetron" orbit will lie in the vicinity of L2 Lagrange point in the anti-Sun direction, at a distance of about 1.5 million km from the Earth.

Data Processing Center (DPC) of Radioastron project is one of the important segments of the ground support of the project. It is a center for collecting and distribution different kinds of information. This is

necessary for the organization of the ground and space radio telescopes observations, control onboard equipment and processing of the data.

The tasks of the scientific Data Processing Center are:

- organization of service information exchange;
- collection and storage of all scientific data;
- processing of scientific information.

Data Processing Center of Radioastron project is a dynamic and scalable system.

DPC has developed and grown very much since Radioastron had been launched. It was determined by many factors. The growth of scientists interest have led to the increasing of the amount of ground telescopes taking part in observations from 5 to 40. And we expected that only 2–3 telescopes could work simultaneously because of the shape and moving of the Earth. Nowadays sometimes up to 30 ground telescopes can work simultaneously with Radioastron. It was expected that only 1 tracking station in Pushchino would take part in the project at the time of Radioastron launch in 2011. Today two tracking stations in Pushchino and Green Bank are working with us. The number of sessions has changed from 20 to 120–140 per month and the duration of the Space Telescope observations has increased too. Information volume per month has grown from 10 TB to 120 TB. Amount of observations increased from 20 to 100–120 and amount of information has grown too.

Our storage has increased from 100 TB to 1 PB on-line and 2 PB off-line.

Technology has changed too. We moved from DAS to NAS. We had to eliminate many bottlenecks. We have expanded Internet channel from 100 Mb/s to 600 Mb/s. We also have expanded 10 GB network to accelerate network interaction.

We have reformed our DPC several times over five years for expanding and optimizing its performance.

After five years of operation, DPC — is several rooms equipped with air conditioning and uninterruptible power supply, video cameras and system of monitoring.

Main components of DPC are:

- on-line storage system for collecting information — 500 TB;
- on-line storage system for data processing — 80 TB;
- on-line storage system for processing results — 160 TB;
- the archive of data on hard drives on 2500 TB;

- archive tapes for 2500 TB;

(the total storage capacity is more than 5 PB!)

- computer system;
- 1 GB/s and 10 Gbit/s network infrastructure, and 600 Mbit/s Internet channel.

We have collected about 2200 TB of information for the five years of observations.

Transferring of large amounts of scientific information over long distances and delivery of information in on-line mode requires a high-speed communication channels. We organized optic lines connecting the processing center and tracking station in Pushchino and tracking headquarter. We also had to organize the delivery of data from the second tracking stations in Green Bank and many ground-based telescopes.

The next VLBI project Millimetron will take much more recourses of computer technique.

Previous scheme of the Data Processing Center for Millimetron project looks like scheme for Radioastron. All information should be collected and analyzed in processing center. We will also have to organize quick exchange of the information between participants of the project. We'll have to collect, store and process all scientific information.

According to preliminary calculations if data transferring speed from Space Telescope is 1.2 GB/s in period 5 years of working from tracking stations and from three ground telescopes we'll receive about 80 PB of data.

We are planning to implement Data Center for storage and processing all of this information.

The first problem we have to solve will be engineering infrastructure:

- appropriate room or special building with protection from leaking and fire-extinguishing system;
- redundant power supply;
- independent power supply (diesel generator);
- uninterruptible power supply;
- air-conditioning on $N + 1$ systems;
- video surveillance and monitoring of all systems;
- room for operators;
- communication channels.

Conclusions:

— The structure and functions of ASC Data Processing Center for Radioastron project are fully adequate to the data processing requirements of the Radioastron Mission.

— When we created Data Processing Center for the Radioastron project, we got experience of working, transferring and storage of large volumes of data. The experience will be useful for our next project "Millimetron".

PUBLICATION AND REUSE OF BIG DATA APPLICATIONS AS SERVICES

Oleg Sukhoroslov

Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow, Russia

Abstract. *The report considers development of domain-specific web services for processing of large volumes of data on high-performance computing resources. The development of these services is associated with a number of challenges, such as integration with external data repositories, implementation of efficient data transfer, management of user data stored on the resource, execution of data processing jobs and provision of remote access to the data. An approach for building big data processing services on the base of Everest platform is presented. The proposed approach takes into account the characteristic features and supports rapid deployment of these services on the base of existing computing infrastructure.*

Keywords: *big data, web services, data transfer, data management, distributed data processing, Hadoop, MapReduce.*

The explosive growth of data, observed in a variety of areas from research to commerce, requires the use of high-performance resources and efficient means for storing and processing large amounts of data. During the last decade, the distributed data processing technologies like Hadoop and Spark are emerged. However, the complexity of the hardware and software infrastructure prevents its direct use by non-specialists, and requires the creation of user-friendly tools to solve particular classes of problems. One way of implementing such tools is the creation of domain-specific services based on the Software as a Service model.

Data-intensive services (DIS), in comparison to conventional computational services with a small amount of data, started to develop recently, so the principles and variants of implementation of these services are poorly understood. There is a lack of best practices for implementation of DIS on the basis of the existing infrastructure for big data processing such as a cluster running Hadoop or Spark platforms which are increasingly used for the analysis of scientific data. Also, little attention is paid to the integration of DIS with existing repositories and data warehouses, including the cloud-based ones, as well as other services. Finally, there is a lack of platforms for implementation and deployment of DIS that would provide ready-made solutions of typical problems encountered when creating this kind of services.

Consider typical requirements to DIS that represent remotely available services for solving a certain class of problems with a large amount of input data. Such services should provide remote interfaces, usually in the form of a web user interface and application programming interface (API). The interface must allow the user to specify the input datasets and parameters of the problem being solved in terms of subject area.

DIS must use high-performance and scalable (normally distributed) implementations of data analysis algorithms, requiring appropriate computing infrastructure for data processing and storage. Such infrastructure is generally represented by one or more computing clusters running Hadoop platform or a similar technology. DIS must translate the user request into one or more computing jobs that are submitted on a cluster and use scalable implementations (e.g., based on MapReduce) of perspective algorithms.

The user must be able to pass arbitrary input data to DIS. If the data is initially located on the user's computer or external storage resource (e.g., a data repository) DIS must implement the transfer of data over a network to the used cluster. When transferring large amounts of data it is important to ensure the maximum transfer rate and automatic failover. Since the process of working with big data is often exploratory, requiring multiple invocations of DIS, the service should support reuse of data loaded to the cluster. In order to optimize the use of network DIS must also cache frequently used datasets on the cluster. Data transfer functions can also be implemented as separate auxiliary services.

Importantly, DIS may operate separately from computing resources used for real data processing. DIS can use multiple resources, that can be situated at different locations. It is also possible that the service uses the resources provided by the user. In such cases it is important for reasons of efficiency to avoid passing the input data from the user to the resource through the service and to transmit the data directly.

In practice, the data analysis is often a multi-step process that requires performing different tasks at different stages of the analysis. In such cases, the results produced by one DIS can be passed as the input to another service. If these services use different resources, there also arises a problem of data transmission between resources. In general DIS should allow the user to download the output to his computer or an external resource, as well as to transfer the data directly to another service. In addition, DIS may provide additional functionality for remote data

preview and visualization. These functions may also be implemented as separate auxiliary services.

DIS must support the simultaneous use by multiple users. This requires the protection of user data, resource distribution between users and isolation of computational processes. In the case of cloud infrastructure, DIS must also manage dynamic allocation and deallocation of resources in the cloud, according to the current load.

Everest [Sukhoroslov, Volkov, Afanasiev, 2015] is a web-based distributed computing platform. It provides users with tools to quickly publish and share computing applications as services. The platform also manages execution of applications on external computing resources attached by users. In contrast to traditional distributed computing platforms, Everest implements the Platform as a Service (PaaS) model by providing its functionality via remote web and programming interfaces. A single instance of the platform can be accessed by many users in order to create, run and share applications with each other. The platform implements integration with servers and computing clusters using an agent that runs on the resource side and plays the role of mediator between the platform and resources. The platform is publicly available online to interested users [<https://everest.distcomp.org/>].

The advantage of using Everest platform to create DIS is the availability of ready-made tools for rapid deployment of computational services and integration with computing resources that do not require a separate installation of the platform. At the same time, since the platform was originally created to support services with a small amount of data, the effective implementation of DIS on the base of Everest requires a number of improvements. In particular, it is necessary to implement support of direct data transfers from external storage to the resource and vice versa, bypassing the platform. In addition, it is required to implement the integration of the agent with the components of Hadoop platform or similar technology used for data storage and processing on the cluster.

Figure 1 presents the proposed scheme of implementation of DIS on the base of Everest platform and existing Hadoop cluster. We plan to add to the agent support for loading data from major types of repositories and storage. Currently the basic support for downloading files via HTTP and FTP protocols, as well as an experimental integration with Dropbox and Dataverse repository are implemented. The downloaded input data is placed in Hadoop Distributed File System (HDFS). The submission

of data processing jobs is performed via the cluster resource manager such as Yet Another Resource Negotiator (YARN). A special adapter was implemented in order to support interaction of Everest agent with YARN, similar in function to the previously created adapters for integration with HPC batch schedulers. Upon a job submission the agent monitors the state of the job and transmits the progress information to the service. On job completion the agent transmits to the service the output files (of small size) and the final status of the job. If the output data produced by the job is of big size, the agent should support direct upload of such data over the network to a specified external storage. The necessary information should be provided by the user when sending the request to the service. At the moment, the upload of output data to the specified FTP server or Dropbox folder is implemented.

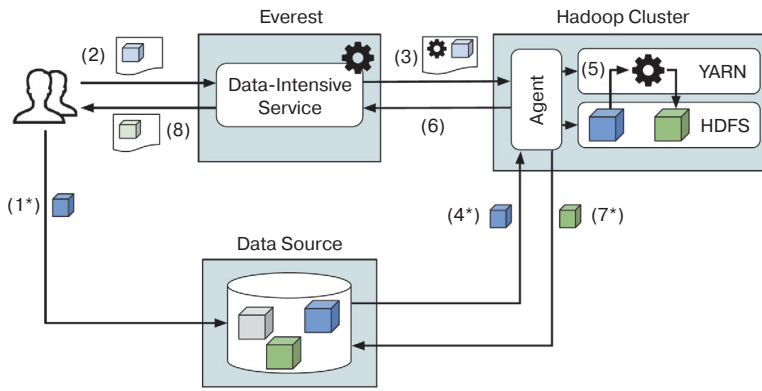


Fig. 1. Implementation of DIS on the base of Everest platform

References

Sukhoroslov O., Volkov S., Afanasiev A.A. Web-Based Platform for Publication and Distributed Execution of Computing Applications. 14th International Symposium on Parallel and Distributed Computing (ISPDC). IEEE, 2015. P. 175–184.

<https://everest.distcomp.org/>.

ANALYSIS OF BIG DATA IN NEUROSCIENCE

Mikhail Ustinin
Keldysh Institute of Applied Mathematics, Moscow, Russia
New York University, New York, USA

Anna Boyko
Keldysh Institute of Applied Mathematics, Moscow, Russia

Abstract. *A new method for the analysis and localization of brain activity has been developed, based on multichannel magnetic field recordings, over minutes, superimposed on the MRI of the individual. Here, a high-resolution Fourier Transform is obtained over the entire recording period, leading to a detailed multi-frequency spectrum. Further analysis implements a total decomposition of the frequency components into functionally invariant entities, each having an invariant field pattern localizable in recording space. The method, addressed as functional tomography, makes it possible to find the distribution of magnetic field sources in space. Here, the method is applied to recordings of spontaneous brain activity in ten healthy adults. The method successfully provides an overall view of brain electrical activity, a detailed spectral description and the localization of sources in anatomical brain space.*

Keywords: *magnetic encephalography, functional tomography, inverse problem solution, alpha rhythm.*

Big Data Analysis

Modern scientific studies are performed by means of new powerful equipment, generating large amounts of detailed data. Magnetic encephalography (MEG) provides an example of a foremost biological technology, comparable with the most sophisticated physical devices. Magnetic encephalographs register magnetic field in hundreds of channels with sampling frequency up to several thousand Hertz. Typical 5 minutes' experiment on the 275 channel device with sampling rate 1200 Hz, provides 100 million field values, so the problem of big data analysis appears a pressing challenge in the MEG technique (Fig. 1).

Many approaches are used to solve various scientific and diagnostic problems of encephalography. Fourier analysis in many implementations can be called the oldest and the most popular of methods used for the brain data analysis. Through the whole history of this method it was connected with difficulties of calculations, so the development of

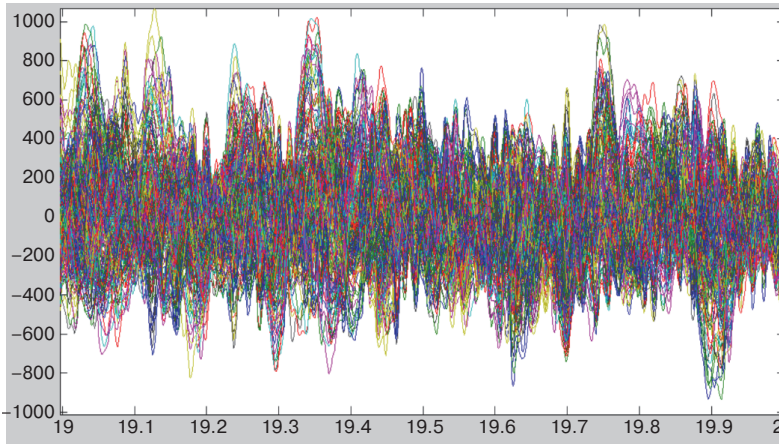


Fig. 1. One second of the raw MEG in 275 channels

the fast Fourier transform (FFT) dramatically advanced the application of the Fourier analysis in many fields, including brain research. Usually in applications of the Fourier analysis to brain studies the spectra are calculated in short (<10 seconds) time windows, based on the well-known property of instability of the brain processes. Typically brain studies register activity in many channels simultaneously for protracted time periods (up to tens of minutes in hundreds of channels). Those registered data are usually processed with two important methodological weaknesses. First weakness is that in time dependence analysis the methods are applied, which were developed for the solitary time series, multichannel recordings are implemented mainly to attempt inverse problem solutions. The second weakness lies in the usage of short time windows (less than 10 seconds), decreasing the resolution of the Fourier transform. Recently the method of precise frequency-pattern analysis to decompose complex systems into functionally invariant entities was proposed [Llinás, Ustinin, 2016; 2014; Llinás et al., 2015]. The method is based on the complete utilization of the long-time series, while the multichannel nature of the data is also completely taken into account, making it possible to implement detailed reconstruction of neuronal circuit activity.

The multichannel Fourier transform calculates a set of spectra for registered functions (Fig. 2). All spectra are calculated for the whole

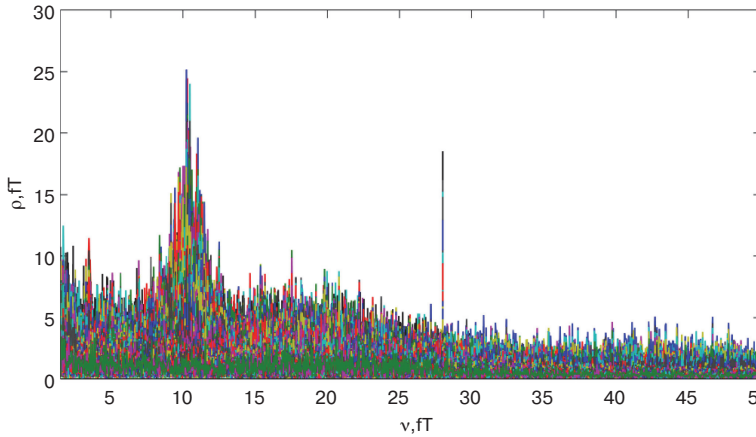


Fig. 2. The multichannel spectrum contains 15 000 frequencies for 5 minutes' experiment

registration time T, which is sufficient to reveal the detailed frequency structure of the system. The step in frequency is equal to $1/T$, thus frequency resolution is determined by the recording time.

The next step of the analysis is to reconstruct time series at each frequency, to extract coherent oscillations and to calculate the magnetic field pattern for those oscillations. Then the inverse problem is solved for each pattern and the solutions are distributed in space, thus providing a functional tomogram of the brain.

Functional tomograms were obtained for alpha rhythm from multichannel MEG data. These functional tomograms demonstrate individual variances of the power spatial distribution, generally corresponding to our present knowledge concerning the alpha rhythm localization in the occipital and posterior parietal lobes (Fig. 3). It can be concluded, therefore, that the functional tomography method, based on magnetic-encephalograms analysis, can determine spontaneous brain activity sources.

A fundamental advantage of this framework lies in the fact, that all recorded data is fully utilized.

Method of functional tomography can be applied to the diagnostics of activity in the whole brain and in broad frequency band, revealing areas of abnormally high or abnormally low activity.

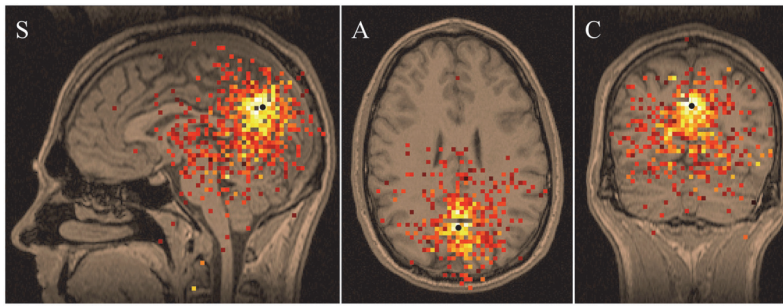


Fig. 3. Functional tomogram of the average alpha rhythm activity of ten subjects, shown over the MRI

Acknowledgement

The study was partly supported by the CRDF Global (USA) (grants CRDF RB1-2027 and RUB-7095-MO-13), by the Russian Foundation for Basic Research (grants No. 16-07-00937, 16-07-01000, 14-07-00636), and by the Program No. I.33П for Fundamental Research of the Russian Academy of Sciences.

References

Llinás R.R., Ustinin M.N. Frequency-pattern Functional Tomography of Magnetoencephalography Data Allows New Approach to the Study of Human Brain Organization // Frontiers in Neural Circuits. 2014. Vol. 8. P. 43.

Llinás R.R., Ustinin M.N. Precise Frequency-Pattern Analysis to Decompose Complex Systems into Functionally Invariant Entities: U.S. Patent. US Patent App. Publ. 20160012011 A1. 2016. January 14.

Llinás R.R., Ustinin M.N., Rykunov S.D. et al. Reconstruction of Human Brain Spontaneous Activity Based on Frequency-pattern Analysis of Magnetoencephalography Data // Frontiers in Neuroscience. 2015. October. Vol. 9.

NEUROBAYES: CONVERGENCE OF BAYESIAN FRAMEWORK AND DEEP NEURAL NETWORKS IN LARGE-SCALE MACHINE LEARNING PROBLEMS

Dmitry Vetrov

National Research University Higher School of Economics,
Moscow, Russia

Abstract. *In the paper we review several important directions on the combination of deep neural networks with Bayesian framework for solving large-scale machine learning problems.*

Keywords: *deep learning, Bayesian framework, variational inference, stochastic optimization.*

First attempts to combine deep learning and Bayesian probabilistic modeling were presented in 2013–2015. Specifically, neural networks were used for approximate Bayesian inference in complex probabilistic models and Bayesian regularization helps to avoid the effect of model overfitting. These works borrow the concept of "evidence" from Bayesian statistics, then introduce a computationally tractable lower bound on the evidence and maximize the bound with respect to neural network parameters. Early results show that the neural-Bayesian fusion outperforms analogous algorithms and have better capabilities for modeling data distributions. As a result, the proposed approach may potentially allow the generation of objects that were always considered to be a product of higher nervous activity (e.g., drawing, handwriting forgery, image captioning, translation etc.). Undoubtedly, further development in this direction will be among the main trends in machine learning for the next couple of years. The most promising areas of application are following.

Bayesian regularization of deep neural networks

One of the ways to prevent machine learning algorithms from overfitting the training sample is to introduce so-called regularization which prohibits the weights of the algorithm from becoming too well-fitted. A possible approach to regularization is Bayesian regularization which

imposes a prior probabilistic distribution of the weights of the algorithm; the learning occurs during the process of Bayesian inference which combines prior restrictions with the impact of the training sample. With the increasing size of modern neural networks, the overfitting problem has also arisen in the problems of deep learning. At the end of 2015, it was shown that one of the popular heuristic procedures for preventing overfitting (so-called dropout) is a rough approximation of the Bayesian regularization of a special kind.

Bayesian regularization of neural networks theoretically makes it possible to perform so-called incremental learning; when new data is coming, the network continues learning instead of learning "from scratch". There are currently no procedures for incremental learning of neural networks. By introducing the prior distribution of the neural network weights and by performing (approximate) Bayesian inference, it is possible to obtain the posterior distribution of the weights which accumulates all the information about previously observed samples. Using this posterior distribution as a new prior distribution when new data arrives allows the network to continue learning without the need for re-using the old data.

Algorithms for building attention maps

One of the most significant results obtained in deep learning in 2015–2016 are the algorithms for building so-called attention maps which help the neural network to "concentrate" on informative fragments in the description of the data (e.g., on particular parts of an image or text). Attention maps have dramatically improved the quality of solving such complex tasks as image caption generation, machine translation etc. The methodology for building attention maps is based on models with latent variables which are a Bayesian mechanism for learning from incomplete data; this methodology is still developing today.

Search for compact representations of neural networks

One of important results in deep learning is the realization of the following fact: whenever there are no restrictions on the training data size (e.g., when training data may be generated like in the AlphaGo system which defeated the Go world champion in 2016), the more neu-

rons and layers the neural network has, the better will be the quality. Modern neural networks have several hundred million parameters and up to thousand layers. Their further growth is constrained by the lack of corresponding RAM sizes on modern PCs, not to mention mobile devices. On the other hand, it is becoming increasingly clear that modern neural network architectures contain much redundancy. One of the ways to eliminate such redundancy is tensor algebra and machinery of tensor decompositions. Its usage allows to compress particular parts of the neural network up to several hundred thousand times almost without any loss in quality and performance.

Stochastic optimization

Any machine learning problem may be reduced to solving a particular optimization problem, e.g., maximization of the likelihood of the training sample correct recognition. With the increasing volumes of training samples and transition to the analysis of big data, it became clear that traditional approaches for optimizing functions that arise in machine learning do not scale (i.e. become extremely ineffective and are not suitable for big data). The solution was a paradigm shift and transition to so-called stochastic optimization techniques which in some cases allowed to find an extremum of the function faster than a single evaluation of the function at a single point. All modern deep learning methods use stochastic optimization. Moreover, the rapid development of deep learning stimulated the development of more effective methods for stochastic optimization. Now several effective methods for stochastic optimization of convex functions are known. However, in the field of deep learning one has to optimize significantly multi-extremal non-convex functions.

Improving procedures for approximate Bayesian inference

The key point, which made the application of Bayesian methods and models in problems with big data (in particular, in deep learning) possible, was the development of scalable procedures for so-called variational Bayesian inference in 2014. First of all, this includes the variational auto-encoder and its numerous modifications proposed in 2015–2016. All of them use the transition from the Bayesian inference

problem to the problem of evidence lower bound optimization (ELBO), which allows to construct an approximate posterior distribution of the parameters interesting to user with the help of stochastic optimization methods. Generally speaking, ELBO is not the only possible variational lower bound. By using the Jensen inequality one can construct an infinite set of various variational lower bounds. At the same time, the more accurate the lower bound is, the better will be the quality of the approximate Bayesian inference.

Development of generative Bayesian models

One of the advantages of the Bayesian approach to the construction of data processing models is the possibility of building complex probabilistic models from the simpler ones. This is possible because the result of the Bayesian inference in one model (posterior distribution of unknown variables) may be used as a prior distribution in another model, and so on. Such complex models are known as probabilistic graphical models and are widely applicable in image/video processing, signal analysis, speech recognition, tracking, social network analysis etc. However, the potential for fitting such complex models have been quite low so far and, in practice, one has to restrict their learning to a relatively poor log-linear class. With the development of deep learning methods, it becomes possible to construct graphical models that use neural networks as building blocks.

References

Kingma D.P., Welling M. Auto-Encoding Variational Bayes. ICLR, 2013.

Rezende D.J., Mohamed Sh., Wierstra D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. ICML, 2014.

Sohl-Dickstein J., Poole B., Ganguli S. Fast Large-scale Optimization by Unifying Stochastic Gradient and Quasi-Newton Methods. ICML, 2014.

Scientific electronic publication

Proceedings of the Russian-French Workshop in Big Data and Applications

Passed for publ. 30.03.2018. Format 60×88/16
Font Newton. 8.0 MB. Publ. sheets 7.0
Publ. No. 2169

National Research University Higher School of Economics
101000, Moscow, Myasnitskaya str., 20
Tel.: +7 (495) 772-95-90*15285