Identifying duplicated ads on property selling and renting websites

V S Tynchenko[1,2]**, V V Kukartsev**[1,2]**, V V Tynchenko**[1,2]**, V V Bukhtoyarov**[1,2], **E A Chzhan**[1]**, V A Kukartsev**[1]**, A A Boyko**[1,2]

[1] Siberian Federal University, 79, Svobodny pr., Krasnoyarsk, 660041, Russia
[2]Reshetnev Siberian State University of Science and Technology, 31, KrasnoyarskyRabochy Av., 660037 Krasnoyarsk, Russia

E-mail: vadimond@mail.ru

Abstract. The article presents a solution for the problem of identifying duplicated ads on property selling websites. This task is formulated in the form of a classification problem: the input parameters are identified then divided into basic and non-basic, as well as a class-forming feature. It is also necessary to consider the preliminary data of processed property objects, which is necessary for proper application of the classification methods. The following is a brief review of chosen modern algorithms for solving classification problems, namely: decision trees, artificial neural networks, logistic regression. As a result of experiments, it was revealed that Artificial neural network gives the most accurate result therefore, this algorithm is suitable for the solution of the stated problem.

1. Introduction

Nowadays there are many sites specializing in property sales and rentals. Using these sites users can pick up accommodation, corresponding to their personal needs without leaving their houses. Sites of this kind simplified the process of property search, but many users are faced with the problem that some ads are duplicated. Usually it happens because some people use sale or rental ads of properties belonging to other people by impersonating the owners. Also, such people can change all the characteristics of the ads, thus deceiving potential tenants and property buyers. The more popular and reliable the site is (according to users), the larger the database of ads is and the harder it is to track and delete stolen ads. According to statistics, about 20% of ads are not genuine.

Avito is an Internet site for advertising of goods and services from individuals and companies and is the second in the world and the first in Russia among classified online advertising. The goods offered for sale on Avito can be new and secondhand. Also, the site publishes advertisements about job vacancies and resumes of applicants. In June 2018, there were more than 35.8 million active advertisements placed on Avito.

For example, users of Avito began to complain that ads placed on the site without their consent also appeared on other sites with loss of quality. Photos were incorrectly cropped, in some cases the descriptions were mixed, i.e. a one-room apartment was priced as a house. In this situation, the potential buyer or tenant cannot be sure of the authenticity of the advertisement being viewed. The solution to this problem is automate search of the duplicated ads using modern data analysis methods.

The task of searching for the duplicated ads can be presented in the form of a classification problem, for which it is necessary not only to make a reasonable choice of the algorithm, but also to correctly present and process the system’s input parameters.

2. Problem statement

In order to build a system that will perform the task of searching for the duplicated ads, it is necessary to determine what parameters they can have. Due to the identified parameters, this task can be considered as a classification problem.

The classification problem is a formalized task, in which there are many objects (situations), divided in some way into classes [1 - 4]. A finite set of classified objects is given. This set is called a sample. Classes of other objects are unknown. The objective is to build an algorithm that can classify an arbitrary object from the original set [5 - 7].

To classify an object means to specify the number (or name) of the class to which the object belongs. The classification of the object is the number or name of the class issued by the classification algorithm as a result of its application to this object [8-10].

One of important cases of classification problems in data processing systems is the case when the classification according to the specified criteria has already been produced, the classes are generated and it is necessary on the basis of known information about the classes' cardinality, their quantity and attributes of their constituent objects compute the static characteristics of classes and identify the relationships among class objects [1, 5 - 7].

In mathematical statistics, classification problems are also called discriminant analysis tasks. In machine learning the problem of classification is solved, in particular, by means of artificial neural networks in the experiment of training with a teacher [3-5, 8-10].

The proposed automated system is a property search portal at which the user can not only find property that satisfies his needs, but can also be sure the viewed ad is unique.

Also, the system protects property owners from copying their ads, because in case of duplication, the published advertisement will not pass the process of moderation which is executed by the developed intellectual subsystem based on the data classification method.

3. Defining system parameters

To determine the authenticity of an ad it is necessary to check the matching characteristics of other ads. The characteristics are divided into two types: basic and non-basic.

The basic characteristics are characteristics of ads that are required when creating an ad. Collectively, certain values of these characteristics indicate the authenticity of an ad.

The non-basic characteristics are ad characteristics that are optional when an ad is created. The values of these characteristics do not affect the authenticity of an ad.

The basic characteristics include:

- Property's photo is the first to check and the main parameter that affects the decision on the authenticity of the ad. To define this parameter, it is necessary to compare photos from the verifiable advertisement with photos from other advertisements and to reveal a percentage of identity.
- Property value is a mandatory characteristic of ads that has a unique value.
- Phone is a mandatory characteristic of ads that helps in identifying the identity of the owner of the property.
- Address is a mandatory characteristic of ads that has a unique value. Indicates the location of the property object.
- Area is a mandatory characteristic of ads that has a unique value. Contains information about the area of property.
- Number of rooms is mandatory characteristic of ads that has a unique value. Contains information about the number of rooms in the property.
- Floor is mandatory characteristic of ads that has a unique value. Indicates the floor in the building where the property is located.

The non-basic characteristics include:

- Description is an optional characteristic of ads, which contains additional information about the property and is filled in any form.
- Type of property is optional characteristic of ads, which contains information about the type of property (apartment, cottage, dacha, commercial).

4. Preliminary data processing of property objects

Although not every characteristic store information that can help to determine the authenticity of the ad, all of the characteristics can be input data for testing the method of searching of the duplicated ads. After reviewing various sources and advertisements, a number of basic and non-basic characteristics were identified. The basic characteristics of a property are: the property's photography, cost, area, address, floor and the owner's phone. The non-basic are: the description of the property, type of the building, number of floors in the building. In order to properly apply the classification method, it is necessary to standardize all the basic characteristics.

In this case, it is proposed to use not the values of the characteristics, but to use the degree of identity of the object's characteristics (OC) with other objects of the group.

The following formula is used to calculate the value of the match coefficient of k-th quantitative characteristic of the i-th object:

$$KC_i^k = \max_{J \neq i} \left(1 - \frac{X_i^k - X_j^k}{X_i^k + X_j^k}\right), \quad (1)$$

where:

- X_i^k – value of the *k*-th quantitative characteristic of the *i*-th object;
- *i, j*– ordinal numbers of objects in the considered set.

For characteristics that do not have quantitative expression, the following formula is applied:

$$KC_k^i = \max_{J \neq i}(B(i,j,k)), \quad (2)$$

where *B(i,j,k)* is a logical function to indicate whether the values are the same characteristics of objects and is calculated by the formula:

$$B(i,j,k) = \begin{cases} 1, X_i^k = X_j^k, \\ 0, X_i^k \neq X_j^k. \end{cases} \quad (3)$$

Among the available characteristics of property objects, the quantitative characteristics are: the object's area, its cost, the degree of similarity of the description and photos. The quality characteristics are: the object's address and the phone number of the ad's publisher.

5. Algorithm selection

The main task of the portal - identification of genuine ads – can be presented as a classification problem. These problems can be solved with such algorithms as: decision trees, artificial neural networks, logistic regression [7 - 11].

A neural network is a system of connected and interacting simple processors (artificial neurons). They are connected to a fairly large network with manageable interaction and are able to perform quite complex tasks together, because neural networks are trained in the process of working [5, 8, 11].

The decision trees method is one of the most popular methods of solving the classification and forecasting problems. When analyzing solutions, the decision tree is used as a visual and analytical decision support tool, where expected values (or expected usefulness) of competing alternatives are calculated [9, 10].

Logistic regression is a method of constructing a linear classifier, which allows to estimate the probability of objects belonging to classes. Logistic regression is useful for situations when you want to be able to predict the presence or absence of a characteristic, or a total based on the values of a set of predicate variables [1 - 3].

6. Experimental research

To select the most suitable classification algorithm for the portal, it is necessary to conduct experimental studies on real data. To do this “Deductor Studio” is used - a platform that allows one to conduct all data analysis. At the initial stage property objects were collected (Table 1). Match factors were calculated for all property values of all objects.

Table 1. Input data for classification (fragment).

Address	Area	Number of rooms	Floor	Phone	Description	Photo	Cost	Authenticity
1	0.922	1	0	1	0.4	0.008	0.2711	No
1	0.94	0	1	1	0.97	0.91	0.96	Yes
1	0.75	1	0	0	0.2	0.07	0.379	No
1	0.6187	0	0	0	0.7	0.05	0.2751	No
1	0.99	1	1	1	0.99	1	1	No
1	0.93	0	1	1	0.96	0.99	0.95	Yes
1	0.95	1	1	1	0.9	1	0.9	No
1	0.91	1	1	1	0.95	0.9	0.9	No
1	0.92	1	1	1	0.92	0.95	0.95	No
1	0.97	0	1	1	0.91	0.93	0.91	Yes

As a result, it is necessary to determine whether the advertisement in question is genuine. Table 2 presents the results of three different classification algorithms:

- Artificial neural network.
- Decision tree.
- Logistic regression.

Table 2. Mating table.

	Artificial neural network		Decision tree		Logistic regression	
	True	False	True	False	True	False
True	173	-	167	6	169	3
False	-	24	6	18	4	21
Probability of correct classification	100%		93,9%		96,44%	

As a result of experiments, it is revealed that artificial neural network gives the most accurate and correct result. In addition, this algorithm has the ability to adapt to changes in the input data. Decision trees and the logistic regression do not meet the stated requirements because the data are not always correctly classified, and small changes in the input data result in a loss of result accuracy.

7. Conclusion

In this work, the solution of identifying duplicate ads on the web-portal is considered. The proposed approach includes an intelligent core for defining ad's uniqueness. It is proposed to use algorithms of data classification: artificial neural networks, decision trees, logistic regression. For their application, initial data were pre-processed. As a result of experiments, it was revealed that artificial neural network gave the most accurate results, therefore, this algorithm was suitable for the solution of the problem.

References

[1] Friedman J, Hastie T and Tibshirani R 2001 The elements of statistical learning *Springer series in statistics* **1** 337-387

[2] Witten I H 2016 *Data Mining: Practical machine learning tools and techniques* (Morgan Kaufmann)

[3] Mitchell T 1997*Machine Learning* (Portland, OR: Science/ Engineering/ Math)

[4] Larose D T and Larose C D 2014 *Discovering knowledge in data: an introduction to data mining* (John Wiley & Sons)

[5] Caudill M and Butler C 1992 *Understanding Neural Networks: Computer Explorations* (Cambridge, MA: the MIT Press)

[6] Milov A V, Tynchenko V S, Kukartsev V V, Tynchenko V V and Antamoshkin O A 2018 Classification of non-normative errors in measuring instruments based on data mining *Advances in Engineering Research* **158** 432-437

[7] Charalambous C 1992 Conjugate gradient algorithm for efficient training of artificial neural networks *IEEE Proceedings* **139(3)** 301-310

[8] Ripley D D 1996 *Pattern recognition and Neural Networks* (Cambridge University Press)

[9] Breiman L 2017 *Classification and regression trees* (Routledge)

[10] Loh W Y 2014 Classification and regression trees *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1(1)** 14-23

[11] Milov A V, Tynchenko V S and Petrenko V E 2018 Algorithmic and software to identify errors in measuring equipment during the formation of permanent joints *2018 International Multi-Conference on Industrial Engineering and Modern Technologies (IEEE)* 8602515