ISSN 1998-0663

№3(33)-2015

БИЗНЕС-ИНФОРМАТИКА

МЕЖДИСЦИПЛИНАРНЫЙ НАУЧНО-ПРАКТИЧЕСКИЙ ЖУРНАЛ

СОДЕРЖАНИЕ

Издатель:
Национальный исследовательский университет «Высшая школа экономики»

Подписной индекс в каталоге агентства «Роспечать» –72315

Выпускается ежеквартально

Журнал включен в Перечень российских рецензируемых научных журналов, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученых степеней доктора и кандидата наук

Главный редактор
А.О. Голосов

Заместитель главного редактора
Д.В. Исаев

Дизайн обложки
С.Н. Борисова

Компьютерная верстка
О.А. Богданович

Администратор веб-сайта
И.И. Хрусталёва

Адрес редакции:
105187, г. Москва, ул. Кирпичная, д. 33
Тел./факс: +7 (495) 771-32-38
http://bijournal.hse.ru
E-mail: bijournal@hse.ru

За точность приведенных сведений и содержание данных, не подлежащих открытой публикации, несут ответственность авторы

При перепечатке ссылка на журнал «Бизнес-информатика» обязательна

Тираж 500 экз.

Отпечатано в типографии НИУ ВШЭ
г. Москва, Кочновский проезд, 3

© Национальный исследовательский университет «Высшая школа экономики»

О ЖУРНАЛЕ

«Бизнес-информатика» – рецензируемый междисциплинарный научный журнал, выпускаемый с 2007 года Национальным исследовательским университетом «Высшая школа экономики» (НИУ ВШЭ). Администрирование журнала осуществляется школой бизнес-информатики НИУ ВШЭ.

Миссия журнала – развитие бизнес-информатики как новой области информационных технологий и менеджмента. Журнал осуществляет распространение последних разработок технологического и методологического характера, способствует развитию соответствующих компетенций, а также обеспечивает возможности для дискуссий в области применения современных информационно-технологических решений в бизнесе, менеджменте и экономике.

Журнал публикует статьи по следующей тематике:

- ✦ анализ данных и интеллектуальные системы
- ✦ информационные системы и технологии в бизнесе
- ✦ математические методы и алгоритмы бизнес-информатики
- ✦ программная инженерия
- ✦ Интернет-технологии
- ✦ моделирование и анализ бизнес-процессов
- ✦ стандартизация, сертификация, качество, инновации
- ✦ правовые вопросы бизнес-информатики
- ✦ принятие решений и бизнес-интеллект
- ✦ моделирование социальных и экономических систем.

В соответствии с решением президиума Высшей аттестационной комиссии Российской Федерации с 2010 года журнал включен в Перечень российских рецензируемых научных журналов, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученых степеней доктора и кандидата наук.

Журнал выпускается ежеквартально и распространяется как в печатном виде, так и в электронной форме.

Журнал «Бизнес-информатика» зарегистрирован в «Роскомнадзоре».
Свидетельство ПИ № ФС 7752404 от 28 декабря 2012 г.

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

ГЛАВНЫЙ РЕДАКТОР

ГОЛОСОВ Алексей Олегович –
кандидат технических наук, Президент компании «ФОРС – Центр разработки»

ЗАМЕСТИТЕЛЬ ГЛАВНОГО РЕДАКТОРА

ИСАЕВ Дмитрий Валентинович –
кандидат экономических наук, доцент кафедры бизнес-аналитики, школа бизнес-информатики, факультет бизнеса и менеджмента, Национальный исследовательский университет «Высшая школа экономики»

ЧЛЕНЫ РЕДКОЛЛЕГИИ

АБДУЛЬРАБ Абиб –
PhD, профессор департамента математики и программной инженерии, Национальный институт прикладных наук, Руан, Франция

АВДОШИН Сергей Михайлович –
кандидат технических наук,профессор, руководитель департамента программной инженерии, факультет компьютерных наук,Национальный исследовательский университет «Высшая школа экономики»

АЛЕСКЕРОВ Фуад Тагиевич –
доктор технических наук, профессор, руководитель департамента математики, факультет экономических наук, Национальный исследовательский университет «Высшая школа экономики»

БАБКИН Эдуард Александрович –
кандидат технических наук, PhD, профессор кафедры информационных систем и технологий, факультет бизнес-информатики и прикладной математики (Нижний Новгород), Национальный исследовательский университет «Высшая школа экономики»

БАЙЕР Алекс –
PhD, Директор KAFAN FX Information Services, Нью-Йорк, США

БАРАНОВ Александр Павлович –
доктор технических наук, профессор, заведующий кафедрой информационной безопасности, школа бизнес-информатики, факультет бизнеса и менеджмента, Национальный исследовательский университет «Высшая школа экономики»

БЕККЕР Йорг –
PhD, проректор, профессор, директор Европейского исследовательского центра в области информационных систем (ERCIS) Мюнстерского университета, Мюнстер, Германия

БЕЛОВ Владимир Викторович –
доктор технических наук, профессор кафедры вычислительной и прикладной математики, факультет вычислительной техники, Рязанский государственный радиотехнический университет

ГРИБОВ Андрей Юрьевич –
кандидат экономических наук, Генеральный директор компании «КиберПлат»

ГРОМОВ Александр Игоревич –
кандидат химических наук, профессор, заведующий кафедрой моделирования и оптимизации бизнес-процессов, школа бизнес-информатики, факультет бизнеса и менеджмента, Национальный исследовательский университет «Высшая школа экономики»

ГУРВИЧ Владимир Александрович –
PhD, приглашенный профессор и исследователь, Центр исследования операций, Ратгерский университет (Университет Нью-Джерси), США

ДЖЕЙКОБС Лоренц –
PhD, профессор медицинского факультета, Университет Цюриха, Швейцария

ЗАНДКУЛЬ Курт –
PhD, заведующий кафедрой информационных систем для бизнеса, институт информатики, факультет информатики и электротехники, Университет Ростока, Германия

ИЛЬИН Николай Иванович –
доктор технических наук, профессор, заместитель начальника Управления специальной связи, Федеральная служба охраны Российской Федерации (ФСО России)

КАЛЯГИН Валерий Александрович –
доктор физико-математических наук, профессор, заведующий кафедрой прикладной математики и информатики, факультет бизнес-информатики и прикладной математики (Нижний Новгород), Национальный исследовательский университет «Высшая школа экономики»

КАМЕННОВА Мария Сергеевна –
кандидат технических наук, директор компании «Логика BPM»

КУЗНЕЦОВ Сергей Олегович –
доктор физико-математических наук, профессор,руководитель департамента анализа данных и искусственного интеллекта, факультет компьютерных наук, Национальный исследовательский университет «Высшая школа экономики»

КУЧЕРЯВЫЙ Евгений Андреевич –
PhD, профессор департамента электроники и коммуникаций, Технологический университет Тампере, Финляндия

МАЛЬЦЕВА Светлана Валентиновна –
доктор технических наук, профессор, заведующий кафедрой инноваций и бизнеса в сфере информационных технологий, руководитель школы бизнес-информатики, факультет бизнеса и менеджмента, Национальный исследовательский университет «Высшая школа экономики»

МЕЙОР Питер –
PhD, заместитель директора консультативной группы по радиокоммуникациям, Международный телекоммуникационный союз (ITU), заместитель руководителя Комиссии ООН по науке и технологиям, Женева, Швейцария

МИРКИН Борис Григорьевич –
доктор технических наук, профессор департамента анализа данных и искусственного интеллекта, факультет компьютерных наук,Национальный исследовательский университет «Высшая школа экономики»

МОТТЛЬ Вадим Вячеславович –
доктор технических наук, профессор кафедры информационной безопасности, факультет кибернетики, Тульский государственный университет

ПАЛЬЧУНОВ Дмитрий Евгеньевич –
доктор физико-математических наук, заведующий кафедрой общей информатики, факультет информационных технологий, Новосибирский государственный университет

ПАРДАЛОС Панайот (Панос) –
PhD, почетный профессор, директор центра прикладной оптимизации, департамент промышленной и системной инженерии, Университет Флориды, США

СИЛАНТЬЕВ Альберт Юрьевич –
доктор технических наук, профессор кафедры информационных бизнес систем, Институт информационных бизнес-систем, Национальный исследовательский технологический университет «МИСиС»

ТАРАТУХИН Виктор Владимирович –
кандидат технических наук, PhD, руководитель научной группы Европейского исследовательского центра в области информационных систем (ERCIS) Мюнстерского университета, Мюнстер, Германия

УЛЬЯНОВ Михаил Васильевич –
доктор технических наук,профессор департамента программной инженерии, факультет компьютерных наук,Национальный исследовательский университет «Высшая школа экономики»

ШАЛКОВСКИЙ Алексей Геннадьевич –
кандидат технических наук, директор Института информационных технологий, Национальный исследовательский университет «Высшая школа экономики»

ISSN 1998-0663

№3(33)-2015

BUSINESS INFORMATICS

INTERDISCIPLINARY ACADEMIC JOURNAL

CONTENTS

Publisher:
National Research University
Higher School of Economics

Subscription index in the «Rospechat» catalogue – 72315

The journal is published quarterly

The journal is included into the list of peer reviewed scientific editions established by the Supreme Certification Commission of the Ministry of Education and Science of the Russian Federation

Editor-in-Chief:
A. Golosov

Deputy Editor-in-Chief
D. Isaev

Design:
S. Borisova

Computer Making-up:
O. Bogdanovich

Website Administration:
I. Khrustaleva

Address:
33, Kirpichnaya str., Moscow,
105187, Russian Federation

Tel./fax: +7 (495) 771-32-38
http://bijournal.hse.ru
E-mail: bijournal@hse.ru

Circulation – 500 copies

Printed in HSE Printing House
3, Kochnovsky proezd, Moscow,
Russian Federation

© National Research University
Higher School of Economics

ABOUT THE JOURNAL

Business Informatics is a peer reviewed interdisciplinary academic journal published since 2007 by National Research University – Higher School of Economics (HSE), Moscow, Russian Federation. The journal is administered by School of Business Informatics.

The mission of the journal is to develop business informatics as a new field within both information technologies and management. It provides dissemination of latest technical and methodological developments, promotes new competences and provides a framework for discussion in the field of application of modern IT solutions in business, management and economics.

The journal publishes papers in the areas of, but not limited to:

- ✦ data analysis and intelligence systems
- ✦ information systems and technologies in business
- ✦ mathematical methods and algorithms of business informatics
- ✦ software engineering
- ✦ Internet technologies
- ✦ business processes modeling and analysis
- ✦ standardization, certification, quality, innovations
- ✦ legal aspects of business informatics
- ✦ decision making and business intelligence
- ✦ modeling of social and economic systems.

Since 2010 the journal is included into the list of peer reviewed scientific editions established by the Supreme Certification Commission of the Ministry of Education and Science of the Russian Federation.

The journal is published quarterly and distributed both in printed and electronic forms.

EDITORIAL BOARD

EDITOR-IN-CHIEF

Dr. Alexey GOLOSOV –
President of FORS Development Center, Russian Federation

DEPUTY EDITOR-IN-CHIEF

Dr. Dmitry ISAEV –
Associate Professor, Department of Business Analytics, School of Business Informatics, Faculty of Business and Management, National Research University Higher School of Economics, Russian Federation

EDITORIAL BOARD

Dr. Habib ABDULRAB –
Professor, Mathematical and Software Engineering Department, National Institute of Applied Sciences – Institut national des sciences appliquées de Rouen (INSA de Rouen), Rouen, France

Dr. Sergey AVDOSHIN –
Professor, Head of School of Software Engineering, Faculty of Computer Science, National Research University Higher School of Economics,Russian Federation

Dr. Fuad ALESKEROV –
Professor, Head of Department of Mathematics, Faculty of Economics, National Research University Higher School of Economics, Russian Federation

Dr. Eduard BABKIN –
Professor, Department ofInformation Systems and Technologies, Faculty of Business Informatics and Applied Mathematics (Nizhny Novgorod), National Research University Higher School of Economics, Russian Federation

Dr. Alex BAYER –
Head of KAFAN FX Information Services, New York, USA

Dr. Alexander BARANOV –
Professor, Head of Department of Information Security Management, School of Business Informatics, Faculty of Business and Management, National Research University Higher School of Economics, Russian Federation

Dr. Jorg BECKER –
Vice Rector, Professor, Director of European Research Center for Information Systems (ERCIS), University of Munster, Germany

Dr. Vladimir BELOV –
Professor, Department of Computational and Applied Mathematics, Faculty of Computer Engineering, Ryazan State Radio Engineering University, Russian Federation

Dr. Andrey GRIBOV –
Director General, CyberPlat Company, Russian Federation

Dr. Alexander GROMOV –
Professor, Head of Department of Modeling and Business Process Optimization, School of Business Informatics, Faculty of Business and Management, National Research University Higher School of Economics, Russian Federation

Dr. Vladimir GURVICH –
Invited Professor and Researcher, Rutgers Center for Operations Research, Rutgers, The State University of New Jersey, USA

Dr. Laurence JACOBS –
Professor, Medical School, University of Zurich, Switzerland

Dr. Kurt SANDKUHL –
Professor, Head of Department of Business Information Systems, Institute of Computer Science, Faculty of Computer Science and Electrical Engineering, University of Rostock, Germany

Dr. Nikolay ILYIN –
Deputy Head, Administration of Special Communication, Federal Security Guard, Russian Federation

Dr.Valery KALYAGIN –
Professor, Head of Department of Applied Mathematics and Informatics, Faculty of Business Informatics and Applied Mathematics (Nizhny Novgorod), National Research University Higher School of Economics, Russian Federation

Dr. Maria KAMENNOVA –
Director, BPM Logic,Russian Federation

Dr. Sergey KUZNETSOV –
Professor,Head of School of Data Analysis and Artificial Intelligence, Faculty of Computer Science,National Research University Higher School of Economics, Russian Federation

Dr. Yevgeni KOUCHERYAVY –
Professor of Department of Electronics and Communication Engineering, Tampere University of Technology, Finland

Dr. Svetlana MALTSEVA –
Professor, Head of Department of Innovation and Business in Information Technologies, Head of School of Business Informatics, Faculty of Business and Management, National Research University Higher School of Economics, Russian Federation

Dr. Peter MAJOR –
Vice-chairman, Radiocommunication Advisory Group of International Telecommunication Union (ITU), vicece-chairman of the UN Commission on Science and Technology for Development (CSTD), Geneva, Switzerland

Dr. Boris MIRKIN –
Professor, School of Data Analysis and Artificial Intelligence, Faculty of Computer Science,National Research University Higher School of Economics, Russian Federation

Dr. Vadim MOTTL –
Professor, Department of Information Security Management, Faculty of Cybernetics, Tula State University, Russian Federation

Dr. Dmitry PALCHUNOV –
Head of Department of General Informatics, Faculty of Information Technologies, Novosibirsk State University, Russian Federation

Dr. Panagote (Panos) PARDALOS –
Distinguished Professor and University of Florida Research Foundation Professor, Director of Center for Applied Optimization, Department of Industrial and Systems Engineering, University of Florida, USA

Dr. Albert SILANTYEV –
Professor, Department of Information Business Systems, Institute of Information Business Systems, National University of Science and Technology «MISIS», Russian Federation

Dr. Victor TARATOUKHIN –
Managing Director European Research Center for Information Systems (ERCIS) Competence Center ERP, Head of ERCIS Lab. Russia, University of Munster, Germany

Dr. Mikhail ULYANOV –
Professor,School of Software Engineering, Faculty of Computer Science,National Research University Higher School of Economics, Russian Federation

Dr. Alexey SHALKOVSKY –
Directorof Institute of Information Technologies, National Research University Higher School of Economics, Russian Federation

SQL QUERY OPTIMIZATION FOR HIGHLY NORMALIZED BIG DATA

Nikolay I. GOLOV
Lecturer, Department of Business Analytics, School of Business Informatics,
Faculty of Business and Management, National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow, 101000, Russian Federation
E-mail: ngolov@hse.ru

Lars RONNBACK
Lecturer, Department of Computer Science, Stocholm University
Address: SE-106 91 Stockholm, Sweden
E-mail: lars.ronnback@anchormodeling.com

This paper describes an approach for fast ad-hoc analysis of Big Data inside a relational data model. The approach strives to achieve maximal utilization of highly normalized temporary tables through the merge join algorithm. It is designed for the Anchor modeling technique, which requires a very high level of table normalization. Anchor modeling is a novel data warehouse modeling technique, designed for classical databases and adapted by the authors of the article for Big Data environment and a massively parallel processing (MPP) database. Anchor modeling provides flexibility and high speed of data loading, where the presented approach adds support for fast ad-hoc analysis of Big Data sets (tens of terabytes).

Different approaches to query plan optimization are described and estimated, for row-based and column-based databases. Theoretical estimations and results of real data experiments carried out in a column-based MPP environment (HP Vertica) are presented and compared. The results show that the approach is particularly favorable when the available RAM resources are scarce, so that a switch is made from pure in-memory processing to spilling over from hard disk, while executing ad-hoc queries. Scaling is also investigated by running the same analysis on different numbers of nodes in the MPP cluster. Configurations of five, ten and twelve nodes were tested, using click stream data of Avito, the biggest classified site in Russia.

Key words: Big Data, massively parallel processing (MPP), database, normalization, analytics, ad-hoc, querying, modeling, performance.

Citation: Golov N.I., Ronnback L. (2015) SQL query optimization for highly normalized Big Data. *Business Informatics*, no. 3 (33), pp. 7–14.

Introduction

Big Data analysis is one of the most popular IT tasks today. Banks, telecommunication companies, and big web companies, such as Google, Facebook, and Twitter produce tremendous amounts of data. Moreover, nowadays business users know how to monetize such data [2]. Various artificial intelligence marketing techniques can transform big customer behavior data into millions and billions of dollars. However, implementations and platforms fast enough to execute various analytical queries over all available data remain the main issue. Until now, Hadoop has been considered the universal solution. But Hadoop has its drawbacks, especially in speed and in its ability to process difficult queries, such as analyzing and combining heterogeneous data [6].

This paper introduces a new data processing approach, which can be implemented inside a relational DBMS. The approach significantly increases the volume of data that can be analyzed within a given time frame. It has been implemented for fast ad-hoc query processing inside the column oriented DBMS Vertica [7]. With Vertica, this approach allows data scientists to perform fast ad-hoc queries, processing terabytes of raw data in minutes, dozens of times faster than this database can normally operate. Theoretically, it can increase the performance of ad-hoc queries inside other types of DBMS, too (experiments are planned within the framework of further research).

The approach is based on the new database modeling technique called Anchor modeling. Anchor modeling was first implemented to support a very high speed of loading new data into a data warehouse and to support fast changes in the logical model of the domain area, such as addition of new entities, new relationships between them, and new attributes of the entities. Later, it turned out to be extremely convenient for fast ad-hoc queries, processing high volumes of data, from hundreds of gigabytes up to tens of terabytes.

The paper is structured as follows. Since not everyone may be familiar with Anchor modeling, Section 1 explains its main principles, Section 2 discusses the aspects of data distribution in an massively parallel processing (MPP) environment, Section 3 defines the scope of analytical queries, Section 4 introduces the main principles of query optimization approach for analytical queries. Section 5 discusses the business impact this approach has on Avito, and the paper is concluded in the final section.

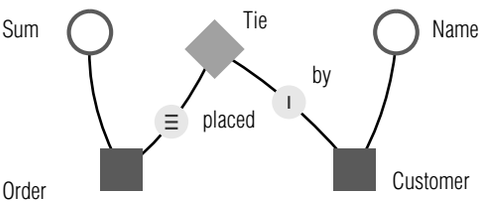
1. Anchor Modeling

Anchor modeling is a database modeling technique, based on the usage of the 6th normal form (6NF) [9]. Modeling in 6NF yields the maximal level of table decomposition, so that a table in 6NF has no non-trivial join dependencies. That is why tables in 6NF usually contain as few columns as possible. The following constructs are used in Anchor modeling:

1. *Anchor*, table of keys. Each logical entity of domain area must have a corresponding Anchor table. This table contains unique and immutable identifiers of objects of a given entity (surrogate keys). Customer and Order are two example anchors. An anchor table may also contain technical columns, such as metadata.

2. *Attribute*, table of attribute values. This table stores the values of a logical entities attributes that cannot be described as entities of their own. The Name of a Customer and the Sum of an Order are two example attributes. An attribute table must contain at least two columns, one for the entity key and one for the attribute value. If an entity has three attributes, three separate attribute tables have to be created.

3. *Tie*, table of entity relationships. For example, which Customer placed a certain Order is typical for a tie. Tie tables must contain at least two columns, one for the first entity key (that of the Customer) and one for the second (that of the related Order).



Anchor		Attribute		Attribute		Tie	
Ord ID	Cust ID	Ord ID	Sum	Cust ID	Name	Ord ID	Cust ID
12	4	12	150	4	Dell	12	4
26	8	26	210	8	HP	26	8
35		35	98			35	

Fig. 1. Anchor modeling sample data model

Ties and attributes can store historical values using principles of slow changing dimensions type 2 (SCD 2) [4]. Distinctive characteristic of Anchor modeling is that historicity is provided by a single date-time field (so-called FROM DATE), not a two fields (FROM DATE – TO DATE), as for Data Vault modeling methodology. If historicity is provided by a single field (FROM DATE), then second border of interval (TO DATE) can be estimated as FROM DATE of next sequential value, or NULL, if the next value is absent.

2. Data distribution

Massive parallel processing databases generally have shared-nothing scale-out architectures, so that each node holds some subset of the database and enjoy a high degree of autonomy with respect to executing parallelized parts of queries. For each workload posed to a cluster, the most efficient utilization occurs when the workload can be split equally among the nodes and

each node can perform as much of its assigned work as possible without the involvement of other nodes. In short, data transfer between nodes should be kept to a minimum. Given arbitrarily modeled data, achieving maximum efficiency for every query is unfeasible due to the amount of duplication that would be needed on the nodes, while also defeating the purpose of a cluster. However, with Anchor modeling, a balance is found where most data is distributed and the right amount of data is duplicated in order to achieve high performance for most queries. The following three distribution schemes are commonly present in MPP databases and they suitably coincide with Anchor modeling constructs.

Broadcast the data to all nodes. Broadcasted data is available locally, duplicated, on every node in the cluster. Knots are broadcasted, since they normally take up little space and may be joined from both attributes and ties. In order to assemble an instance with knotted attributes or relationships represented through knotted ties without involving other nodes, the knot values must be available locally. Thanks to the uniqueness constraint on knot values, query conditions involving them can be resolved very quickly and locally, to reduce the row count of intermediate result sets in the execution of the query.

Segment the data according to some operation that splits it across the nodes. For example, a modulo operation on a hash of the surrogate identifier column in the anchor and the attribute tables could be used to determine on which node data should end up. Using the same operation on anchors and attributes keeps an instance of an ensemble and its history of changes together on the same node. Assembling an instance in part or in its entirety can therefore be done without the involvement of other nodes. Since anchors and attributes retain the same sort order for the surrogate identifier, no sort operations are needed when they are joined.

Project the data with a specified sort order and specified segmentation. If table has multiple columns, if it is segmented by one column, and is joined by another column, than this join operation will require redistribution of data across all nodes. MPP databases support such joins with an ability to create *projection* – full copy of a table, segmented and sorted by another column. There can be multiple projections of a single table, according to business needs. Ties have one projection for each role in the tie, where each segmentation is done over and ordered by the surrogate identifier representing the corresponding role. This ensures that all relationships that an instance takes part in can be resolved without the involvement of other nodes, and that joining a tie will not require any sort operations. Therefore, tie will require no more than one additional projection.

Usage of single date-time field for historicity (SCD2) gives significant benefit – it does not require updates. New values can be added exclusively by inserts. This feature is very important for column-based databases, that are designed for fast select and insert operations, but are extremely slow for delete and update operations. Single date-time field historicity also requires efficient analytical queries to estimate closing date for each value. Distribution strategy, described above, guarantees, that all values for single anchor are stored on a single node and properly sorted, so closing date estimation can be extremely efficient.

3. Analytical queries

Databases can be utilized to perform two types of tasks: OLTP tasks or OLAP tasks. OLTP means online transaction processing: huge number of small insert, update, delete, select operations, where each operation processes small chunk of data. OLTP tasks are typical for operational databases. OLAP means online analytical processing: relatively small number of complex analytical select queries, where each query processes significant part of data (or whole data). Given article is focused on OLAP queries to Big Data.

Analytical queries may include three types of subtasks:

1. Combine dataset according to query. Big Data is stored as tables, which have to be combined (joined) according to given conditions.

2. Filter dataset according to query. Query can contain conditions on the data: single column conditions, column comparisons and sampling conditions (for example «top 10 percent of user accounts according to total number of payments of each user account»).

3. Calculate aggregates over filtered datasets.

4. Efficient plans for analytical queries

SQL can describe almost all types of analytical queries. It is one of the main advantages of SQL and relational databases: if analytical query can be formulated using SQL, than relational database can process it and return correct result. Main problem lies in the optimization of the query execution time.

In this paper, queries are analyzed from a Big Data analyst's point of view. Analysts want to get their answers as fast as possible. Query with suboptimal execution plan can be processed tens of times slower than optimal one. If query processes Big Data, inefficient plan can work hours and days, while efficient one can return result in minutes.

This section contains main principles of optimization analytical queries to Big Data in MPP environment, obtained by Avito. Growth of Avito data warehouse enabled its analysts to study various approaches to query optimization and chose most efficient ones, able to speed queries to Big Data in an MPP environment from hours to minutes. These researches were launched at the initial phase of DWH development, because there were concerns about speed of OLAP queries to a Anchor Model. On the further phases, those results were applied to analysis of both, normalized data (Anchor modeling) and denormalized data (data marts) with equal effectiveness.

Here is a list of main risks of query execution plan in a MPP environment:

✦ **Join spill.** According to Section 3, analytical tasks can require joins. Joins can be performed via Hash Join, Nested Loop Join and Sort-Merge join algorithms. Nested loop join is inefficient for OLAP queries (it's best for OLTP) [3]. Sort phase of sort-merge algorithm join can be inefficient for MPP environment, so if source date is not sorted, query optimizer prefers hash join algorithm. Hash join is based on storing left part of the join in RAM. If RAM is not enough, query will face a join spill. Join spill (term can differ in various DBMS-s) mean that some part of query execution plan requires too much RAM, and source data have to be separated into N chunks (small enough to fit into available RAM) to be processed consequently and at the end - to be combined together. Join spill reduces maximum RAM utilization, but increases disk I/O, a slower resource. Following diagram illustrates, why a condition reducing a table on one side of the join may not reduce the number of disk operations for the table on the other side of the join. Therefore, disk I/O operations can be estimated according to *optimistic* (concentrated keys) or *pessimistic* (spread out keys) scenarios. In the optimistic one, the number of operations is the same as in a RAM join, whereas in the pessimistic one the whole table may need to be scanned for each chunk, so disk I/O can increase N times. *According to modern servers(>100Gb of RAM for a node), this risk is actual when left part of the join contains over one billion of rows.* Hash join for tables of millions of rows is safe.

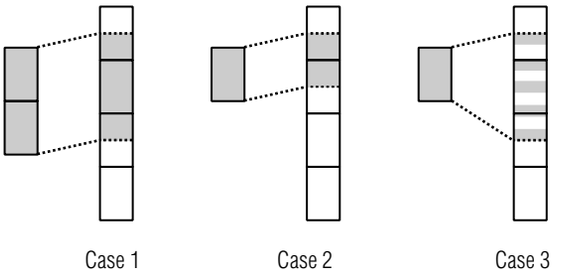


Fig. 2. Three cases of inner join

Figure above demonstrates three cases of inner join. First case: smaller table contain two disk pages of keys, correspondent keys in bigger table occupy three disk pages. Second case: smaller table reduced twice, to one disk page, correspondent keys in bigger table are concentrated, so to read them one need to scan only two disk pages instead of three. Third case: smaller table reduced twice, but correspondent keys are not concentrated but distributed (see gaps), so one need to read same number of disk pages, as before (three).

✦ Group by spill. Group by phase of the query can also be executed by Hash Group algorithm, and it also can be spilled on disk, as a join phase. *This risk is actual if there are more than hundreds of millions of groups in a group by phase.*

✦ Resegmentation/broadcast. If two tables are joined together, and they are joined and segmented over nodes of MPP cluster using different keys, than joining may be performed either by resegmentation of one table across all nodes, or by broadcasting all data of one table across all nodes (see Section 2). *This risk is actual if tables in question contains billions of rows.*

✦ **Wrong materialization**. Materialization strategy is extremely important for column-based databases [8]. Materialization strategy defines process of combining column data together (in a column based databases they are stored separately). It is extremely important for join queries. There can be early materialization (read columns, combine columns, than apply filter), or late materialization (read filtered column, apply filter, read other columns according to filtered rows, combine). According to [1], it is more efficient to select late materialization strategy. But if a query is complex, query optimizer can make a mistake. If query requires join of tables A, B and C, and has a filtering condition to tables A and B, late materialization may lead to loading of filtered values from A, B in parallel, all values from C, and join them together. This strategy is efficient if condition on table A and B reduce data a lot, while join with tables B and C don't reduce data. Otherwise, if each condition takes ¼ of A and B, and conditions are uncorrelated, than join-

ing filtered tables A and B will result in 1/16 part of data, so early materialization can significantly reduce disk I/O of tables B and C.

Risks, listed above, can be avoided by using modified query execution plans, based on following principles:

✧ **Maximum merge join utilization.** Sort-Merge join algorithm is a most effective for non-selective analytical queries among such algorithms as Hash Join, Nested Loop Join and Sort-Merge join, especially if analyzed data are already sorted on join key, so only merge join operation is required, without sort operations [3].

✧ **Single time disk page access.** Query plan with a join spill has awful characteristics in terms of pessimistic number of disk page access operations because many disk pages are accessed multiple times during plan execution. To create plan with optimal characteristics each required disk page must be accessed no more than once.

✧ **Massive temporary table utilization.** Two principles, mentioned above, cannot be guaranteed for all ad-hoc queries to real-world Big Data databases, created according to 1/2/3 normal form. Table of dozen columns cannot be sorted on each column (without duplicating this table dozen times). Also it cannot be guaranteed that all keys in joined tables are concentrated, to equalize pessimistic and optimistic disk page operation count (see *fig. 2*). But this principles can be guaranteed in a Anchor Modeling data model, if each data subset, retrieved after each step of filtering, is stored as a temporary table. Temporary tables can be sorted on specific column, to optimize further merge join. Also those temporary tables contain concentrated keys, only keys, required for given ad-hoc query. This feature helps to avoid double disk page access. Concerning speed of creating temporary table for large data set: modern HDDs provide equal speed for sequential read and sequential write. **Important.** Given algorithms don't require creation of indexes of any kind. Only storing sorted rows as a temporary table, so this operation can be as fast, as saving a data file on HDD. Sorting is enough for efficient Merge Join.

Efficient plans for analytical queries, utilizing principles, listed above, can be created using following algorithm:

0. Assume that query is based on joining tables $T_1,...T_N$, filtered according to complex conditions $C_1^1,...C_N^1,...C_1^N,...C_N^N$, there C_j^i is a condition on table j, which can be estimated only when i tables are already loaded. So, C_j^i is single table filtering condition. C_j^i is a combination of all conditions on table j.

1. Estimate data reduction rate for each table

$$\frac{R(C_j^1(T_j))}{R(T_j)},$$

where $R(T_j)$ means number of rows. Chose the tables with best reduction rate to be the first one (assume that it is table T_1).

2. Take k_1 tables, identically segmented with table T_1. Assume that those tables are $T_1,...T_{k_1}$. Join all k_1 tables, using all applicable filtering conditions, using merge join algorithm, consequently, starting from table T_1. So table T_1 has to be scanned and filtered first, than only corresponding rows from T_2 has to be scanned and filtered according to conditions C_2^1 and C_2^2, than the same for tables $T_3,...T_{k_1}$.

3. Estimate data reduction rate for remaining tables

$$\frac{R\left(C_j^{k1+1}\left(T_j\right)\right)}{R\left(T_j\right)},$$

using data tables, joined and filtered on step 2. Chose the tables with best reduction rate to be the next one (assume that it is table T_{k1+1}).

4. Save results of joining and filtering tables from step 2 inside a temporary table T_1^{new}, identically segmented with table T_{k1+1}.

5. Replace tables $T_1,...T_{k_1}$ of initial list with a table T_1^{new}.

6. Return to step 2 with reduced list of tables.

7. It list of tables contain single table, perform final aggregations, save results.

Given algorithm has important drawbacks:

1. It requires a lot of statistics about source tables, about data reduction rate of conditions, for step 1. In some cases first estimation of statistics can require more efforts, than analytical query itself.

2. Statistics estimation, as well as creation of temporary tables with dedicated segmentation, has to be implemented manually, because current version of query optimizers and query processing engines can not do it. Avito used Python implementation.

Drawback 1 can be avoided using estimates from previously calculated queries, continuously. First processing of brand new query can be time-consuming.

Drawback 2 can be avoided for identically segmented tables, because algorithm ends on step 2. That is why Anchor Modeling data model for MPP environment is especially favorable for approach. Step 2 does not require implementation of additional logic, it can be performed using SQL, database query execution engine and some hints.

5. Business applications

The approach described in Section 4 has been implemented for ad-hoc querying in a Big Data (data warehouse) solution at Avito. Based in Moscow, Avito is Russias fastest growing e-commerce site and portal, «Russia's Craiglist». The Avito data warehouse is implemented according to the Anchor modeling methodology. It contains ≈ 200 Anchors, ≈ 600 Attributes, and ≈ 300 Ties. Using this model it loads, maintains, and presents many types of data. The greatest volume can be found in the part loaded from the click stream data sources. Click stream data are loaded every 15 minutes, with each 15-minute batch containing 5 mln. rows at the beginning of 2014 and 15 mln. at the beginning of 2015.

Each loaded row contained approximately 30 columns at the beginning of 2014 and approximately 70 columns at the beginning of 2015. Each column is a source for a separate Anchor, Attribute or Tie table. The current size of the Avito data warehouse has been limited to 51Tb for licensing reasons. It contains years of consistent historical data from various sources (back office, Google DFP/AdSense, MDM, accounting software, various aggregates), and a rolling half year of detailed data from click stream.

The presented approach for ad-hoc queries was successfully implemented by the Avito BI team in less than a year. It provides analysts at Avito with a tool for fast (5-10-2060 minutes) analysis of near-real time Big Data, according to various types of tasks, such as direct marketing, pricing, CTR prediction, illicit content detection, and customer segmentation.

Conclusions

A high degree of normalization has the advantage of flexibility. The resulting data models can be extended easily and in a lossless manner. Anchor modeling also enables parallel loading of all entities, their attributes, and links with or without historization of each link and attribute. The supposed drawback of a highly normalized data model is slow ad-hoc reporting, which is why Inmon [4] recommends combining normalized models for centralized repositories with denormalized data marts. The paper, however, demonstrates that while Anchor modeling may require some sort of denormalization for single-entity analysis, it can yield very high performance for queries that involve multiple linked entities, and particularly so when there is a risk of join spill (lack of RAM).

The experiments carried out showed benefits of the given approach for the simplified case of two linked entities. Realworld business cases sometimes require three, four or even a dozen linked entities in the same ad-hoc query. Such cases multiply the risk of join spill occurring in some step, and amplify its negative effect. If number of joins, tables and filtering conditions increases, query RAM requirements estimation accuracy decreases. Query optimizer can significantly overestimate RAM requirements and cause an unnecessary join spill with all associated drawbacks. The approach from Section 4 has been in use at Avito for over a year, for ad-hoc and regular reporting, and even for some near-real time KPIs. It has demonstrated stability in terms of execution time and resource consumption, while ordinary queries often degraded after months of usage because of data growth.

One can wonder if the approach from Section 4 can be applied for implementations based on other techniques than Anchor modeling. The answer is yes, it can be applied, but with additional work. The strictly determined structure of Anchor modeling enables Avito to have a fully automated and metadata-driven implementation of the given approach. To conclude, the approach, while not being uniquely tied to Anchor modeling, suits such implementations very well and compensates for the commonly believed drawback of normalization in the perspective of ad-hoc querying. ■

References

1. Abadi D.J., Myers D.S., DeWitt D.J., Madden S.R. (2007) Materialization strategies in a Column-Oriented DBMS. Proceedings of the *23rd IEEE International Conference on Data Engineering (ICDE 2007), April 15-20, 2007, Istanbul, Turkey*, pp. 466–475.
2. Chen M., Mao S., Liu Y. (2014) Big Data: A survey. *Mobile Networks and Applications*, no. 19, pp. 171–209.
3. Graefe G. (1999) The value of merge-join and hash-join in SQL Server. Proceedings of the *25th International Conference on Very Large Data Bases (VLDB 1999), September 7-10, 1999, Edinburgh, Scotland*, pp. 250–253.
4. Inmon W.H. (1992) *Building the Data Warehouse*, Wellesley, MA: QED Information Sciences.
5. Kalavri V., Vlassov V. (2013) MapReduce: Limitations, optimizations and open issues. Proceedings of the *12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom 2013)*, July 16-18, 2013, Melbourne, Australia, pp. 1031–1038.

6. Knuth D. (1998) *The art of computer programming. Volume 3: Sorting and searching*. 2nd edition, Boston, MA: Addison–Wesley.
7. Lamb A., Fuller M., Varadarajan R., Tran N., Vandiver B., Doshi L., Bear C. (2012) The Vertica analytic database: C-store 7 years later. Proceedings of *the VLDB Endowment*, vol. 5, no. 12, pp. 1790–1801.
8. Shrinivas L., Bodagala S., Varadarajan R., Cary A., Bharathan V., Bear C. (2013) Materialization strategies in the Vertica analytic database: Lessons learned. Proceedings of the *29th IEEE International Conference on Data Engineering (ICDE 2013), April 8-11, 2013, Brisbane, Australia*, pp. 1196–1207.
9. Ronnback L., Regardt O., Bergholtz M., Johannesson P., Wohed P. (2010) Editorial: Anchor Modeling – agile information modeling in evolving data environments, *Data and Knowledge Engineering*, vol. 69, no. 12, pp. 1229–1253.

ОПТИМИЗАЦИЯ SQL-ЗАПРОСОВ ДЛЯ ВЫСОКОНОРМАЛИЗОВАННЫХ БОЛЬШИХ ДАННЫХ

Н.И. ГОЛОВ
преподаватель кафедры бизнес-аналитики,
школа бизнес-информатики, факультет бизнеса и менеджмента,
Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20
E-mail: ngolov@hse.ru

Л. РОНБАК
преподаватель факультета компьютерных наук, Университет Стокгольма
Адрес: SE-106 91 Stockholm, Sweden
E-mail: lars.ronnback@anchormodeling.com

В данной статье описывается подход для быстрого анализа больших данных в реляционной модели данных. Целью данного подхода является достижение максимального использования высоконормализанных временных таблиц, объединяемых посредством алгоритма соединения слиянием (merge join algorithm). Подход был разработан для методологии Anchor Modeling, предполагающей крайне высокий уровень нормализации таблиц. Anchor Modeling – это новейшая методология построения хранилищ данных, разработанная для классических баз данных и адаптированная для задач больших данных и MPP (массивно-параллельных) баз данных авторами статьи. Anchor Modeling обеспечивает гибкость расширения и высокую скорость загрузки данных, в то время как представленный подход к оптимизации запросов дополняет методологию возможностью «на лету» проводить быстрый анализ больших выборок данных (десятки Тб).

В статье описаны и оценены различные подходы к оптимизации планов выполнения запросов для колоночных и обычных (строчных) баз данных. Представлены и сопоставлены результаты теоретических оценок и практических экспериментов на реальных данных, проведенных на платформе колоночной массивно-параллельной (MPP) базы данных HP Vertica. Результаты сравнения демонстрируют, что подход особенно эффективен для случаев нехватки доступной оперативной памяти, в результате чего оптимизатору запросов базы данных при обработке аналитических запросов приходится переходить от наиболее оптимального режима обработки в оперативной памяти (in-memory) к режиму подкачки с жесткого диска. Также изучен вопрос масштабирования нагрузки. Для этого один и тот же анализ производился на кластерах массивно-параллельной СУБД Вертика, состоящих из разного количества серверов. Были испытаны конфигурации из пяти, десяти и двенадцати серверов. Для анализа применялись данные типа «поток кликов» – обезличенные данные о кликах пользователей Авито, крупнейшего российского сайта объявлений.

Ключевые слова: большие данные, массивно-параллельная обработка (MPP), база данных, нормализация, аналитика, аналитика «на лету», запросы, моделирование, производительность.

Цитирование: Golov N.I., Ronnback L. SQL query optimization for highly normalized Big Data // Business Informatics. 2015. No. 3 (33). P. 7–14.

Литература

1. Abadi D.J., Myers D.S., DeWitt D.J., Madden S.R. Materialization strategies in a Column-Oriented DBMS // Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE 2007), April 15-20, 2007, Istanbul, Turkey. IEEE, 2007. P. 466–475.
2. Chen M., Mao S., Liu Y. Big Data: A survey // Mobile Networks and Applications. 2014. No. 19. P. 171–209.
3. Graefe G. The value of merge-join and hash-join in SQL Server // Proceedings of the 25th International Conference on Very Large Data Bases (VLDB 1999), September 7-10, 1999, Edinburgh, Scotland. 1999. P. 250–253.
4. Inmon W.H. Building the Data Warehouse. Wellesley, MA: QED Information Sciences, 1992. 272 p.
5. Kalavri V., Vlassov V. MapReduce: Limitations, optimizations and open issues // Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom 2013), July 16-18, 2013, Melbourne, Australia. IEEE, 2013. P. 1031–1038.
6. Knuth D. The art of computer programming. Volume 3: Sorting and searching. 2nd edition. Boston, MA: Addison–Wesley, 1998. 782 p.
7. The Vertica analytic database: C-store 7 years later / A. Lamb [et al] // Proceedings of the VLDB Endowment. 2012. Vol. 5, No. 12. P. 1790–1801.
8. Shrinivas L., Bodagala S., Varadarajan R., Cary A., Bharathan V., Bear C. Materialization strategies in the Vertica analytic database: Lessons learned // Proceedings of the 29th IEEE International Conference on Data Engineering (ICDE 2013), April 8-11, 2013, Brisbane, Australia. IEEE, 2013. P. 1196–1207.
9. Editorial: Anchor Modeling – agile information modeling in evolving data environments / L. Ronnback [et al] // Data and Knowledge Engineering. 2010. Vol. 69, No. 12. P. 1229–1253.

CREDIT SCORING BASED ON SOCIAL NETWORK DATA

Alexey A. MASYUTIN
Post-graduate student, School of Data Analysis and Artificial Intelligence,
Faculty of Computer Science, National Research University Higher School of Economics

Address: 20, Myasnitskaya Street, Moscow, 101000, Russian Federation
E-mail: alexey.masyutin@gmail.com

Social networks accumulate huge amounts of information, which can provide valuable insights on people's behavior. In this paper, we use social data from Vkontakte, Russia's most popular social network, to discriminate between the solvent and delinquent debtors of credit organizations. Firstly, we present the datacenter architecture for social data retrieval. It has several functions, such as client matching, user profile parsing, API communication and data storing. Secondly, we develop two credit scorecards based exclusively on social data. The first scorecard uses the classical default definition: 90 days delinquency within 12 months since the loan origination. The second scorecard uses the classical fraud definition as falling into default within the first 3 months. Both scorecards undertake WOE-transformation of the input data and run logistic regression afterwards. The findings are as follows: social data better predict fraudulent cases rather than ordinary defaults, social data may be used to enrich the classical application scorecards. The performance of the scorecards is at the acceptable level, even though the input data used were exclusively from the social network. As soon as credit history (which usually serves as input data in the classical scorecards) is not rich enough for young clients, we find that the social data can bring value to the scoring systems performance. The paper is in the area of interest of banks and microfinance organizations.

Key words: credit scoring, social networks, probability of default, social data, Vkontakte.

Citation: Masyutin A.A. (2015) Credit scoring based on social network data. *Business Informatics*, no. 3 (33), pp. 15–23.

1. Introduction

Social networks are an inexhaustible source of information about people. According to its financial filings at the time of its IPO which took place in 2012, Facebook stored around 111 megabytes of photos and videos for each of its users [1], whose number now exceeds a billion. That adds up to 100 petabytes of personal information.

The data are mostly not structured and chaotic by type, but they allow one to get to know their clients much deeper. Properly processed and then structured information about users brings value to online stores, recruitment agencies, banks and many other business-to-client and business-to-business companies. E-commerce is able to learn more about the behavior and preferences of thecustomers, and banks can determine the credit rating of the borrowermore accurately.

For example, when assessing a person's income, they can use information about the places they visit, as well as the countries theytravelto. The places and countries are classified according to price category. Further formed scorecards are lists of places and countries with ratios of their influence on the level of creditworthiness. For each country, the scorecard will be unique.

However, there are several issues concerning social data implementation. First, the companies have to construct solid business processes, which would adopt data-driven decision-making, before they can extract value from the flow of social data. The cornerstone of such processes is the data retrieval mechanism and architecture of data storage. Here companies face a dilemma: to build up the database within local IT department, or to outsource the social data retrieval to the third parties (SaaS). The latter case is very similar to

credit history bureau query, when banks request information on the applicants.

The second issue is related to the legal aspects of personal data processing. Obviously, when dealing with personal data from social networks,companies should learn to do it legally, so as not to be subject to prosecution. The laws vary from country to country and can dramatically influence the ability to use the social data.

Thirdly, the company must have enough competencies to perform the data analysis. The techniques are versatile machine learning algorithms, ranging from logistic regression to sophisticated data mining techniques. Again, the company can outsource this process as soon as there are many consulting analytics (e.g. SAS).

We have to mention that there is not much academic literature covering the use of social data in credit scoring [5]. However, there is a number of papers analyzing the social networking from the sociological point of view [6]. There were also several attempts to bind the activities within social networks (such as Twitter) with movements in stock prices [8]. The scarcity of academic research is explained by the fact that the use of social data in scoring started no more than 3-4 years ago. Moreover, the accumulation of particular results concerning the use of social data takes place in the business area, rather than academic area. The companies that launch such «social data» projects aren't likely to focus on derivation of universal laws or theoretical results. Besides, the companies are running their activities in a highly competitive environment and are not interested in sharing the knowledge at instance, whichis actually implied by the academic style of research. At least, we can list some companies whose domain is social data retrieval, aggregation and customer analytics for credit organizations: Wonga (USA), Kreditech (Germany), Big Data Scoring (Estonia, Finland), Lenddo (Philippines, Columbia), SOCSCOR (Russia), Crediograph (Ukraine).

In our analysis we will examine the predictive power of the social data from Vkontakte social network (also VK). It is the number one in Russia by visitors per month. Its monthly audience makes a total of over 50 million users.

This paper consists of four parts. The first one is introduction. The second one describes the typical architecture for social data retrieval and integration with bank databases. It will be based on an example of one Russian bank whose name we cannot disclose due to the confidential reasons. The third part involves data description used for default prediction, discusses the variables relevant to the default event and considers prediction accuracy. The fourth part is conclusion.

2. Social Data Retrieval and Storage

2.1. Social network objects

First of all, we define the set of notions we use when talking about social data. A profile is a set of properties describing a user of a social networking service. It can include their name, patronymic name, last name, date of birth, place of residence, work and study, interests, communities, friends and feed.

A feed is a part of a user's profile which contains events describing the user: messages, what they are fond of ('likes'), communities, applications installed etc.

Open data are a part of a user's profile, which is accessible without loggingin and can be retrieved automatically.

Accessible data are, meanwhile, another part of a user's profile, which can be accessed from other users' accounts and which can be retrieved automatically on a regular basis using someone's else account. It does not require other users' personal involvement and there is no limit on the number of pattern messages sent from this user's account. In addition to this, accessible data should be structured, e.g. the feed should contain XML or JSONmarkings.

Parsing is the process of matching a linear sequence of lexemes (words, tokens) of a natural or formal language with its formal grammar.

2.2. Retrieval process and datacenter architecture

The basic tool for data retrieval is the so-called API (application programming interface). In effect, it is a set of ready-to-use classes, procedures, functions, structures and constants offered by an application (library, service) to be used in outer software products. A social network user must allow the access to their personal data within the social network. This usually happens when the user installs an API application, it can be a game or a discount offer. The application is granted rights to perform API requests to the user profile and activities at any time.

The architecture of the solution is shown below in the *Fig. 1.*

MDM stands for Master Data Management system, a core banking information system. Apparently, the credit applicant must be identified with Vkontakte user. The identification is carried out by a combination of parameters: first and last name, date of birth, current city, city of origination, e-mail, or exact social network id (if the applicant gives such information via the application).

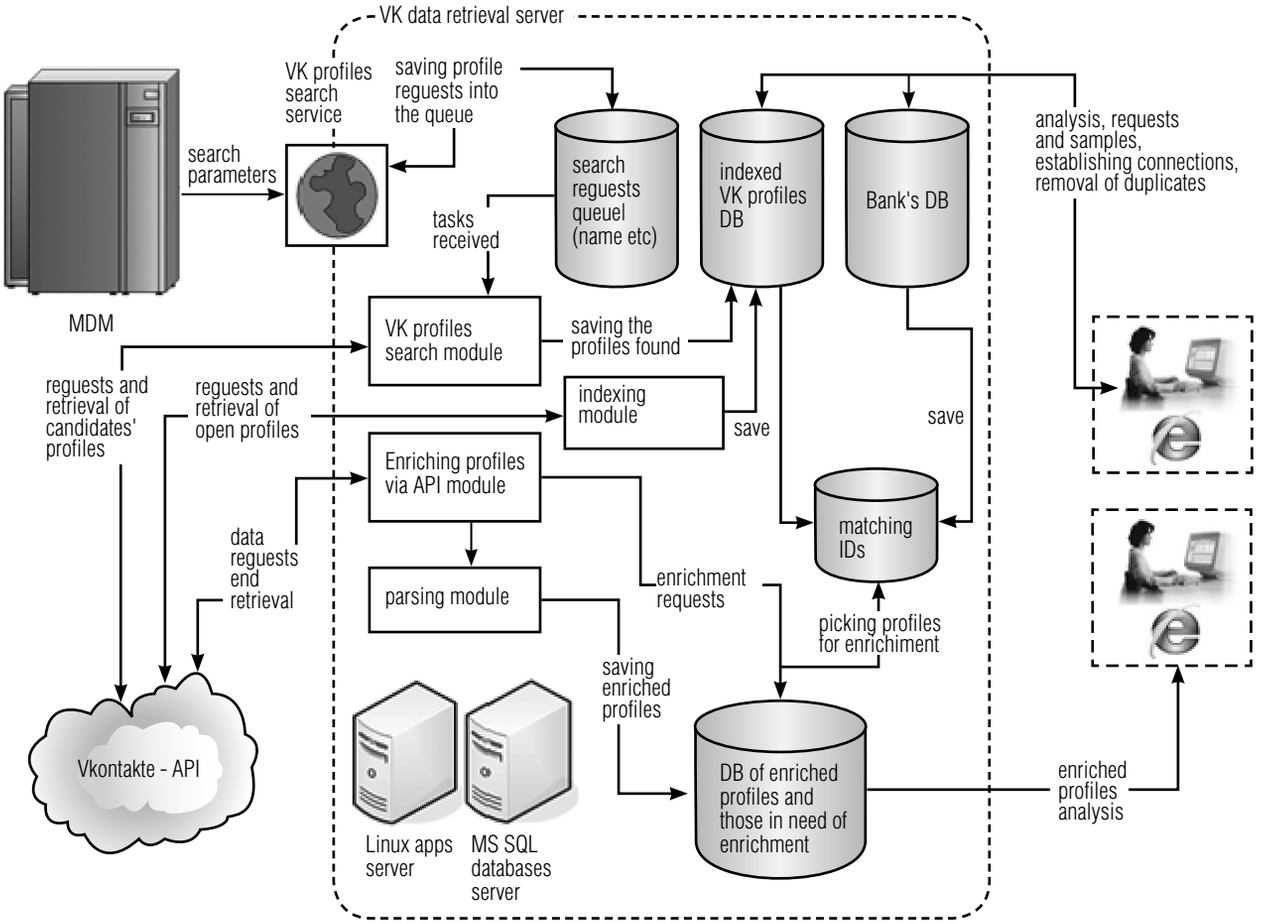


Fig. 1. The datacenter architecture for social data retrieval and storage

The architecure of the datacenter is aimed to accomplish the following tasks:

✦ Downloading all Vkontakte profiles,

✦ Appending newly created profiles on a regular basis,

✦ Downloading detailed data (subscriptions, communities, friends, wallposts) of Vkontakte profiles with the span of 3 years,

✦ Regular downloading detailed data of VK profiles (subscriptions, communities, friends, new posts on the wall since the last upload).

Additionaly, it provides an interface to browse, search and compare the downloaded data to clients' data from the core banking system. Downloading and updating is performed in astreaming mode according to the chosen set of profiles, the updating period can vary from several days to several weeks, i.e. the downloading is not performed online.

The solution has following volume indicators:

✧ Downloading and storage of all Vkontakte profiles for indexing purposes (about 250 million),

✧ Downloading detailed data on 300,000 Vkontakte profiles.

The system is able to adjust to scale for downloading a 5 times larger number of profiles without significant changes in its architecture and functions.

3. Data Description and Default Prediction

The event of default in retail banking is defined as more than 90 days of delinquency within the first 12 months after the loan origination. Defaults are divided into fraudulent cases and ordinary defaults. The default is told to be a fraudulent case when delinquency starts at one of the three first months. It means that when submitting a credit application, the borrower did not even intend to payback. Otherwise, the default is ordinary when the delinquency starts after the first three months on book.

That is why scorecards are usually divided into fraud and application scorecards. In fact the only difference is

the target variable definition, while the sets of predictors and the data mining techniques remain the same. The default cases are said to be «bad», and the non-default cases are said to be «good».

We consider a dataset of 27540 microfinance loans, originated in 2012. This segment is characterized by small-cash high-margin loans with short credit term (3-6 months). The further details of the data source are not presented due to the non-disclosure agreement. We develop both scorecards and examine their accuracy via out-of-sample validation. The validation process requires calculation of performance metrics (ROC-curve and Gini coefficient) of the model based on the data sample that was retrieved from the same parent population but was not used to develop the model itself (validation set). This approach allows the user to check for accuracy and stability of the model. The size of the validation set we used was 30% from the parent population.

The analytics was carried out using SAS Enterprise Miner, the most spread analytical software in the banking sphere.

The mathematical architecture of the scorecard is based on a logistic regression, which takes the transformed variables as input. The transformation of the initial variables we use is WOE-transformation [3]. It is wide-spread in credit scoring, to apply such a transformation to the input variables as soon as it accounts for non-linear dependencies and provides certain robustness coping with potential outliers. The aim of the transformation is to divide each variable into no more than k categories. At step 0, all the continuous variables are binned into 20 quantiles, the nominal and ordinal variables are left as they are. Now, when all the variables are categorized, we compute the odds ratio for each category. Then for each predictor variable $X_i (i = 1 \ldots n)$ we merge non-significant (chi-square statistics based on differences in odds) categories.

1. If X_i has 1 category only, stop and set the adjusted p-value to be 1.

2. If X_i has k categories, go to step 7.

3. Else, find the allowable pair of categories of X_i (an allowable pair of categories for ordinal predictor is two adjacent categories, and for nominal predictor is any two categories) that is least significantly different (i.e., most similar). The most similar pair is the pair whose test statistic gives the largest p-value with respect to the dependent variable Y.

4. For the pair having the largest p-value, check if its p-value is larger than a user-specified alpha-level α merge. If it does, this pair is merged into a single compound category. Then a new set of categories of X_i is formed. If it does not, then if the number of categories is less or equal to k go to step 6, else merge two categories with highest p-value.

5. Go to step 2.

6. (Optional) Any category having too few observations (as compared with a user-specified minimum segment size) is merged with the most similar other category as measured by the largest of the p-values.

7. The adjusted p-value is computed for the merged categories by applying Bonferroni adjustments [4].

Having accomplished the abovementioned steps, we acquire categorized variables instead of the continuous ones. When each variable $X_i (i = 1 \ldots n)$ is finally binned into a certain number of categories (k_i), we are able to calculate the odds for each category j ($j = 1 \ldots k_i$), the weight of evidence for each category and, as a result, information value for each variable:

$$odds_{ij} = \frac{\%\, goods_{ij}}{\%\, bads_{ij}}$$

$$WOE_{ij} = ln(odds_{ij}) \qquad (1)$$

$$IV_i = \sum_{j=1}^{k_i} (\%\, goods_{ij} - \%\, bads_{ij}) \cdot WOE_{ij}.$$

Information value can also be used in feature selection (e.g. variables with information value less than 0.02 are often rejected, as soon as they do not reveal predictive power). More details on WOE-transformation are provided in [7].

The role of the WOE-transformation is that, instead of initial variables, the logistic regression receives WOE as input. So, each input variable is a discrete transformed variable, which takes values of WOE. When estimating the logistic regression, the usual maximum likelihood was applied.

3.1. Ordinary default scorecard

A scorecard is a set of rules. Each rule gives a number of scorepoints to the applicant. Then the scorepointsare summed up to form the final score. The higher the score, the less the probability of default. The variables were selected from a set of about 100 calculated parameters. The criteria for includinga variable into the scorecard was Information value no less than 0.2 and correlation analysis within variables themselves (so as to rule out multi-collinearity).

Table 1.

Ordinary Default

VAR CODE	ORDINARY DEFAULT SCORECARD		
	Variable name	Value range	Scorepoint
Grouped Levels for x_{11}	Marital status	'In love', 'Engaged'	11
		'Single'	19
		Missing	23
		'Married'	27
Grouped Levels for x_{69}	Political views	'Communistic', 'Monarchic', 'Socialistic'	16
		'Indifferent', 'Liberal', Missing	24
Age	Age (in years)	age< 25	5
		25<= age< 28	12
		28<= age< 37	21
		37<= age< 52	44
		52<= age	66
Sex	Sex	Female	27
		Male	15
x_1	Number of days since last visit	x1< 1	26
		1<= x1< 3	30
		3<= x1< 37	20
		37<= x1< 149	17
		149<= x1	11
x_{30}	Number of days since the first post	x30< 265	20
		265<= x30< 399	15
		399<= x30< 1330.5	23
		1330.5<= x30	36
x_{76}	Number of job places	0, Missing	21
		>=1	29

We use the ROC analysis to evaluate the performance of the scorecard. The ROC analysis provides the set of feasible accuracy measures, which form an ROC curve and reflect the accuracy of the classifier. In terms of classification, the scorecard labels the applicant as positive (high probability of default) when its score is less than the cutoff (decision boundary). The accuracy is defined by the probability of the two possible errors. The first one is to approve the loan for a bad applicant (false negative), and the second one is to reject a good applicant (false positive). This gives us the two dimensions for the ROC curve: sensivity (vertical axis) and one minus specificity (horizontal axis):

$$Sensivity = \frac{TruePositive}{Positive}; \quad (2)$$

$$Specificity = \frac{TrueNegative}{Negative} \Rightarrow$$

$$\Rightarrow 1 - Specificity = \frac{FalseNegative}{Negative}. \quad (3)$$

The ROC curves for train and validation set are presented below in *Figures 2* and *3*:

Fraudulent case scorecard can be found below in *Table 2.*

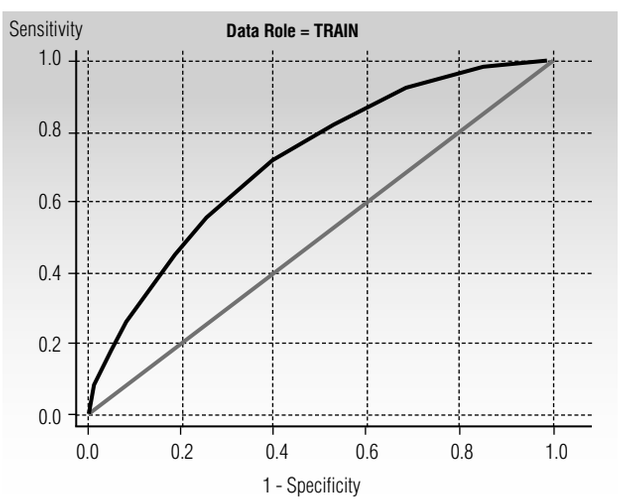


Fig. 2. ROC-curve for ordinary default (training)

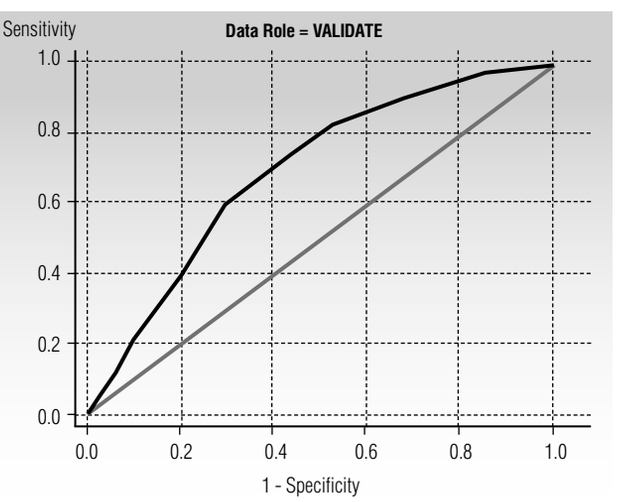


Fig. 3. ROC-curve for ordinary default (validation)

Table 2.

Fraudulent Case

VAR CODE	FRAUDULENT CASE SCORECARD		
	Variable name	Value range	Scorepoint
Age	Age (in years)	age< 29	3
		29<= age< 33	11
		33<= age< 39	25
		39<= age< 47	42
		47<= age	62
Sex	Sex	Male	4
		Female	26
x_1	Number of days since last visit	x1< 2	21
		2<= x1< 3	0
		3<= x1< 10	22
		10<= x1< 1565	10
		1565<= x1	-4
x_{11}	Marital status	'In love', 'All is difficult', 'In couple'	4
		'Single'	11
		'Engaged'	14
		'Married'	24
x_{21}	Number of subscriptions	x21< 2, _MISSING_	19
		2<= x21< 3	17
		3<= x21< 7	20
		7<= x21< 16	11
		16<= x21	4
		1330.5<= x30	8
x_{30}	Number of days since the first post	x30< 497	26
		497<= x30< 710	21
		710<= x30< 1324	14
		1324<= x30	13
x_{39}	Number of user's posts with photos	x39< 1	13
		1<= x39< 2	18
		2<= x39< 4	15
		4<= x39< 44	11
		44<= x39	19
x_{41}	Number of user's posts with video	x41< 1, Missing	15
		1<= x41< 2	17
		2<= x41< 3	13
		3<= x41< 16	12
		16<= x41	5
x_{51}	Number of children	x51< 1, Missing	12
		1<= x51< 2	13
		2<= x51	15
x_{65}	Major things in life	'Career and Money', 'Entertainment', 'Fame and Influence'	2
		'Beauty and Art', 'Research and Science', Missing	16
x_{66}	Major qualities in people	'Kindness and Honesty', 'Humor and Lust', 'Health and Beauty'	19
		'Courage and Perseverance', 'Mind and Creativity'	14

The corresponding ROC analysis can be found below in *Figures 4* and *5.*

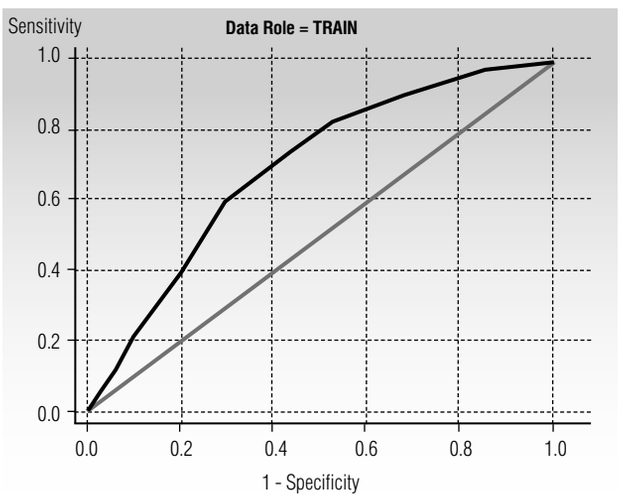


Fig. 4. ROC-curve for fraudulent cases (training)

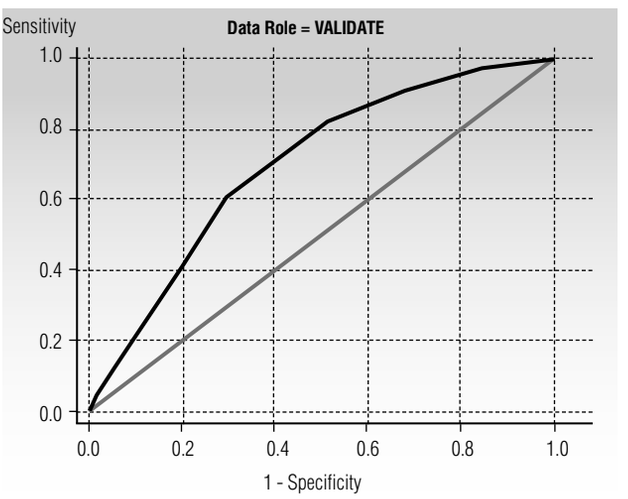


Fig. 5. ROC-curve for fraudulent cases (validation)

The ROC analysis show that both models are of mild predictive power and of high stability. Gini coefficients are 36% (32%) for train (validation) sample for ordinary default scorecard. The corresponding Gini coefficients for fraudulent case scorecard are 47% (39%). Of course, the main contribution was made by Age and Sex variables, which are known even without social network. However, if other variables are excluded, the Gini coefficients fall by 7-8%.

From our point of view, the result that fraudulent case scorecard showed better performance can be explained in the following way. When the borrower initially intends to receive a loan and not to pay any installments right from the start, of course, he or she tries to disclose less information in the social network. Missing values of predictors' parameters tend to receive lower scorepoints. As far as the interests in life are concerned, the person who is obsessed with 'Entertainment' or 'Fame' is, to our mind, more likely to try to get easy money by receiving a loan with no intent to pay it back.

If we compare the Gini coefficients of our models to the ones of the classical application scoring models (usually 45-55%, see [2]), we find that the latter are more accurate. That brings the understanding that the social data indeed brings value and increases discriminative power,however, it should not be used standalone. The social data should enrich the usual set of application variables and contribute to the higher discriminative power of banking scoring. Such performance metrics as Gini coefficient are not the only way to estimate the efficiency of social data implementation. In fact, the most important criterion is the profit increase due to the model implementation. However, the final effect depends on the particular bank or microfinance organization that is considering the possibility to use social data. Basically, a 7-8%Gini increase may lead to 0.5-1% decrease in default frequency, which is the key performance indicator of the risk management system in the bank. That is why we have to understand when the project concerning social data implementation is going to bring payoff, and when it will become just a burden. At the moment, we see two situations when the credit scoring based on social data is applicable. The first one is the microfinance segment. The application flow is traditionally very risky in this segment, and classical scorecards tend to reject all the applications. However, many microfinance organizations compensate that risk with extraordinarily high rates. That is why the companies still have to obtain additional information to provide at least any discrimination between «bad» and, say, «very bad» applicants. Moreover, given that young people tend to be riskier and they do not have sufficient credit history, the social data starts being the only data source except for the loan application form. The second situation that can be favorable for social data implementation is when the bank has a large client base and huge amount loans in its portfolio. The reasoning is quite straightforward. Given that fixed costs of social data center launching and its support are high, the decrease in default frequency will be sufficient to compensate those costs only when the loan portfolio volume is large.

The real life examples of successful implementations, from our point of view, are still to come, as soon as many banks and microfinance organizations start pilot social dataprojects. As far as Russia is concerned, there are several consultant agencies such as Digital Society Laboratory, SOCSCOR, Double Data, SAS, SocioHub etc. that provide scoring service for banks and the microfinance segment. However, the details of such projects are still rarely disclosed, and the discussions that take place within professional banking conferences, as a rule, are held in a more general way.

4. Conclusion

In this paper we described the schema of social data retrieval in banking sphere. The social network we considered is Vkontakte, which is the largest in Russia. Then we examined the value that the data can bring to the credit scoring. We developed two scorecards (fraudulent case and ordinary default) based solely on the social data. The findings are the social data can indeed show acceptable discriminative power, especially in the case of fraud scoring.

As an area for further research, one can test whether there is an increase in Gini coefficient, when the social data predictors are added to the bureau of credit history scoring (e.g. Equifax) with the information on behavioral variables of the applicant. ■

References

1. Tucker P. (2013) *Has Big Data made anonymity impossible?* MIT Technology Review. Available at: http://www.technologyreview.com/news/514351/has-big-data-made-anonymity-impossible/ (accessed 5 November 2014).

2. Thomas L., Edelman D., Crook J. (2002) Credit scoring and its applications. *Monographs on Mathematical Modeling and Computation*, SIAM: Pliladelphia, pp. 107–117.
3. Wu J., Coggeshall St. (2012) *Foundations of predictive analytics*, CRC Press.
4. Hochberg Y. (1988) A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, vol. 75, no. 4, pp. 800–802.
5. Skiba S.A., Loiko V.I. (2013) Social'nyj skoring [Socialscoring]. *Scientific Journal of KubGAU*, no.7 (91), pp. 1258–1275. Available at: http://ej.kubagro.ru/2013/07/pdf/89.pdf(accessed 05 November 2014). (in Russian)
6. D'Andrea, Alessia et al. (2009) An overview of methods for virtual social network analysis. *Computational Social Network Analysis: Trends, Tools and Research Advances*, Springer, pp. 3–25.
7. SAS Institute Inc. (2012). *Developing credit scorecards using credit scoring for SAS® Enterprise Miner™* 12.1. Cary, NC: SAS Institute Inc.
8. Porshnev A., Redkin I. (2014) Analysis of Twitter users' mood for prediction of gold and silver prices in the stock market. *Communications in Computer and Information Science*, no. 436, pp. 190–197.

КРЕДИТНЫЙ СКОРИНГ НА ОСНОВЕ ДАННЫХ СОЦИАЛЬНЫХ СЕТЕЙ

А.А. МАСЮТИН
аспирант, департамент анализа данных и искусственного интеллекта, факультет компьютерных наук, Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20
E-mail: alexey.masyutin@gmail.com

Социальные сети аккумулируют значительное количество информации, которая позволяет получать дополнительные сведения о поведении людей. В данной работе мы используем данные наиболее посещаемой социальной сети «Вконтакте», чтобы выделять сегменты неплатежеспособных клиентов банка. Во-первых, мы представляем архитектуру центра хранения и обработки данных из социальных сетей. Он включает в себя инструменты для соотнесения реального клиента и его виртуального профиля в социальной сети, парсинг профилей социальной сети, получение данных об активности пользователя через API, и наконец, само хранилище данных. Во-вторых, на исторических данных мы разрабатываем две скоринговые карты, основанные исключительно на данных активности клиента в социальных сетях. Первая карта прогнозирует событие обычного дефолта – выхода на просрочку по ссуде более 90 дней за первые 12 месяцев с момента получения кредита. Вторая скоринговая карта прогнозирует событие мошеннического дефолта. Обе карты используют WOE-трансформацию входящих данных и затемприменяют логистическую регрессию по преобразованным данным. В результате данные социальных сетей лучше прогнозируют случаи мошеннических дефолтов, в отличие от обычных случаев просрочки. Качество скоринговых карт находится на приемлемом уровне, что подтверждается ROC-анализом и коэффициентами Джини. Поскольку классические скоринговые системы во многом опираются на кредитную историю клиента, которая зачастую отсутствует у молодых заемщиков, мы считаем, что данные социальных сетей могут служить их заменой. Таким образом, данные социальных сетей могут быть использованы для обогащения классических скоринговых систем банков и микрофинансовых организаций.

Ключевые слова: кредитный скоринг, социальные сети, вероятность дефолта, социальные данные, Вконтакте.

Цитирование: Masyutin A.A. Credit scoring based on social network data // Business Informatics. 2015. No. 3 (33). P. 15–23.

Литература

1. Tucker P. Has Big Data made anonymity impossible? MITTechnologyReview, 2003. [Электронный ресурс]: http://www.technologyreview.com/news/514351/has-big-data-made-anonymity-impossible/ (дата обращения 05.11.2014).
2. Thomas L., Edelman D., Crook J. Credit scoring and its applications / Monographs on Mathematical Modeling and Computation. SIAM: Pliladelphia, 2002. P. 107–117.
3. Wu J., Coggeshall St. Foundations of predictive analytics, CRC Press, 2012.
4. Hochberg Y.A sharper bonferroni procedure for multiple tests of significance // Biometrika. 1988. Vol.75, No. 4. P. 800–802.
5. Скиба С.А., Лойко В.И. Социальный скоринг// Научный журнал КубГАУ. 2013. № 7 (91). С. 1258–1275) [Электронный ресурс]: http://ej.kubagro.ru/2013/07/pdf/89.pdf (дата обращения 05.11.2014).
6. D'Andrea, Alessia et al.An overview of methods for virtual social network analysis // Computational Social Network Analysis: Trends, Tools and Research Advances. Springer, 2009. P. 3–25.
7. Developing credit scorecards using credit scoring for SAS® Enterprise Miner™ 12.1. Cary, NC: SAS Institute Inc., 2012.
8. Porshnev A., Redkin I. Analysis of Twitter users' mood for prediction of gold and silver prices in the stock market // Communications in Computer and Information Science. 2014. No. 436. P. 190–197.

СНИЖЕНИЕ РАЗМЕРНОСТИ МНОГОМЕРНЫХ ПОКАЗАТЕЛЕЙ С НЕЛИНЕЙНО ЗАВИСИМЫМИ КОМПОНЕНТАМИ

Е.Р. ГОРЯИНОВА
кандидат физико-математических наук,
доцент департамента математики, факультет экономических наук,
Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20
E-mail: el-goryainova@mail.ru

Ю.А. ШАЛИМОВА
студентка магистратуры, факультет экономических наук,
Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20
E-mail: july.shalimova@yandex.ru

При решении задачи сжатия многомерного вектора показателей используют методы факторного анализа, одним из которых является метод максимального правдоподобия (ММП). В системе коррелированных количественных показателей он позволяет выявить некоррелированные общие факторы, которые без существенной потери информации могут представлять исходные показатели. Нахождение общих факторов проводится с помощью специального представления корреляционной матрицы наблюдаемых признаков. Однако коэффициент корреляции не определен для признаков, представленных в номинальной шкале, а для признаков, имеющих нелинейный характер зависимости, не может служить измерителем силы связи. Для таких ситуаций традиционные методы факторного анализа оказываются малоэффективными.

В статье предложены две модификации ММП, использующие в качестве мер связи признаков ранговые коэффициенты корреляции Спирмена и коэффициенты Крамера. Для сравнения качества сжатия традиционного и двух адаптированных ММП проведен численный эксперимент. С помощью метода Монте-Карло смоделированы 12-мерные векторы, состоящие из четырех независимых трехмерных подвекторов, координаты которых имеют зависимости линейного и нелинейного типа. Установлено, что из трех рассмотренных методов только адаптированный метод, использующий коэффициенты Крамера, способен верно объединить в общий фактор показатели, связанные немонотонным типом зависимости. С другой стороны, в тех случаях, когда зависимость между признаками носит монотонный характер, этот метод менее эффективен, чем два других. Для демонстрации работоспособности указанных методов на реальных данных представлено решение задачи снижения размерности динамики относительного прироста потребительских цен в 2008-2014 годах для группы продовольственных товаров.

Ключевые слова: факторный анализ, общие факторы, метод максимального правдоподобия, корреляционная матрица, матрица нагрузок, коэффициент ранговой корреляции Спирмена, коэффициент Крамера.

Цитирование: Горяинова Е.Р., Шалимова Ю.А. Снижение размерности многомерных показателей с нелинейно зависимыми компонентами // Бизнес-информатика. 2015. № 3 (33). С. 24–33.

Введение

При изучении сложных объектов исследователи пытаются описать их большим числом показателей. Как правило, это приводит к тому, что среди собранных данных имеются группы показателей, которые характеризуют одно и то же свойство объекта и поэтому являются зависимыми, а также малоинформативные показатели, которые не несут в себе существенной информации об объектах. Статистический анализ таких массивов становится затруднительным и может приводить к неверным результатам. Поэтому возникает необходимость описать наблюдаемые показатели меньшим числом интегративных показателей, сохранив при этом как можно больше важной информации об объектах.

Основная идея факторного анализа состоит в том, что структура связей между анализируемыми признаками может быть объяснена тем, что эти признаки зависят от меньшего числа других непосредственно неизмеряемых показателей, называемых общими факторами. Классическая модель факторного анализа, описанная в работе [1], предполагает, что каждая наблюдаемая переменная представляется в виде линейной комбинации некоррелированных общих факторов и одного частного фактора, оказывающего влияние только на данную переменную. Основная задача факторного анализа состоит в оценивании матрицы нагрузок, элементами которой являются корреляции между исходными показателями и общими факторами, оценивании дисперсий частных факторов и интерпретации общих факторов. Решение этой задачи позволяет в рамках факторной модели удовлетворительно воспроизводить корреляции между наблюдаемыми показателями.

Наиболее распространенными методами решения этой задачи являются метод главных факторов [2, 3], метод минимальных остатков [4] и метод максимального правдоподобия (ММП) [5]. Так, согласно методу главных факторов, требуется провести оценивание дисперсий частных факторов, а затем применить процедуры компонентного анализа [6] к редуцированной корреляционной матрице, из элементов главной диагонали которой вычтены найденные оценки дисперсий. Принцип оценивания матрицы нагрузок методом минимальных остатков основан на минимизации суммы квадратов разностей между выборочными корреляциями и корреляциями, воспроизводимыми факторной моделью с фиксированным числом факторов. В ММП предполагается, что общие и частные факторы имеют гауссовское распределение, а оценками нагрузок являются те значения, при которых достигается максимум функции правдоподобия элементов выборочной корреляционной матрицы при фиксированном числе общих факторов. Оценивание числа общих факторов в двух последних методах проводится с помощью последовательного применения хи-квадрат тестов. Заметим, что методы факторного анализа используют в качестве мер связи коэффициенты корреляции исходных показателей. Однако на практике нередко возникают задачи, в которых показатели являются зависимыми, но некоррелированными. Например, в работе [7] установлена квадратичная зависимость между вероятностью дефолта и размером активов банка. Кроме того, многие показатели в социологических и психологических исследованиях измеряются в номинальной шкале, и коэффициент корреляции для этих величин не определен. Таким образом, если компоненты вектора показателей имеют зависимости нелинейного характера или измерены в различных шкалах, то процедура снижения размерности такого вектора требует корректировки.

Объектом исследования данной работы являются методы (в частности, ММП) снижения размерности в модели факторного анализа, а предметом исследования – адаптация методов сжатия для векторов с нелинейной структурой зависимости компонент. Предлагаемая нами модификация заключается в том, что в качестве оценки неизвестной корреляционной матрицы будут использоваться матрицы коэффициентов ранговой корреляции Спирмена и матрицы коэффициентов Крамера. С помощью компьютерного моделирования будет показано, что адаптированный ММП является более эффективным для решения задачи снижения размерности многомерного вектора с нелинейно зависимыми компонентами.

Данная работа имеет следующую структуру. В разделе 1 представлена модель факторного анализа и традиционный ММП, используемый в факторном анализе. В разделе 2 описаны адаптированные ММП и процедура компьютерного моделирования случайных векторов с линейно и нелинейно зависимыми компонентами. В разделе 3 проведен сравнительный анализ качества сжатия смоделированных векторов. В разделе 4 с помощью рассмотренных методов решена задача снижения размерности показателей изменения относительного

прироста потребительских цен в 2008-2014 году для группы продовольственных товаров.

1. Задача факторного анализа

Пусть $X = (X_1,\ldots,X_r)^T$ – r-мерный вектор наблюдаемых показателей у каждого из n объектов. Обозначим вектор стандартизированных наблюдений через $\vec{x} = (x_1,\ldots,x_r)^T$, где

$$x_i = \frac{X_i - \overline{X}_{i\cdot}}{s_i},\ \overline{X}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} X_{ij},\ s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}_{i\cdot}\right)^2.$$

Согласно канонической модели факторного анализа вектор $\vec{x}$ представляется в виде

$$\vec{x} = L\vec{f} + \vec{\varepsilon}, \qquad (1)$$

где L – детерминированная матрица $r \times k$, $k < r$, $\vec{f} = (f_1,\ldots,f_k)^T$ – случайный вектор центрированно-нормированных некоррелированных общих факторов, $\vec{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_r)^T$ – случайный вектор центрированных частных факторов, таких что, коэффициенты корреляции $\rho(\varepsilon_i,\varepsilon_j) = 0$; $\rho(\varepsilon_i,f_m) = 0$; $i, j = 1, ..., r$; $m = 1, ..., k$.

Из формулы (1) следует, что ковариационная матрица C вектора $\vec{x}$ удовлетворяет соотношению

$$C = LL^T + V, \qquad (2)$$

где V – диагональная матрица размера $r \times r$ с диагональными элементами $D\varepsilon_i = v_i$, а элементы l_{ij}, $i = 1, ..., r$; $j = 1, ..., k$ матрицы L являются коэффициентами корреляции между признаками x_i и факторами f_j, то есть $l_{ij} = \rho(x_i, f_j)$. По этой причине L называют матрицей нагрузок.

Предположим дополнительно, что вектор общих факторов $\vec{f} \sim N(0, I)$, I – единичная матрица размера $k \times k$, а $\vec{\varepsilon} \sim N(0, V)$.

Основная задача факторного анализа состоит в оценивании матрицы нагрузок L и дисперсий v_i, $i = 1, ..., r$. Выше было сказано о том, что разработано несколько методов решения этой задачи. Поскольку в данной работе моделируются гауссовские показатели, для решения задачи будет использован оптимальный в этой ситуации ММП, дающий асимптотически эффективные оценки указанных параметров [3].

Традиционно в качестве оценок элементов матрицы C используются выборочные ковариации, построенные по результатам n наблюдений вектора $\vec{x} = (x_1,\ldots,x_r)^T$. Обозначим через A матрицу выборочных ковариаций с элементами

$$a_{ij} = \frac{1}{n}\sum\nolimits_{m=1}^{n} x_{im}x_{jm},\ i,j = 1,..., r.$$

Следуя ММП, для оценивания l_{ij} и v_i, $i = 1, ..., r$; $j = 1, ..., k$ нужно выписать совместную плотность элементов матрицы A, прологарифмировать ее и найти те значения l_{ij} и v_i, при которых достигается максимальное значение логарифмической функции правдоподобия. Как показано в работе [8], решение этой задачи сводится к нахождению собственных векторов матрицы $V^{-1}(A - V)$, найти которые можно с помощью итерационной процедуры. Соответствующий итерационный алгоритм был реализован нами в среде Matlab и подробно описан в работе [9]. Отметим, что поскольку $\vec{x}$ – стандартизированный, то ковариационная матрица C является корреляционной, а A – выборочной корреляционной матрицей. Вообще говоря, ММП позволяет выбирать в качестве матрицы C как ковариационную матрицу, так и корреляционную.

Заметим, что L и f в формуле (1) определяются с точностью до вращения, цель которого в получении качественной интерпретации факторов. Наиболее распространенными методами вращения являются варимакс и квартимакс [10].

Еще одной проблемой при решении задачи факторного анализа является выбор числа общих факторов k. Существует несколько способов решения этой задачи, как теоретически обоснованных, так и эмпирических. Если в факторном анализе применяется ММП, то определение числа общих факторов основывается на проверке статистической гипотезы о том, что число общих факторов равно заданной величине k. Тестовая статистика отношения правдоподобия при сделанных предположениях имеет распределение хи-квадрат.

2. Адаптация ММП для нелинейно зависимых показателей

Как показано в предыдущем разделе, модель факторного анализа предполагает, что значения признаков линейно зависят от общих факторов, а в качестве меры связи самих признаков используются коэффициенты корреляции. Если же признаки связаны нелинейной зависимостью или измеряются в номинальной шкале, то коэффициент корреляции теряет свою информативность как измеритель силы связи. Поэтому в качестве мер связи таких признаков надо использовать другие коэффициенты, например, коэффициент ранговой корреляции Спирмена [11] или коэффициент Крамера [12].

Коэффициентом ранговой корреляции Спирмена ρ_{YZ} случайных величин Y и Z, построенным по наблюдениям (Y_1, Z_1), ..., (Y_n, Z_n), называется статистика

$$\rho_{YZ} = \frac{\sum_{m=1}^{n}\left(R_m - \frac{n+1}{2}\right)\left(S_m - \frac{n+1}{2}\right)}{\sqrt{\sum_{m=1}^{n}\left(R_m - \frac{n+1}{2}\right)^2 \sum_{m=1}^{n}\left(S_m - \frac{n+1}{2}\right)^2}},$$

в которой R_m – ранг элемента Y_m в выборке Y_1, ..., Y_n, а S_m – ранг элемента Z_m в выборке Z_1, ..., Z_n.

Отметим, что ρ_{YZ} может служить оценкой степени монотонной зависимости между величинами Y и Z [13]. Обозначим через P матрицу с элементами ρ_{ij}, $1 \le i, j \le r$, где $\rho_{ij} = \rho_{x_i x_j}$ – ранговый коэффициент корреляции Спирмена показателей x_i и x_j.

Дадим определение коэффициента Крамера для наблюдений (Y_1, Z_1), ..., (Y_n, Z_n) двумерного вектора (Y, Z). Для этого разобьем область V_Y возможных значений величины Y на l непересекающихся интервалов $\Delta_{Y,i}$, $i = 1, ..., l$, так, что $\bigcup_{i=1}^{l}\Delta_{Y,i} = V_Y$, а область V_Z возможных значений величины Z на s непересекающихся интервалов $\Delta_{Z,j}$, $j = 1, ..., s$, так, что $\bigcup_{j=1}^{s}\Delta_{Z,j} = V_Z$. Пусть n_{ij} – число пар выборки (Y_1, Z_1), ..., (Y_n, Z_n), попавших в прямоугольник $\Delta_{Y,i} \times \Delta_{Z,j}$, $i = 1, ..., l, j = 1, ..., s$.

Обозначим $n_{i.} = \sum_{j=1}^{s} n_{ij}$, а $n_{.j} = \sum_{i=1}^{l} n_{ij}$.

Тогда коэффициент Крамера определяется как

$$k_{YZ} = \sqrt{\frac{\widehat{\chi}^2_{YZ}}{n \cdot min\{(l-1),(s-1)\}}}, \text{ где}$$

$$\widehat{\chi}^2_{YZ} = n\sum_{i=1}^{l}\sum_{j=1}^{s}\frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{n_{i.}n_{.j}} -$$

статистика критерия хи-квадрат. В работе [12] показано, что коэффициент Крамера, принимающий значения в интервале [0,1], может служить мерой, характеризующей силу связи между признаками Y и Z. Обозначим через K матрицу с элементами k_{ij}, $1 \le i, j \le r$, где $k_{ij} = k_{x_i x_j}$ — коэффициент Крамера показателей x_i и x_j.

Рассмотрим следующие две модификации ММП. Назовем «модификацией 1» адаптированный ММП, в котором матрица выборочных коэффициентов корреляции A заменяется матрицей коэффициентов Спирмена P, и, соответственно, «модификацией 2» – адаптированный ММП, в котором матрица A заменена матрицей коэффициентов Крамера K. Наше предположение состоит в том, что при наличии монотонных, но нелинейных зависимостей между компонентами вектора $\vec{x}$ задачу выделения общих факторов эффективнее решать, используя модификацию 1, а при наличии нелинейных немонотонных связей – модификацию 2. Это предположение проверяется на тестовых данных с помощью обширного численного эксперимента.

В рамках эксперимента 12-мерные векторы $\vec{x} = (x_1, \ldots, x_{12})^T$ были сгенерированы таким образом, чтобы компоненты вектора образовывали 4 независимые группы по 3 признака в каждой группе. При этом признаки первой группы сильно коррелированны между собой, признаки второй группы связаны «зашумленной» функциональной зависимостью линейного типа, признаки третьей группы связаны «зашумленной» функциональной зависимостью нелинейного монотонного типа, а признаки четвертой группы – «зашумленной» функциональной зависимостью немонотонного типа.

Принцип моделирования коррелированных величин базируется на использовании следующего свойства, доказанного в работе [9]. Если случайные величины Y и W независимы и имеют конечные дисперсии, а величина $Z = \alpha W + Y$, то коэффициент корреляции $\rho_{ZW} = \rho$ величин Z и W связан с константой α соотношением

$$\alpha = \sqrt{\frac{\rho^2}{1-\rho^2}} \cdot \sqrt{\frac{DY}{DW}} sign(\rho). \qquad (3)$$

Теперь с помощью встроенного в Matlab датчика генерируется стандартная нормальная случайная величина $x_1 \sim N(0; 1)$; затем, используя соотношение (3), генерируется $x_2 \sim N(0; 1)$, такая что $\rho_{x_1 x_2} = 0,7$; затем $x_3 \sim N(0; 1)$, такая что $\rho_{x_3 x_2} = 0,7$.

Принцип генерации второй, третьей и четвертой групп следующий. Пусть случайные величины α_1, α_2, α_3 имеют усеченное стандартное нормальное распределение, а величины ε_1, ..., $\varepsilon_9 \sim N(0; 1)$. Тогда значения признаков x_4, ..., x_{12} вычисляются по следующим формулам:

$$x_4 = \alpha_1 + \varepsilon_1, \quad x_5 = f(\alpha_1) + \varepsilon_2, \quad x_6 = f(f(\alpha_1)) + \varepsilon_3,$$
$$x_7 = \alpha_2 + \varepsilon_4, \quad x_8 = g(\alpha_2) + \varepsilon_5, \quad x_9 = g(g(\alpha_2)) + \varepsilon_6,$$
$$x_{10} = \alpha_3 + \varepsilon_7, \quad x_{11} = h(\alpha_3) + \varepsilon_8, \quad x_{12} = h(h(\alpha_3)) + \varepsilon_9,$$

где функция $f(\cdot)$ — линейная функция, $g(\cdot)$ — нелинейная монотонная функция, $h(\cdot)$ — нелинейная функция. Реализации значений пар признаков для каждой из четырех групп объема 10 000 представлены на *рис. 1.*

Помимо указанных модификаций ММП потребовалось применить другой способ определения числа общих факторов, так как критерий, основанный на статистике отношения правдоподобия,

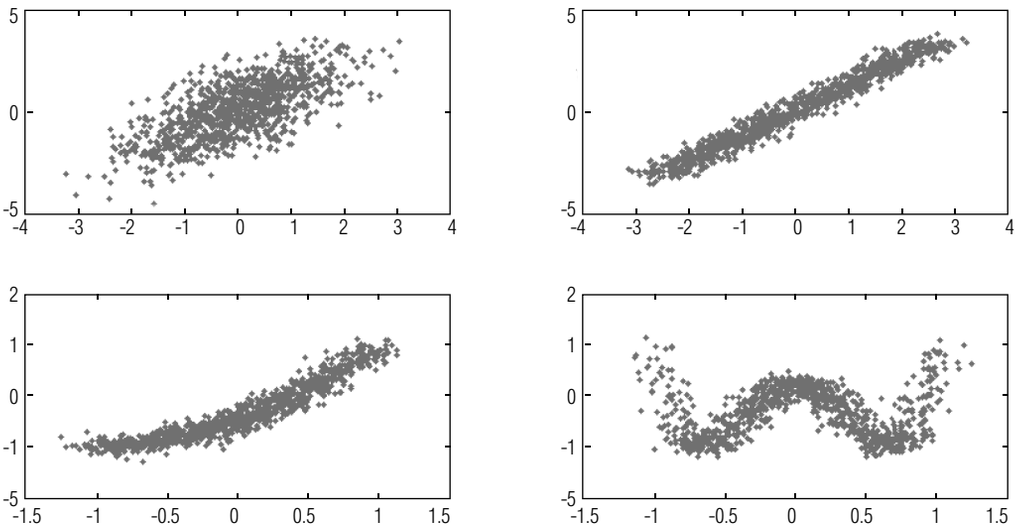


Рис. 1. Реализации признаков в группах демонстрационных данных

оказался неработоспособным на моделированных данных. Этот факт объясняется тем, что тестовая статистика имеет распределение хи-квадрат в случае гауссовских наблюдений, а компоненты x_7, ..., x_{12} сгенерированного вектора $\vec{x}$ являются нелинейными преобразованиями гауссовских случайных величин и, следовательно, не являются гауссовскими. Поэтому для определения числа общих факторов нами был реализован следующий эмпирический метод.

На первом шаге применяется ММП с числом общих факторов равным числу признаков. Затем для полученной матрицы нагрузок L вычисляются коэффициенты

$$\mu_j = \sqrt{\sum_{i=1}^{r} l_{ij}^2}, j = 1,\ldots,r. \qquad (4)$$

Каждый из коэффициентов μ_j показывает количество суммарного среднеквадратического отклонения признаков, которое объясняется добавлением j-го фактора к уже имеющимся $j-1$ факторам $f_1,\ldots,f_{j-1}$. В случае нормированных признаков положим число общих факторов равным k, если $\mu_k \geq 1$, а $\mu_{k+1} < 1$. На втором шаге запускается алгоритм ММП с выбранным числом факторов. Обоснование такого способа выбора приведено в работе [9].

3. Сравнительный анализ традиционного и адаптированных ММП

Перейдем к представлению результатов сжатия вектора $\vec{x} = (x_1,\ldots,x_{12})^T$, структура которого описана в разделе 2. Последовательно применим к демонстрационным данным все три метода с максимальным числом общих факторов равным 12. Значения μ_1, ..., μ_{12}, вычисленные по формуле (4), для традиционного ММП и двух его модификаций представлены на *рис. 2, 3* и *4* соответственно.

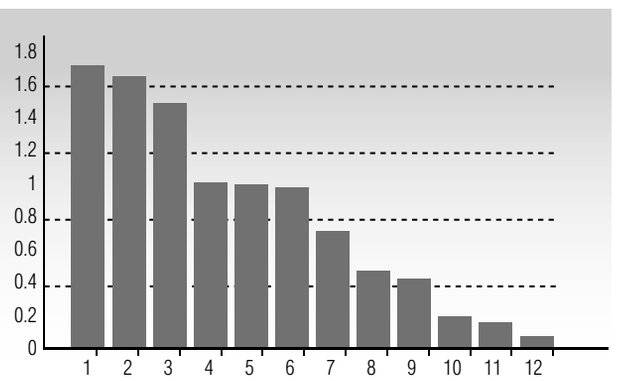


Рис. 2. Приращения объясняемого среднеквадратического отклонения для факторов с первого по двенадцатый для традиционного метода

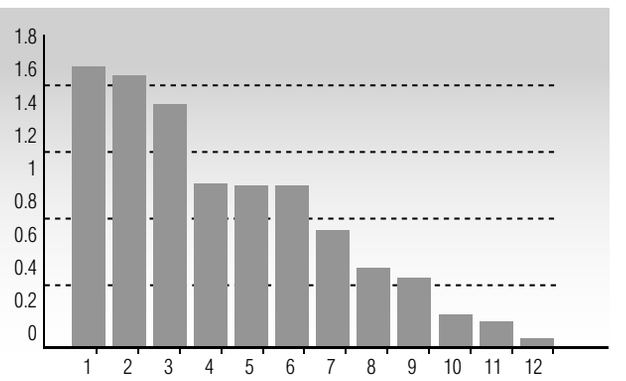


Рис. 3. Приращения объясняемого среднеквадратического отклонения для факторов с первого по двенадцатый для модификации 1

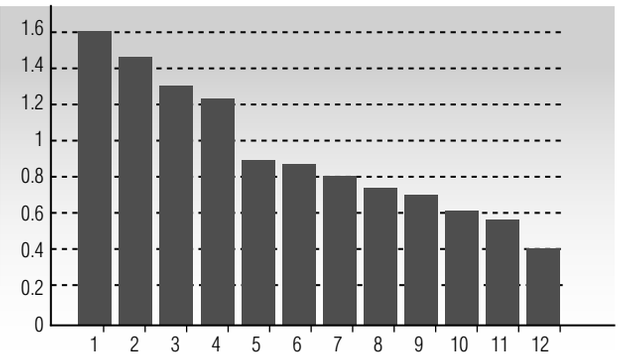


Рис. 4. Приращения объясняемого среднеквадратического отклонения для факторов с первого по двенадцатый для модификации 2

Согласно *рис. 2* и *3*, значение $\mu_j > 1$ получено при $j = 1, 2, 3$ и $\mu_j \approx 1$ при $j = 4, 5, 6$. Поэтому для традиционного ММП и модификации 1 считаем число общих факторов $k = 6$. Матрица нагрузок традиционного ММП имеет следующий вид:

-0.0022	-0.0042	**0.7144**	0.0057	-0.0010	0.0077
-0.0047	0.0077	**0.9845**	-0.0001	0.0009	-0.0001
0.0003	-0.0046	**0.7079**	-0 0045	0 0180	-0.0007
0.0033	**0.9842**	0.0009	0.0000	-0.0018	0.0012
0.0020	**0.9957**	-0.0001	0.0011	0.0004	-0.0004
0.0028	**0.9989**	-0.0006	-0.0003	0.0001	-0.0000
-0.9379	0.0017	-0.0040	0.0039	-0.0042	-0.0046
-0.9994	0.0018	-0.0001	-0.0001	-0.0000	0.0000
-0.9217	0.0068	-0.0047	-0.0029	0.0013	0.0052
0.0016	0.0099	-0.0023	-0.0160	-0.0232	**0.8265**
-0.0136	-0.0026	0.0041	**0.8483**	-0.1786	0.0078
0.0130	0.0097	0.0158	-0.1914	**-0.8366**	-0.0208

Видно, что этот метод правильно определяет общие факторы, соответствующие группе признаков с монотонным нелинейным типом зависимости (высокие нагрузки этих признаков на первый фактор выделены в столбце 1), группе с линейным типом зависимости (высокие нагрузки этих признаков на второй фактор выделены в столбце 2) и группе сильно коррелированных признаков (высокие нагрузки этих признаков на третий фактор выделены в столбце 3). Признаки x_{10}, x_{11}, x_{12} имеют высокие нагрузки на шестой, четвертый и пятый факторы соответственно. Таким образом, традиционный метод выделяет в отдельные факторы признаки связанные немонотонным типом зависимости.

Матрица нагрузок модификации 1 имеет следующий вид:

0.0045	-0.0070	**0.6911**	0.0080	-0.0020	0.0067
0.0099	-0.0081	**0.9782**	-0.0024	-0.0008	0.0004
0.0133	-0.0100	**0.6984**	0.0017	-0.0033	0.0118
0.9850	-0.0016	-0.0024	0.0002	-0.0013	0.0022
0.9964	-0.0015	0.0005	-0.0007	0.0002	-0.0004
0.9989	-0.0005	-0.0005	0.0002	0.0000	-0.0000
-0.0102	**-0.9792**	-0.0026	-0.0003	0.0008	-0.0014
-0.0094	**-0.9787**	-0.0047	0.0003	-0.0024	0.0005
-0.0094	**-0.8860**	-0.0084	0.0078	-0.0063	0.0061
0.0012	-0.0013	0.0164	0.0018	-0.0174	**-0.7782**
-0.0030	-0.0074	-0.0096	**-0.8918**	-0.0138	-0.0005
-0.0029	0.0211	-0.0081	0.0196	**-0.8486**	0.0112

Этот способ также верно выделяет три группы зависимых признаков – признаки с линейным типом зависимости (первый фактор), признаки с монотонным нелинейным типом зависимости (второй фактор) и сильно коррелированные признаки (третий фактор). Как и традиционный ММП, модификация 1 не выявляет четвертую группу признаков, связанных немонотонным типом зависимости.

На *рис. 4* видно, что больше единицы оказались значения μ только для четырех факторов. Матрица нагрузок для модификации 2 при $k = 4$ имеет следующий вид:

0.0343	-0.0386	0.1032	**-0.4380**
0.0338	-0.0503	0.1406	**-0.6847**
0.0300	-0.0499	0.1020	**-0.4317**
0.7777	0.0080	-0.0028	0.0003
0.9019	0.0168	-0.0108	0.0059
0.9278	0.0209	-0.0091	0.0044
0.0415	**-0.7596**	-0.0391	0.0163
0.0413	**-0.8061**	-0.0485	0.0172
0.0376	**-0.6744**	-0.0318	0.0138
0.0305	-0.0502	**0.4291**	0.0461
0.0388	-0.0682	**0.7326**	0.1241
0.0373	-0.0638	**0.5750**	0.0875

Из трех рассмотренных способов только этот способ верно выделяет все четыре группы зависимых признаков. Так, в первый фактор выделены признаки с линейным типом зависимости, во второй – признаки с нелинейным монотонным типом зависимости, в третий – признаки с немонотонным типом зависимости, а в четвертый – сильно коррелированные признаки. Однако следует заметить, что нагрузки для групп показателей с монотонными типами связи ниже, чем у двух предыдущих методов.

Отметим, что на других смоделированных данных аналогичной структуры представленный эмпирический метод определения числа факторов продемонстрировал адекватные результаты. Применение к матрицам нагрузок методов вращения не внесло существенных изменений.

При попытке задать в традиционном ММП и в модификации 1 число общих факторов $k = 4$ были получены матрицы нагрузок, у которых в четвертый фактор выделялась лишь одна из компонент четвертого подвектора.

4. Пример с реальными данными

Продемонстрируем работу трех рассмотренных методов на реальных данных. Для демонстрации эффективной работы методов сжатия многомерных признаков хотелось выбрать такие показатели, чтобы наличие зависимости между ними было в значительной степени предсказуемо из соображений здравого смысла. Мы выбрали еженедельные средние потребительские цены на некоторые продукты питания за период с января 2008 г. по апрель 2014 г. В данном случае признаками являются цены на конкретные товары, а наблюдениями – цены на товары в фиксированные моменты времени. Согласно модели факторного анализа, наблюдения за каждым признаком должны быть независимы и одинаково распределены. Но, поскольку цены на товары растут с течением времени, то в качестве реализации X_{ij} i-го признака для j-го наблюдения будем рассматривать не саму цену i-го товара в момент времени j (обозначим ее c_{ij}), а величину относительного прироста цены, т.е.

$$X_{ij} = \frac{c_{ij} - c_{i(j-1)}}{c_{i(j-1)}}.$$

В качестве признаков были выбраны относительные приросты цен на следующие товары: говядина, сосиски и сардельки, колбаса полукопченая и варено-копченая, колбаса вареная I сорта, говядина и свинина тушеная консервированная, масло сливочное, сметана, творог жирный, сыры сычужные твердые и мягкие, мука пшеничная, хлеб и булочные изделия из пшеничной муки. Еженедельные средние потребительские цены на эти продукты за указанный период взяты с сайта Федеральной службы государственной статистики (*www.gks.ru*). Понятно, что первые пять продуктов образуют «мясную» группу, следующие четыре продукта – «молочную» группу, а последние два продукта – «мучную» группу.

Применим последовательно все три способа сжатия к имеющимся данным. Для определения числа общих факторов вычислим для каждого метода коэффициенты μ_1, ..., μ_{11} по формуле (4). Для обеих модификаций значения больше единицы имели первые три коэффициента, поэтому число общих факторов $k = 3$. У традиционного ММП близким к единице оказался и μ_4, что вызывает некоторые сомнения относительно включения четвертого фактора. Мы приняли решение о включении трех факторов. Отметим, что в отличие от моделированных данных, для реальных данных потребовалось применить методы вращения нагрузочной матрицы. Это позволило существенно улучшить интерпретируемость результатов каждого из трех методов. Поэтому опустим представление матриц нагрузок, полученных до процедуры вращения.

Матрица нагрузок традиционного ММП после вращения имеет следующий вид:

0.7837	0.0763	-0.1608
0.7953	0.0406	-0.2426
0.8927	0.0537	-0.0971
0.4513	-0.0657	0.1389
0.6518	-0.0541	0.0374
-0.0711	0.0205	-0.6946
0.0975	-0.4298	-0.4831
0.2239	-0.0078	-0.8595
0.0544	0.2281	-0.7201
-0.0472	-0.9172	-0.0053
0.0166	-0.7206	0.1819

Как и ожидалось, признаки отчетливо объединились в три группы. Первый фактор объединяет продукты «мясной» группы, второй – «мучной» группы, а третий – «молочной». Однако из общей картины несколько выбиваются строки, соответ-

ствующие приросту цен на вареную колбасу (строка 4) и сметану (строка 7). Видно, что прирост цен на колбасу имеет существенно меньшую нагрузку на «мясной» фактор, чем остальные признаки из этой группы, а прирост цен на сметану имеет немалую нагрузку 0,429 и в «мучной» группе.

Матрица нагрузок модификации 1 после вращения имеет следующий вид:

0.0273	-0.6796	-0.0774
0.1401	-0.7700	-0.0513
0.0854	-0.7521	-0.0426
0.0247	-0.7064	-0.1153
0.0110	-0.7383	-0.1250
0.7496	-0.0413	-0.0142
0.7817	-0.1850	-0.0820
0.8146	-0.1559	-0.1012
0.7728	0.1204	0.1601
0.0432	-0.0625	-0.7701
-0.0180	-0.1533	-0.6487

Этот способ также позволяет явно выделить три фактора, соответствующих «молочной» (первый фактор), «мясной» (второй фактор) и «мучной» (третий фактор) группам. Но, в отличие от результатов традиционного ММП, четвертая и седьмая строки, соответствующие вареной колбасе и сметане, мало отличаются от других строк своих групп. То есть, разбиение строк на группы «похожести» оказывается более четким, чем в традиционном методе.

Матрица нагрузок модификации 2 после вращения имеет следующий вид:

0.4365	0.1265	–0.1382
0.5407	0.1425	–0.0716
0.5426	0.1227	–0.1277
0.5381	0.1130	–0.1235
0.5009	0.1265	–0.1350
0.1331	0.5229	–0.1304
0.1839	0.5642	–0.0820
0.1573	0.5590	–0.0389
0.0704	0.4679	–0.2211
0.1558	0.1338	–0.4246
0.1625	0.1216	–0.5062

Этот способ также правильно выделяет три фактора, причем картина разбиения признаков на похожие группы достаточно отчетливая. Однако все признаки имеют на «свои» факторы меньшие нагрузки, чем в двух предыдущих матрицах.

Заключение

В данной работе рассмотрена задача снижения размерности многомерного вектора показателей. При решении этой задачи применен традиционный ММП и две модификации этого метода, использующие в качестве мер связи признаков ранговые коэффициенты корреляции Спирмена (модификация 1) и коэффициенты Крамера (модификация 2). Для сравнения качества сжатия этими методами проведен численный эксперимент, в ходе которого сгенерированы 12-мерные случайные векторы, состоящие из четырех независимых подвекторов. При этом компоненты первого подвектора являлись сильно коррелированными, компоненты второго – связанными «зашумленной» функциональной зависимостью линейного типа, компоненты третьего – связанными «зашумленной» функциональной зависимостью монотонного нелинейного типа, а компоненты четвертого – немонотонной «зашумленной» функциональной зависимостью. Оказалось, что традиционный ММП достаточно хорошо выделяет в общие факторы коррелированные признаки и признаки, связанные зависимостями линейного и монотонного типа. Однако этот метод не способен выделить в единую группу признаки, связанные немонотонной зависимостью. Модификация 1 показала аналогичные результаты, и только модификация 2 правильно выделила все четыре группы связанных признаков. Это объясняется тем, что коэффициенты Крамера, использованные в модификации 2, основаны на статистике критерия хи-квадрат, который является состоятельным против любого вида альтернатив о зависимости случайных величин. Критерии же, основанные на выборочном коэффициенте корреляции, используемом в качестве меры связи признаков в традиционном методе, или на ранговом коэффициенте Спирмена, используемом в модификации 1, являются состоятельными лишь против альтернатив о линейной или монотонной зависимости признаков соответственно. Однако универсальность коэффициента Крамера имеет и негативную сторону: его применение при выявлении линейных и монотонных зависимостей менее эффективно, чем использование коэффициента корреляции.

Рассмотренные методы показали адекватные результаты в практической задаче снижения размерности вектора относительного прироста цен на продовольственные товары. Поскольку все три способа сжатия выделили одинаковые факторы, следует признать, что истинные зависимости между показателями имеют монотонный характер. Наиболее четкую структуру матрицы нагрузок показала модификация 1. Этот факт, по-видимому, говорит о том, что существенный вклад в вариацию признаков вносят частные факторы, а коэффициенты Спирмена, как более робастные оценки истинных коэффициентов корреляции, лучше улавливают наличие линейной зависимости зашумленных данных, чем выборочные коэффициенты корреляции. ■

Литература

1. Anderson T.W., Rubin H. Statistical inference in factor analysis // Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Vol. 5. Berkeley: University of California Press. 1956. P. 111–150.
2. Harman H. Modern factor analysis. Chicago: University of Chicago Press, 1960. 469 p.
3. Прикладная статистика: Классификации и снижение размерности / С.А. Айвазян и [др.]; под ред. С.А. Айвазяна. М.: Финансы и статистика, 1989. 607 с.
4. Harman H., Jones W. Factor analysis by minimizing residuals (minres) // Psychometrika. 1966. Vol. 31, No. 3. P. 351–369.
5. Lawley D., Maxwell A.F. Factor analysis as astatistical method. London: Butterworths, 1963. 145 с.
6. Лагутин М.Б. Наглядная математическая статистика. М.: Бином. Лаборатория знаний, 2007. 472 с.
7. Карминский А.Н., Костров А.В. Моделирование вероятности дефолта российских банков: расширенные возможности // Журнал Новой Экономической Ассоциации. 2013. № 1. С. 64–86
8. Ивченко Г.И., Медведев Ю.И. Введение в математическую статистику. М.: ЛКИ, 2010. 600 с.
9. Горяинова Е.Р., Шалимова Ю.А. Снижение размерности показателей смешанной структуры / Препринт WP7/2014/08, серия WP7 «Математические методы анализа решений в экономике, бизнесе и политике». М.: ИД ВШЭ, 2014. – 40 с.
10. Факторный, дискриминантный и кластерный анализ / Дж.-О. Ким и [др.]. М.: Финансы и статистика, 1989. 216 с.
11. Kendall M.G. Rank correlation methods. London: Griffin, 1970. 272 p.
12. Cramer G. (1961) Matematical methods of statistics. NY: Princeton, 1961. 575 p.
13. Горяинова Е.Р., Панков А.Р., Платонов Е.Н. Прикладные методы анализа статистических данных. М.: ИД ВШЭ, 2012. 310 с.

REDUCING THE DIMENSIONALITY OF MULTIVARIATE INDICATORS CONTAINING NON-LINEARLY DEPENDENT COMPONENTS

Elena R. GORYAINOVA
Associate Professor, Department of Mathematics, Faculty of Economic Sciences, National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow, 101000, Russian Federation
E-mail: el-goryainova@mail.ru

Julia A. SHALIMOVA
Graduate Student, Faculty of Economic Sciences, National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow, 101000, Russian Federation
E-mail: july.shalimova@yandex.ru

To solve the problem of reduction of the multidimensional vector of indicators methods of factor analysis are used. One of them is the maximum likelihood method (MLM). It allows to identify uncorrelated common factors among the set of correlated quantitative indicators. The uncorrelated common factors can represent initial indicators without significant loss of information. Common factors are detected using a special representation of the correlation matrix of the observed indicators. However, the correlation coefficient is not defined for the characteristics measured in a nominal scale. In addition, it cannot serve as a measure for the strength of the coupling indicators with nonlinear dependence. Traditional methods of factor analysis are ineffective for such situations. Two MLM modifications are proposed in the paper. They use the rank Spearman correlation coefficients and Cramer coefficients as measures of relationship between variables. 12-dimensional vectors with their coordinates dependent on each other with linear and nonlinear dependency were simulated, using the Monte Carlo method. Then a comparative analysis of the effectiveness of the traditional MLM and the two proposed modifications of the MLM was carried out for these data. It is shown that only adapted method that uses the Cramer coefficients is able to combine correctly the indicators related with nonmonotonic dependency in the common factor. On the other hand, this method has a lower efficiency than the other two methods in the cases where the dependency between variables is linear or monotonic. To demonstrate the efficiency of these methods on real data, the task of reducing the dimension of the dynamics of the relative consumer price growth in the years 2008-2014 for a group of food products has been solved.

Key words: factor analysis, common factors, the maximum likelihood method, correlation matrix, matrix of loadings, Spearman rank correlation coefficient, Cramer coefficient.

Citation: Goryainova E.R., Shalimova Ju.A. (2015) Snizhenie razmernosti mnogomernyh pokazatelei s nelineino zavisimymi komponentami [Reducing the dimensionality of multivariate indicators containing non-linearly dependent components]. *Business Informatics*, no. 3 (33), pp. 24–33 (in Russian).

References

1. Anderson T. W., Rubin H. (1956) Statistical inference in factor analysis. Proceedings of the *Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 5. Berkeley: University of California Press, pp. 111–150.
2. Harman H. (1960) *Modern factor analysis*. Chicago: University of Chicago Press.
3. Ajvazyan S.A., Buhshtaber V.M., Enukov I.S., Meshalkin L.D. (1989) *Prikladnaya statistika: Klassifikaciya i snizenie razmernosti* [Applied statistics: Classification and reducing the dimension]. Moscow: Finansy i statistika (in Russian).
4. Harman H., Jones W. (1966) Factor analysis by minimizing residuals (minres), *Psychometrika*, vol. 31, no. 3, pp. 351–369.
5. Lawley D., Maxwell A.F. (1963) *Factor analysis as astatistical method*. London: Butterworths.
6. Lagutin M.B. (2007) *Naglyadnaya matematicheskaya statistika* [Visual mathemathical statistics]. Moscow: Binom. Laboratoria znanij (in Russian).
7. Karminsky A., Kostrov A. (2013) Modelirovanie verojatnosti defolta rossijskih bankov: rasshirennye vozmozhnosti [Modeling the default probabilities of Russian banks: Extended abillities], *Journal of the New Economic Association*, no. 1 (17), pp. 64–86 (in Russian).
8. Ivchenko G.I., Medvedev Yu.I. (2010) *Vvedenie v matematicheskuju statistiku* [Introduction to mathematical statistics]. Moscow: LKI (in Russian).
9. Goryainova E., Shalimova Ju. (2014) *Snizhenie razmernosti pokazatelej smeshannoj struktury* [Reduction of dimensionality for the indicators that have a mixed structure]. Working paper WP7/2014/8. Moscow: HSE (in Russian).
10. Kim J.-O., Mueller C.U., Klecka C. (1989) *Faktornij, diskriminantnij i klasternij analiz* [Factor, discriminant and cluster analysis]. Moscow: Finansy i statistika (in Russian).
11. Kendall M.G. (1970) *Rank correlation methods*, London: Griffin.
12. Cramer G. (1961) *Matematical methods of statistics*. NY: Princeton.
13. Goryainova E.R., Pankov A.R., Platonov E.N. (2012) *Prikladnye metody analiza statisticheskih dannyh* [Applied methods of statistical data analysis]. Moscow: HSE (in Russian).

КЛАСТЕРНЫЙ АНАЛИЗ ВИЗУАЛЬНОГО ВОСПРИЯТИЯ СТРУКТУРЫ ДАННЫХ

В.В. ЛАПТЕВ
кандидат искусствоведения, доцент кафедры инженерной графики и дизайна Института металлургии, машиностроения и транспорта, Санкт-Петербургский политехнический университет Петра Великого
Адрес: 195251, г. Санкт-Петербург, ул. Политехническая, д. 29
E-mail: laptevsee@yandex.ru

П.А. ОРЛОВ
старший преподаватель кафедры инженерной графики и дизайна Института металлургии, машиностроения и транспорта, Санкт-Петербургский политехнический университет Петра Великого; Исследователь Университета Восточной Финляндии
Адрес: 195251, г. Санкт-Петербург, ул. Политехническая, д. 29
E-mail: paul.a.orlov@gmail.com

Структуры данных являются распространенными показателями в среде управления бизнес-проектами. Инфографика как особое направление коммуникационного дизайна предусматривает ряд графических способов, позволяющих визуализировать информацию такого рода. Применение каждого из имеющихся типов диаграмм сопряжено с определенными ограничениями, связанными с особенностями визуального восприятия и семиотической спецификой. Из-за недостаточной степени изученности был выбран тип структурной диаграммы – потоковая диаграмма Сэнкей, которая часто используется в бизнес-процессах для представления структуры данных. Для выявления методов оценки формы графического образа визуализации структуры данных был проведен эксперимент, в котором в качестве стимула выступала 4-потоковая диаграмма. Результаты глазодвигательной активности человека фиксировались с помощью системы видеоокулографии или ай-трекера.

В качестве метода анализа были приняты иерархические дивизимные алгоритмы, работающие с универсальным кластером, состоящим из всех зрительных фиксаций, с последующим пошаговым разбиением его на меньшие части. Было обнаружено, как минимум, четыре кластера, основанных на координатах. В найденной модели присутствовал «входной» кластер и «выходная группа кластеров» и явно определился центральный кластер зрительных фиксаций. При дальнейшем увеличении числа кластеров картина менялась в сторону большей детализации. Очевидно, что прослеживается определенный нарратив при рассматривании диаграммы, выявляющий последовательность «движения» потока, от целого к его структурным частям. В итоге кластерная алгоритмизация их анализа позволяет перевести визуальную интерпретацию структур числовых данных в круг задач поддержки принятия решений, решаемых с помощью программных средств.

Ключевые слова: кластерный анализ, инфографика, визуализация данных, структура данных, диаграмма, видеоокулография, ай-трекер.

Цитирование: Лаптев В.В., Орлов П.А. Кластерный анализ визуального восприятия структуры данных // Бизнес-информатика. 2015. № 3 (33). С. 34–43.

Введение

Управление информацией, которая подразумевает построение структур различного рода и сценариев их восприятия, является важной прикладной задачей. Структуры данных являются распространенными показателями в среде управления бизнес-проектами. Вопросы, связанные с их визуальной интерпретацией, являются актуальными в связи с вариативностью формы представления. До сих пор существуют проблемы визуализации числовых данных, возникающие при выборе соответствующих типов диаграмм. Так, один и тот же пример структуры данных может быть представлен в различных паттернах: брусковых, секторных, плоскостных или потоковых диаграммах. Правильный выбор формы основывается не только на учете контекста и семантических связей между числовым массивом и его графическим образом, но и на удобочитаемости графика и простоте его восприятия. Это касается визуализации не только структуры числовых данных, но и классификационных или иерархических схем с количественным анализом.

1. Изучение вопросов формообразования при визуализации структуры данных

Графическое представление таких результатов, а значит, и определение формы сообщения, относится к вопросам коммуникативного дизайна. Инфографика, как его составная часть, рассматривается как возможный объект прикладной информатики [1, 2]. Необходимо рассмотреть особенности формирования структурных диаграмм с точки зрения удобства визуального восприятия и точности анализа представляемых данных. В работе предлагаются методы определения эффективности выбора формы визуального образа, основанные на кластерном анализе глазодвигательной активности человека с помощью системы видеоокулографии или ай-трекера (*eye-tracker*). Это позволяет не только включить в исходные параметры точность и скорость определения параметров, но и зарегистрировать направление взгляда наблюдателя, длительность фиксации и протяженность саккад.

Первоначально исследования, касающиеся определения эффективности вида структурных диаграмм, рассматривали в качестве оппозиции дилемму «что лучше: столбик или сектор?», оставляя за скобками другие формы представления структуры. Данные виды формы графического образа являлись основными для представления структуры числовых данных. В многочисленных работах исследовались вопросы применения брусковых (столбиковых и полосовых) и секторных диаграмм [3–6]. В основе проводимых экспериментов лежала оценка скорости и точности определения структуры (в процентах от целого). Собранные данные анализировались с помощью статистических методов оценки количественных отклонений показаний испытуемых с предъявляемыми стимулами. На основании этого было подтверждено предположение, что точность и скорость оценки структуры зависит, в первую очередь, от пропорционального состава долей, а уже во вторую – от вида диаграммы. Например, при визуализации соотношения 50% и 50% предпочтение отдавалось секторной диаграмме, для иных размеров долей – брусковой и т. п.

Продолжение дискуссии о соответствии формы диаграммы структуре числовых данных велась попрежнему на основе экспериментальных данных о точности и скорости визуального восприятия. Однако во главу угла был поставлен статистический метод качественной оценки распознавания, когда респондент отмечал соотношение долей и их совокупностей: «меньше, больше, равно» [7]. Эти методы использовались при расширении поля исследований визуальной структуры данных на другие типы формы графического образа: линейные и брусковые [8], брусковые и плоскостные [9], секторные, брусковые, кольцевые, плоскостные диаграммы [10]. Кроме того, в последнее время из-за популярности оперативных диаграмм управления бизнес-процессами (*dashboards*) значительное внимание стало уделяться исследованиям принципов визуализации сложных иерархических структур [11, 12]. При этом следует отметить, что изучению восприятия потоковых диаграмм (или диаграмм Сэнкей), которые зачастую входят в состав графических комплексов контроля, уделяется недостаточное внимание.

2. Условия применения потоковых диаграмм

Возникновение потоковых диаграмм относится к середине XIX в., когда возникла потребность в количественной информации о трафике для постройки новых дорог, мостов, каналов и предприятий. Требовались разнообразные данные о физических, технических, политических и геостратегических условиях в различных точках государства. Кроме того, возникла необходимость учитывать распре-

деление и мобильность людей и капитала. Такая корреляция экономических ресурсов и демографии была бы наиболее наглядна на картах, где с помощью графического языка выражалась бы количественная информация.

Примером такого подхода служит британский атлас (*Atlas to accompany the second report of the railway commissioners,* 1838). Его автором был железнодорожный инженер Генри Харнесс (*Henry Drury Harness*), член комиссии по изучению железных дорог. В 1837 г. он проиллюстрировал доклад комиссии серией плоскостных и потоковых картодиаграмм, представляющих распределение населения в городах Великобритании и соответствующее перемещение товаров и пассажиров железных дорог. Способ линейного изображения количественных показателей мог наглядно показать перемещение на карте того или иного экономического объекта: пассажиров, грузов, капитала, электроэнергии и т. п. Потоковые диаграммы могли соединять точки на карте прямыми, но чаще в качестве оси абсцисс использовались определенные топографические линии: реки, морские пути, железные или шоссейные дороги, трубопроводы, высоковольтные линии.

Наиболее известными статистиками прошлого, широко применявшими этот метод на практике, были французский инженер Шарль Минар (*Charles Joseph Minard*) и бельгийский железнодорожный инженер Альфред Бельпер (*Alfred Jules Belpaire*). В 1845 г. Минар показал возможность потоковых диаграмм на примере пассажирского трафика между городами Дижон и Мюлуз. Посредством толщины линии он выразил количественные показатели, которые были перенесены в координатную систему, где ось абсцисс выполняли железные дороги. Каждый миллиметр толщины означал тысячи перевезенных пассажиров. В русских экономических картах потоковые диаграммы с «масштабными полосками» начал применять И.Ф. Борковский в 1870-х гг. в отчетах экспедиции, снаряженной Вольным экономическим и Русским географическим обществами для исследования хлебной торговли и производства в России.

Потоковые диаграммы использовались для визуализации связей с количественными характеристиками не только на картах, но и на блок-схемах процесса, иерархических графах и т.п. В них ширина линий пропорциональна количеству потока, визуализирующего, например, баланс, переводы между процессами или структуру затрат. Такой специфический тип потоковой диаграммы получил название «диаграмма Сэнкей» (*sankey diagram*), которое происходит от имени Мэтью Сэнкея (*Matthew Henry Phineas Riall Sankey*), ирландского инженера XIX в. Он в 1898 г. использовал этот способ графического представления информации для сравнения эффективности использования энергии парового двигателя.

При визуализации структуры целого потоковые диаграммы расставляют визуальные акценты на динамике передачи данных, т.е. на потоках внутри системы. Они выявляют доминирующие части, полезны в поиске «слабого звена», показывают балансы показателей в системе. К подобному способу визуализации структуры можно отнести и графики параллельных координат, которые очень близки по графическому образу к потоковым диаграммам и служат для количественной характеристики связей. Здесь, как и в потоковых диаграммах, происходит не только разделение целого на доли, но и их визуальное обособление. Разделение долей столбика или полосы происходит и в так называемой структурной диаграмме водопада (*waterfall chart*). В общем, потоковую диаграмму можно интерпретировать как динамическую модификацию брусковой диаграммы. Следует отметить, что уже проводилось исследование вопроса восприятия структурных диаграмм с объединенными и разделенными долями [13]. В результате было получены данные о том, что элементы визуальной структуры могут в определенной степени предсказуемо влиять на семантическую интерпретацию данных, которая выходит за пределы простого считывания данных. Это необходимо учитывать при разработке и оценке методов визуализации. Однако методы, в основе которых лежит анализ точности и скорости восприятия данных или качественная (но приблизительная) оценка структуры, не рассматривают набор данных в целом на основе элементов визуального дизайна. Для выяснения того, как пользователь оценивает композицию визуализации, должна быть принята во внимание глазодвигательная активность человека: направление взгляда, длительность фиксаций и протяженность саккад.

3. Постановка эксперимента

Для определения условий выбора формы структурной диаграммы была выдвинута следующая гипотеза: «Паттерн зрительных фиксаций рассма-

тривания имеет связь с структурой потоковой диаграммы в условиях соответствующей задачи». Для уточнения гипотезы введем вопрос исследования: «Является ли семантическая основа потоковой диаграммы детерминирующим фактором для положения фиксаций взора?» Другими словами, проверяется, будут ли фиксации взора испытуемого группироваться в левой части диаграммы («входная часть» – целое) и в правой части («выходная часть» – доли); будет ли переходная фаза потоковой диаграммы («центральная часть») оставлена без внимания.

Человеческий глаз постоянно (за исключением некоторых фаз сна) находится в движении. Принято различать в глазодвигательной активности определенные фазы: дрейфы, фиксации, варианты саккад, нистагмы [14–16]. Интерес представляют зрительные фиксации – дрейф, медленное, плавное перемещение глаза в небольшой зоне и саккады – скачкообразные движения высокой скорости, при которых резко изменяется позиция глаза. Считается, что зрительная информация обрабатывается в момент фиксации [17]. Проверка гипотезы потребовала регистрации зрительных фиксаций, для чего в настоящей работе была использована технология видеоокулографии.

Для получения первоначальных данных и апробации математических алгоритмов было принято решение остановиться на типе «*case study*» [18–20] и пригласить одного испытуемого, не ознакомленного с целями исследования. Испытуемому в случайном порядке был предъявлен стимульный материал в виде структурных диаграмм, состоящих из четырех потоков. Данная инфографика демонстрировала процентное разделение инвестиций по отдельным проектам. На экране монитора с разрешением вывода 1280 на 1024 пикселей в центре испытуемый видел потоковую диаграмму размером 500 на 500 пикселей, с направлением потоков слева направо (*рис. 1*). В нижней части экрана расположены интерактивные элементы для выбора ответа. На втором и третьем варианте выбора показаны фиксации взора испытуемого. Примером правильного ответа будет служить 4-й вариант

Испытуемому предлагалось сделать выбор из пяти вариантов ответа. Каждый вариант представлял собой вертикальный набор процентного соотношения веса выходного потока к входному. Таким образом, испытуемому необходимо было соотнести цифровые значения с шириной (весом) выходных потоков или частей структуры.

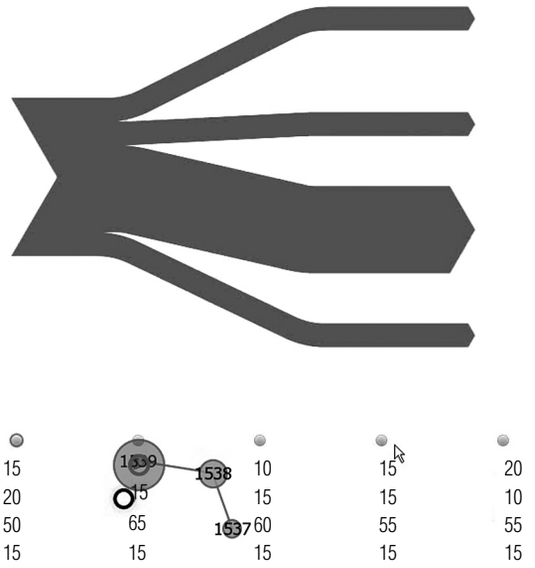


Рис. 1. Пример стимульного материала: четырехпотоковая диаграмма Сэнкей, демонстрирующая процентное разделение инвестиций по отдельным проектам

Испытуемый производил выбор с помощью компьютерного манипулятора типа мышь. Подразумевалось, что он является уверенным пользователем компьютера и задание с точки зрения человеко-компьютерного взаимодействия не является для него новым. Выбор производился кликом мышки по соответствующему варианту. Значения нерелевантных результатов подбирались таким образом, чтобы не было явного ошибочного результата, вызывающего отторжение варианта.

В процессе решения задачи глазодвигательная активность регистрировалась на оборудовании айтрекер *SMI RED250*. Кроме того, автоматически фиксировалась правильность ответа и длительность решения задачи. Для проведения эксперимента было разработано программное обеспечение на базе платформы *NetBeans*. После совершения выбора испытуемому предъявлялся стимул другой природы на 15 секунд, после чего снова предъявлялся стимул с четырехпотоковой диаграммой. Общее количество полезных стимулов – 30 штук. Однако, анализировались только те задания, с которыми испытуемый справился верно, общим числом 25 штук.

4. Анализ векторов кластеризации

Анализ глазодвигательной активности является нетривиальной задачей, в ходе решения которой требуется учесть несколько факторов одновре-

менно. Это координатное положение точки взора, протяженность саккады, длительность фиксации, размер зрачка и другие. Кластерный анализ – технология группирования объектов в ранее неизвестные группы. Он отличается от дискриминантного анализа тем, что не известны ни число кластеров, ни их характеристики. Проблема определения числа кластеров является одной из основных нерешенных до настоящего времени задач кластерного анализа. В рамках наших условий задача осложняется еще и тем, что необходимо определить число векторов кластеризации.

В качестве метода анализа были приняты иерархические дивизимные алгоритмы, работающие с универсальным кластером, состоящим из всех зрительных фиксаций, с последующим пошаговым разбиением его на меньшие части. «Удобство таких методов состоит в том, что процесс деления можно в любой момент остановить. Три наиболее популярных дивизимных метода – бисекция *k*-средних, бисекция главной компоненты и концептуальный кластер-анализ» [21, с. 10]. С помощью кластерного анализа была предпринята успешная попытка группировки результатов по назначенным признакам.

Для математического анализа были взяты данные только правильно решенных задач. Причем фиксации учитывались только в области диаграмм и не учитывались в области выбора ответа. В качестве аппарата кластеризации был выбран метод *k*-средних с многомерным вектором (под вектором кластеризации мы понимаем параметры глазодвигательной активности, которые выбираются для анализа).

Для определения качества векторов кластеризации были рассмотрены два алгоритма: *Davies Bouldin* [22, 23] и *Average Within Distance* [24]. Анализ оценки качества кластеризации с разным количеством кластеров приведен на *рис. 2*, где представлена визуализация расчетов для нескольких наборов векторов кластеризации:

– ALL 4 – для вектора, основанного на «координате X», «координате Y», «Длительности фиксации» и «Размере зрачка»;
– PUPIL – для вектора, основанного на «координате X», «координате Y» и «Размере зрачка»;
– DURATION – для вектора, основанного на «координате X», «координате Y» и «Длительности фиксации»;
– XY – для вектора, основанного на «координате X» и «координате Y»;

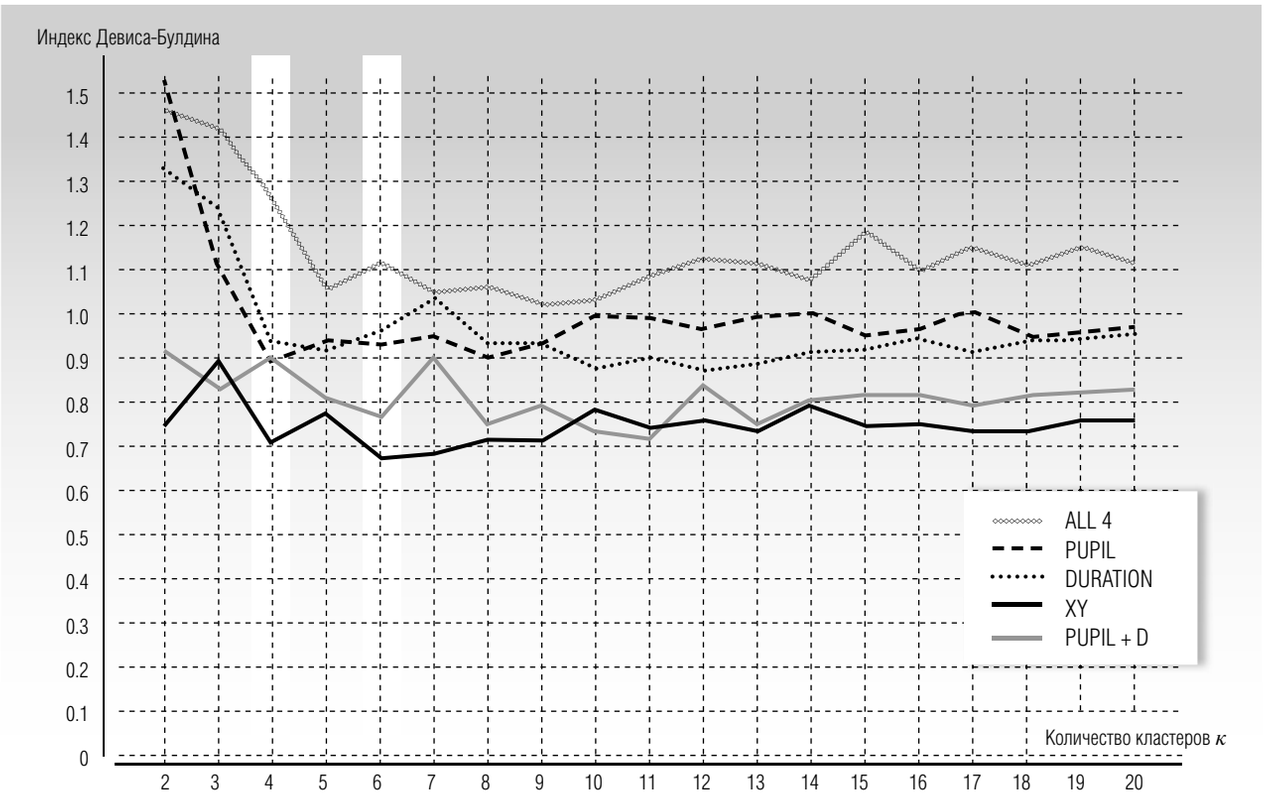


Рис. 2. Показатели качества кластеризации при количестве кластеров от 2 до 20 по алгоритму *Davies Bouldin*

– PUPIL+D – для вектора, основанного на «Длительности фиксации» и «Размере зрачка» (без «координаты X» и «координаты Y»).

График показывает, что наибольший интерес, в зависимости от состава вектора кластеризации, представляет группировка в четыре и шесть кластеров. По индексу Девиса-Булдина нас интересуют точки с наименьшим показателем. Однако для дальнейших исследований были также взяты другие размеры кластеризации: 3, 5, 7 и 8. Для большей наглядности и интерпретации предполагаемых результатов рассмотрен вектор кластеризации, содержащий два аргумента: «координату X» и «координату Y».

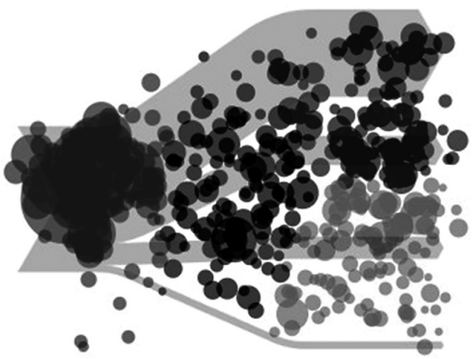


Рис. 3. Визуализация кластерной модели соотнесенной с фиксациями взора

Рабочим был принят вектор параметров для кластеризации, включающий в себя «координату X» и «координату Y». Визуализация результата при четырехкластерном разбиении представлена на *рис. 3* с демонстрацией стимульного материала одного из заданий. При этом точки фиксации показаны со всех 30 стимулов и наложены друг на друга.

5. Сравнение кластерных моделей

Для визуальной оценки кластеризации при использовании моделей с разным числом кластеров были построены соответствующие диаграммы визуализации (*рис. 4 и 5*). При разбиении на четыре кластера можно увидеть, что одна группа фиксаций сконцентрирована во входящей части стимула (потоковой диаграммы) с левой стороны, а две «выходящие» группы располагаются в его правой части. Однако обнаруживается устойчивый кластер, который не принимался во внимание при первоначальной постановке задачи: это четвертая группа, которая располагается в центральной части стимула.

При сравнении двух вариантов кластеризации можно отметить, что кластеры при группировке на шесть частей во входной части диаграммы ведут себя объяснимо. Положение нижнего левого кластера можно интерпретировать условиями эксперимента: в нижней части под диаграммой располагались варианты ответа, с которыми испытуемый должен был сверяться. Несмотря на то, что при обработки данным методом часть стимула с вариантами ответа была отсечена, результат этой сверки все же был получен в виде отдельного кластера. Особый интерес представляет шестой кластер, который располагается в правой части диаграммы, превратив два кластера на выходе в три.

На *рис. 5*, где представлены варианты моделей с 3, 5, 7 и 8 группировками, видно, что при кластеризации отсекаются вспомогательные фиксации, которые принадлежат задаче переключения внимания на сравнение с ответом.

Для проверки качества моделей было проведено тестирование моделей. В качестве алгоритма проверки

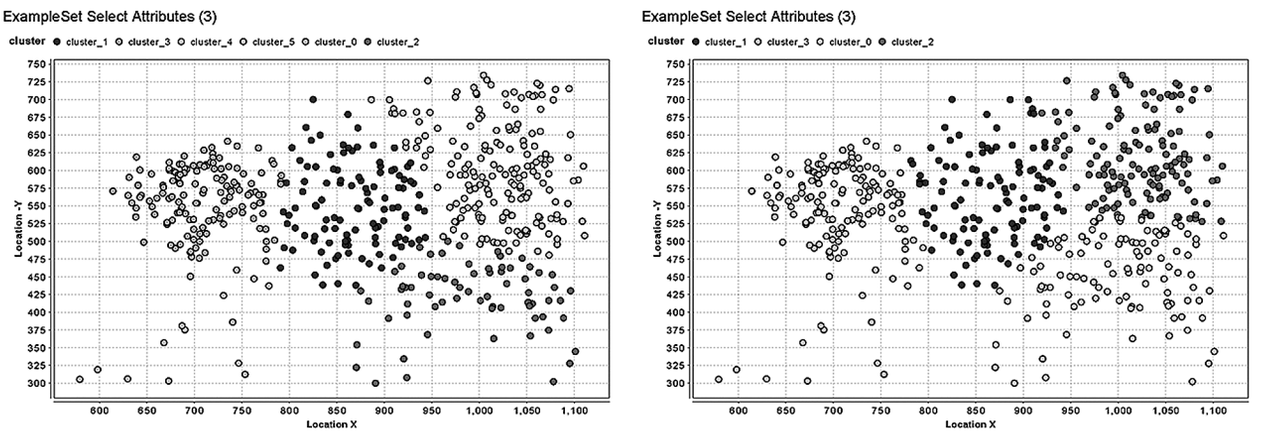


Рис. 4. Диаграммы расположения фиксаций при группировке в 4 и 6 кластеров

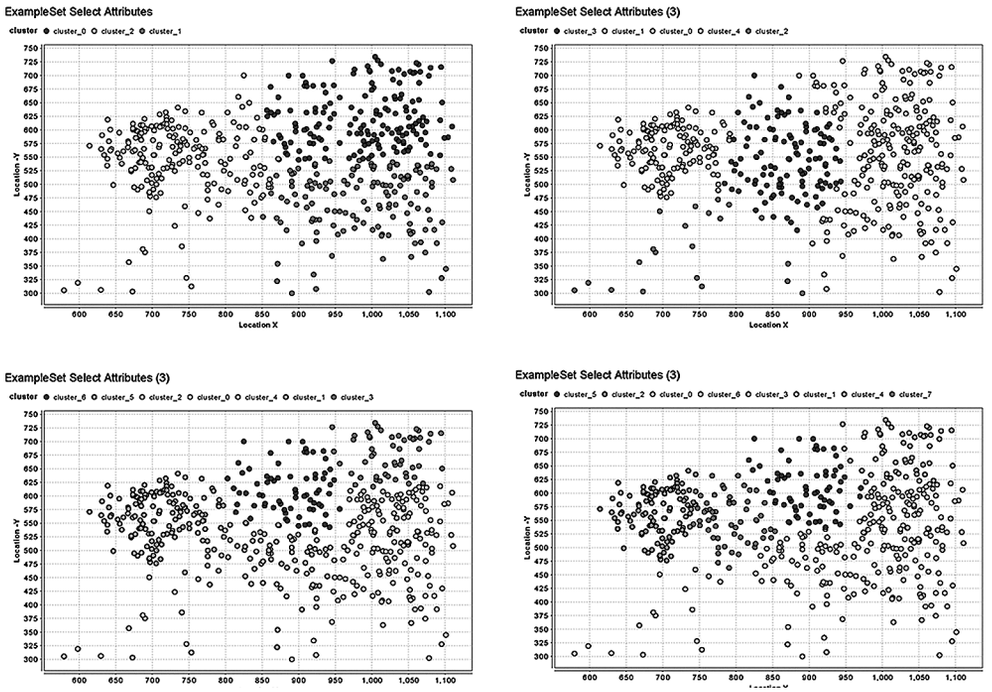


Рис. 5. Кластерные модели с 3, 5, 7 и 8 группировками

был выбран алгоритм *Map Clustering on Labels* из пакета *RapidMiner Core* [25]. Суть его работы заключается в следующем: используя как входной параметр кластерную модель с данными, он оценивает соответствие между данной моделью и моделью прогнозирования. Он настраивает входную модель и оценивает наиболее подходящие пары «элемент – кластер». Результат имеет вероятностный атрибут предсказания, который является производным от атрибута кластера.

Таблица 1.

Проверка кластерной модели с группировкой на четыре кластера

	true cluster_1	true cluster_3	true cluster_0	true cluster_2	class precision
pred. cluster_1	**117**	0	0	0	100%
pred. cluster_3	0	**151**	0	0	100%
pred. cluster_0	0	0	**109**	0	100%
pred. cluster_2	0	0	0	**135**	100%
class recall	100%	100%	100%	100%	

Результаты тестирования моделей с четырьмя и шестью кластерами приведены в *табл. 1* и *2*. Общая точность предсказания пар элементов для 4- и 6-кластерной модели составляют соответственно 100% и 98,44%. Значения точности для моделей кластеризации представлены в *табл. 3*.

Таблица 2.

Проверка кластерной модели с группировкой на шесть кластеров

	true cl._1	true cl._3	true cl.r_4	true cl._5	true cl.r_0	true cl._2	class precision
pred. cluster_1	**111**	0	0	0	0	1	99,11%
pred. cluster_3	0	**138**	0	0	0	0	100%
pred. cluster_4	0	0	**70**	1	0	0	98,59%
pred. cluster_5	0	0	0	**104**	0	4	96,30%
pred. cluster_0	1	1	0	0	**15**	0	88,24%
pred. cluster_2	0	0	0	0	0	**66**	100%
class recall	99,11%	99,28%	100%	99,05%	100%	92,96%	

Таблица 3.

Точность предсказания соотнесения элемента с кластером

Параметр\Число кластеров модели	3	4	5	6	7	8
accuracy	100%	100%	97,85%	98,44%	93,16%	90,43%

Из *табл. 3* видно, что кластеризация до шести групп включительно дает приемлемые результаты для анализа, но при увеличении числа кластеров вероятность верного предугадывания падает. Инте-

ресным представляется то, что при четырех кластерах модель демонстрирует 100% результат.

Заключение

В результате анализа визуальной оценки кластерной модели можно сделать заключение, что кластеры располагаются в соответствии со структурой стимула, т.е. четырехпотоковой диаграммы Сэнкей. Принято считать, что глазодвигательная активность детерминирована текущей задачей или общей целью [26]. Исходя из этого, полученный результат может быть интерпретирован только единственным способом: не стимульный материал сам по себе влияет на глазодвигательную активность, а задача, которую решает при этом испытуемый. Однако, по условиям эксперимента, стимульный материал является частью задачи и не может быть исключен. В таком случае решение прямой задачи по извлечению информации из потоковой диаграммы имеет связь с паттерном зрительных фиксаций. Алгоритмизация анализа кластерных моделей позволяет перевести визуальную интерпретацию структур числовых данных в круг задач поддержки принятия решений, решаемых с помощью программных средств.

Гипотеза относительно двух групп фиксаций во входной и в выходной части потоковой диаграммы не подтвердилась. Было обнаружено как минимум четыре кластера, основанных на координатах взгляда и длительности фиксации. В найденной модели присутствовал «входной» кластер и «выходная группа кластеров», также явно определился и центральный кластер зрительных фиксаций. При дальнейшем увеличении числа кластеров картина меняется в сторону большей детализации. Каждая часть структуры сравнивается с представленной таблицей данных последовательно от верхней правой части и далее вниз. Этим обусловлена правая «выходная группа кластеров». Также осуществляется сравнение каждой части с целым, представляемым «входным» кластером. Центральный кластер интерпретирует сравнение частей между собой в динамической части, когда разделение на части уже есть, но еще незначительное. В этом случае потоки композиционно объединяются в единый блок.

Дополнительное значение диаграмме дает обозначенное стрелками направление движения слева направо. Очевидно, что прослеживается определенный нарратив при рассматривании диаграммы, выявляющий последовательность «движения» потока от целого к его структурным частям. Вероятно, зритель включается в ее «потоковый» смысл. Для подтверждения данного вывода требуется дополнительные эксперименты, для анализа результатов которых можно использовать кластерный метод. ■

Литература

1. Лаптев В.В. Инфографика: основные понятия и определения // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Гуманитарные и общественные науки. 2013. № 4 (184). С. 180–187.
2. Орлов П.А. Инфографика и программирование. СПб.: Эйдос, 2013. 351 с.
3. Eells W.C. The relative merits of circles and bars for representing component parts // Journal of the American Statistical Association. 1926. Vol. 21. P. 119–132.
4. von Huhn R. A discussion of the Eells' experiment // Journal of the American Statistical Association. 1927. Vol. 22, No. 160. P. 31–36.
5. Croxton F.E., Stein H. Graphic comparisons by bar, squares, circles, and cubes // Journal of the American Statistical Association. 1932. Vol. 27, No. 177. P. 54–60.
6. Cleveland W.S., McGill R. Graphical perception: Theory, experimentation, and application to the development of graphical methods // Journal of the American Statistical Association. 1984. Vol. 79, No. 387. P. 531–554.
7. Spence I., Lewandowsky S. Displaying proportions and percentages // Applied Cognitive Psychology. 1991. No. 5. P. 61–77.
8. Zacks J., Tversky B. Bars and lines: A study of graphic communication // Memory & Cognition. 1999. Vol. 27, No. 6. P. 1073–1079.
9. Heer J., Bostock M. Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design // Proceedings of ACM Human Factors in Computing Systems (CHI 2010). April 10–15, 2010, Atlanta, USA. P. 203–212.
10. Kosara R., Ziemkiewicz C. Do Mechanical Turks dream of square pie charts? // Proceedings of Beyond Time and Errors: Novel Evaluation Methods for Information Visualization (BELIV). ACM Press, 2010. P. 373–382.
11. Ziemkiewicz C., Kosara R. Preconceptions and individual differences in understanding visual metaphors // Proceedings of the Eurographics/IEEE-VGTC Symposium on Visualization (EuroVis 2009). June 10-12, 2009, Berlin, Germany. 2009. Vol. 28. No. 3. P. 911–918.

12. Kosara R., Bendix F., Hauser H. Parallel sets: Interactive exploration and visual analysis of categorical data // Proceedings of the IEEE Symposium on Information Visualization 2005 (InfoVis 2005). October 23-25, 2005, Minneapolis, MN, USA. 2005. P. 133–140.
13. Ziemkiewicz C., Kosara R. Implied dynamics in information visualization // Proceedings of the 2010 International Conference on Advanced Visual Interfaces (AVI 2010), May 25-29, 2010, Rome, Italy. 2010. P. 215–222.
14. Ярбус А.Л. Роль движений глаз в процессе зрения. М.: Наука, 1965. 165 с.
15. Гиппенрейтер Ю.Б. Движения человеческого глаза. М.: МГУ, 1978. 256 с.
16. Барабанщиков В.А., Милад М.М. Методы окулографии в исследовании познавательных процессов и деятельности. М.: Ин-т психологии РАН, 1994. 87 с.
17. Величковский В.М. Когнитивная наука: основы психологии познания. Т. 2. М.: Академия, 2006. 432 с.
18. Ma H.-H. An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median // Behavior Modification. 2006. No. 30 (5). P. 598–617.
19. Crosbie J. Interrupted time-series analysis with brief single-subject data // Journal of Consulting and Clinical Psychology. 1993. No. 61 (6). P. 966–974.
20. The use of single-subject research to identify evidence-based practice in special education / R.N. Horner [et al.] // Exceptional Children. 2005. No. 71. P. 165–179.
21. Алескеров Ф.Т., Белоусова В.Ю., Егорова Л.Г., Миркин Б.Г. Анализ паттернов в статике и динамике. Часть 1: Обзор литературы и уточнение понятия // Бизнес-информатика. 2013. № 3 (25). С. 3–18.
22. Davies D.L., Bouldin D.W. A cluster separation measure // Pattern Analysis and Machine Intelligence. IEEE Transactions. 1979. No. 2. P. 224–227.
23. Ray S., Turi R.H. Determination of number of clusters in k-means clustering and application in colour image segmentation // Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT 1999), December 27-29, 1999, Calcutta, India. 1999. P. 137–143.
24. Petrovic S. A comparison between the Silhouette index and the Davies-Bouldin index in labeling IDS clusters // Proceedings of the 11th Nordic Workshop of Secure IT Systems, October 19-20, 2006, Linkoping, Sweden. 2006. P. 53–64.
25. Graczyk M., Lasota T., Trawinski B. Comparative analysis of premises valuation models using KEEL, RapidMiner, and WEKA // Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems. Berlin, Heidelberg: Springer, 2009. P. 800–812.
26. Бернштейн Н.А. Очерки о физиологии движений и физиологии активности. М.: Наука, 1990. 496 с.

CLUSTER ANALYSIS OF VISUAL PERCEPTION OF DATA STRUCTURE

Vladimir V. LAPTEV
Associate Professor, Department of Engineering Graphics and Design, Institute of Metallurgy, Mechanical Engineering and Transport, Peter the Great St.Petersburg Polytechnic University
Address: 29, Politekhnicheskaya Street, St. Petersburg, 195251, Russian Federation
E-mail: laptevsee@yandex.ru

Paul A. ORLOV
Senior Lecturer, Department of Engineering Graphics and Design, Institute of Metallurgy, Mechanical Engineering and Transport, Peter the Great St.Petersburg Polytechnic University; PhD candidate University of Eastern Finland
Address: 29, Politekhnicheskaya Street, St. Petersburg, 195251, Russian Federation
E-mail: paul.a.orlov@gmail.com

Data structures are common indicators in the fields of management and business. Infographics (serious graphics), a special area of Communication Design, provides a number of graphical ways to visualize this type of data. The application of each available chart type corresponds to certain limitations, which are associated with features of visual perception and semiotic aspects. In our study, we chose the Sankey flow diagram because of an insufficient degree of

scrutiny. This type of diagram is often used to represent data structure in business processes. We built an eye-tracking study to identify the methods of assessment forms of graphical image of data structure visualization. In our experiment, we used a 4-flow Sankey diagram as a stimulus.

Hierarchical divisive algorithms were taken as the method of analysis. This method works with a universal cluster consisting of all gaze fixation, followed by step partitioning it into smaller pieces. It has been found that there are at least four clusters based on the coordinates. In the present model, we found an «input» cluster and an «output cluster group» and clearly defined the central cluster of gaze fixations. Increasing the number of clusters changes the picture in the direction of greater detail. We show a certain narrative that is traced when viewing charts. This narrative identifies the sequence of flows «movement» from the whole to its structural parts. As a result, cluster analysis allows the visual interpretation of numerical data structures in a range of tasks to support decision making that can be solved by software.

Key words: cluster analysis, infographics, data visualization, data structure, diagram, eye tracking, eye tracker.

Citation: Laptev V.V., Orlov P.A. (2015) Klasternyi analiz vizual'nogo vosprijatija structury dannyh [Cluster analysis of visual perception of data structure]. *Business Informatics*, no. 3 (33), pp. 34–43 (in Russian).

References

1. Laptev V.V. (2013) Infografika: osnovnye ponjatija i opredelenija [Infographics: Basic concepts and definitions]. *St. Petersburg State Polytechnical University Journal. Humanities and Social Sciences,* no. 4 (184), pp. 180–187 (in Russian).
2. Orlov P.A. (2013) *Infografika i programmirovanie* [Infographics and programming]. St. Petersburg: Jejdos (in Russian).
3. Eells W.C. (1926) The relative merits of circles and bars for representing component parts. *Journal of the American Statistical Association*, vol. 21, pp. 119–132.
4. von Huhn R. (1927) A discussion of the Eells' experiment. *Journal of the American Statistical Association*, vol. 22, no. 160, pp. 31–36.
5. Croxton F., Stein H. (1932) Graphic comparisons by bar, squares, circles, and cubes. *Journal of the American Statistical Association,* vol. 27, no. 177, pp. 54–60.
6. Cleveland W., McGill R. (1984) Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association,* vol. 79, no. 387, pp. 531–554.
7. Spence I., Lewandowsky S. (1991) Displaying proportions and percentages. *Applied Cognitive Psychology,* vol. 5, pp. 61–77.
8. Zacks J., Tversky B. (1999) Bars and lines: A study of graphic communication. *Memory & Cognition,* vol. 27, no. 6, pp. 1073–1079.
9. Heer J., Bostock M. (2010) Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. Proceedings of *ACM Human Factors in Computing Systems (CHI 2010), April 10–15, 2010, Atlanta, GA, USA,* pp. 203–212.
10. Kosara R., Ziemkiewicz C. (2010) Do Mechanical Turks dream of square pie charts? Proceedings of *Beyond Time and Errors: Novel Evaluation Methods for Information Visualization (BELIV)*. ACM Press, pp. 373–382.
11. Ziemkiewicz C., Kosara R. (2009) Preconceptions and individual differences in understanding visual metaphors. Proceedings of the *Eurographics/ IEEE-VGTC Symposium on Visualization, (EuroVis 2009). June 10-12, 2009, Berlin, Germany*, vol. 28, no. 3, pp. 911–918.
12. Kosara, R., Bendix F., Hauser H. (2005) Parallel sets: Interactive exploration and visual analysis of categorical data. Proceedings of the *IEEE Symposium on Information Visualization 2005 (InfoVis), (InfoVis 2005). October 23-25, 2005, Minneapolis, MN, USA*, pp. 133–140.
13. Ziemkiewicz C., Kosara R. (2010) Implied dynamics in information visualization. Proceedings of the *2010 International Conference on Advanced Visual Interfaces (AVI 2010), May 25-29, 2010, Rome, Italy*, pp. 215–222.
14. Jarbus A.L. (1965) *Rol' dvizhenij glaz v processe zrenija* [The role of eye movements in vision]. Moscow: Nauka (in Russian).
15. Gippenrejter Ju.B. (1978) *Dvizhenija chelovecheskogo glaza* [Movement of the human eye]. Moscow: MGU (in Russian).
16. Barabanshhikov V.A., Milad M.M. (1994) *Metody okulografii v issledovanii poznavatel'nyh processov i dejatel'nosti* [Oculography methods in the study of cognitive processes and activities]. Moscow: Institut psihologii RAN (in Russian).
17. Velichkovskij B.M. (2006) *Kognitivnaja nauka: osnovy psihologii poznanija* [Cognitive science: foundations of cognitive psychology]. Moscow: Akademija (in Russian).
18. Ma H.-H. (2006). An alternative method for quantitative synthesis of single-subject researches: percentage of data points exceeding the median. *Behavior Modification*, vol. 30 (5), pp. 598–617.
19. Crosbie J. (1993) Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology*, no. 61 (6), pp. 966–974.
20. Horner R.H., Carr E.G., Halle J., McGee G., Odom S., Wolery M. (2005) The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, no. 71, pp. 165–179.
21. Aleskerov F.T., Belousova V.Yu., Egorova L.G., Mirkin B.G. (2013) Analiz patternov v statike i dinamike. Chast' 1: Obzor literatury i utochnenie ponjatija [Methods of pattern analysis in statics and dynamics. Part 1: Examples of application for social and economic processes analysis]. *Business Informatics*, no. 3 (25), pp. 3–18 (in Russian).
22. Davies D. L., Bouldin D. W. (1979) A cluster separation measure. *Pattern Analysis and Machine Intelligence. IEEE Transactions*, no. 2, pp. 224-227.
23. Ray S., Turi, R. (1999) Determination of number of clusters in k-means clustering and application in colour image segmentation. Proceedings of the *4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT 1999), December 27-29, 1999, Calcutta, India,* pp. 137–143.
24. Petrovic S. (2006) A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. Proceedings of the *11th Nordic Workshop of Secure IT Systems, October 19-20, 2006, Linkoping, Sweden,* pp. 53–64.
25. Graczyk M., Lasota T., Trawinski B. (2009) Comparative analysis of premises valuation models using KEEL, RapidMiner, and WEKA. *Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems,* Berlin, Heidelberg: Springer, pp. 800–812.
26. Bernshtejn N.A. (1990) *Ocherki o fiziologii dvizhenij i fiziologii aktivnosti* [Essays about the physiology of movements and physiology of activity]. Moscow: Nauka (in Russian).

ОБ ЭФФЕКТИВНОСТИ РАСПОЗНАВАНИИ ЛИЦ С ПОМОЩЬЮ ЛИНЕЙНОГО ДИСКРИМИНАНТНОГО АНАЛИЗА И МЕТОДА ГЛАВНЫХ КОМПОНЕНТ

А.В. МОКЕЕВ
старший преподаватель кафедры информационных систем,
факультет экономики и предпринимательства,
Южно-Уральский государственный университет
Адрес: 454080, г. Челябинск, пр. Ленина, д. 76
E-mail: gr.smk@mail.ru

В.В. МОКЕЕВ
доктор технических наук, заведующий кафедрой информационных систем,
факультет экономики и предпринимательства,
Южно-Уральский государственный университет
Адрес: 454080, г. Челябинск, пр. Ленина, д. 76
E-mail: mokeyev@mail.ru

Рассматривается решение задачи распознавания лиц с помощью метода главных компонент (МГК) и линейного дискриминантного анализа (ЛДА). Главная идея подхода МГК+ЛДА состоит в том, что, во-первых, изображение лица проецируется из исходного векторного пространства в подпространства лица главных компонент, во-вторых, для получения линейного классификатора используется линейный дискриминантный анализ. В работе исследуется эффективность подхода МГК+ЛДА для случая, когда изображения лиц не проходят предварительную обработку (масштабирование, поворот, центрирование, выравнивание яркости). Эффективность подхода МГК и ЛДА исследуется на изображениях лиц базы ORL. Показывается, что при увеличении числа изображений в классе учебной выборки, повышается точность распознавания лиц. При небольшом числе изображений для повышения качества распознавания лиц предлагается расширять учебную выборку изображениями, полученными путем масштабирования и поворота исходных изображений. При обработке больших наборов изображений для вычисления главных компонент предлагается использовать методы линейной конденсации и синтеза главных компонент. Метод синтеза главных компонент базируется на разбиении исходного множества изображений на небольшие наборы изображений, получении собственных векторов этих наборов (частных решений) и вычислении собственных векторов исходного набора на основе частных решений. Метод линейной конденсации использует понижение порядка матриц, позволяющее достаточно точно вычислять собственные векторы, собственные значения которых находятся в заданном интервале. Показано, что методы линейной конденсации и синтеза главных компонент позволяют существенно снизить трудоемкость построения классификатора при использовании подхода на МГК+ЛДА, не снижая точности распознавания лиц.

Ключевые слова: распознавание изображений, анализ главных компонент, линейный дискриминантный анализ, метод линейной конденсации, база данных ORL, синтез главных компонент.

Цитирование: Мокеев А.В., Мокеев В.В. Об эффективности распознавании лиц с помощью линейного дискриминантного анализа и метода главных компонент // Бизнес-информатика. 2015. № 3 (33). С. 44–54.

Введение

В настоящее время методы анализа данных активно развиваются в направлении обработки больших объемов данных. Источниками возникновения больших данных являются непрерывно поступающие данные с измерительных устройств (аудио и видео-регистрации и т.п.). Обработка изображений сегодня широко используется в системах безопасности для идентификации людей по изображениям лиц, мониторинга состояния технических объектов.

Существует большое количество методов и подходов, использующихся в системах распознавания лиц [1]. Среди них можно выделить метод главных компонент (МГК), линейный дискриминантный анализ (ЛДА), скрытые марковские модели (СММ), вейвлеты Габора. При использовании скрытых марковских моделей [2] для решения задачи распознавания лиц для каждого класса лиц вычисляется своя скрытая марковская модель. Далее для неизвестного образа запускаются все имеющиеся модели, и среди них ищется та, которая выдаёт самый близкий результат. Недостатком такого подхода является то, что скрытые марковские модели не обладают хорошей различающей способностью, т.к. алгоритм обучения максимизирует отклик на свои классы, но не минимизирует отклик на другие классы. Методы распознавания, основанные на использовании вейвлетов Габора [3,4], показывают высокую эффективность. Фильтры Габора используются на стадии предобработки для формирования вектора Габор-особенностей изображения лица. Метод вейвлетов Габора устойчив к изменениям в освещении, поскольку не использует напрямую значения оттенков серого каждого пикселя, а извлекает особенности.

Метод главных компонент является одним из наиболее популярных при решении задач распознавании лиц. Реконструкция лиц с помощью МГК впервые была сделана в работе [5]. Метод распознавания, известный как метод собственных лиц, определяет пространство признаков, которое уменьшает размерность оригинального пространства данных. Однако, методы распознавания, основанные на МГК, страдают от двух ограничений, а именно: слабой дискриминационной силы и больших вычислительных затрат. Недостатки могут быть преодолены путем добавления ЛДА [6]. ЛДА ищет такое преобразование изображений, которое бы минимизировало внутриклассовые и максимизировало межклассовые различия набора изображений. Суть совместного применения МГК и ЛДА заключается в том, что сначала с помощью МГК сокращается размерность изображений, а затем в рамках ЛДА выполняется преобразование, которые лучше всего отличают один класс изображений от другого [7].

Как правило, процесс распознавания лиц начинается с процедуры нормализации (масштабирование, центрирование, отсечение фона, выравнивание яркости). Нормализация изображений требует дополнительных вычислений, что не играет существенной роли при построении классификатора, но на этапе распознавания новых изображений в режиме реального времени может оказать негативное воздействие, так как нормализация каждого нового изображения приводит к временным затратам.

В данной работе рассматривается решение задачи распознавания лиц на основе МГК и ЛДА без использования процедуры нормализации лиц. Распознавание изображений на основе МГК и ЛДА начинается с построения классификатора, который затем используется при распознавании неизвестных изображений. При построении классификатора используется учебная выборка ненормализованных изображений. Однако в этом случае хорошее качество распознавания лиц может быть достигнуто при увеличении числа изображений в классе лиц, что приводит к росту вычислительных затрат на создание классификатора. Трудоемкость расчета главных компонент с увеличением числа изображений растет по кубическому закону. Поэтому необходимы методы, снижающие трудоемкость построения классификаторов при большом числе изображений. В работах [8, 9] предлагаются новые подходы (метод линейной конденсации и метод синтеза главных компонент), снижающие трудоемкость вычисления главных компонент больших наборов изображений. Одним из способов снижения трудоемкости построения классификатора при большом числе изображений (>> 1000) в учебной выборке является двумерный метод главных компонент (2ДМГК) [10], в котором главные компоненты ищутся по строкам и столбцам. Основная идея 2ДМГК заключается в том, чтобы найти оптимальные проекции строк без преобразования изображения в вектор. Недостатком двумерного метода главных компонент является то, что он требует больше главных компонент для описания изображения, чем одномерный метод главных компонент. Таким образом, двумерному методу главных компонент необходимо больше памяти для хранения изображений и больше затрат времени при их классификации. В случае, когда число изображений намного больше числа строк/столбцов изображения использование

2ДМГК может приводить к снижению трудоемкости вычисления главных компонент.

В данной работе рассматривается технология построения классификаторов с использованием методов МГК и ЛДА без нормализации изображений. Использование таких классификаторов может снизить затраты на этапе распознавания неизвестных изображений, что особенно важно при работе системы распознавания лиц в реальном времени. Для снижения трудоемкости построения классификаторов в случае, когда учебная выборка содержит большое число изображений, предлагается использовать методы линейной конденсации и синтеза главных компонент.

1. Теоретический базис: МГК плюс ЛДА

Пусть имеется набор изображений, каждый из которых описывается вектором x_i^k, где i – номер изображений ($i = 1, 2, 3, ..., M_k$, $k = 1, 2, ..., K$). Размерность вектора x_i^k равняется числу пикселей изображения (N). Таким образом, весь набор изображений можно представить в виде матрицы $\mathbf{X}$, столбцами которой являются векторы x_i^k. Размерность пространства изображений определяется произведением $N \times M$.

Обозначим матрицу отцентрированных изображений как $\mathbf{X^0}$. Столбцами $\mathbf{X^0}$ являются векторы

$$\overline{x}_i^k = x_i^k - m, \text{ где}$$

$$m = \frac{1}{M}\sum_k^K \sum_i^{M_k} x_i^k -$$

среднеарифметический вектор изображений,

$$M = \sum_{k=1}^{K} M_k.$$

Подход МГК+ЛДА состоит из двух этапов: на первом этапе применяется МГК для уменьшения размерности от N до p и получения матрицы $\mathbf{V}_{pca}$, которая формируется из собственных векторов уравнения:

$$(\mathbf{A} - \lambda \mathbf{I}) v_0^{pca} = 0, \qquad (1)$$

где $\mathbf{A}^* = 1/M \, \mathbf{X^0} (\mathbf{X^0})^T$ – ковариационная матрица размерностью $N \times N$, индекс «T» означает транспонирование матрицы, $\mathbf{I}$ – единичная матрица, v_0^{pca} – собственный вектор, λ – собственное значение.

В связи с тем, что матрица $\mathbf{A}^*$ имеет высокий порядок, вычисление собственных векторов представляет существенные трудности. Поэтому более эффективно вычислять главные компоненты $\mathbf{V}_{pca}$ по матрице собственных векторов $\mathbf{U}_{pca}$ [9], которые определяются путем решения уравнения:

$$(\mathbf{A} - \lambda \mathbf{I}) u_0^{pca} = 0, \qquad (2)$$

где $\mathbf{A} = 1/M \, (\mathbf{X^0})^T \mathbf{X^0}$ – матрица Грамма размерностью $M \times M$, u_{0i}^{pca} – собственный вектор.

Следует отметить, что собственные значения уравнений (1) и (2) совпадают. В связи с тем, что M существенно меньше N, можно значительно снизить трудоемкость вычисления матрицы $\mathbf{V}_{pca}$.

На втором этапе, применяется ЛДА с целью нахождения таких линейных комбинаций признаков, которые наилучшим образом разделяют классы изображений лиц. Целью ЛДА является получение матрицы преобразования $\boldsymbol{Wlda}$, которая минимизирует внутриклассовое и максимизирует межклассовое расстояние в пространстве признаков

$$\mathbf{W}_{lda} = arg\max_{\mathrm{w}} \frac{\left|\mathbf{W}^T \mathbf{V}_{pca}^T \mathbf{A}_b \mathbf{V}_{pca} \mathbf{W}\right|}{\left|\mathbf{W}^T \mathbf{V}_{pca}^T \mathbf{A}_\omega \mathbf{V}_{pca} \mathbf{W}\right|} = arg\max_{\mathrm{w}} \frac{\left|\mathbf{W}^T \mathbf{A}_b' \mathbf{W}\right|}{\left|\mathbf{W}^T \mathbf{A}_\omega' \mathbf{W}\right|}. \qquad (3)$$

Здесь $\mathbf{A}_b$ – ковариационная матрица межклассовых различий, $\mathbf{A}_\omega$ – ковариационная матрица внутриклассовых различий [6].

Столбцами матрицы $\mathbf{W}_{lda}$ являются собственные векторы w_{0i}^{lda}, которые получаются в результате решения уравнения:

$$(\mathbf{A}_b' - \lambda \mathbf{A}_\omega') w_0^{lda} = 0. \qquad (4)$$

Задача (4) является обобщенной задачей собственных значений. Для решения этой задачи предлагается использовать обобщенный метод Якоби, который не требует обращений матрицы $\mathbf{A}_\omega$ [11]. В результате решения (4) определяется матрица дискриминантных компонент, столбцами которой являются собственные векторы уравнения (4) с наибольшими собственными значениями.

2. Распознавание ненормализованных изображений лиц

Экспериментальное исследование эффективности подхода МГК + ЛДА проводится с использованием изображений лиц, собранных в базе данных ORL. База ORL содержит изображения 40 человек, каждый из которых описывается 10 различными изображениями. На *рис. 1* приведены примеры изображений лиц из базы данных ORL.

Для исследования качества распознавания используется процедура кросс-валидации, усред-

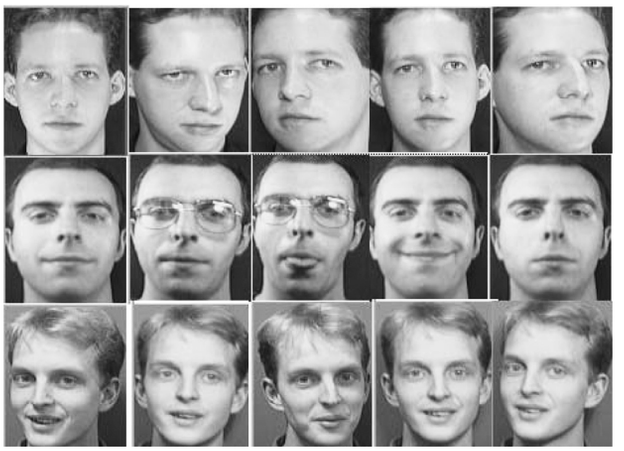


Рис. 1. Примеры лиц, выделенных из изображений базы данных ORL

няющая коэффициенты распознавания, полученные при различном делении базы изображений на учебные и тестовые наборы. Эксперименты проводятся для учебных наборов, содержащих L (2, 4, 8) изображений в каждом классе базы ORL, которые выбираются случайно. Все оставшиеся изображения составляют тестовую выборку. Таким образом, учебные наборы состоят из (Lx40) изображений, а тестовые наборы – из (400-Lx40) изображений. Не существует перекрытия между тестовым и учебным наборами. Для повышения точности оценки выполняются десять различных делений изображений на учебные и тестовые наборы, а полученные в ходе экспериментов коэффициенты распознавания изображений лиц усредняются. На *рис. 2* показаны усредненные коэффициенты распознавания людей по изображениям лиц тестовой выборки в зависимости от числа главных компонент для случаев,

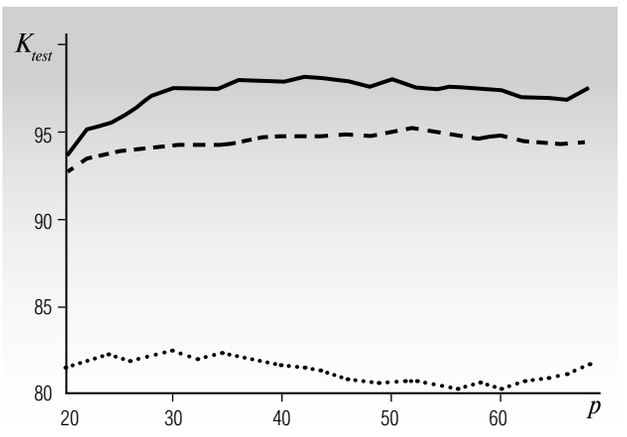


Рис. 2. Усредненные коэффициенты распознавания изображений тестовой выборки базы ORL в зависимости от числа главных компонент: L = 8 (сплошная линия); L = 4 (штриховая линия); L = 2 (пунктирная линия)

когда учебная выборка содержит 2 изображения в классе (штрих-пунктирная линия), 4 изображения в классе (штриховая линия) и 8 изображений в классе (сплошная линия) соответственно. Отметим, что коэффициенты распознавания, показанные на *рис. 2*, получены для ненормализованных изображений лиц.

В работе [12] представлены результаты распознавания лиц базы ORL различными методами для случаев, когда число изображений в классе меняется от 2 до 9. Изображения лиц предварительно масштабируются, центрируются и поворачиваются так, чтобы центры глаз были на горизонтальной линии. Усредненные коэффициенты распознавания, полученные методами МГК, ЛДА, DLA (Discriminative Locality Alignment [13]), NDLPP (Null space Discriminant Locality Preserving Projections[14]), RLPDA (Regularized Locality Preserving Discriminant Analysis [12]) представляются в *табл. 1.* Сравнение коэффициентов распознавания, представленных на *рис. 2,* с результатами, полученными другими методами, показывают хорошую точность распознавания подхода МГК плюс ЛДА даже для ненормализованных изображений, когда число изображений в классе достаточно большое. Если число изображений невелико, то можно увеличить число изображений в классах учебной выборки путем масштабирования и поворота исходных изображений.

Таблица 1.

Результаты сравнения точности распознавания различных алгоритмов

Метод	Число изображений в классе		
	2	4	8
МГК	69,6	83,6	94,5
ЛДА	80,1	91,5	96,3
DLA	73,3	92,6	98,8
NDLPP	83,0	94,7	98,3
RLPDA	80,7	94,8	98,6
МГК+ЛДА	82,2	94,2	98,2

В табл. 2 представлены коэффициенты распознавания, полученные с помощью классификаторов, построенных на расширенных учебных выборках. Исходные выборки с 2, 4 и 8 изображениями в классе расширяются путем добавления изображений, полученных из исходных путем масштабирования и поворота. В скобках указано число изображений

в классе, полученное в результате расширения выборки. Дополнительные изображения получаются путем уменьшения/увеличения исходных изображений на 5% и/или поворота изображений по часовой или против часовой стрелки на 4°.

Как видно из *таблицы*, расширение учебной выборки дает наибольший эффект в случае, когда число изображений в классе небольшое. Это означает, что расширение учебной выборки оригинальными, а не производными изображениями представляет более эффективный путь повышения качества классификаторов.

Таблица 2.

Коэффициенты распознавания различных алгоритмов

Способ расширения учебной выборки	Число исходных изображений в классе		
	2	4	8
Масштабирование	85,6 (6)	95,8 (12)	98,6 (24)
Поворот	85,4 (6)	94,6 (12)	98, 7 (24)
Масштабирование и поворот	85,4 (18)	96,5 (36)	98,3 (72)

На *рис. 3* показаны усредненные коэффициенты распознавания лиц тестовой выборки для вариантов:

а) учебная выборка содержит 2 изображения в классе ($L = 2$) (пунктирная линия);

б) учебная выборка ($L = 2$), дополненная изображениями, полученными путем масштабирования (увеличения или уменьшения) исходных изображений на 5% (штриховая линия);

в) учебная выборка ($L = 2$), дополненная изображениями, полученными путем поворота исходных изображений по часовой или против часовой стрелки на 4° (штрих-пунктирная линия);

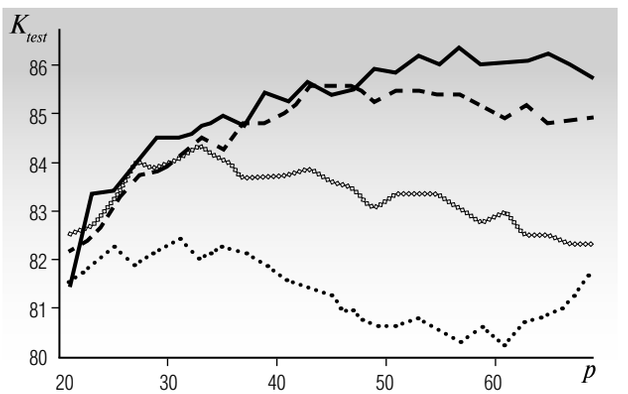


Рис. 3. Усредненные коэффициенты распознавания изображений тестовой выборки базы ORL при расширении учебной выборки

г) учебная выборка ($L = 2$), дополненная изображениями, полученными путем масштабирования (увеличения или уменьшения) исходных изображений на 5% и поворота по часовой или против часовой стрелки на 4° (сплошная линия).

Как видно из *рисунка*, увеличение числа изображений в классе приводит к росту качества распознавания, при этом в случае двух изображений в классе максимальные значения коэффициентов распознавания получаются для 30 главных компонент, а в случае, когда число изображений в классе расширяется до 18, коэффициенты распознавания достигают своих максимальных значений уже при 55 главных компонент. Таким образом, чем больше изображений в классе, тем больше главных компонент потребуется для построения качественного классификатора.

Проведенные исследования показывают, что при построении классификатора на основе учебной выборки из ненормализованных изображений качество распознавания лиц увеличивается с ростом числа изображений в классе. Если число изображений в классе невелико, то учебная выборка может быть дополнена изображениями, полученными путем поворота и масштабирования исходных изображений. Таким образом, даже для распознавания сравнительно небольшого числа людей, размер учебной выборки может быть достаточно большим, и вычисление главных компонент может потребовать существенных вычислительных ресурсов.

3. Вычисление главных компонент больших наборов изображений

Время вычисления главных компонент зависит в основном от трудоемкости решения уравнения (2), которая существенно растет с увеличением порядка матриц. В случае если число изображений в наборе велико, то порядок матриц становится большим и определение главных компонент требует значительных вычислительных затрат. Обзор методов решения задачи собственных значений можно найти в работе [9]. Из существующих методов можно выделить метод Хаусхолдера, который позволяет выполнять требуемые преобразования быстрее, чем метод Гивенса или метод Якоби, так как требует выполнения меньшего числа, хотя и более трудоемких операций. Однако, с помощью метода Хаусхолдера вычисляются все собственные значения и соответствующие им собственные векторы, в то время как при вычислении главных компонент требуется срав-

нительно небольшое число собственных векторов с наибольшими собственными значениями.

Для решения неполной задачи собственных значений больших матриц предлагаются метод синтеза главных компонент [8] и метод линейной конденсации [9]. Метод синтеза главных компонент базируется на разбиении исходного множества изображений на небольшие наборы изображений, получении собственных векторов этих наборов (частных решений) и вычислении собственных векторов исходного набора на основе частных решений. Точность метода зависит от размера наборов изображений (m_k) и числа собственных векторов, образующих частное решение l_k ($l_k < m_k$). При уменьшении l_k размеры результирующей матрицы, которая используется для вычисления собственных векторов, уменьшаются, однако при этом растут погрешности собственных векторов.

Метод линейной конденсации реализуется в виде алгоритма блочно-ортогональной конденсации. В работе [9] описывается алгоритм многоуровневой линейной конденсации для вычисления собственных значений в интервале от 0 до μ_2 и соответствующих им собственных векторов. Алгоритм включает пять шагов. Первый шаг представляет процедуру многоуровневого понижения порядка матриц, которая начинается с того, что все признаки сортируются в порядке убывания диагональных коэффициентов ковариационной матрицы **A**. На каждом уровне процедуры понижения порядка матриц выбирается группа признаков с минимальными диагональными коэффициентами. Решение об исключении выбранных признаков принимается при выполнении условия:

$$\mu_{min} > k_c \mu_2. \qquad (5)$$

Здесь $\mu_{min} = 1/\lambda_{max}$ – наименьшее собственное значение блока удаляемых переменных $\mathbf{A}_{ss}^k$, k_c – параметр отсечения, μ_2 – верхняя граница диапазона вычисляемых собственных значений (диапазона конденсации).

Если условие (5) выполняется, то осуществляется понижение порядка матрицы. Условие (5) может быть переписано в виде $\lambda_{max} < \lambda_2 / k_c$. Как известно, сумма диагональных коэффициентов матрицы равна сумме собственных значений. Поэтому набор удаляемых признаков формируется из признаков, которым соответствуют наименьшие диагональные коэффициенты матрицы **A**. По сути, делается предположение о том, что чем меньше сумма собственных значений λ, тем меньше максимальное собственное значение анализируемой матрицы. В процессе понижения порядка матрицы **A** последовательно получаются матрицы $\mathbf{A}_1$, $\mathbf{A}_2$, ..., $\mathbf{A}_k$, где k — уровень понижения порядка матрицы **A**. С увеличением k уменьшается порядок матрицы $\mathbf{A}_k$.

Эффективность алгоритма многоуровневой линейной конденсации зависит от того, насколько сильно понижается порядок матриц, однако, степень понижения не всегда оказывается достаточно высокой. В работе [9] демонстрируется быстродействие алгоритма многоуровневой линейной конденсации на примере вычисления 67 главных компонент наборов различной размерности (от 500 до 5000). Однако дальнейшие исследования показали, что при расчете большого числа главных компонент (несколько сотен) не удается существенно снизить порядок матриц. Это обусловлено тем, что собственные значения в районе верхней границы μ_2 лежат очень плотно и не удается выбрать исключаемые признаки, не нарушая условия (5).

Для преодоления этих трудностей предлагается использовать алгоритм блочно-ортогональной конденсации, который также представляет собой многоуровневый процесс понижения порядка матрицы. Матрица также делится на блоки, однако, при выборе удаляемых признаков блок признаков (кандидатов на удаление) приводится к диагональной форме с помощью ортогонального преобразования.

Пусть уравнение (2) на k-ом уровне понижения порядка матриц имеет вид:

$$(\mathbf{I}_k - \mu \mathbf{A}_k)\, u_{k-1}^{pca} = 0, \qquad (6)$$

где $\mu = 1/\lambda$, u_{k-1}^{pca} представляет вектор u_0^{pca} на k-1 уровне понижения порядка матриц. Представим матрицу $\mathbf{A}_k$ и вектор u_{k-1}^{pca} в форме:

$$\mathbf{A}_k = \begin{bmatrix} \mathbf{A}_{bb}^k & \mathbf{A}_{bs} \\ \mathbf{A}_{sb}^k & \mathbf{A}_{ss}^k \end{bmatrix},\ u_{k-1}^{pca} = \begin{Bmatrix} u_b^{k-1} \\ u_s^{k-1} \end{Bmatrix}.$$

Здесь индекс b относится к удерживаемым признакам, и индекс s – к удаляемым признакам.

Диагонализация матрицы удаляемых признаков осуществляется с помощью ортогонального преобразования:

$$\mathbf{P}_s^T \mathbf{A}_{ss}^k \mathbf{P}_s = \Sigma_{ss}. \qquad (7)$$

Матрица $\mathbf{P}_s$ является ортогональной матрицей и состоит из собственных векторов матрицы $\mathbf{A}_{ss}^k$, полученных при решении уравнения:

$$(\mathbf{I}_s - \mu \mathbf{A}_{ss}^k)\, p_s = 0. \qquad (8)$$

Таким образом,

$$u_k^{pca} = \begin{bmatrix} \mathbf{I}_b & 0 \\ 0 & \mathbf{P}_s \end{bmatrix} \begin{Bmatrix} u_b^{k-1} \\ u_s^{k-1} \end{Bmatrix} = \mathbf{P} u_{k-1}^{pca}. \quad (9)$$

Отметим, что $\mathbf{P}_s$ является ортогональной матрицей, поэтому справедливо следующее соотношение:

$$\mathbf{P}_s^T \mathbf{I}_s \mathbf{P}_s = \mathbf{P}_s^T \mathbf{P}_s = \mathbf{I}_s \quad (10)$$

Подставляем соотношение (9) в уравнение (6) и далее, умножая справа на матрицу $\mathbf{P}^T$, с учетом выражения (10) получаем:

$$(\mathbf{I}_k - \mu \mathbf{A}_k^*)\, u_k^{pca} = 0, \quad (11)$$

где $\mathbf{A}_k^* = \mathbf{P}^T \mathbf{A}_k \mathbf{P}$.

С учетом блочного деления и соотношения (9) матрица $\mathbf{A}_k$ может быть представлена в виде:

$$\mathbf{A}_k^* = \begin{bmatrix} \mathbf{I}_b & 0 \\ 0 & \mathbf{P}_s \end{bmatrix}^T \begin{bmatrix} \mathbf{A}_{bb}^k & \mathbf{A}_{bs} \\ \mathbf{A}_{sb}^k & \mathbf{A}_{ss}^k \end{bmatrix} \begin{bmatrix} \mathbf{I}_r & 0 \\ 0 & \mathbf{P}_s \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{bb}^k & \mathbf{A}_{bs}^k \mathbf{P}_s \\ \mathbf{P}_s^T \mathbf{A}_{sb}^k & \Sigma_{ss} \end{bmatrix}.$$

Если диагональные коэффициенты матрицы Σ_{ss} упорядочены по убыванию, то обратная величина первого диагонального коэффициента матрицы Σ_{ss} будет равна наименьшему собственному значению блока удаляемых признаков (μ_{min}). Проверка условия (5) позволяет определить, можно ли удалять признаки, входящие в блок удаляемых признаков. Если удалять признаки нельзя, то условие (5) проверяется для второго диагонального коэффициента. По сути это означает, что мы уменьшаем блок удаляемых признаков на единицу и принимаем решение об удалении уменьшенного блока признаков. Если условие (5) не выполняется, то мы снова уменьшаем блок удаляемых признаков на единицу, т.е. условие (5) проверяется для третьего диагонального коэффициента матрицы Σ_{ss} и т.д. Если условие (5) выполняется, то это означает, что мы определили блок признаков, которые можно удалять. Если условие (5) не выполняется и число удаляемых признаков равно нулю, то это означает, что процесс понижения порядка закончен.

Метод линейной конденсации и метод синтеза главных компонент дают приближенное решение задачи собственных значений. Главные компоненты могут формироваться как на основе приближенного, так и точного решения задачи собственных значений. При этом формируются подпространства для описания изображений, которые отличаются друг от друга. Различие между подпространствами можно оценить через расстояние между подпространствами. Пусть v_{0i}^* – собственные векторы, вычисленные методом линейной конденсации или методом синтеза главных компонент, v_{0i}^{pca} – собственные векторы, полученные методом Хаусхолдера. Все собственные векторы имеют единичную длину. Расстояние между подпространствами одинаковой размерности определяется через синус наибольшего главного угла между ними. Наибольший главный угол между подпространствами $L\{v_{0i}^{pca}\}$ и $L\{v_{0i}^*\}$ определяется следующим образом:

$$cos\theta = \max_i \max_j (v_{0i}^{pca})^T v_{0j}^*.$$

Расстояние между подпространствами равно

$$dist\left(v_0^{pca}, v_0^*\right) = \sqrt{1 - cos^2\theta} = sin\,\theta.$$

В методе линейной конденсации точность вычисления собственных векторов регулируется с помощью параметра отсечения. Исследование влияния величины параметра отсечения на расстояние между подпространствами $L\{v_{0i}^{pca}\}$ и $L\{v_{0i}^*\}$ выполняется на наборе из 320 изображений базы ORL. Собственные векторы вычисляются с использования алгоритма блочно-ортогональной конденсации с различными значениями параметра отсечения и методом Хаусхолдера.

В *табл. 3* представлено изменение синуса наибольшего угла между подпространствами, образованными 40 главными компонентами, полученными методом Хаусхолдера и алгоритмом блочно-ортогональной конденсации при различных значениях параметра отсечения.

Таблица 3.

Расстояние между подпространствами главных компонент, вычисленных алгоритмом блочно-ортогональной конденсации и методом Хаусхолдера

Размер блока	Параметр отсечения			
	1,1	1,5	2	2,5
20	0,00237	0,00111	0,00012	0,00024
40	0,00198	0,00070	0,00029	0,00012
60	0,00318	0,00140	0,00042	0,00018

Как видно из *таблицы*, расстояние между подпространствами, образованными точным и приближенным решением задачи собственных значений, сокращается с увеличением параметра отсечения. Однако, и при малых значениях параметра отсечения расстояние между подпространствами небольшое.

Точность вычисления собственных векторов методом синтеза главных компонент зависит от раз-

мера частного набора (блока) изображений (m_k) и количества собственных векторов блоков (l_k). Вместо числа собственных векторов блока l_k будем использовать его относительное значение $\overline{l_k} = l_k / m_k$. Исследования точности выполняются с использованием набора из 320 изображений базы ORL. Методом синтеза главных компонент вычисляются 40 собственных векторов для различных значений m_k и $\overline{l_k}$. В *табл. 4* представлены значения расстояния (синуса наибольшего угла) между подпространствами, которые образованы собственными векторами, полученными методом Хаусхолдера и методом синтеза главных компонент.

Таблица 4.

Расстояние между подпространствами главных компонент, вычисленных методом синтеза главных компонент и методом Хаусхолдера

Размер блока m_k	Относительное число собственных векторов блока $\overline{l_k}$ (%)					
	12,5	25	37,5	50	62,5	75
20	0,0337	0,0232	0,0184	0,0125	0,0097	0,0068
40	0,0345	0,0254	0,0183	0,0131	0,0099	0,0073
80	0,0326	0,0266	0,0206	0,0159	0,012	0,0076

Из *таблицы* видно, что метод синтеза главных компонент позволяет вычислять собственные векторы, которые образуют подпространства, которые отличаются от подпространства, полученного с помощью метода Хаусхолдера, на незначительное расстояние.

Исследование влияния точности вычисления главных компонент на качество распознавания лиц выполняется на базе данных ORL. Собственные векторы вычисляются с использованием метода линейной конденсации, метода синтеза главных компонент и метода Хаусхолдера. При этом изображения, хранящиеся в базе данных, делятся на обучающую и тестовую выборки. Для исследования точности распознавания лиц используется процедура кросс-валидации, усредняющая коэффициенты распознавания, полученные по различным учебным выборкам. Обучающая выборка формируется из 8 изображений каждого класса базы данных ORL, которые выбираются случайно. Все оставшиеся изображения составляют тестовую выборку. Были проведены по 10 экспериментов с главными компонентами, полученными методом Хаусхолдера, методом линейной конденсации и методом синтеза главных компонент. Качество распознавания в каждом эксперименте оценивается по коэффициенту распознавания лиц тестовой выборки. В результате обработки результатов экспериментов получаются средние значения коэффициентов распознавания тестовой выборки, которые представлены в *табл. 5.*

Таблица 5.

Коэффициенты распознавания

Число главных компонент	Линейная конденсация	Синтез главных компонент	Метод Хаусхолдера
30	94,9	94,5	95,1
35	94,9	94,3	94,6
40	95,8	95,8	95,4
45	96,3	96,5	96,4
50	96,3	95,8	96,0
55	95,6	96,0	95,8
60	96,1	96,4	96,4
65	96,1	96,1	96,3
70	96,4	96,6	96,3
75	96,1	96,8	96,4

Из *таблицы* видно, что при использовании метода линейной конденсации коэффициент распознавания достигает величины 96,4, что соответствует точности распознавания для случая, когда главные компоненты вычисляются методом Хаусхолдера. В случае применения метода синтеза главных компонент точность распознавания немного выше. Таким образом, вычисление собственных векторов с использованием методов линейной конденсации и синтеза главных компонент хотя и дает приближенное решение, но не снижает качество распознавания изображений лиц. В ряде экспериментов главные компоненты, вычисленные методами линейной конденсации и синтеза главных компонент, обеспечивают более высокое качество распознавания.

Для демонстрации быстродействия методов линейной конденсации и синтеза главных компонент проводятся эксперименты, в ходе которых вычисляются 290 главных компонент наборов изображений. Затраты на вычисление главных компонент сравниваются со временем вычислений методом Хаусхолдера. Исследования выполняются на изображениях лиц базы данных FERET. В *табл. 6* показывается зависимость относительного времени вычисления 290 главных компонент методом синтеза главных компонент и методом линейной конденсации для наборов с разным числом изображений.

Относительное время рассчитывается как отношение времени вычисления главных компонент методом линейной конденсации или методом синтеза главных компонент ко времени, затраченным методом Хаусхолдера при нахождении собственных векторов. В *таблице* использованы следующие обозначения: *ТК* – время вычисления главных компонент методом линейной конденсации, *ТС* – время вычисления главных компонент методом синтеза главных компонент, *ТХ* – время вычисления главных компонент методом Хаусхолдера. Как видно из таблицы, оба метода превосходят по скорости вычислений метод Хаусхолдера, однако метод синтеза главных компонент демонстрирует более высокое быстродействие.

Таблица 6.

Относительное время вычисления главных компонент

Число изображений	$T_c / T_к$	$T_к / T_х$
1089	0,63	0,92
2059	0,36	0,6
4080	0,13	0,29
5793	0,1	0,26

Заключение

Рассмотрено решение задачи распознавания изображений лиц с помощью линейного дискриминантного анализа и метода главных компонент. Построение классификаторов предлагается выполнять с использованием учебной выборки ненормализованных изображений. Проведено исследование точности распознавания изображений ненормализованных лиц, выделенных из изображений базе данных ORL. Показано, что подход МГК плюс ЛДА дает хорошую точность распознавания даже для изображений лиц не прошедших процедуру нормализации. При этом увеличение числа изображений в классе учебной выборки повышает точность распознавания лиц. Если число изображений в классе невелико, то предлагается расширять учебную выборку изображениями, полученными путем масштабирования и поворота исходных изображений. Для снижения трудоемкости вычисления главных компонент наборов с большим числом изображений предлагается использовать метод линейной конденсации и метод синтеза главных компонент. Показано, что эти методы позволяют существенно снизить трудоемкость расчетов, не снижая точности распознавания лиц. ■

Литература

1. Vijayakumari V. Face recognition techniques: A survey // World Journal of Computer Application and Technology. 2013. No. 1 (2). P. 41–50.
2. Nefian A.V., Hayes M.H. Hidden Markov models for face detection and recognition // IEEE Transcations on Pattern Analysis and Machine Intelligence. 1999. No. 1. P. 141–145.
3. Serrano A., de Diego I.M., Conde C., Cabello E. Recent advances in face biometrics with Gabor wavelets: A review // Pattern Recognition Letters. 2010. No. 31 (5). P. 372–381.
4. Shen L., Bai L. A review on Gabor wavelets for face recognition // Pattern Analysis and Applications. 2006. No. 9 (2–3). P. 273–292.
5. Kirby M., Sirovich L. Application of the KL procedure for the characterization of human faces // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1990. Vol. 12, No. 1. P. 103–108.
6. Etemad K. Chellappa R. Discriminant analysis for recognition of human face images // Journal of the Optical Society of America. 1997. Vol. 14, No. 8. P. 1724–1733.
7. Belhumeur P.N., Hespanha J.P., Kriegman D.J. Eigenfaces vs fisherfaces: Recognition using class specific linear projection // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1997. Vol 19, No. 7. P. 711–720.
8. Мокеев А.В. О точности и быстродействии метода синтеза главных компонент // Бизнес-информатика. 2010. № 3 (13). С. 65–68.
9. Мокеев В.В. О повышение эффективности вычислений главных компонент в задачах анализа изображений // Цифровая обработка сигналов. 2011. №4. С. 29–36
10. Щеголева Н.Л., Кухарев Г.А. Применение алгоритмов двумерного анализа главных компонент для задач распознавания изображений лиц // Бизнес информатика. 2011. № 4 (18). С. 31–38.
11. Мокеев В.В., Томилов С.В. О решении проблемы выборки малого размера при использовании линейного дискриминантного анализа в задачах распознавания лиц // Бизнес информатика. 2013. № 1 (23). С 37–43.

12. Gu X., Gong W., Yang L. Regularized locality preserving discriminant analysis for face recognition // Neurocomputing. 2011. No. 74 (17). P. 3036–3042.
13. Zhang, T., D. Tao, X. Li, et al. Patch alignment for dimensionality reduction // IEEE Transaction on Knowledge and Data Engineering. 2009. No. 21 (9). P. 1299–1313.
14. Yang L., Gong Wj., Gu X., et al. Null space discriminant locality preserving projections for face recognition // Neurocomputing. 2008. No. 71 (16). P. 3644–3649.

ON EFFICIENCY OF FACE RECOGNITION USING LINEAR DISCRIMINANT ANALYSIS AND PRINCIPAL COMPONENT ANALYSIS

Andrey V. MOKEYEV
Senior Lecturer, Department of Information Systems,
Faculty of Economics and Entrepreneurship, South Ural State University
Address: 76, Lenina prospect, Chelyabinsk, 454080, Russian Federation
E-mail: gr.smk@mail.ru

Vladimir V. MOKEYEV
Head of Department of Information Systems,
Faculty of Economics and Entrepreneurship, South Ural State University
Address: 76, Lenina prospect, Chelyabinsk, 454080, Russian Federation
E-mail: mokeyev@mail.ru

The solution of the face recognition problem by means of principal component analysis (PCA) and linear discriminant analysis (LDA) is being considered. The main idea of this approach is that firstly, we project the face image from the original vector space to a face subspace via PCA, secondly, we use LDA to obtain a linear classifier. In the paper, the efficiency of the PCA+LDA approach to face recognition without preliminary processing (scaling, rotation, translating) is investigated. Research shows that the higher the number of images in a class of teach sample, the higher the face recognition rate. When the number of images is small, face recognition performance can be improved by expanding the training set using the images received by scaling and rotating of initial images. The efficiency of PCA+LDA approach is investigated on the images of ORL database. When processing large sets of images, methods of linear condensation and principal component synthesis are suggested to calculate the main components. The principal component synthesis method is based on splitting an initial image set into small sets of images, obtaining eigenvectors of these sets (particular solutions) and calculation of eigenvectors of an initial image based on particular solutions. The linear condensation method is based on the decrease of an order of matrix allowing to calculate pretty exactly eigenvectors whose eigenvalues are located in the preset interval. It is shown that linear condensation and principal component synthesis methods allow to decrease significantly the processing time of building a classifier by PCA+LDA approach, without reducing face recognition rate.

Key words: face recognition, principal component analysis, linear discriminant analysis, linear condensation method, database ORL, principal component synthesis.

Citation: Mokeyev A.V., Mokeyev V.V. (2015) Ob jeffektivnosti raspoznavanii lic s pomoshh'ju linejnogo diskriminantnogo analiza i metoda glavnyh komponent [On efficiency of face recognition using linear discriminant analysis and principal component analysis]. *Business Informatics*, no. 3 (33), pp. 44–54 (in Russian).

References

1. Vijayakumari V. (2013) Face recognition techniques: A survey. *World Journal of Computer Application and Technology*, no. 1 (2), pp. 41–50.
2. Nefian A.V., Hayes M.H. (1999) Hidden Markov models for face detection and recognition. *IEEE Transcations on Pattern Analysis and Machine Intelligence*, no. 1, pp. 141–145.
3. Serrano A., de Diego I.M., Conde C., Cabello E. (2010) Recent advances in face biometrics with Gabor wavelets: A review. *Pattern Recognition Letters*, no. 31 (5), pp. 372–381.
4. Shen L., Bai L. (2006) A review on Gabor wavelets for face recognition. *Pattern Analysis and Applications*, no. 9 (2–3), pp. 273–292.
5. Kirby M., Sirovich L. (1990) Application of the KL procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103–108.
6. Etemad K. Chellappa R. (1997) Discriminant analysis for recognition of human face images // *Journal of the Optical Society of America*, vol. 14, no. 8, pp. 1724–1733.
7. Belhumeur P.N., Hespanha J.P., Kriegman D.J. (1997) Eigenfaces vs fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720.
8. Mokeev A.V. (2010) O tochnosti i bystrodejstvii metoda sinteza glavnyh komponent [On accuracy and performance principal component synthesis method]. *Business Informatics*, no. 3 (13), pp. 65–68 (in Russian).
9. Mokeev V.V. (2011) O povyshenie jeffektivnosti vychislenij glavnyh komponent v zadachah analiza izobrazhenij [On high performance of novel method of principal components computation for image analysis problem]. *Cifrovaja obrabotka signalov*, no. 4, pp. 29–36 (in Russian).
10. Shhegoleva N.L., Kuharev G.A. (2011) Primenenie algoritmov dvumernogo analiza glavnyh komponent dlja zadach raspoznavanija izobrazhenij lic [Algorithms 2DPCA for face recognition]. *Business Informatics*, no. 4 (18), pp. 31–38 (in Russian).
11. Mokeev V.V., Tomilov S.V. (2013) O reshenii problemy vyborki malogo razmera pri ispol'zovanii linejnogo diskriminantnogo analiza v zadachah raspoznavanija lic [On solution of small sample size problem with linear discriminant analysis in face recognition]. *Business Informatics*, no. 1, pp. 37–43 (in Russian).
12. Gu X., Gong W., Yang L. (2011) Regularized locality preserving discriminant analysis for face recognition. *Neurocomputing*, no. 74 (17), pp. 3036–3042.
13. Zhang, T., D. Tao, X. Li, et al (2009) Patch alignment for dimensionality reduction. *IEEE Transaction on Knowledge and Data Engineering*, no. 21 (9), pp. 1299–1313.
14. Yang L., Gong Wj., Gu X., et al (2008) Null space discriminant locality preserving projections for face recognition. *Neurocomputing*, no. 71 (16), pp. 3644–3649.

КРИТЕРИАЛЬНАЯ ОЦЕНКА ВОЗМОЖНОСТИ АВТОМАТИЗАЦИИ БИЗНЕС-ПРОЦЕССОВ ПРЕДПРИЯТИЙ МАЛОГО БИЗНЕСА НА ПЛАТФОРМЕ ПУБЛИЧНОГО ОБЛАКА

М.А. АНИКАНОВА
специалист по облачным решениям, Департамент по работе со средними и малыми организациями и партнерами, Microsoft Россия
Адрес: 121614, г. Москва, ул. Крылатская, д. 17/1
E-mail: v-maanik@microsoft.com

А.Ф. МОРГУНОВ
кандидат технических наук, доцент кафедры корпоративных информационных систем, школа бизнес-информатики, факультет бизнеса и менеджмента, Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20
E-mail: amorgunov@hse.ru

Статья посвящена исследованию возможностей и перспективности автоматизации бизнес-процессов предприятий малого бизнеса с помощью SaaS-приложений, размещенных в публичном облаке. Одним из основных достоинств, присущих облачным решениям, можно считать упрощение ИТ-инфраструктуры при высоком уровне ее масштабируемости и богатой функциональности. Для внедрения облачных аналогов таких «тяжелых» локальных решений, как ERP- или CRM-системы, не требуется больших финансовых инвестиций и временных затрат за счет более простой и гибкой платформы, поддержка которой требует значительно меньших усилий. Это, в свою очередь, дает возможность ИТ-персоналу переориентироваться на более значимые для бизнеса проекты. Одним из наиболее важных достоинств подобного типа решений является тот факт, что большая часть затрат на ИТ переходит из статьи капитальных расходов в операционные, позволяя не изымать значительные финансовые ресурсы из оборота компании.

Несмотря на то, что стоимость SaaS-приложений значительно меньше, чем единовременные затраты на внедрение локальных решений, цена ошибки при принятии решения о построении ИТ-инфраструктуры, в том числе и на базе SaaS продуктов, для малых предприятий по-прежнему велика, так как ее перестроение потребует существенных дополнительных затрат и может оказаться критичным для бюджета организации. Поэтому в статье рассматривается набор критериев, позволяющих на этапе планирования ИТ-инфраструктуры малого предприятия определить целесообразность использования технологических возможностей приложений, размещенных в публичном облаке. Все разработанные критерии разделены на три основных группы: функциональные, финансово-экономические и технические. Все они подробно описаны и ранжированы по своей важности с помощью метода экспертной оценки признанных специалистов в сфере облачных технологий и ИТ в целом. С использованием полученных количественных значений критериев выведена формула, в соответствии с которой можно получить показатель, оценивающий целесообразность перевода конкретного бизнес-процесса компании малого бизнеса на облачную платформу.

Материалы статьи могут быть интересны как специалистам, занимающимся внедрением информационных систем, так и руководителям предприятий малого бизнеса, для оптимизации расходов на ИТ.

Ключевые слова: SaaS приложения, публичное облако, критерий, малое предприятие, бизнес-процесс.

Цитирование: Аниканова М.А., Моргунов А.Ф. Критериальная оценка возможности автоматизации бизнес-процессов предприятий малого бизнеса на платформе публичного облака // Бизнес-информатика. 2015. № 3 (33). С.55–64.

Введение

Тема облачных решений как инновационного вида оптимизации бизнес-процессов обрела популярность только в последние годы, несмотря на то, что концепция облака как такового уже достаточно давно используется многими компаниями для решения прикладных задач, иногда даже без понимания самого термина, не так давно введенного в обиход. Различные модели предоставления облачных услуг, виды организации доступа к ним и тому подобные детали часто вызывают непонимание у лиц, принимающих решения. Поэтому, несмотря на достаточно очевидные преимущества облачного подхода для бизнеса, они принимают решения в пользу привычной локальной ИТ-инфраструктуры и, таким образом, экспертиза ИТ и групп разработки остаются привязанными к текущей среде. Однако в тех компаниях, которые все-таки решили воспользоваться этими преимуществами, возникают вопросы о том, какие именно бизнес-процессы целесообразно выводить за рамки организации, используя облачные ресурсы, и по каким критериям выбирать эти бизнес-процессы, чтобы достичь максимальной эффективности их автоматизации и сократить затраты на ИТ. Особенно актуальны эти вопросы для предприятий малого бизнеса, так как цена ошибки здесь высока.

Цель данной статьи состоит в том, чтобы сформулировать критерии выбора бизнес-процессов малого бизнеса для перевода на платформу публичного облака в виде SAAS-приложений.

Исследуя возможности облачных решений, мы будем в основном говорить о публичных облаках. Несмотря на то, что, по данным IDC, российские компании тратят на частное облако больше, чем на публичное, публичному облаку как концепции массива ресурсов, отличного от привычной инфраструктуры, прогнозируют высокий уровень роста и все большее распространение в России [1]. С точки зрения возможностей использования облачных решений для автоматизации прикладных решений бизнеса, постановка данного ограничения будет оправдана, поскольку для компаний среднего и особенно малого бизнеса использование публичного облака намного более перспективно и выгодно – в основном из-за того, что развертывание частного облака требует больших инвестиций в ИТ за счет необходимости построения собственного центра обработки данных, обеспечения его программной платформой, постоянной поддержкой функционирования системы частного облака и других затрат. К тому же публичное облако по сравнению с частным обладает преимуществом повышенной масштабируемости и функциональной гибкости построенной ИТ-инфраструктуры.

1. Использование SaaS-решений на платформе публичных облаков компаниями малого бизнеса

Несмотря на распространенность и привычность термина «малый бизнес», понимание принципов разделения бизнеса на группы в соответствии с размером компаний всегда было темой споров и неопределенности. В связи с отсутствием общепринятой мировой терминологии, различные эксперты трактуют понятия, связанные с масштабами бизнеса, по-разному, что часто вносит некоторую неясность в этом вопросе. Для того, чтобы система координат была наиболее объективна, трактовку понятия «малое предпринимательство» было решено взять из законодательства РФ. Понятие и характеристика компаний малого бизнеса в России прописаны в Федеральном Законе №209-ФЗ от 24.07.2007 и определяются как часть совокупности малых и средних предприятий [2]. Основным критерием разделения бизнеса по размеру является количество работающих сотрудников за отчетный период. Так, если средняя численность работников за календарный год больше 15 (компания с меньшим количеством работников называется микропредприятием), но не превышает 100 человек, то компания может быть отнесена к малому предпринимательству.

Также критерием считается годовая выручка компании: с 1 января 2013 г. согласно Постановлению Правительства РФ от 9 февраля 2013 г. №101 «О предельных значениях выручки от реализации товаров (работ, услуг) для каждой категории субъектов малого и среднего предпринимательства» для малых предприятий определено ограничение от 60 до 400 млн. рублей [3].

На текущий момент общепринятой или хотя бы известной сегментации среднего и малого бизнеса (СМБ) и, в частности, малого бизнеса по принципу принятия ИТ-решений не существует. Однако определение профиля потенциального потребителя ИТ-продуктов и услуг для вендоров становится все более приоритетным, особенно в сфере СМБ, на который смещается фокус. Недавние исследования рынка показали, что основной мотив покупки

программного обеспечения (ПО) для компаний малого и среднего бизнеса в России – это появление новых задач, для решения которых нужно новое ПО (61%), что связано с нацеленностью бизнеса на рост и расширение (44%). Чаще всего ПО воспринимается малыми предприятиями как конкурентное преимущество, так как ключевой проблемой бизнеса большинство компаний назвало именно конкуренцию (14%). Малый и средний бизнес ценит возможность подстроить ПО «под себя» (61%), причем в основном они готовы платить за решение своих проблем (62%) [4].

Если говорить в общем, то более трех четвертей компаний (77%) имеют те или иные проблемы, решаемые с помощью облачных технологий. Это показывает, насколько высок потенциал у облачных решений в данном сегменте бизнеса [4].

Что касается текущего состояния ИТ-инфраструктуры в компаниях малого бизнеса и использования облачных решений, то можно рассмотреть ситуацию на примере отношения компаний малого бизнеса в России к использованию серверов (*рис. 1*) [5].

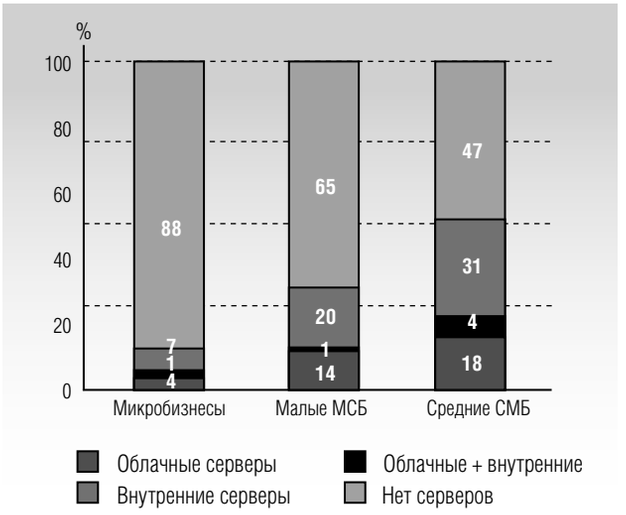


Рис. 1. Доля пользователей облачных серверов среди российских компаний малого и среднего бизнеса (2012 г.) [5]

Из диаграммы видно, что чем меньше размер компании, тем больше вероятность, что серверы не используются в принципе. Однако рассматривая специфику малого бизнеса, заметим, что из тех компаний, которые используют серверы, уже сейчас большая часть использует облачные решения. Использование гибридной схемы для малых предприятий не несет больших экономических преимуществ и скорее связано с обеспечением требований безопасности, а также с попыткой избежать «революции» в подходе к построению ИТ-инфраструктуры. Однако если смотреть в целом, то 21% всех малых компаний на данный момент так или иначе использует облачные сервисы для решения бизнес-задач.

По данным аналитического агентства Parallels, в 2012 году объем рынка облачных решений в сегменте малого и среднего бизнеса в России составил 15,6 млрд. руб. [5]. По данным того же агентства, рынок будет расти в среднем примерно на 34%, достигая 37,7 млрд. рублей к 2015 году [5]. Такой рост во многом будет осуществляться не только за счет увеличения спроса на уже существующий ассортимент облачных услуг, но и за счет расширения этого ассортимента. Это объясняется тем, что облачные технологии находятся на стадии активного развития и пока еще не покрывают все потребности прикладных задач бизнеса.

В сфере публичных облаков изначально существует три основных модели (или уровней) обслуживания [6].

Первый уровень, – платформа как услуга (*Platform-as-a-Service, PaaS*), – представляет собой модель обслуживания в сфере облаков, в которой пользователю предоставляется возможность использования облачной инфраструктуры для размещения собственного любого базового программного обеспечения, позволяющего впоследствии размещать на нем любые существующие приложения или разрабатывать новые.

Второй уровень, – инфраструктура как услуга (*Infrastructure-as-a-Service IaaS*), – представляет собой модель обслуживания, когда облачная инфраструктура используется для самостоятельного управления различными предоставляемыми вычислительными ресурсами, например, для обработки или хранения данных, или установки и запуска любого программного обеспечения, начиная от операционных систем и заканчивая прикладным и платформенным программным обеспечением.

Наконец, третий уровень, – программное обеспечение как услуга (*Software-as-a-Service, SaaS*), – это единственная модель обслуживания, которая предоставляется непосредственно конечному пользователю, чем принципиально отличается от вышеописанных двух моделей.

В дальнейшем мы будем говорить об использовании SaaS-решений на платформе публичных облаков.

Одним из основных достоинств, присущих облачным решениям, можно назвать упрощение ИТ-инфраструктуры и оптимизация управления ею. Во-первых, для внедрения облачных аналогов таких тяжелых локальных решений, как ERP- или CRM-системы, не требуется больших усилий ИТ-персонала и временных затрат за счет более простой и автоматизированной платформы. Во-вторых, функций, необходимых для эффективной работы ИТ-инфраструктуры, но требующих настройки, в облачных решениях становится все меньше, поэтому от ИТ-специалистов компании все чаще требуются лишь поддержка текущего состояния ИТ и элементарные начальные настройки.

Существенным пунктом, относящимся к затратам, является цена подписки на облачные SaaS продукты. Несмотря на то, что цена такой услуги очевидно меньше, чем стоимость локальной установки, компании часто пугает необходимость постоянно платить за облачные услуги и, соответственно, исходящая из этого привязка к провайдеру. Многие компании опасаются задержек в выделении бюджета на ИТ, из-за которых при использовании облачных решений может остановиться работа на предприятии. Кроме того, при переходе на использование облачных продуктов решение поменять облачную ИТ-инфраструктуру обратно на локальную, скорее всего, станет достаточно болезненным и затратным шагом. По этим причинам многие компании стараются избежать необходимости периодически оплачивать ИТ-услуги, пытаясь таким способом обеспечить непрерывность бизнеса.

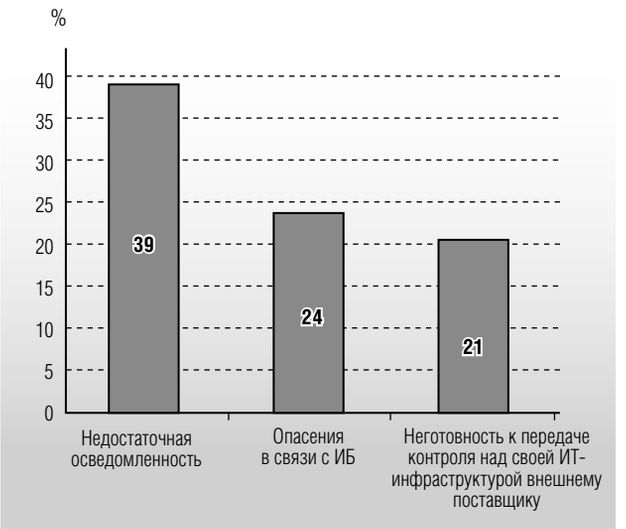


Рис. 2. Результаты опроса российских компаний в части барьеров для распространения облачных вычислений (2010 г.) [7]

Среди факторов, препятствующих переходу к облачным решениям, респонденты выделили следующие (*рис. 2*).

Как видно из рисунка, основным барьером для использования облачных решений является недостаток знаний о технологиях – тот фактор, устранить который, с одной стороны, сложно, а с другой – легко, повышая осведомленность компаний с помощью маркетинга.

2. Критерии выбора бизнес-процессов для перевода в SaaS-приложения на базе публичных облаков

Оценивая собственные бизнес-процессы и принимая решение по переводу их на платформу публичного облака, компания должна обратить внимание на совокупность факторов, так как ориентация лишь на один из них может привести к негативному результату.

В ходе анализа литературы, интервьюирования специалистов в сфере облачных технологий, а также обобщения накопленного опыта по переводу бизнес-процессов малого бизнеса на платформу публичного облака были выделены характеристики бизнес-процессов, которые соотносятся с платформенными характеристиками облачных ИТ-решений и, соответственно, делают эти бизнес-процессы потенциальными для перевода в SaaS-приложения.

Все критерии были разделены на три основных группы: функциональные, финансово-экономические и технические. Рассмотрим подробнее каждую из трех групп.

2.1. Функциональные критерии

2.1.1. Необходимость мультипользовательского доступа к ресурсам бизнес-процесса

Во многих компаниях, в том числе малых, большинство бизнес-процессов затрагивают не одного работника, а нескольких, формируя потребность организовать возможность многопользовательского доступа к ресурсам. Примером такого процесса является любая цепочка операций в сфере документооборота, когда один документ проходит процедуры согласования, визирования и т.п., вовлекая в процесс многих сотрудников организации. Данный критерий оценивается с точки зрения требований к бизнес-процессу в части количества пользователей,

которым необходим доступ к ресурсам. Сложность реализации мультипользовательского доступа возрастает с количеством задействованных пользователей, поэтому чем больше это количество, тем удобнее использовать облачные технологии для решения данной задачи.

2.1.2. Необходимость обеспечения мобильного доступа и доступа с различных видов устройств к ресурсам бизнес-процесса

Динамика рынка малого бизнеса требует от предпринимателей постоянного доступа к корпоративным ресурсам. Сейчас это требование актуально для большинства бизнес-процессов, однако остаются определенные виды, к которым с точки зрения безопасности мобильный доступ будет запрещен внутрикорпоративными политиками или требованиями законодательства (например, доступ к персональным данным определенной категории). Данный критерий может оцениваться по бинарной системе, в виде «мобильный доступ нужен / не нужен».

2.1.3.Тип бизнес-процесса и его важность для бизнеса

В любой компании управляющие и поддерживающие процессы (т.е. процессы, не оказывающие непосредственное влияние на основное производство) автоматизируются легче: в них в наименьшей степени проявляется специфика бизнеса и поэтому глубокая кастомизация программных решений не требуется. Если же говорить об особенностях малого бизнеса, то «неспецифичность» второстепенных процессов выражена здесь особенно ярко: в компаниях с количеством сотрудников от 15 до 100 человек процессы управления кадрами, безопасностью или документооборотом представлены в наиболее типовом виде. Так как облачные решения по модели обслуживания SaaS имеют ограниченные возможности кастомизации и в принципе не предназначены для глубокой настройки, то такие второстепенные процессы имеют наибольший потенциал для перевода в облачный режим. Критичность для бизнеса или важность бизнес-процесса также необходимо учитывать при принятии решения о его переводе на облачную платформу. Если бизнес-процесс имеет высокую важность, то он в большей степени требует оптимизации, обеспечения максимальной надежности, непрерывности и безопасности. При этом важно, чтобы подобный бизнес-процесс стабильно и гарантированно поддерживался информационными системами. Таким образом, данная характеристика может быть переформулирована как критичность непрерывности функционирования бизнес-процесса.

2.1.4. Стандартизированность и повторяемость бизнес-процесса

Данный критерий определяет, насколько бизнес-процесс может быть назван типовым внутри организации и вне ее. В большинстве компаний малого сегмента бизнес-процессы не являются уникальными, особенно в части второстепенных процессов. Оценка данного критерия влияет на степень кастомизации конечного ИТ-решения: в какой степени он должен быть адаптирован для конкретной организации и, соответственно, быть уникальным. С другой стороны, если процесс внутри компании каждый раз протекает немного по-разному, то за счет большого количества сценариев и сложной логики его автоматизация будет осуществляться намного сложнее (для данного критерия необходим метод экспертной оценки).

2.1.5. Обеспеченность безопасности бизнес-процесса

Данный критерий является скорее ограничивающим использование облачных решений, чем открывающим новые возможности их применения. Он связан именно с инфраструктурными требованиями, которые подразделяются на три основных группы: законодательные требования, требования регуляторов и внутрикорпоративные требования. Каждая из этих групп требований касается различных видов обеспечения безопасности корпоративных данных, включая условия хранения, условия обработки и передачи персональных данных, требования к сертификациям и соответствию другим требованиям.

2.2. Финансово-экономические критерии

2.2.1. Возможность сокращения итоговых затрат на бизнес-процесс при использовании платформы публичного облака

Поскольку чаще всего основной целью автоматизации бизнес-процесса является уменьшение общей стоимости его обеспечения, его высокая стоимость бизнес-процесса является первым при-

знаком необходимости его оптимизации и автоматизации. Использование облачных SaaS-решений как инструмента такой автоматизации открывает широкий круг возможностей сокращения затрат. Например, становится возможным сократить время работы вовлеченных сотрудников за счет оптимизации процесса, что сокращает затраты на оплату труда. Используя методы финансовой оценки бизнеса, можно попытаться сравнить затраты на бизнес-процесс в текущей ситуации и при внедрении облачного решения, с учетом стоимости самого проекта внедрения. Если имеет место определенная экономия, то такой перевод будет целесообразным.

2.2.2. Условия выделения ИТ-бюджета: регулярность, прогнозируемость

Возможность оплаты сервиса по мере его использования важна для малого бизнеса. Эта возможность переводит затраты на ИТ из капитальных затрат в операционные, что позволяет более гибко управлять небольшим оборотом компании и вкладывать деньги в более приоритетные направления, достигая поставленных бизнес-целей. В то же время компании необходимо выделять определенную (пусть и небольшую сумму) на ИТ регулярно, на постоянной основе.

2.3. Технические критерии

2.3.1. Уровень и динамика нагрузки бизнес-процесса

Данный технический критерий тесно связан с функциональным критерием необходимости обеспечения многопользовательского доступа к ресурсам бизнес-процесса: именно количество пользователей, задействованных в бизнес-процессе, а также интенсивность их работы формируют уровень нагрузки бизнес-процесса. Для понимания динамики нагрузки аналитики компании или внешние аудиторы при описании бизнес-процесса «как есть» строят графики зависимости используемых вычислительных ресурсов от времени. Такие графики могут иметь различный вид, однако с точки зрения использования для автоматизации бизнес-процессов облачных решений есть несколько типов динамики нагрузки, наиболее подходящих для перевода на облачную платформу.

Включение и выключение. Данный тип характеризуется резким включением и выключением рабочих нагрузок на бизнес-процесс (*рис. 3*). Примером такой динамики могут служить пакетные задачи, во время которых простаивают избыточные ресурсы. В случае подобных интервальных нагрузок компания терпит издержки на поддержание ненужных вычислительных ресурсов, особенно критичные для малого бизнеса.

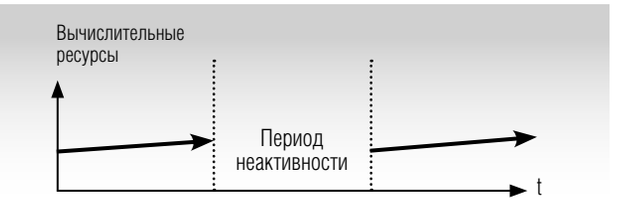


Рис. 3. Динамика нагрузки на бизнес-процесс вида «включение и выключение»

Высокий темп роста. Вторым типом динамики загрузки является ее быстрый рост (*рис. 4*). Высокий темп роста характерен для многих бизнес-процессов и чаще всего связан с ростом бизнеса в целом и расширением круга бизнес-потребностей. Однако имеют место ситуации, когда спрос на бизнес-процесс растет внутри организации, например, за счет увеличения количества сотрудников. Поддержка масштабируемости и, в частности, быстрого роста – в принципе, достаточно сложная задача для ИТ во многом из-за невозможности предоставления оборудования и других ресурсов в короткие сроки. Для малого бизнеса такая задача может стать практически невыполнимой, так как кроме расширения ИТ-обеспечения бизнес-процессов она может потребовать усложнения всей ИТ-инфраструктуры.

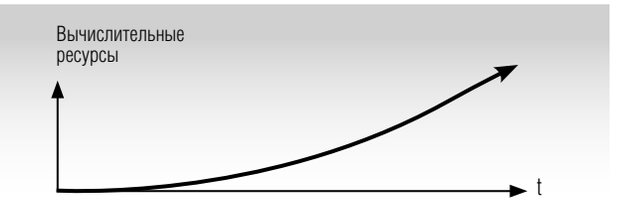


Рис. 4. Динамика нагрузки на бизнес-процесс вида «высокая скорость роста»

Непредсказуемый всплеск. Третьим типом динамики является незапланированная нагрузка на бизнес-процесс (*рис. 5*). Внезапное увеличение пользовательского спроса на один процесс влияет на производительность всей инфраструктуры на пике нагрузки. В этом случае малому бизнесу чаще всего невозможно обеспечить дополнительные вычислительные ресурсы, так как задача «запастись» инфраструктурой заранее, особенно в условиях ограниченного бюджета на ИТ, трудновыполнима.

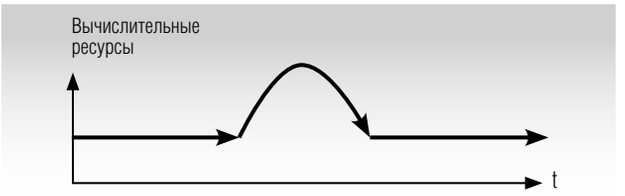


Рис. 5. Динамика нагрузки на бизнес-процесс вида «непредсказуемый всплеск»

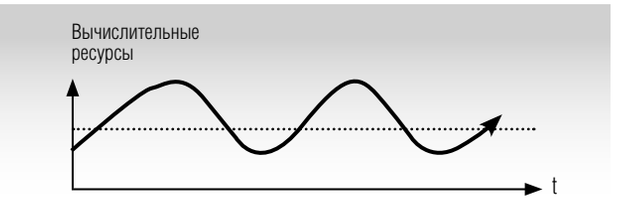


Рис. 6. Динамика нагрузки на бизнес-процесс вида «циклические пики нагрузки»

Циклические пики нагрузки. Предсказуемый всплеск как тип динамики нагрузки является характерным для служб с микросезонными трендами (*рис. 6*). Подобная пиковая нагрузка связана с периодическим увеличением спроса на бизнес-процесс, но, несмотря на предсказуемость такой динамики, ее обеспечение требует значительного усложнения ИТ-инфраструктуры, ограниченного небольшими ИТ-бюджетами малого бизнеса и человеческими ресурсами. Важно, что в моменты пиковой нагрузки менее эффективно обеспечиваются другие бизнес-процессы, часто также критичные для бизнеса, таким образом, снижается уровень общей производительности инфраструктуры для поддержания бизнеса. В случае же обеспечения дополнительных мощностей для поддержания микросезонных пиков образуется лишняя инфраструктура, которая периодически простаивает.

2.3.2. Технологическое обеспечение бизнес-процессов

Бывают случаи, когда при автоматизации бизнес-процессы требуют специализированных решений или аппаратного обеспечения, а также некоторых необходимых для работы технических параметров, таких как время отклика системы на запрос, скорость работы, доступ к сторонним ресурсам или возможность интеграции с информационными системами в ИТ-инфраструктуре компании. Все эти параметры достигаются за счет построения технической инфраструктуры, отвечающей всем требованиям и обеспечивающей работу ИТ-решений. Иногда подобные требования накладывают определенные ограничения на использование SaaS-решений на базе публичного облака, так как многие технические характеристики в нем не имеют возможности внешней корректировки. Таким образом, при наличии подобных ограничений выбор может быть сделан в пользу локальных аналогов SaaS-решения.

3. Количественная оценка критериев

Для получения весовых характеристик каждого из критериев была проведена их экспертная оценка. В ее рамках каждый эксперт по 10-балльной шкале оценивал важность каждого из разработанных критериев и его влияние на выбор бизнес-процессов для перевода на платформу публичного облака. Результаты ранжирования критериев приведены в *табл. 1.*

Таблица 1.

Ранжирование критериев по результатам экспертной оценки

Критерий	Коэффициенты значимости	Значения критериев
Возможность сокращения итоговых затрат на бизнес-процесс при использовании платформы публичного облака (x1)	0,14	0 или 1
Тип бизнес-процесса и его важность для бизнеса (x2)	0,13	от 0 до 1
Обеспеченность безопасности бизнес-процесса (x3)	0,12	0 или 1
Уровень и динамика нагрузки на бизнес-процесс (x4)	0,11	от 0 до 1
Необходимость мультипользовательского доступа к ресурсам бизнес-процесса (x5)	0,11	0 или 1
Необходимость обеспечения мобильного доступа и доступа с различных видов устройств к ресурсам бизнес-процесса (x6)	0,11	0 или 1
Технологическая обеспечение бизнес-процесса (x7)	0,0974	0 или 1
Условия выделения ИТ-бюджета: регулярность и прогнозируемость (x8)	0,0921	0 или 1
Стандартизированность и повторяемость бизнес-процесса (x9)	0,0905	от 0 до 1

Для определения возможности использования полученных результатов и степени согласованности экспертов были рассчитаны коэффициенты конкордации и критерий согласованности Пирсона. Коэффициент конкордации составил 0,074, что говорит о слабой степени согласованности экспертов, однако критерий Пирсона показал, что коэффициент конкордации – случайная величина, поэтому полученные результаты имеют смысл и могут использоваться в исследовании. С использованием количественной оценки разработанных критериев можно получить формулу, в результате расчета которой можно получить число *y* (от 0 до 1), показывающее целесообразность перевода бизнес-процесса на облачную платформу:

$$y=\begin{cases} 0{,}14\cdot x_1+0{,}13\cdot x_2+0{,}12\cdot x_3+0{,}11\cdot x_4+0{,}11\cdot x_5+ \\ +0{,}11\cdot x_6+0.0974\cdot x_7+0.0921\cdot x_8+0.0905\cdot x_9; \\ \quad при\ x_3\neq 0 и\ x_7\neq 0 \\ \quad 0,\ при\ x_3=0\ или\ x_7=0 \end{cases} \quad (1)$$

Для принятия решения необходимо количественно оценить каждый из разработанных критериев и, воспользовавшись формулой (1), получить значение *y*. Если это значение лежит в интервале от 0 до 0,3, то бизнес-процесс лучше не автоматизировать с помощью облачных SaaS-решений. Если значение *y* находится в пределах от 0,3 до 0,65, то имеет смысл рассмотреть гибридное решение, объединяющее облачные сервисы и локальные продукты. Если же *y* находится в пределах от 0,65 до 1, то представляется целесообразным полностью перевести бизнес-процесс на облачную платформу.

Заключение

В данной статье предложена система критериев выбора бизнес-процессов для перевода на платформу публичного облака в виде SaaS-приложений, которая может помочь компаниям малого бизнеса в принятии решений в области построения ИТ-инфраструктуры. Описана взаимосвязь характеристик бизнес-процессов и типовых платформенных характеристик, построенных на технологиях публичного облака ИТ-решений, позволяющая дать ответ на вопрос о выборе бизнес-процессов для перевода на SaaS-платформу. ■

Литература

1. Прохоров А. Russia cloud services market 2014–2018 forecast and 2013 analysis // IDC Russia [Электронный ресурс]: http://idcrussia.com/ru/research/published-reports/55531-russia-cloud-services-market-2014-2018-forecast-and-2013-analysis/2-abstract (дата обращения 27.08.2015).
2. Федеральный закон от 24 июля 2007 г. №209-ФЗ «О развитии малого и среднего предпринимательства в Российской Федерации» (с изменениями и дополнениями) // Система ГАРАНТ [Электронный ресурс]: http://base.garant.ru/12154854/ (дата обращения 20.12.2014).
3. Постановление Правительства РФ от 9 февраля 2013 г. №101 «О предельных значениях выручки от реализации товаров (работ, услуг) для каждой категории субъектов малого и среднего предпринимательства» // Система ГАРАНТ [Электронный ресурс]: http://www.garant.ru/hotlaw/federal/456590/ (дата обращения 20.12.2014).
4. Исследование практики принятия IT-решений в малых и средних российских компаниях (по заказу ООО «Майрософт Россия»). М.: ИнФОМ, 2014.
5. Лаврентьева Т. Определены самые востребованные облачные услуги в России // CNews [Электронный ресурс]: http://www.cnews.ru/top/2012/12/17/opredeleny_samye_vostrebovannye_oblachnye_uslugi_v_rossii_512922 (дата обращения 20.12.2014).
6. Маланин В. Баланс ресурсов и процессов // Intelligent enterprise (Корпоративные системы). 2013. №2 (248). С. 38–40.
7. Булусов А. ИТ-директора боятся «облаков» // CNews аналитика [Электронный ресурс]: http://www.cnews.ru/reviews/free/infrastructure2009/articles/survey.shtml (дата обращения 20.12.2014).

CRITERIAL EVALUATION OF THE POSSIBILITY OF SMALL BUSINESSES BUSINESS PROCESS AUTOMATION ON PUBLIC CLOUD PLATFORM

Maria A. ANIKANOVA
Cloud Solutions Specialist, Small and Medium Solutions and Partners Unit, Microsoft Russia
Address: 17/1, Krylatskaya Street, Moscow, 121614, Russian Federation
E-mail: v-maanik@microsoft.com

Alexander F. MORGUNOV
Associate Professor, Department of Corporate Information Systems,
School of Business Informatics, Faculty of Business and Management,
National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow, 101000, Russian Federation
E-mail: amorgunov@hse.ru

The article is dedicated to the research of the possibility and viability of small companies business processes automation using public cloud SaaS applications. One of the fundamental advantages of cloud solutions is IT infrastructure simplification along with high-level scalability and rich functionality. Cloud counterparts of such «heavy» on-premise software as ERP or CRM systems do not require large financial investments and time expenditures, due to a more simple and agile platform, which support requires less effort, giving IT specialists an opportunity to reorient at more important projects. One of the most significant advantages of such solutions is the fact that the major part of IT expenses can be converted from capital to operational costs, which gives small business companies the possibility not to withdraw a big amount of money from corporate cash flow.

The cost of SaaS applications is much lower than one-time expenditure on on-premise products implementation. However, the cost of error for small organizations on the stage of decision-making concerning IT infrastructure construction and management (including SaaS-based architecture) is still high, for any further IT infrastructure changes will require significant additional costs and can turn out to be critical for the company's budget. That is why the set of criteria, which allows companies to define the expediency of public cloud applications technological possibilities usage on the stage of small business IT-infrastructure planning, is observed in the article. All developed criteria are divided into three main groups: functional, financial & economic, and technical; they are described in details separately, and then ranged according to their importance using expert evaluation method, involving recognized IT experts. The formula was developed using quantitative estimations, which helped to derive the specific index, evaluating reasonability of concrete business processes automation with the help of public cloud SaaS applications.

The article materials will prove to be of interest to information systems integration specialists and small business decision makers in order to estimate optimization of IT costs.

Key words: SaaS applications, public cloud, criterion, small business, business process.

Citation: Anikanova M.A., Morgunov A.F. (2015) Kriterial'naja ocenka vozmozhnosti avtomatizacii biznes-processov predprijatij malogo biznesa na platforme publichnogo oblaka [Criterial evaluation of the possibility of small businesses business process automation on public cloud platform]. *Business Informatics*, no. 3 (33), pp. 55–64 (in Russian).

References

1. Prokhorov A. (2014) Russia cloud services market 2014–2018 forecast and 2013 analysis. *IDC Russia*. Available at: http://idcrussia.com/ru/research/published-reports/55531-russia-cloud-services-market-2014-2018-forecast-and-2013-analysis/2-abstract (accessed 27 August 2015).
2. Federal Law of Russian Federation, 24 July 2007, No. 209-FZ «O razvitii malogo i srednego predprinimatel'stva v Rossijskoj Federacii» [On development of small and medium entrepreneurship in Russian Federation]. *GARANT System*. Available at: http://base.garant.ru/12154854/ (accessed 20 December 2014) (in Russian).

3. Resolution of the Government of Russian Federation, 9 February 2013, No. 101 «O predel'nyh znachenijah vyruchki ot realizacii tovarov (rabot, uslug) dlja kazhdoj kategorii sub'ektov malogo i srednego predprinimatel'stva» [On maximum values of revenue from sales of goods (works, services) for each of the category of small and medium business]. *GARANT System*. Available at: http://www.garant.ru/hotlaw/federal/456590/ (accessed 20 December 2014) (in Russian).
4. InFOM, Microsoft Russia (2014) *Issledovanie praktiki prinjatija IT-reshenij v malyh i srednih rossijskih kompanijah* [Research of IT decision making practices in Russian small and medium companies]. Moscow: InFOM (in Russian).
5. Lavrentyeva T. (2012) Opredeleny samye vostrebovannye oblachnye uslugi v Rossii [The most essential in Russia cloud services are identified]. *CNews*. Available at: http://www.cnews.ru/top/2012/12/17/opredeleny_samye_vostrebovannye_oblachnye_uslugi_v_rossii_512922 (accessed 20 December 2014) (in Russian).
6. Malanin V. (2013) Balans resursov i processov [The balance between resources and processes]. *Intelligent enterprise (Korporativnye sistemy)*, no. 2 (248), pp. 38–40 (in Russian).
7. Bulusov A. (2009) IT-direktora bojatsja «oblakov» [IT directors are afraid of clouds]. *CNews Analytics*. [Available at:]: http://www.cnews.ru/reviews/free/infrastructure2009/articles/survey.shtml (accessed 20 December 2014) (in Russian).

УПРАВЛЕНИЕ СТОИМОСТЬЮ ПОСТАВОК ЗАПАСНЫХ ЧАСТЕЙ ДЛЯ ПОСЛЕПРОДАЖНОГО ОБСЛУЖИВАНИЯ СЛОЖНЫХ ТЕХНИЧЕСКИХ ИЗДЕЛИЙ

С.М. ЯМПОЛЬСКИЙ
кандидат технических наук,
доцент кафедры бизнес-аналитики, факультет бизнеса и менеджмента,
Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20
E-mail: syampolsky@hse.ru

А.С. ШАЛАМОВ
доктор технических наук, научный сотрудник отдела статистических проблем информатики и управления, Институт проблем информатики Российской академии наук
Адрес: 119333, г. Москва, ул. Вавилова, д. 44
E-mail: a-shal5@yandex.ru

А.П. КИРСАНОВ
доктор технических наук, профессор кафедры бизнес-аналитики, школа бизнес-информатики, факультет бизнеса и менеджмента, Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20
E-mail: ki@hse.ru

Е.В. ОГУРЕЧНИКОВ
старший преподаватель кафедры бизнес-аналитики, факультет бизнеса и менеджмента, Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20
E-mail: eogurechnikov@hse.ru

В статье рассматриваются вопросы управления стоимостью жизненного цикла технических изделий в части, касающейся организации и осуществления мероприятий по поставкам запасных частей в рамках их послепродажного обслуживания.

Рассмотрен вариант модели сети Петри, которая описывает причинно-следственные связи между событиями, связанными с планированием и управлением поставками на основе использования вероятностной аналитической модели послепродажного обслуживания технических изделий и программного комплекса анализа рисков по технико-экономическим критериям. Результатом работы таких моделей является планирование приемлемого баланса между стоимостью и качеством изделий и его текущее обеспечение, в том числе путем учета и минимизации финансовых рисков.

Приведен пример автоматизированного планирования поставок запасных частей. Динамика изменения количества технических изделий, находящихся в эксплуатации, представляется в интегрированном графическом виде, дающем возможность прогнозировать коэффициент исправности изделий, обусловленный наличием исправных изделий на складе заказчика и производительностью ремонтных органов.

Обосновано применение метода освоенного объема для анализа рисков отклонения от плана выполнения поставок запасных частей. Отслеживание освоенного объема финансовых средств позволило прогнозировать как успешность завершения поставок запасных частей, так и риски отклонения от намеченных сроков и бюджета.

Рассмотрен пример автоматизированного анализа рисков. Оценка степени соответствия затрат бюджетным характеристикам осуществляется с помощью показателя эффективности, который используется для анализа качества функционирования соответствующих подразделений заказчика и корректировки их дальнейшей работы. Для выбранного года показатель эффективности определяется и оптимизируется при заданном коэффициенте исправности для каждого заказчика в процессе автоматизированного планирования поставок запасных частей.

Предлагаемый подход является достаточно универсальным, что предопределяет возможность его применения для решения задач управления стоимостью жизненного цикла продукции и услуг в различных организационно-технико-экономических системах.

Ключевые слова: послепродажное обслуживание, сеть Петри, автоматизированное планирование, поставка запасных частей, метод освоенного объема, технико-экономические критерии эффективности, стоимость поставок, анализ рисков, бюджет по завершении проекта.

Цитирование: Ямпольский С.М., Шаламов А.С., Кирсанов А.П., Огуречников Е.В. Управление стоимостью поставок запасных частей для послепродажного обслуживания сложных технических изделий // Бизнес-информатика. 2015. № 3 (33). С. 65–73.

Введение

Важным аспектом качества проекта по созданию сложных технических изделий является степень отражения в документации проекта требований, предъявляемых к изделию [5]. Требования достижения баланса между значениями показателей качества и стоимости изделия оказывают существенное влияние на управление конфигурацией как самого изделия, как и компонентов его послепродажного обслуживания (ППО). Учет интересов производителя и заказчика изделий в части ППО – это предмет компромисса, который должен достигаться в процессе подготовки соответствующего договора и быть зафиксированным в нем.

В соответствии с современными требованиями рынка главными проблемами проектирования изделий являются:

✦ разработка эффективных алгоритмов управления конфигурацией, обеспечивающих достижение требуемого баланса стоимости и интегральных характеристик качества изделия;

✦ разработка модели стоимости жизненного цикла (СЖЦ) изделия, которая должна быть чувствительной к конкретным управляющим решениям и определять итог всех затрат.

Управление СЖЦ технических изделий предполагает:

1. При разработке и проектировании – выбор такой конструкции, а также системы технического обслуживания и ремонта, при которых стоимость жизненного цикла изделий будет минимальна.

2. При ведении тендерной работы поставщиком – планирование СЖЦ и предоставление данных заказчику (например, данные расчета на первые 2 года и прогноз на 10–20 лет).

3. На этапе эксплуатации парка изделий:

✧ оптимальное управление бюджетом, включающее планирование на заданный период и адекватное распределение бюджета между задачами ППО;

✧ мониторинг данных (в режиме реального времени) о фактически выполненных работах, использованных ресурсах и произведенных затратах;

✧ качественный и количественный анализ рисков отклонения от утвержденного бюджета;

✧ воздействие на факторы, вызывающие отклонения от запланированной стоимости работ ППО и обеспечивающие их завершение в рамках утвержденного бюджета.

Место системы ППО в общей структуре интегрированной информационной среды показано на *рис. 1.*

Особое место в процессе ППО занимают вопросы поставки запасных частей (ЗЧ) и современные технологии автоматизированного управления ими, что обеспечивает поддержание требуемого уровня конкурентоспособности эксплуатируемых изделий.

Для специалистов организаций, осуществляющих ППО технических изделий, помимо моделей и инструментальных информационных средств оптимального планирования процессов на основе использования вероятностной аналитической модели [2], большой практический интерес представляют модели автоматизированного монито-

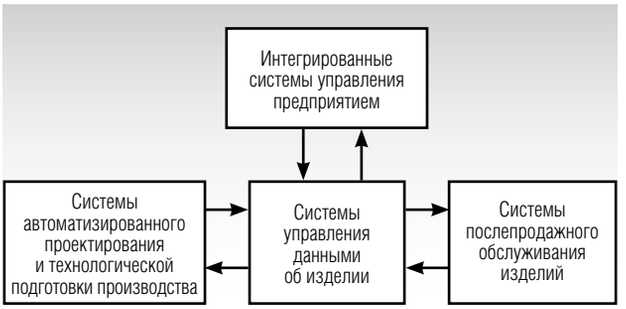


Рис. 1. Место систем ППО в общей структуре интегрированной информационной среды

ринга совокупности данных, на основе которых осуществляется управление мероприятиями по осуществлению поставок ЗЧ. Результатом работы таких моделей является планирование приемлемого баланса между стоимостью и качеством изделий и его текущее обеспечение, в том числе, путем учета и минимизация финансовых рисков. Необходимость управления рисками диктуется не столько соответствующими стандартами, сколько потребностями самого бизнеса, не желающего терять значительные финансовые средства по незначительным поводам.

В настоящей работе рассматривается подход по управлению СЖЦ на этапе эксплуатации, включающий использование модели сети Петри для анализа стоимости поставок ЗЧ. Модель сети Петри описывает причинно-следственные связи между событиями, связанными с планированием и управлением поставками, на основе использования вероятностной аналитической модели ППО технических изделий и применения разработанного программного комплекса анализа рисков отклонения от плана поставок по технико-экономическим критериям.

1. Модель сети Петри для анализа стоимости поставок запасных частей

Представим модель для анализа стоимости поставок ЗЧ в виде сети Петри с одноцветными фишками (*рис. 2*). Причинно-следственные связи, реализуемые в рамках этой модели, обусловливают выполнение стандарта требований по управлению СЖЦ технических изделий в части, касающейся контроля стоимости поставок при проведении послепродажного обслуживания.

Позиции сети интерпретируются следующим образом:

- P_1 – анализ текущих поставок ЗЧ;
- P_2 – определение вероятности возникновения риска;
- P_3 – оценка степени влияния риска;
- P_4 – оценка величины риска;
- P_5 – построение дерева решений;
- P_6 – анализ результатов;
- P_7 – планирование поставок запасных частей;
- P_8 – разработка плана реагирования на риски;
- P_9 – проведение мониторинга рисков.

Переходы сети интерпретируются следующим образом:

- t_1 – запрос на проведение стоимостного анализа поставок ЗЧ;
- t_2 – запрос на проведение качественной оценки рисков;
- t_3 – запрос на обработку данных;
- t_4 – запрос на проведение количественной оценки рисков;
- t_5 – запрос на проведение анализа результатов;
- t_6 – запрос на проведение планирования поставок ЗЧ;
- t_7 – запрос на проведение мониторинга рисков.

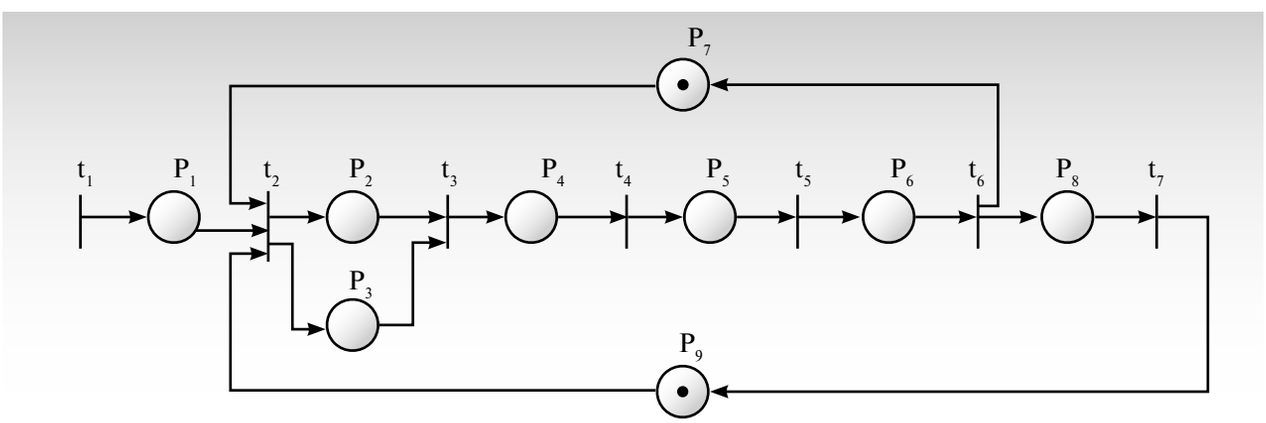


Рис. 2. Модель сети Петри для анализа стоимости поставок запасных частей

Работа данной модели предусматривает предварительное формирование оптимального по стоимости плана поставок на заданном периоде эксплуатации изделий, а также сбор исходной информации по идентификации рисков аналогичных проектов (начальная разметка сети Петри).

В ходе работы модели осуществляется расчет стоимости работ по поставкам и сравнение полученных результатов с плановыми показателями. Результат сравнения позволяет контролировать эффективность выполнения плана поставки и вероятность риска отклонения от него.

Риски отклонения от плана поставки ЗЧ вызывают действия, предпринимаемые в связи с результатами их качественной и количественной оценки, объединенные причинно-следственными связями.

Качественный анализ предполагает обнаружение рисков, исследование их особенностей, выявление последствий их реализации в форме экономического ущерба и раскрытие источников информации относительно каждого вида риска.

Критерии качественного анализа могут учитывать затраты, официальные и предписанные требования, социально-экономические аспекты и факторы внешней среды, интересы заказчика, приоритеты и иные исходные данные для оценки.

Результат процесса качественного анализа – это оценка величины риска на основе вероятности возникновения и степени влияния на результаты поставки ЗЧ, а также градация рисков по возможным последствиям.

Количественный анализ выполняется для рисков, которые были выявлены и квалифицированы в результате качественного анализа. Наиболее распространенным методом количественного анализа рисков является анализ дерева решений, которое описывает рассматриваемую ситуацию с учетом каждой из имеющихся возможностей выбора и возможного сценария реализации риска.

По итогам проведения качественного и количественного анализа рисков вырабатывается четкое представление о стратегиях, используемых для реагирования на каждый вид риска.

Мониторинг выполнения плана поставки ЗЧ позволяет прогнозировать как успешность завершения плана, так и степень влияния рисков, связанных с отклонением от намеченных сроков и объемов поставки.

2. Вероятностная аналитическая модель послепродажного обслуживания технических изделий

При автоматизированном планировании поставок запасных частей в рамках послепродажного обслуживания технических изделий целесообразно использование вероятностной аналитической модели [2], которая позволяет прогнозировать, в частности, динамику изменений количества ЗЧ на складе заказчика и обеспечивать планирование бюджета для поддержания заданного уровня исправности эксплуатируемых изделий. При этом планируются и осуществляются мероприятия по созданию необходимых запасов ЗЧ и их возобновлению в соответствии с перечнем поставки и принятой политикой пополнения.

В этих целях создается и используется информационно-аналитический программный комплекс (ИАК) прогнозирования и оптимизации плановых показателей, позволяющий, в том числе, определить оптимальные необходимые объемы поставок ЗЧ и их периодичность.

ИАК обеспечивает определение минимального по стоимости плана поставок на заданном периоде эксплуатации технических изделий [3].

Исходные данные модели системы ППО, на основе которой функционирует ИАК, включают:

- ✦ параметры надежности технических изделий;
- ✦ параметры всех видов обслуживания и ремонта технических изделий;
- ✦ сроки службы изделий;
- ✦ интенсивность использования запасных частей;
- ✦ значимость запасных частей;
- ✦ сроки поставки ЗЧ.

Результатом автоматизированного планирования стоимости поставок, выполненного с помощью ИАК, является определение величины базовых плановых затрат (БПЗ) на осуществление мероприятий по поставкам ЗЧ для каждого заказчика – технического центра (ТЦ) (*табл. 1*), при заданных возможностях ремонта технических изделий и при обеспечении заданного коэффициента исправности (*рис. 3*).

Коэффициент исправности (КИ) характеризует вероятность того, что при использовании в установленных условиях изделие окажется работоспособным в произвольно выбранный момент времени в установившемся процессе эксплуатации [3].

Таблица 1.

Значения БПЗ мероприятий по обеспечению поставок ЗЧ (по годам)

Год		2015	2016	2017	2018
Заказчик	КИ				
ТЦ №1	0,92	5 165 105,8	0,0	7 184 314,6	7 564 019,1
ТЦ №2	0,94	6 772 436,4	7 531 571,2	0,0	7 824 760,4
ТЦ №3	0,96	6 374 436,6	0,0	7 975 710,9	7 511 443,3
Итог:		**18 311 978,8**	**7 531 571,2**	**15 160 025,5**	**22 900 222,8**

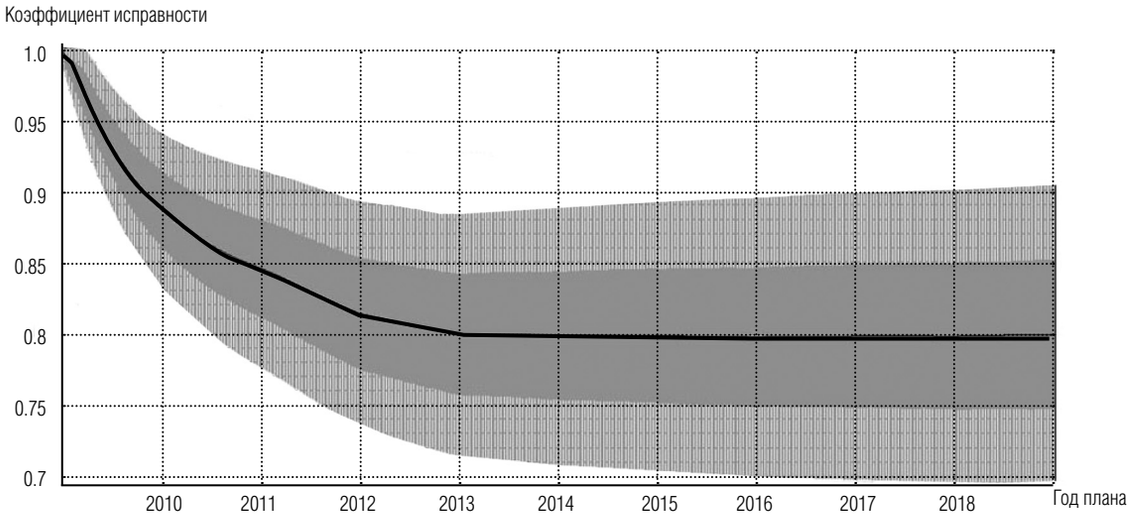


Рис. 3. Динамика изменения коэффициента исправности с доверительным интервалом

На *рис. 3* показана сглаженная кривая прогнозируемой динамики статистических оценок значений коэффициента исправности, соответствующая оптимальной по стоимости программе поставок, в среднем удовлетворяющей требуемому уровню исправности КИ = 0,8 (отмечен штриховой линией).

Доверительный интервал оценки КИ принимается равным $m \pm 2\sigma$, где m и $\sigma(t)$ – математическое ожидание и среднекадратическое отклонение значения оценки коэффициента исправности [2].

3. Программа для анализа рисков отклонения от плана поставок

Для анализа рисков отставания от плана поставок ЗЧ предлагается использовать метод освоенного объема. Данный метод позволяет эффективно организовать информационное сопровождение процесса поставок, а также измерение и контроль эффективности его выполнения. Постоянное отслеживание освоенного объема финансовых средств позволяет прогнозировать как успешность завершения поставок ЗЧ, так и риски отклонения от намеченных сроков, бюджета и т.д. [1].

Реализацию метода освоенного объема для анализа рисков отставания от плана целесообразно осуществлять с помощью информационной системы 1С «Предприятие».

В настоящее время существует разработанная на платформе 1С информационно-аналитическая система для сопровождения технических изделий на послепродажных стадиях жизненного цикла (*www.appius.ru*), которая решает, в частности, следующие задачи:

- ✦ создание паспорта изделия;
- ✦ регистрация передачи нового изделия заказчику;
- ✦ регистрация поступления изделий на доработку и в различные виды ремонта;
- ✦ формирование отчетов по анализу отказов и рекламаций изделий.

Кроме того, в конфигурацию информационной системы 1С «Предприятие» может быть включена

дополнительная программа (программный комплекс), которая будет функционировать совместно с существующей информационно-аналитической системой, и работа которой позволит анализировать ход выполнения контракта по поставкам ЗЧ конкретному заказчику с оценкой степени критичности полученных результатов.

В работе [4] представлена программа, позволяющая проводить сравнение ежемесячной плановой стоимости запланированных работ и ежемесячной плановой стоимости выполненных работ по поставкам ЗЧ. На основании этих значений была выполнена стоимостная оценка процесса поставки ЗЧ за заданный период времени и получены такие показатели, как величина отклонения от календарного плана поставок и индекс отклонения от календарного плана.

В качестве источника вероятности возникновения рисков было использовано значение коэффициента исправности, полученного в результате моделирования случайных процессов эксплуатации изделий, осуществляемого при автоматизированном планировании поставок ЗЧ, а индекс отклонения от календарного плана, в свою очередь, являлся характеристикой степени влияния рисков на результаты поставки.

Результатом дальнейших исследований явилась разработка программного комплекса (ПК), работа которого основана на модели сети Петри (*рис. 2*). ПК позволяет получать информацию о выполненных заказах *(табл. 2)* с использованием системы запросов к соответствующим документам информационной системы 1С «Предприятие».

Конечным результатом выполнения расчетов с помощью ПК является стоимостная оценка комплекса мероприятий по обеспечению поставки ЗЧ.

Оценке подвергается степень соответствия произведенных затрат бюджетным характеристикам, вычисляемая в виде показателя эффективности (ПЭВ) по формуле:

$$ПЭВ = \frac{БПЗ - БСВР}{БПЗ - ФСВР}, \text{ где:}$$

БПЗ – величина базовой стоимости запланированного комплекса мероприятий за рассматриваемый период времени (по плановым ценам);

БСВР – величина базовой стоимости выполненных мероприятий за тот же период времени (по плановым ценам);

ФСВР – величина фактической стоимости выполненных мероприятий по существующим ценам.

Полученные значения данного показателя анализируются следующим образом:

✦ если $ПЭВ < 1$, то имеет место экономия финансовых средств на проведение работ по поставкам ЗЧ;

✦ если $ПЭВ = 1$, то ход работ по поставкам ЗЧ соответствует плану;

✦ если $ПЭВ > 1$, то имеет место перерасход финансовых средств на проведение работ по поставкам ЗЧ.

Следует отметить, что *ПЭВ* характеризует не только эффективность планирования работ, но и косвенным образом отражает качество прогнозирования цен на рынке услуг. Данный показатель может быть использован для анализа качества функционирования соответствующих подразделений заказчика и корректировки их дальнейшей работы.

В качестве периода времени для проведения стоимостного анализа поставок запасных частей выбирается один год. Для выбранного года показатель *БПЗ* (показатель стоимости) определяется и оптимизируется при заданном коэффициенте исправности (интегрированном показателе эксплуатационно-технического качества) для каждого заказчика в процессе автоматизированного планирования поставок *ЗЧ*, выполненного с помощью *ИАК* (*табл. 1*).

Таблица 2.

Результат стоимостного анализа эффективности мероприятийпо организации поставок запасных частей

Заказчик	КИ	БПЗ	БСВР	ФСВР	ПЭВ
ТЦ №1	0,92	5 165 105,8	2 155 435,2	2 909 112,6	1,33
ТЦ №2	0,94	6 772 436,4	2 265 165,4	2 070 572,6	0,96
ТЦ №3	0,96	6 374 436,6	2 465 145,6	2 558 902,3	1,02
Итог:		**18 311 978,8**	**9 895 416,8**	**11 138 587,5**	

Заказчиками запасных частей являются технические центры (*ТЦ)*, функциональным назначением которых является техническое обслуживание и ремонт сложных технических изделий. Показатель эффективности характеризует уровень риска, связанный с отклонением от стоимости поставок ЗЧ. Очевидно, что наиболее конкурентоспособными являются заказчики, чьи риски оказываются наименьшими.

Заключение

В предложенной работе в рамках решения проблемы управления СЖЦ сложных технических изделий представлен подход к автоматизированному управлению стоимостью поставок ЗЧ для послепродажного обслуживания сложных технических изделий. Данный подход включает:

✦ оптимальное управление бюджетом, в частности, планирование его показателей на заданном периоде эксплуатации;

✦ мониторинг стоимости ППО, который производится путем непрерывного сбора информации о фактически выполненных работах, использованных ресурсах и произведенных затратах на послепродажных стадиях жизненного цикла технических изделий, а также последующем воздействии на факторы, вызывающие отклонения от запланированной стоимости работ ППО и обеспечивающие их завершение в рамках утвержденного бюджета.

В рамках реализации этого подхода предложено использование:

✧ типового ИАК для определения оптимальной величины базовых плановых затрат мероприятий по поставкам ЗЧ для каждого заказчика при заданном коэффициенте исправности на заданном промежутке времени;

✧ модели сети Петри для анализа стоимости поставок в режиме реального времени, описывающей причинно-следственные связи между событиями, учитываемыми в работе вероятностной аналитической модели ППО технических изделий;

✧ метода освоенного объема для анализа рисков из-за отставания от плана мероприятий по поставкам ЗЧ;

✧ программного комплекса для автоматизированного анализа указанных рисков из-за отставания от плана выполнения поставок запасных частей;

✧ показателя (коэффициента) эффективности, характеризующего выполнение плана на заданном промежутке времени на основе данных, вычисленных с использованием вышеприведенных методов, моделей, алгоритмов и программных комплексов.

Реализацию данного подхода предлагается осуществлять на основе платформы 1С, для включения в имеющуюся конфигурацию информационной системы 1С и совместного функционирования с существующей информационно-аналитической системой – для сопровождения технических изделий в режиме о*n-line* на послепродажных стадиях жизненного цикла.

Дальнейшее развитие представленного подхода заключается в построении и разработке алгоритмов по корректировке оптимального плана поставок, с целью снижения угроз рисков для достижения целей данного проекта.

Предлагаемый подход является достаточно универсальным, что предопределяет возможность его применения для решения задач управления стоимостью жизненного цикла продукции и услуг в различных организационно-технико-экономических системах. ■

Литература

1. Масловский В.П. Управление проектами. Красноярск: ИПК, СВУ, 2008. 179 с.
2. Синицын И.Н., Шаламов А.С. Лекции по теории систем интегрированной логистической поддержки. М.: Торус Пресс, 2011. 615 с.
3. Шаламов А.С. Интегрированная логистическая поддержка. М.: Университетская книга, 2008. 463 с.
4. Ямпольский С.М., Шаламов А.С. Автоматизированное управление поставками запасных частей на основе технологий функционального и математического моделирования процессов // Логистика и управление цепями поставок. 2014. №6 (65). С. 34–40.
5. A Guide to the Project Management Body of Knowledge (PMBOK Guide). Fifth edition. Newtown Square, PA: Project Management Institute, 2013. 589 p.

COST MANAGEMENT FOR THE SUPPLY OF SPARE PARTS FOR AFTER-SALES SERVICE OF COMPLEX TECHNICAL PRODUCTS

Sergey M. YAMPOLSKY
Associate Professor,
Department of Business Analytics, School of Business Informatics,
Faculty of Business and Management,
National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow, 101000, Russian Federation
E-mail: syampolsky@hse.ru

Anatoly S. SHALAMOV
Researcher, Department of Statistical Problems of Informatics and Management,
Institute of Informatics Problems, Russian Academy of Sciences
Address: 44, Vavilova Street, Moscow, 119333, Russian Federation
E-mail: a-shal5@yandex.ru

Alexander P. KIRSANOV
Professor, Department of Business Analytics,
School of Business Informatics, Faculty of Business and Management,
National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow, 101000, Russian Federation
E-mail: ki@hse.ru

Eugene V. OGURECHNIKOV
Senior Lecturer, Department of Business Analytics,
School of Business Informatics, Faculty of Business and Management,
National Research University Higher School of Economic
Address: 20, Myasnitskaya Street, Moscow, 101000, Russian Federation
E-mail: eogurechnikov@hse.ru

The article considers the issues of technical product life cycle management in the field of spare parts delivery organization and management within the framework of after-sales service.

It provides an examination of a Petri net model, describing the cause-effect relations between events that are linked to delivery planning and management, based on a probabilistic analytical model for after-sales service of technical products and a program-based risk analysis system based on technical and economic criteria. The result of a given model's performance is planning of an acceptable balance between the cost and quality of products and their current maintenance, which includes detection and minimization of financial risks.

An example that illustrates automated planning of spare parts delivery is given. Dynamics of operated technical products' quantity variation is represented in the integrated graphic type, providing an opportunity to predict an average factor of technical product's serviceability, determined both by a number of serviceable technical products in a warehouse of the customer and productivity of repair agencies.

The earned value method application is proved to be an effective tool for risk analysis of schedule variance in the field of spare parts delivery. Monitoring of the earned value of finances permits to forecast not only the probability of successful completion of spare parts delivery, but also the risks of both cost and schedule variance.

An example of automated risk analysis is provided. Estimated coincidence degree of actual cost and planned value is calculated by means of the effectiveness index, which is used to analyze the quality of customer's subdivisions performance and to correct further functioning. For a selected year, the effectiveness index can be defined and optimized for the predetermined serviceability factor, assigned for every customer during the process of automated planning of spare parts delivery.

The approach presented in the article can be considered quite universal, which predetermines an opportunity to apply it in order to provide solutions for product and service life cycle management problems in various organizational technical and economic systems.

Key words: after-sales service, Petri net model, automated planning, spare parts delivery, earned value method, technical and economic criteria of efficiency, delivery cost, automated risk analysis, budget at completion.

Citation: Yampolsky S.M., Shalamov A.S., Kirsanov A.P. , Ogurechnikov E.V. (2015) Upravlenie stoimost'ju postavok zapasnyh chastei dlya posleprodazhnogo obsluzhivaniya slozhnyh tehnicheskih izdelii [Cost management for the supply of spare parts for after-sales service of complex technical products]. *Business Informatics*, no. 3 (33), pp.65–73 (in Russian).

References

1. Maslovsky V. (2008) *Upravlenie proektami* [Project management]. Krasnoyarsk: IPK, SVU (in Russian).
2. Sinicin I., Shalamov A. (2011) *Lekcii po teorii sistem integrirovannoi logisticheckoi poddergki* [Lectures on the theory of integrated logistics support systems]. Moscow: Torus Press (in Russian).
3. Shalamov A. (2008) *Integrirovannaya logisticheckaya poddergka* [Integrated logistic support]. Moscow: University book (in Russian).
4. Yampolsky S., Shalamov A. (2014) Avtomatizirovannoe upravlenie postavkami zapasnyh chastej na osnove tehnologij funkcional'nogo i matematicheskogo modelirovanija processov [Automated management of spare parts deliveries based on functional and mathematical process modelling technologies]. *Logistics and supply chain management*, no. 6 (65), pp. 34–40 (in Russian).
5. Project Management Institute (2013) *A Guide to the Project Management Body of Knowledge (PMBOK Guide)*. Fifth edition, Newtown Square, PA: Project Management Institute.

ДАТА-ЦЕНТРЫ И АКТИВЫ ПРЕДПРИЯТИЯ

Р.Р. СУХОВ
Финансовый управляющий, АНО «Институт «Аптайм»
Адрес: 125315, г. Москва, Большой Коптевский проезд, д. 6
E-mail: r.sukhov@uptimetechnology.ru

М.Б. АМЗАРАКОВ
Директор, АНО «Институт «Аптайм»
Адрес: 125315, г. Москва, Большой Коптевский проезд, д. 6
E-mail: m.amzarakov@uptimetechnology.ru

Е.А. ИСАЕВ
кандидат технических наук, профессор,
заведующий кафедрой управления информационными системами
и цифровой инфраструктурой, школа бизнес-информатики,
факультет бизнеса и менеджмента,
Национальный исследовательский университет «Высшая школа экономики»;
заведующий лабораторией, Физический институт имени П.Н. Лебедева,
Российская академия наук (РАН)
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20
E-mail: eisaev@hse.ru

С.В. МАЛЬЦЕВА
доктор технических наук, профессор, заведующая кафедрой инноваций и бизнеса
в сфере информационных технологий, школа бизнес-информатики,
факультет бизнеса и менеджмента,
Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20
E-mail: smaltseva@hse.ru

В статье рассматриваются взаимосвязь и взаимовлияние центра обработки данных (ЦОД, дата-центр) и активов предприятия. Цель статьи – дать представление о том, как дата-центр может оказывать влияние на активы компании и их конечную стоимость. Отражены аспекты, важные для понимания причин интереса компаний к правильному формированию объекта инвестирования и последующего учета этих инвестиций как существенной части активов предприятия. Приведено обоснование того, что в составе некоторых предприятий дата-центр сам по себе является важным активом, а в некоторых бизнес-моделях представляет собой ключевой актив компании.

Отталкиваясь от определений терминов «активы» и «дата-центр» в статье рассмотрены варианты участия дата-центра в бизнесе предприятии и его влияния на конечную стоимость компании через стоимость ее активов.

Рассмотрены примеры того, как дата-центр, по сути, становится предметом производства на крупных предприятиях, бизнес которых построен на хранении, обработке информации и предоставлении услуг по доступу к этой информации. Приведены примеры такого рода предприятий из различных секторов экономики.

Затронуты вопросы государственного регулирования, касающиеся требований по созданию дата-центров, необходимых для выполнения регулирующих функций государства. Рассмотрены вопросы корпоративной безопасности предприятия и влияние дата-центра на сохранность информации. Также уделено внимание вопросу опосредованного влияния дата-центра на стоимость активов предприятия, через повышение надежности обработки данных, повышение уровня безопасности хранимых и обрабатываемых данных и, как следствие, влияние на рыночную стоимость предприятия как бизнеса посредством повышения доверия потребителей.

Ключевые слова: дата-центр, активы предприятия, рыночная стоимость акций, риск, бизнес-процесс.

Цитирование: Сухов Р.Р., Амзараков М.Б., Исаев Е.А., Мальцева С.В. Дата-центры и активы предприятия // Бизнес-информатика. 2015. № 3 (33). С.74–79.

Введение

Активы (от лат. «activus» – «действенный») представляют собой совокупность имущества и денежных средств, принадлежащих предприятию, фирме или компании (здания, сооружения, машины и оборудование, материальные запасы, банковские вклады, ценные бумаги, патенты, авторские права и другая собственность, имеющая денежную оценку). В широком смысле, это любые ценности, обладающие денежной стоимостью. Активы принято делить на материальные (осязаемые) и нематериальные (неосязаемые); к последним относятся интеллектуальные продукты, патенты, долговые обязательства других предприятий а также особые права на использование ресурсов [1].

Дата-центр – это инженерно-технический комплекс, оборудованный необходимым оборудованием для безопасной, длительной и непрерывной эксплуатации ИТ-систем и телекоммуникационного оборудования с надлежащими характеристиками энергоснабжения, климатических параметров и безопасности [2]. В общем случае в состав дата-центра также входят серверные (телекоммуникационные) комнаты, либо отдельные серверные стойки в небольших компаниях.

Обобщенно можно выделить следующие факторы, мотивирующие предприятия использовать дата-центры в своей основной деятельности:

✦ выполнение норм законодательства и/или технических норм;

✦ необходимость обеспечения конфиденциальности информации;

✦ защита от несанкционированных действий третьих лиц;

✦ минимизация финансовых потерь от простоев бизнеса;

✦ устранение перебоев в выполнении деловых операций;

✦ требования к специальным условиям эксплуатации со стороны производителей ИТ-оборудования;

✦ сложные математические вычисления, направленные на аналитические исследования больших объемов данных, математическое моделирование сложных процессов, а также осуществление финансовых и биржевых операций в реальном масштабе времени;

✦ защита от конкурентов и внешних недоброжелателей, включая предотвращение мошенничества и краж, вредительства со стороны персонала, а также защиту от хакерских атак и промышленного шпионажа.

С точки зрения структуры активов предприятия дата-центр может играть различные роли:

1. Дата-центр является частью структуры ИТ-активов предприятия, которые, в свою очередь, направлены на обеспечение работы пользователей и бизнес-процессов предприятия. Дата-центр, как правило, является вспомогательным подразделением предприятия;

2. Дата-центр является частью структуры активов предприятия, относящихся к средствам производства. Предприятие строит свой бизнес с использованием дата-центров, однако дата-центр – это ресурс, а не источник дохода;

3. Дата-центр является ключевым активом и основным средством производства предприятия. Предоставление в пользование ресурсов дата-центра является базовой бизнес-моделью компании – владельца такого дата-центра, а сам дата-центр является основным источником дохода.

Для определенных категорий бизнеса наличие собственного дата-центра, а также их минимальное количество регулируется на законодательном уровне. Подавляющее большинство крупных и средних компаний имеют собственные или арендованные дата-центры (под арендованным дата-центром понимается часть большого дата-центра, оказывающего коммерческие услуги многим клиентам).

В зависимости от функции дата-центра на предприятии и, соответственно, его места в структуре активов требования к дата-центру отличаются. Рассмотрим более подробно каждый из функциональных блоков.

1. Дата-центр как часть структуры ИТ-активов предприятия

Наличие дата-центра в составе ИТ-активов характерно для компаний, у которых основная деятельность связана, например, с производством, торговлей, банковскими или страховыми услугами. Компания использует дата-центр для размещения телекоммуникационного оборудования и/или ИТ-систем, необходимых для обработки, хранения, передачи создаваемых и получаемых в процессе жизнедеятельности предприятия данных. Таким образом, в данном случае дата-центр является объектом, поддерживающим основную деятельность предприятия.

Наличие дата-центра позволяет компании улучшить управляемость бизнеса, поддерживать необходимый уровень деловой активности, обеспечивать конфиденциальность корпоративных и личных данных, обеспечивать исполнение технологических и корпоративных процессов, обеспечить выполнение требований регуляторов и государственных органов.

Появление дата-центра как выделенной производственной единицы организационной структуры предприятия сигнализирует о достаточной зрелости компании, поскольку необходимость централизации ИТ-ресурсов и, как следствие, создание специализированного сопровождения этих ресурсов в составе дата-центра является логичным процессом развития ИТ-культуры компании по мере ее роста. Грамотно реализованный дата-центр позволяет обеспечить необходимый уровень доступности как внутренних, так и внешних ИТ-сервисов компании, что положительно сказывается на ее способности выполнять основные бизнес-функции. Дата-центр в такой модели может быть как собственным (полностью принадлежащим предприятию на основании права собственности), так и арендованным, т.е. принадлежащим предприятию на основании права пользования.

Пример: Коммерческий банк. Для успешной работы банка, особенно если он работает в розничном сегменте, наличие дата-центра является обязательным, поскольку иначе оказывать банковские услуги населению будет невозможно. Сложное специализированное программное обеспечение требует высокопроизводительного серверного оборудования, что, в свою очередь, требует качественного энергоснабжения и охлаждения. Высокие требования по сохранности и защищенности данных накладывают свои требования на физические и информационные средства защиты информации. Реализовать весь комплекс мероприятий, не имея дата-центра, практически невозможно.

Кроме того, в соответствии с регулирующими документами Банка России [3] требуется обеспечивать непрерывность деятельности и (или) восстановление деятельности в случае возникновения нестандартных и чрезвычайных ситуаций. По сути, требуется иметь еще один дата-центр, расположенный в месте, отличном, от местоположения основного дата-центра.

На основе данного примера можно сделать вывод, что наличие, как минимум, двух дата-центров является обязательным требованием для банковской деятельности.

Также можно сказать, что применение дата-центра в составе ИТ-активов предприятия позволяет снижать риски владения активами [3], снижать стоимость владения активами, снижать стоимость предоставляемых клиентам услуг, выполнять предъявляемые к компании требования [4], обеспечивать конкурентоспособность предприятия. Все перечисленные факторы положительно влияют на повышение стоимости оценки активов компании, что в конечном итоге повышает ее капитализацию.

Вывод: Для многих направлений бизнеса дата-центр являются очень значимым элементом, поскольку его качество в значительной степени влияет на успешность бизнеса в целом. Дата-центр, с одной стороны, является важным элементом в составе активов бизнеса, а с другой стороны – активно влияет на стоимость других активов компании и общей стоимости бизнеса в целом.

2. Дата-центр как часть средств производства предприятия

Рассмотрим ситуацию, когда дата-центр является инфраструктурным объектом, на базе которого у компании – его владельца построен собственный бизнес. В подавляющем большинстве случаев бизнес таких компаний строится на предоставлении потребителям услуг более высокого уровня, чем физические ресурсы дата-центра. К таким высокоуровневым услугам можно отнести предоставление в пользование виртуальных ИТ-ресурсов или вычислительной мощности по запросу, а также услуги, построенные на Интернет-запросах и Интернет-посещениях (поисковые системы, социальные сети и т.п.).

В данном случае дата-центр является важным, зачастую ключевым элементом основной деятельности предприятия. В такого рода бизнес-моделях дата-центр с равной вероятностью может быть как собственным, так и взятым в аренду. В большой степени на выбор способа владения дата-центром влияет масштаб бизнеса компании: чем крупнее компания, тем более вероятным является принятие решения о собственном дата-центре.

Пример 1: Компания, оказывающая услуги облачной бухгалтерии. В своей бизнес-модели такая компания нуждается в ресурсах дата-центра для того, чтобы обеспечить доступность своего сервиса и иметь возможность хранить и обрабатывать данные безопасно и надежно. Для того, чтобы реализовывать основной сервис, компании достаточно воспользоваться услугами дата-центра на коммерческой основе.

В данном примере дата-центр, как важный элемент оказания услуги, очень сильно влияет как на надежность работы компании в целом, так и на оценку стоимости компании, в частности. Причинами этого являются:

✧ выбор поставщика услуги дата-центра. Использование надежного поставщика услуги в аутсорсинговой модели позволяет иметь высокий уровень доступности услуги, что повышает доверие клиентов к конечной услуге. Это, в свою очередь, позволяет продавать услуги по более высокой цене;

✧ влияние на оценку бизнеса. При проведении оценки бизнеса в случаях раундов вторичного финансирования, листинга на биржах и т.п., качество поставщиков компании оказывает важную роль на конечную оценку компании в целом. Т.е. надежный и известный поставщик услуги дата-центра повышает доверие инвесторов и, как следствие этого, стоимость активов в частности и компании в целом.

Пример 2: Компания – крупный поисковый сервис. В данной бизнес-модели дата-центр необходим для хранения информации об индексах и другой служебной информации. В данном примере у компании есть собственный дата-центр и, скорее всего, не один. Необходимость иметь собственный дата-центр продиктована очень глубокой взаимосвязью ИТ, телекоммуникационных сервисов и инженерных систем дата-центра в бизнес-процессах компании.

Дата-центр в данном примере является очень существенным фактором, влияющим на эффективность работы компании. Более того, бизнес такой компании очень сильно зависит от качества работы дата-центра. Это объясняется следующими соображениями:

✦ дата-центр, как инженерно-технический объект, не приносит прибыли владельцам бизнеса, однако расходы на содержание дата-центра ложатся на себестоимость услуг, которые оказывает компания своим клиентам. Поэтому компания заинтересована снижать издержки эксплуатации дата-центра, обеспечивая при этом необходимый уровень доступности ИТ, поскольку иначе произойдет снижение выручки компании, что является недопустимым;

✦ способность дата-центра соответствовать требованиям бизнеса (например, масштабируемость, энергоэффективность, энерговооруженность) является залогом успешного развития компании, снимая, либо, наоборот, создавая ограничения для бизнеса.

✦ Способность компании создать оптимальную конфигурацию дата-центра создает важные предпосылки для достижения минимальных эксплуатационных издержек дата-центров в компании в течение всех лет эксплуатации дата-центра.

Вывод: В совокупности вышеперечисленные факторы влияют на конечную оценку стоимости активов компании, поскольку дата-центр влияет на себестоимость и, как следствие, на маржинальность оказываемых компанией услуг. Между тем, необходимо отметить, что дата-центр в таких компаниях все равно не является ключевым активом предприятия, хотя и очень важным. Этим объясняется то, что многие крупные Интернет-компании начинали развитие бизнеса, пользуясь арендованными дата-центрами.

3. Дата-центр как ключевой актив предприятия

Отдельно необходимо сказать о предприятиях, для которых дата-центр является ключевым активом предприятия. В мире вообще, и в России, в частности, активно развивается целая отрасль – коммерческие услуги дата-центра. В данном случае дата-центр, как правило, является собственностью предприятия, либо арендуется полностью как инженерный комплекс, включающий здание и прилегающую территорию.

Для компаний, которые работают в данном сегменте бизнеса, дата-центр является ключевым ак-

тивом, который позволяет быть участником рынка коммерческих услуг дата-центров. По своей роли в таких компаниях дата-центр является источником дохода.

Основной услугой, которая предлагается клиенту, является услуга аренды части ресурсов дата-центра. Обобщенно данная услуга называется «аренда стойко-места», «аренда стойки», «colocation» и т.п. В действительности смысловое значение гораздо шире, так как в данное понятие входит предоставление в пользование сразу нескольких ресурсов компании: электроэнергии, кондиционирования воздуха, резервированного электроснабжения и т.д.

Компании данного сегмента инвестируют средства в строительство дата-центров, спроектированных специально с учетом возможности оказания услуг множеству клиентов. Крупные компании обладают дата-центрами в больших количествах. На мировом рынке присутствуют операторы, владеющие десятками дата-центров по всему миру.

Вывод: Поскольку в данном случае дата-центр является ключевым ресурсом, то активы компании в большей степени зависят от его качества и надежности. В итоге рыночная стоимость таких компаний определяется параметрами надежности и непрерывности работы дата-центров и фактически является их производной.

Заключение

Таким образом в настоящей статье рассмотрены взаимосвязь современных компаний, применяющих ИТ-технологии в своих бизнес-процессах, и дата-центров, посредством которых обеспечивается качественная работа ИТ-ресурсов.

В статье выделены три наиболее типовые фокус-группы предприятий, являющихся пользователями услуг дата-центров, а также показаны различия в степени интеграции и зависимости успешности ведения бизнеса от работоспособности дата-центров.

В качестве выводов показаны явно прослеживающиеся тенденции и взаимосвязи между качеством дата-центров и конечной стоимостью активов компаний. Активы могут быть выражены через акционерную стоимость, стоимость акций, либо рыночную капитализацию. Степень интеграции дата-центра в бизнес-процессы компаний демонстрирует прямую связь с рыночной стоимостью компаний. ■

Литература

1. Райзберг Б.А., Лозовский Л.Ш., Стародубцева Е.Б. Современный экономический словарь. М.: ИНФРА-М, 2015. 512 с.
2. Amzarakov M.B., Sukhov R.R., Isaev E.A. Modular data center: The holistic view // Business Informatics. 2014. No. 3 (29). P. 7–14.
3. Положение Банка России от 16 декабря 2003 г. № 242-П «Об организации внутреннего контроля в кредитных организациях и банковских группах» // Информационно-правовой портал ГАРАНТ [Электронный ресурс]: http://base.garant.ru/584330/ (дата обращения: 29.03.2015).
4. ГОСТ Р ИСО/МЭК 27005-2010. Информационная технология. Методы и средства обеспечения безопасности. Менеджмент риска информационной безопасности // Открытая база ГОСТов [Электронный ресурс]: http://standartgost.ru/g/ГОСТ_Р_ИСО/МЭК_27005-2010 (Дата обращения: 29.03.2015).

DATA CENTERS AND ASSETS OF A COMPANY

Rafael R. SUKHOV
Finance Manager, INO Uptime Technology
Address: 6, Bolshoy Koptevskiy proezd, Moscow, 125315, Russian Federation
E-mail: r.sukhov@uptimetechnology.ru

Maxim B. AMZARAKOV
Director, INO Uptime Technology
Address: 6, Bolshoy Koptevskiy proezd, Moscow, 125315, Russian Federation
E-mail: m.amzarakov@uptimetechnology.ru

Eugene A. ISAEV
Professor, Head of Department of Information Systems and Digital Infrastructure Management, School of Business Informatics, Faculty of Business and Management, National Research University Higher School of Economics;
Head of Laboratory, P.N. Lebedev Physical Institute, Russian Academy of Sciences
Address: 20, Myasnitskaya Street, Moscow, 101000, Russian Federation
E-mail: eisaev@hse.ru

Svetlana V. MALTSEVA
Professor, Head of Department of Innovation and Business in Information Technologies, School of Business Informatics, Faculty of Business and Management, National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow, 101000, Russian Federation
E-mail: smaltseva@hse.ru

The paper is focused on data centers and assets of a company, including their relationship and interaction. The purpose of the article – to give an idea of how the Data Center may have an impact on the company's assets and their final value. The aspects that are important for understanding the reasons of companies' interest in formation of the investment object and subsequent accounting of such investments as a significant part of a company's assets are discussed. Justification of the fact that in some enterprises a data center itself is an important asset, and in some business models – a key asset of a company is provided.

Relying on the definitions of «assets» and «data center» terms the variants of participation of a data center in the business of an enterprise and its influence on the company's final value through the company's assets are discussed in the article.

The article presents examples of how a data center becomes the subject of production in large enterprises whose business is based on the storage, processing and delivery of information services related with access to this information. Some examples of such companies representing different industries are provided.

The issues of statutory regulation on the requirements related with establishment of data centers for the purposes of performing regulatory functions are considered. The questions of corporate security and the impact of data centers on information safety are discussed.

Certain attention is paid to indirect influence of a data center on a company's assets value by improving data reliability, improving security of stored and processed data and as a result the impact on the market value of an enterprise as a business through increasing of consumers' confidence.

Key words: data center, assets of a company, market value of shares, risk, business process.

Citation: Sukhov R.R., Amzarakov M.B., Isaev E.A., Maltseva S.V. (2015) Data-centry i aktivy predprijatija [Data centers and assets of a company]. *Business Informatics*, no. 3 (33), pp.74–79 (in Russian).

References

1. Raizberg B.A., Lozovsky L.Sh., Starodubtseva E.B. (2015) *Sovremennyj jekonomicheskij slovar'* [Modern economics dictionary]. Moscow: INFRA-M (in Russian).
2. Amzarakov M.B., Sukhov R.R., Isaev E.A. (2014) Modular data center: The holistic view. Business Informatics, no. 3 (29), pp. 7–14.
3. *Bank of Russia Regulation, 16 December 2003, no. 242-P «Ob organizacii vnutrennego kontrolja v kreditnyh organizacijah i bankovskih gruppah»* [On internal control organization in credit institutions and banking groups]. GARANT System. Available at: http://base.garant.ru/584330/ (accessed 29 March 2015) (in Russian).
4. R ISO/MEC 27005-2010. *Informacionnaja tehnologija. Metody i sredstva obespechenija bezopasnosti. Menedzhment riska informacionnoj bezopasnosti* [Information technology. Security methods and instruments. Information security risk management]. Open base of Russian standards. Available at: http://standartgost.ru/g/ГОСТ_Р_ИСО/МЭК_27005-2010 (accessed 29 March 2015) (in Russian).

АВТОРАМ

Представляемая для публикации статья должна быть актуальной, обладать новизной, отражать постановку задачи (проблемы), описание основных результатов исследования, выводы, а также соответствовать указанным ниже правилам оформления.

Текст должен быть тщательно вычитан автором, который несет ответственность за научно-теоретический уровень публикуемого материала.

Материалы представляется в электронном виде по адресу:

bijournal@hse.ru.

ТРЕБОВАНИЯ К ОФОРМЛЕНИЮ СТАТЕЙ

ТЕКСТ СТАТЬИ представляется в редакцию в электронном виде (в формате MS Word, версия 2003 или выше).

ОБЪЕМ. Ориентировочный объем статьи составляет 20-25 тысяч знаков (с пробелами).

ШРИФТ, ФОРМАТИРОВАНИЕ, НУМЕРАЦИЯ СТРАНИЦ

ШРИФТ – Times New Roman, кегль набора – 12 пунктов, полуторный интервал, форматирование по ширине. Нумерация страниц – вверху по центру, поля: левое – 2,5 см, верхнее, нижнее и правое – по 1,5 см.

НАЗВАНИЕ СТАТЬИ приводится на русском и английском языках. Название статьи должно быть информативным и раскрывать содержание статьи.

СВЕДЕНИЯ ОБ АВТОРАХ приводятся на русском и английском языках и включают следующие элементы:

✦ фамилия, имя, отчество всех авторов полностью

✦ должность, звание, ученая степень каждого автора

✦ полное название организации – места работы каждого автора в именительном падеже, полный почтовый адрес каждой организации (включая почтовый индекс)

✦ адрес электронной почты каждого автора.

АННОТАЦИЯ К СТАТЬЕ представляется на русском и английском языках.

✦ Объем – 200-300 слов.

✦ Аннотация должна быть информативной (не содержать общих слов).

✦ Аннотация должна отражать основное содержание статьи и быть структурированной (следовать логике описания результатов в статье).

✦ Структура аннотации: предмет, цель, метод или методологию проведения исследования, результаты исследований, область их применения, выводы.

✦ Метод или методологию проведения исследований целесообразно описывать в том случае, если они отличаются новизной или представляют интерес с точки зрения данной работы. В аннотациях статей, описывающих экспериментальные работы, указывают источники данных и характер их обработки.

✦ Результаты работы описывают предельно точно и информативно. Приводятся основные теоретические и экспериментальные результаты, фактические данные, обнаруженные взаимосвязи и закономерности. При этом отдается предпочтение новым результатам и данным долгосрочного значения, важным открытиям, выводам, которые опровергают существующие теории, а также информации, которая, по мнению автора, имеет практическое значение.

✦ Выводы могут сопровождаться рекомендациями, оценками, предложениями, гипотезами, описанными в статье.

✦ Сведения, содержащиеся в названии статьи, не должны повторяться в тексте аннотации. Следует избегать лишних вводных фраз (например, «автор статьи рассматривает...»).

✦ Исторические справки, если они не составляют основное содержание документа, описание ранее опубликованных работ и общеизвестные положения, в аннотации не приводятся.

✦ В тексте аннотации следует употреблять синтаксические конструкции, свойственные языку научных и технических документов, избегать сложных грамматических конструкций.

✦ В тексте аннотации следует применять значимые слова из текста статьи.

КЛЮЧЕВЫЕ СЛОВА приводятся на русском и английском языках. Количество ключевых слов (словосочетаний) – 6-10. Ключевые слова или словосочетания отделяются друг от друга точкой с запятой.

ФОРМУЛЫ. При наборе формул, как выключных, так и строчных, должен быть использован редактор формул MS Equation. В формульных и символических записях греческие (русские) символы, а также математические функции записываются прямыми шрифтами, а переменные аргументы функций в виде английских (латинских) букв – наклонным курсивом (пример «cos a», «sin b», «min», «max»). Нумерация формул – сквозная (по желанию авторов допускается двойная нумерация формул с указанием структурного номера раздела статьи и, через точку, номера формулы в разделе).

РИСУНКИ (графики, диаграммы и т.п.) могут быть оформлены средствами MS Word или MS Excel. Ссылки на рисунки в тексте обязательны и должны предшествовать позиции размещения рисунка. Допускается использование графического векторного файла в формате wmf/emf или cdr v.10. Фотографические материалы предоставляются в формате TIF или JPEG, с разрешением изображения не менее 300 точек на дюйм. Нумерация рисунков – сквозная.

ТАБЛИЦЫ оформляются средствами MS Word или MS Excel. Нумерация таблиц – сквозная.

СПИСОК ЛИТЕРАТУРЫ составляется в соответствии с требованиями ГОСТ 7.0.5-2008. Библиографическая ссылка (примеры оформления размещены на сайте журнала http://bi.hse.ru/). Нумерация библиографических источников – в порядке цитирования. Ссылки на иностранную литературу – на языке оригинала без сокращений.

СПИСОК ЛИТЕРАТУРЫ ДЛЯ АНГЛОЯЗЫЧНОГО БЛОКА оформляется в соответствии с требованиями SCOPUS (примеры оформления размещены на сайте журнала http://bi.hse.ru/). Для транслитерации русскоязычных наименований можно воспользоваться сервисом http://translit.ru/.

ЛИЦЕНЗИОННЫЙ ДОГОВОР

Для размещения полнотекстовых версий статей на сайте журнала с авторами заключается лицензионный договор о передаче авторских прав.

Плата с авторов за публикацию рукописей не взимается.