ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

UDC 519.2

R. T. KASUMOVA

Azerbaijan National Academy of Sciences, Institute of Information Technology, Baku, Azerbaijan

ON INTELLECTUAL ANALYSIS OF DOMAIN NAME REGISTRATION DATA

Анотація: Робота присвячена формуванню бази знань системи доменних імен, що відповідають інтересам Азербайджанської Республіки. З цією метою в роботі розробляється сховище даних для обробки великого обсягу реєстраційних даних доменних імен, проводиться кластеризація цих даних і генеруються правила для видобутку нових знань.

Ключові слова: домен, DNS, адміністратор, регістратор, регістрант, кластеризація.

Abstract: The paper is devoted to formation of knowledge base of the domain names system, serving to interests of the Azerbaijan Republic. According to this purpose, the data warehouse is developed for processing great volume of registration data of domain names, clustering data is performed, and rules are generated for extraction of new knowledge.

Key words: domain, domain name system, administrator, registrar, registrant, clustering, categorical data.

Introduction

As a global telecommunications network of information and calculation resources, Internet creates a global information space, serves as a physical basis for World Wide Web and a variety of systems (protocols) of data transfer. Domain name system is used for addressing of requests in Internet. Domain is a spatial region of domain names and is characterized by independence of subdomain allocation, inclusion of information systems in domain structure, availability of special information systems (DNS-servers) containing data on domain names, allocated in the domain, and executes the function of organization of domain name space [1]. Domain name is an identifier of a domain and (or) information system, possesses a unique structure conditioned as: limited set of symbols, a name that identifies the domain, which contains domain name and as a voluntary part domain label or host name unique within the limits of the domain, which is contained in the domain name.

Domain name carries out the functions of identification, individualization and addressing [2]. Analyses demonstrate that, domain names are also used as means for conduction of unfit and unethical competition. One of the examples of unethical use of Internet, is use of famous trademarks, service trademarks, place of origins of commodities, as well as brand names in domain names [3].

Relevance

Currently, deficiencies existing in the field of domain name registration, absence of transparency in regards to domain name registration process, violation of domain name registration rules by the registrars, purchase and use of domain name (in irresponsible manner) with the purpose of its further sale, non-existence of a single policy against invaders of domain names (cyber squatters, phishing etc), as well as software, which allow to conduct an accurate analysis of registration data about domain names collected in DNS (Domain Name System).

Listed problems make conduction of scientific analysis of domain name registration data collected in DNS servers necessary. Considering the dynamics of increasing of number of domain names, it is possible to conduct processing and analysis of this information collected from thousands of domain names, obtain new knowledge, detect regularities and make necessary decisions.

As intellectual analysis of registration data can be the solutions reason of such issues as forecasting, making operative, effective and analytical decisions in domain field, definition of facts, evaluation of the real condition of domain market, research domain name monitoring problems etc.

Objective

The objective of the article is development of data storage for processing of a large volume of domain name registration data, conduct clustering of this data and generates rules for gaining new knowledge.

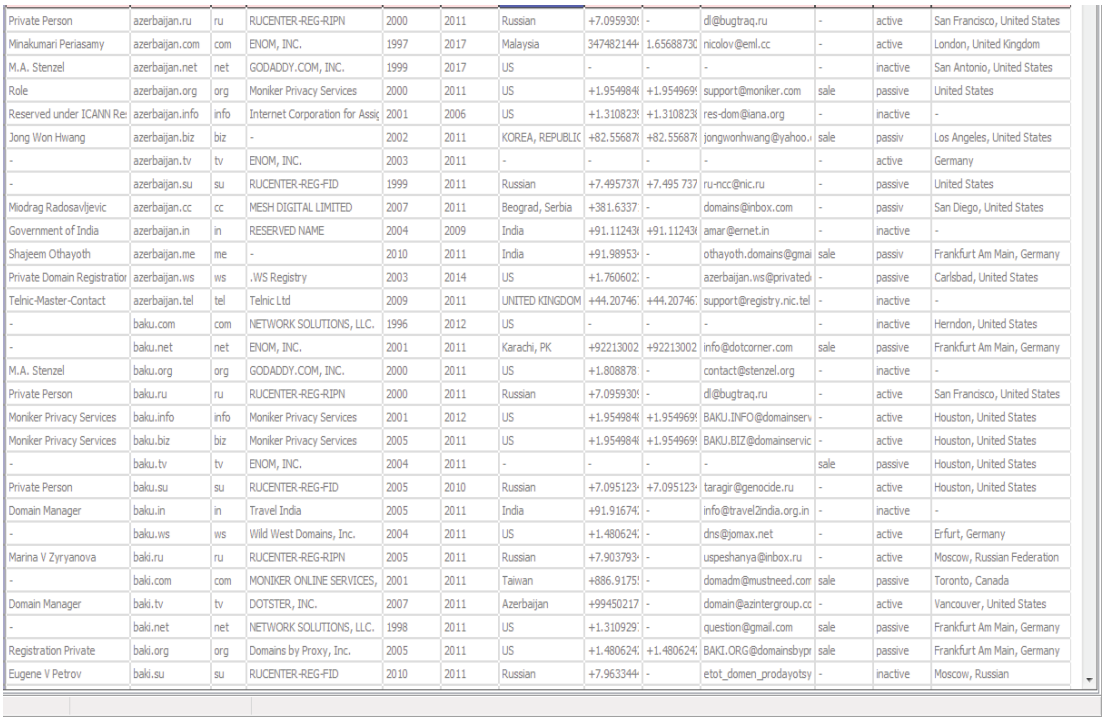
Tasks

Following tasks are formulated in research purpose:

1) Processing of domain name registration data in storage, clustering of this data using CLOPE algorithm;
2) Generation of rules using Magnum Opus v.5.4.1. Program for each cluster and decision making.

Problem Solution

Domain registration information include: domain name, registrar, name, address, admin-o, admin-c, organization, created, updated, free-date, phone, e-mail, nserver, type, source, paid-till etc [4] (pic. 1).



Private Person	azerbaijan.ru	ru	RUCENTER-REG-RIPN	2000	2011	Russian	+7.095930	-	dl@bugtraq.ru	-	active	San Francisco, United States
Minakumari Periasamy	azerbaijan.com	com	ENOM, INC.	1997	2017	Malaysia	347482144	1.6568873(	nicolov@eml.cc	-	active	London, United Kingdom
M.A. Stenzel	azerbaijan.net	net	GODADDY.COM, INC.	1999	2017	US	-	-	-	-	inactive	San Antonio, United States
Role	azerbaijan.org	org	Moniker Privacy Services	2000	2011	US	+1.954984(	+1.954969(	support@moniker.com	sale	passive	United States
Reserved under ICANN Re	azerbaijan.info	info	Internet Corporation for Assig	2001	2006	US	+1.310823(	+1.310823(	res-dom@iana.org	-	inactive	-
Jong Won Hwang	azerbaijan.biz	biz	-	2002	2011	KOREA, REPUBLIC	+82.556871	+82.556871	jongwonhwang@yahoo.(	sale	passiv	Los Angeles, United States
-	azerbaijan.tv	tv	ENOM, INC.	2003	2011	-	-	-	-	-	active	Germany
-	azerbaijan.su	su	RUCENTER-REG-FID	1999	2011	Russian	+7.495737(	+7.495 737	ru-ncc@nic.ru	-	passive	United States
Miodrag Radosavljevic	azerbaijan.cc	cc	MESH DIGITAL LIMITED	2007	2011	Beograd, Serbia	+381.6337	-	domains@inbox.com	-	passiv	San Diego, United States
Government of India	azerbaijan.in	in	RESERVED NAME	2004	2009	India	+91.11243(	+91.11243(	amar@ernet.in	-	inactive	-
Shajeem Othayoth	azerbaijan.me	me	-	2010	2011	India	+91.98953-	-	othayoth.domains@gmai	sale	passiv	Frankfurt Am Main, Germany
Private Domain Registratior	azerbaijan.ws	ws	.WS Registry	2003	2014	US	+1.760602:	-	azerbaijan.ws@privated	-	passive	Carlsbad, United States
Telnic-Master-Contact	azerbaijan.tel	tel	Telnic Ltd	2009	2011	UNITED KINGDOM	+44.20746:	+44.20746:	support@registry.nic.tel	-	inactive	-
-	baku.com	com	NETWORK SOLUTIONS, LLC.	1996	2012	US	-	-	-	-	inactive	Herndon, United States
-	baku.net	net	ENOM, INC.	2001	2011	Karachi, PK	+92213002	+92213002	info@dotcorner.com	sale	passive	Frankfurt Am Main, Germany
M.A. Stenzel	baku.org	org	GODADDY.COM, INC.	2000	2011	US	+1.808878	-	contact@stenzel.org	-	inactive	-
Private Person	baku.ru	ru	RUCENTER-REG-RIPN	2000	2011	Russian	+7.095930:	-	dl@bugtraq.ru	-	active	San Francisco, United States
Moniker Privacy Services	baku.info	info	Moniker Privacy Services	2001	2012	US	+1.954984(	+1.954969(	BAKU.INFO@domainserv	-	active	Houston, United States
Moniker Privacy Services	baku.biz	biz	Moniker Privacy Services	2005	2011	US	+1.954984(	+1.954969(	BAKU.BIZ@domainservic	-	active	Houston, United States
-	baku.tv	tv	ENOM, INC.	2004	2011	-	-	-	-	sale	passive	Houston, United States
Private Person	baku.su	su	RUCENTER-REG-FID	2005	2010	Russian	+7.095123-	+7.095123-	taragir@genocide.ru	-	active	Houston, United States
Domain Manager	baku.in	in	Travel India	2005	2011	India	+91.91674:	-	info@travel2india.org.in	-	inactive	-
-	baku.ws	ws	Wild West Domains, Inc.	2004	2011	US	+1.480624:	-	dns@jomax.net	-	active	Erfurt, Germany
Marina V Zyryanova	baki.ru	ru	RUCENTER-REG-RIPN	2005	2011	Russian	+7.903793-	-	uspeshanya@inbox.ru	-	active	Moscow, Russian Federation
-	baki.com	com	MONIKER ONLINE SERVICES,	2001	2011	Taiwan	+886.9175!	-	domadm@mustneed.com	sale	passive	Toronto, Canada
Domain Manager	baki.tv	tv	DOTSTER, INC.	2007	2011	Azerbaijan	+99450217	-	domain@azintergroup.cc	-	active	Vancouver, United States
-	baki.net	net	NETWORK SOLUTIONS, LLC.	1998	2011	US	+1.310929:	-	question@gmail.com	sale	passive	Frankfurt Am Main, Germany
Registration Private	baki.org	org	Domains by Proxy, Inc.	2005	2011	US	+1.480624:	+1.480624:	BAKI.ORG@domainsbypr	sale	passive	Frankfurt Am Main, Germany
Eugene V Petrov	baki.su	su	RUCENTER-REG-FID	2010	2011	Russian	+7.963344-	-	etot_domen_prodayotsy	-	inactive	Moscow, Russian

Picture 1 – Domain registration data example

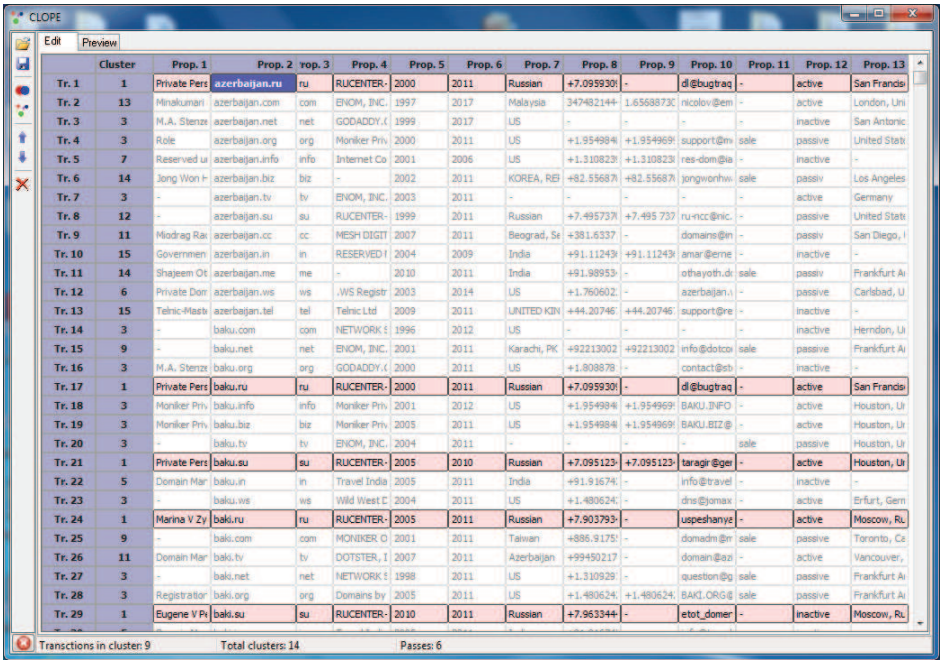
As domain registration data mainly consist of categorical data (which cannot be regulated in space), application of traditional algorithms for objects’ clustering is ineffective. Clustering – is a fundamental data analysis and Data Mining task that groups together similar objects. On modern level, clustering is frequently used as the first step at data analysis [5, 6].

High dimensionality (thousand fields) and large volume (hundred thousand and millions of records) of data base tables, complexity of metrics definition for calculation of distance among categorical data, very low productivity at pair-wise comparison of distance between points (k-means) at each iteration procedure on large record arrays, and sometimes even inapplicability require application of scaled algorithms of categorical data clustering. In most algorithms, metrics based on Euclidian distance concept is used as objects’ proximity metrics $d(x,y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}$, where $x=(x_1,x_2,\ldots,x_n)$, $y=(y_1,y_2,\ldots,y_n)$, thresholds – λ are given for cluster setting, if $d(x,y)<\lambda$, then $x,y\in A_i$, $i=\overline{1,c}$. But given metrics is not always effective, as it obliges clusters to have a spherical form that is not inherent to them. Consequently, known k-means clustering algorithms cannot achieve a satisfying result.

Currently, a variety of clustering algorithms were proposed for working with categorical data. But, they do not always meet abovementioned requirements. LargeItem is considered as one of the effective algorithms, which is based on optimization of some global criteria.

CLOPE algorithm, proposed in 2002 by a group of Chinese scientists, allows clustering task solution of not only categorical, but also any transactional data [7]. It provides a higher productivity and better quality of clustering in comparison with LargeItem algorithms and many other hierarchical algorithms. The key is that all features of objects are measured in nominal scale. However, before launching CLOPE, data must be brought to normalized form. It can have a form of a binary matrix, as in associative rules, as well as being a biunivocal mapping between a set of unique objects of the table and a set of whole numbers. CLOPE is easily counted and interpreted. During its operation, algorithm saves a small amount of information on each cluster in RAM and requires a minimal number of data set scanning. This allows its application for clustering of huge volumes of categorical data. CLOPE automatically selects a quantity of clusters; moreover it is regulated by a single parameter – repulsion coefficient [8].

Thus in reviewed case, CLOPE is one of the effective algorithms, based on which is the idea of maximization of global criteria – cost function Profit (C), which increases the proximity of transactions in clusters through increasing the cluster histogram parameter (pic.2).



Picture 2 –Domain Registration Data Clustering

Formula for calculation of global criteria – cost function looks like following: for given transaction base $D=\{t_1,t_2,...t_n\}$ and r is the repulsion coefficient, to find such a splitting of $C=\{C_1,C_2,...,C_k\}$ that

$$Profit(C,r)\to \max \qquad (1)$$

Where r regulates the level of similarity of transactions within the cluster –bigger is r, larger is the final number of clusters.

Cost function formula:

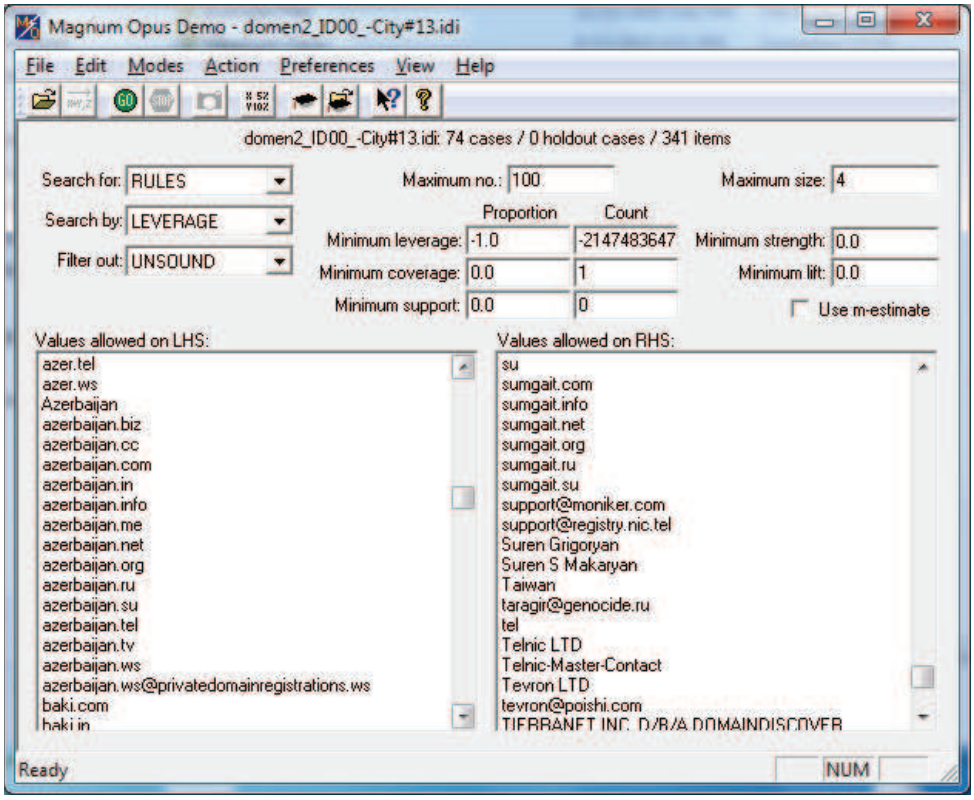
$$Profit(C,r)=\frac{\sum_{i=1}^{k}\frac{S(C_i)}{W(C_i)^r}\times|C_i|}{\sum_{i-1}^{k}|C_i|}, \qquad (2)$$

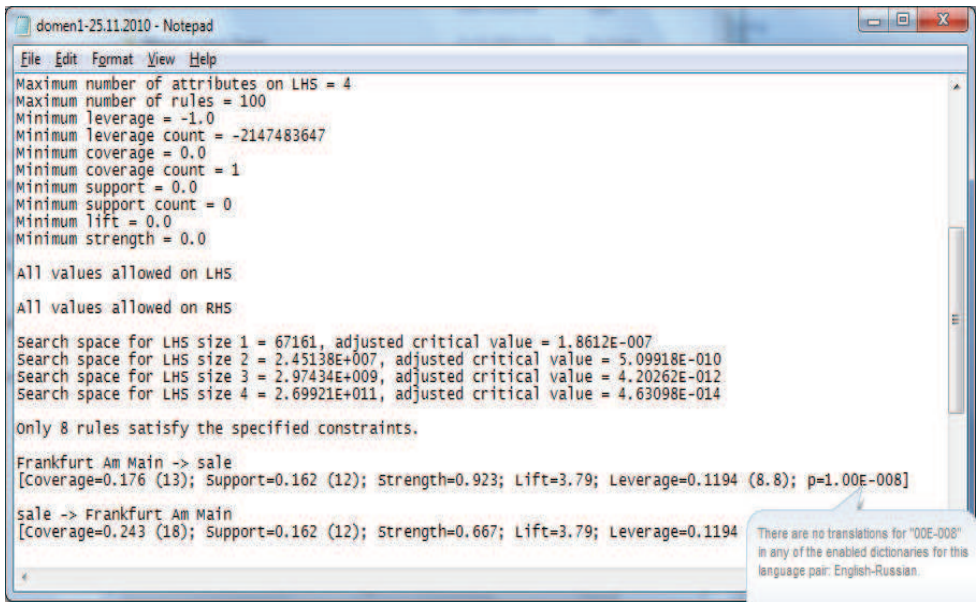
where $S(C_i)$ – is the total number of entry of i objects in clusters C_i, $W(C_i)$- is the quantity of unique objects C_i, $|C_i|$ - is the quantity of objects in i^{th} cluster, k – is the number of clusters, r - is the positive real number larger than 1.

After completion of clustering stage, we generate rules for each cluster using Magnum Opus v.5.4.1 program (http://www.giwebb.com), which is given case detects associations within the data framework of one cluster (pic. 3). Magnum Opus is easy-to-use and finds new associations in data. Associative rules are the results f the operation of his program.

In accordance in [9], associative rule is determined following way: let’s assume that $I=\{i_1,i_2,\ldots,i_n\}$ - a set of elements (binary attributes called elements). Any subset I is called an itemset of elements or simply itemset. Number of elements in the set is called its length. Let’s assume that D - is a set of transactions, where each transaction T, $T\subseteq I$ - represents a binary vector, where $T.i_k=1$, if element $i_k, k=\overline{1,n}$ is present in the transaction, otherwise $T.i_k=0$ ($T.i_k$ – is the value of k^{th} attribute). Let’s assume that X is a set of elements from I. Let’s say that, transaction T, $T\subseteq I$ contains X, i.e. $X\subseteq T$, if $\forall i_k\in X$, then $T.i_k=1$. Associate rule is $X\to Y,[c,s]$ type implication, where $X\subset I, Y\subset I, X\bigcap Y=\varnothing$, $c(confidence)$ - is

the confidence of the rule, s $(support, significan\ ce)$ - is the support of the rule. Support of the rule - s is the percentage of transactions from D, containing $X \bigcup Y$, i.e. $s(X \to Y) = s(X \bigcup Y)$; and confidence of the rule demonstrates that c percentage of transactions from D containing X also contains Y, i.e. $c(X \to Y) = s(X \bigcup Y) / s(X)$. Abovementioned definition is an actual definition of Boolean (binary) associative rule, as it is only reviewed if $i_k, k = \overline{1,n}$ element is present in the transaction or not. In fact, transactions of data bases usually contain numerical and categorical data types.





Picture 3 – Rules generated for each cluster using Magnum Opus v.5.4.1 program

For example, experiments conducted with the domain name system knowledge base , serving the interests of the Republic of Azerbaijan, demonstrated that the majority of generic high level domains – gTLDs with the names of geographical locations of the country were purchased by foreign citizens living outside our republic in 1997-2001 years.

Conclusion

Today, undoubtedly, success of any research field directly depends on its capability of extracting and creating knowledge. With expansion of the commercial use of Internet capabilities, given sphere of human activity is the maximally demanded activity field for gaining profit as for right holders as well as for perpetrators. Usage of such inseparable attribute of presence in the Web as domain name opened new capabilities for lawful, as well as illegal use of such habitual commercial instruments, such as trade mark and brand name. For this reason, detection and replenishment of knowledge in developed domain name system knowledge base can become one of the main research directions in struggle against problems in registration qualification and use of domain names.

Reference

1. Серго А.Г. Доменные имена. М:. «Бестселлер», 2006, с. 368.
2. Милютин З. Ю. Правовой статус доменного имени //Патенты и лицензии. – 2005. № 6. - с. 19-23.
3. Калятин В. О. Проблема конфликта прав на доменные имена с правами на иные средства индивидуализации. // Юридический мир. – 2001. № 5. - с. 19-28
4. Венедрюхин А.А. Доменные войны. СПб.: Питер, 2009, с. 224.
5. Дюк В., Самойленко А. Data Mining: Учебный курс. – СПб: Питер, 2001, с. 368.
6. Алгулиев Р.М., Касумова Р.Т., Алекперова И.Я. О современных концепциях, поддерживающих принятие решений //Известия Национальной Академии Наук Азербайджана. Серия физико-математических и технических наук. – 2005, №2, с. 70-75.
7. Yang, Y., Guan, H., You. J. CLOPE: A fast and Effective Clustering Algorithm for Transactional Data /KDD '02 Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. July 23-26, 2002, Edmonton, Alberta, Canada, pp. 682-687.
8. Wang, K., Xu, C., Liu, B. Clustering transactions using large items. In Proc. CIKM'99, Kansas, Missouri, 1999.
9. Agrawal R., Imielinski T., Swami A. Mining association rules between sets of items in large databases. / In Proceedings of the ACM SIGMOD Conference on Management of Data, Washington D.C., May 1993, pages 207-216.

Відомості про авторів

Касумова Рена Тофік кизи – голова відділення Азейбаржанської національної академії наук, інституту інформаційних технологій, вул. Агаєва, 9, Баку, Азейбаржан. (99412) 439-85-48, kasumova-rena@rambler.ru.