МРНТИ 20.53.21
УДК 004.042
DOI: https://doi.org/10.51889/2020-1.1728-7901.56

WEB APPLICATION FOR PROCESSING A LARGE AMOUNT OF DATA IN THE FIELD OF BUSINESS

Zhanibek Zh.A.[1], Balakayeva G.T.[1]

[1] Al-Farabi Kazakh National University, Almaty, Kazakhstan

Abstract

Big Data is one of the main drivers of the formation of information and communication technologies in modern conditions of high-tech production. The ever-growing capabilities of analyzing a large amount of information are currently significantly changing the business environment and the business processes that take place in it. The use of Big Data technologies can play a significant role in the innovative development of the digital economy in the near future. In today's world, where information is updated at an incredible speed and comes from a variety of sources, companies have to work with huge amounts of information and data. Big Data technologies allow you to collect, store, structure and analyze large amounts of information. This helps the management of the company to find patterns and causal relationships between various factors and use this advantage to obtain positive results. The article is devoted to the study of the basic concepts associated with Big Data, the basics and principles of working with methods and approaches of Big Data. Particular attention is paid to methods of processing these types of information using the data preprocessing method. This method was used for specific examples and the corresponding results were obtained.

Keywords: Big Data, information, preprocessing, Web application, data processing, scaling data, normalization.

Аннотация

Ж.А. Жәнібек [1], Г.Т. Балакаева [1]
[1]Казахский Национальный Университет им. аль-Фараби, г. Алматы, Казахстан

ВЕБ-ПРИЛОЖЕНИЕ ДЛЯ ОБРАБОТКИ БОЛЬШОГО ОБЪЕМА ДАННЫХ В СФЕРЕ БИЗНЕСА

Big Data считается одним из ключевых драйверов развития справочно-коммуникационных технологий. Регулярно возрастающие способности рассмотрения значительного числа данных в наше время значимым способом меняют сферу предпринимательства. Применение технологий Big Data способно исполнить важную значимость во инновационном формировании числовой экономики в недалекой перспективе.

В современном мире, где информация обновляется с невообразимой скоростью и поступает из самых различных источников, фирмам приходится трудиться с большими массивами сведений и данных. Big Data дают возможность коллекционировать, хранить, структурировать и анализировать большие объемы информации. Это может помочь управлению компании отыскивать закономерности и причинно-следственные связи меж разными причинами и применить это превосходство для получения положительных результатов. Статья посвящена изучению основных понятий Big Data, основы и принципы работы. Особое внимание уделяется способам обработки информации с использованием метода предварительной обработки данных. Данный метод был использован для конкретных примеров и были получены соответствующие результаты.

Ключевые слова: большие данные, информация, предварительная обработка данных, Веб приложение, обработка данных, масштабирование данных, нормализация.

Аңдатпа

Ж. А. Жәнібек [1], Г. Т. Балакаева [1]
[1]аль-Фараби атындағы Қазақ Ұлттық Университеті, Алматы қ., Қазақстан

БИЗНЕС САЛАСЫНДАҒЫ КӨПТЕГЕН МӘЛІМЕТТЕРДІ ӨҢДЕУГЕ АРНАЛҒАН ВЕБ-ҚОСЫМША

Big Data қазіргі заманғы жоғары технологиялық өндірісте ақпараттық-коммуникациялық технологияларды қалыптастырудың негізгі қозғаушы күштерінің бірі болып табылады. Көптеген ақпараттарды талдаудың үнемі өсіп келе жатқан мүмкіндіктері қазіргі уақытта іскерлік ортаны және ондағы бизнес-процестерді айтарлықтай өзгертеді. Big Data технологияларын пайдалану жақын болашақта цифрлық экономиканың инновациялық дамуында маңызды рөл атқара алады. Ақпарат керемет жылдамдықпен жаңартылатын және әртүрлі көздерден алатын қазіргі әлемде компаниялар үлкен көлемде ақпаратпен және мәліметтермен жұмыс істеуге мәжбүр. Үлкен деректер технологиясы үлкен көлемдегі ақпаратты жинауға, сақтауға, құрылымдауға және талдауға мүмкіндік береді. Бұл компания басшылығына әртүрлі факторлар арасындағы себептер мен байланыстарды табуға және оң нәтиже алу үшін осы артықшылықты пайдалануға көмектеседі. Мақала үлкен мәліметтермен байланысты негізгі ұғымдарды, үлкен деректердің әдістері мен тәсілдерімен жұмыс істеу негіздері мен принциптерін зерттеуге арналған. Деректерді алдын-ала өңдеу әдісін қолдана отырып, ақпараттың осы түрлерін өңдеу әдістеріне ерекше көңіл бөлінеді. Бұл әдіс нақты мысалдар үшін қолданылды және тиісті нәтижелер алынды.

Түйін сөздер: үлкен деректер, ақпараттар, деректерді алдын-ала өңдеу, Веб қосымша, мәліметтерді өңдеу, деректерді масштабтау, қалыпқа келтіру.

In modern conditions of the formation and development of the information society in various sectors of the economy, a huge amount of data is created and accumulated. In business, industrial field, the volume of technological information, media data necessary for enterprise management is constantly increasing [2, p. 171]. New programs, services and tools based on the use of information and communication technologies appear. As a result of the digitalization of the economy, the need for information products and services is growing. To meet customer needs, companies have to process and analyze colossal amounts of data, varying degrees of structure and from various sources. Thus, the accumulated information becomes a strategically important asset, the effectiveness of the management of which significantly affects the results of the enterprise [3]. In recent years, humanity has produced more information than in the entire history of its existence. Every year, the amount of information in the world increases by an average of 40% [4]. The growth of data volumes is accompanied by the advent of software and hardware that provide storage, processing, calculation and analysis of a large amount of information. The cost of storing information at the same time decreased, which affected the ability to collect more data and analyze factors unrelated to each other. The human brain cannot detect such patterns as the computer notes, producing completely unexpected causal and quantitative relationships [1, p. 56].

As a result of the combination of these two processes - the growing need for business to collect, store, analyze large volumes of data and the creation of technical tools that can efficiently process data with minimal costs, an interesting and promising area of technology development called Big Data has appeared.

In this regard, we decided to develop a web application where entrepreneurs can publicly download their databases for data processing and make various queries.

There are several methods of processing large amounts of data. In our case, we chose the preprocessing method.

Real world data is usually:

• Incomplete- the absence of certain attributes or their values of interest or containing only aggregate data.

• Contain noise- there are errors or outliers in the values.

• Inconsistent - contain inconsistencies in codes or names.

Tasks of data preprocessing:

• Data cleansing - filling in missing values, detecting and deleting distorted data and outliers.

• Data Integration- Use multiple databases, data cubes, or files.

• Data Transformation- Normalization and Aggregation.

• Data reduction- reduction in volume, but obtaining the same or similar analytical results.

• Data discretization- part of data reduction, replacement of numerical attributes with nominal ones.

• Clear text - delete embedded characters that may interfere with data alignment, such as embedded tab characters in a tab-delimited file, embedded new lines that can split records, etc.

The following are some steps for pre-processing data from a Data Mining perspective.

Scaling data

Input variables must be scaled, that is reduced to a single range of change. The need for scaling is due to several reasons. After encoding information with inputs and outputs, dissimilar quantities varying in different ranges. It is desirable to bring all input variables to a single range and normalize (the maximum absolute value of the input variables should not exceed unity). Otherwise, errors caused by variables that vary over a wide range will have a stronger effect than errors from variables that vary over a narrow range. By providing a change in each input variable within the same range, we will ensure equal influence of each.

Therefore, the input variables, as a rule, are scaled, so that the variables change in the range of variation of the function, as a rule, [0,1] or [−1,1]. In practice, you can not strictly maintain a single range of input data, but scaling the input data simplifies the work.

Each input variable is scaled independently of the other variables.

The scale of the input and output variables is not related. With a known range of variation of the variable, it is advisable to use linear scaling [5]. For example, for each input variable, linear scaling has the form

$$t_i = \frac{(x_{i-} x_{\min})(b-a)}{x_{\max} - x_{\min}} + a$$

where x_i is the input variable, t_i is the converted input variable supplied to the network input, $[a, b]$ is the allowable range of input variables, for example, [−1,1]; $[x_{min}, x_{max}]$ - the range of variation of the input variable. Output variables are scaled similarly.

Normalization

Of all the existing distributions, the most popular is the normal distribution, which is due, first of all, to the fact that normal (with a normal distribution) observations are quite easy and convenient to investigate. Let us consider the transformations that make it possible to obtain approximately normal from the available data.

Logarithmic conversion. Often, data with positive values have a distribution with positive asymmetry that resembles a log-normal distribution, x^2 or a γ-distribution. If the random variable *X* has a logarithmically normal distribution, then its logarithm will be normal, therefore, using the logarithmic transformation allows you to obtain approximately normally distributed values only in cases where the distribution of the quantity *X* is qualitatively similar to the logarithmically normal distribution.

If the values of the random variable *X* lie in the interval (α, β), then the values of the normalized variable

$$Y = \ln \frac{\alpha - X}{X - \beta}$$

can vary from −∞ to + ∞. Thus, it is possible that Y can be approximately normal.

The use of this design is recommended for studying the correlation coefficient. Correlation coefficient r, calculated from a sample of *n* pairs $(x_i,\ y_i)$.

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{((\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2)((\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2)^{\frac{1}{2}}}$$

whose values lie in the interval (−1,1). The sample distribution is, as a rule, highly skewed and its exact shape depends on the value ρ of the correlation coefficient in the original population.

Converted Statistics:

$$z = \frac{1}{2}\ln \frac{1+r}{1-r}$$

has an "almost" normal sample distribution with a mathematical expectation

$$\frac{1}{2}\ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$$

and dispersion (approximately)

$$\frac{1}{n-3}$$

This transformation greatly simplifies the question of the accuracy of r as an estimate of ρ. For completeness, we give one more fact.

If r is the correlation coefficient of a sample of n independent pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, where x_i is observation of a normally distributed random variable *X*, and y_i are observations of a normally distributed random variable *Y*, which are independent of each other, then the selective distribution of statistics

$$\sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}$$

Student's distribution will be with $n-2$ degrees of freedom.

Normalizing transformation of the distribution x^2. The distribution x^2 is very popular and there are fairly accurate tables of the values of this distribution, but in many cases, it is more convenient to work with an approximately normal function of x^2. For this purpose, x^2- variable with ν degrees of freedom can be transformed as follows [5].

For sufficiently large ν, for example, ν > 100, the variable

$$X = \sqrt{2x^2} - \sqrt{2\nu - 1}$$

is approximately distributed according to the standard normal law, but even at ν∈(30, 100) the approximation is quite good.

The best result is the conversion

$$X = \frac{\sqrt[3]{x^2/\nu} - (1 - 2/9\nu)}{\sqrt{2/9\nu}}, \nu > 30$$

Conversion using the probability integral. Generally speaking, any continuous random variable can be precisely normalized by the transformation of the probability integral. Let *F(X)* be a distribution function of *X* at x, then the transformed variable *Z=F(X)* will have a normal distribution of (0,1). If $\Phi(y)$ is a distribution function of a standard normal variable at y, then the random variable $\Phi(Y)$ will be uniformly distributed over (0,1), hence the transformation $X \rightarrow Y$

$$\Phi(Y) = F(Y)$$

or

$$Y = \Phi^{-1}(F(X))$$

converts *X* to a standard normal variable.

The “Sabyrzhan” and “Toimart” supermarkets’ database was taken as an example. The data collects all the products where each product is described by variables. Our task was to build a model that would output query results using the preprocessing method.

There was a request for what kinds of bakery products are sold more (Figure 1).

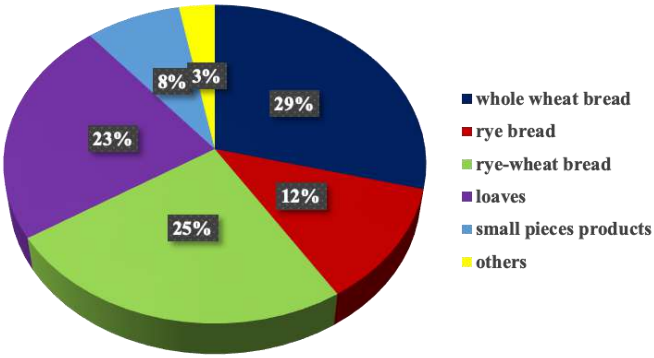

Figure 1. Chart of preferences of the population in bakery products

Figure 1 presents a chart of preferences of “Sabyrzhan” and “Toimart” supermarkets in bakery products. The diagram shows that the consumption of wheat bread occupies a significant place in the total consumption of bread.

In the course of the study came the verdict that one of the key issues is assessing the effectiveness of the Big Data project. First, these technologies can dramatically reduce the cost and time of analyzing a large amount of information and prepare information in the shortest possible time for operational and management decisions. Second, the use of Big Data enables the personalization of services in the B2B and B2C markets. The main thing is to learn how to properly process and analyze the data received, turning information into an asset and strategic resource for the development of the organization.

References:

1 Майер-Шенбергер В., Кукьер К.. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим; пер. с англ. Инны Гайдюк. – М.: Манн, Иванов и Фербер, 2014. – p.240

2 Волкова Ю.С. Большие Данные в современном мире // Концепт. Т. 2016. – №11. – p. 171-175

3 Big Data: проблема, технология, рынок [An electronic resource]. – 2019. – URL: http://compress.ru/Article.aspx?id=22725 (Date of the application: 02.12.2017)

4 Tech Pro Research [An electronic resource]. – 2019. – URL: http://www.techproresearch.com/topic/big-data/ (Date of the application: 24.11.2019)

5 Preprocessing [An electronic resource]. – 2014. – URL: http://pzs.dstu.dp.ua/DataMining/preprocessing/index.html (Date of the application: 05.12.2019)

6 Balakayeva, G. T., Phillips, C., Darkenbayev, D. K., & Turdaliyev, M. (2019). Using NoSQL for processing unstructured big data. News of the National Academy of Sciences of the Republic of Kazakhstan, Series of Geology and Technical Sciences, 6(438), 12-21. doi:10.32014/2019.2518-170X.151