About the Identification of Categorical Registration Data of Domain Names in the Data Warehouse

Rena Gasimova
Institute of Information Technology of ANAS, Baku, Azerbaijan
kasumova-rena@rambler.ru

***Abstract*— This work is dedicated to formation of data warehouse for processing of a large volume of registration data of domain names. For this purpose, fuzzy record comparison algorithms are for clearing of registration data of domain names reviewed in this work. Also, identification method of domain names registration data for data warehouse formation is proposed.**

Keywords*— *domain name system; registrar; registrant; category data; data warehouse; data clearing; fuzzy search algorithms; Damerau-Levenstein distance; decision tree

I. Introduction

Modern approach to automation of decision making support is based on use of data warehouse (DW) concept. Rampant development of information technologies, and data collection, storage and procession means in particular, allows to collect vast volumes of data which require analyzing. DW provides analysts, executives and top managers with capability to study large volumes of interdependent data using fast interactive information reflection on different levels of detailing from different points of view in accordance with notion of the user on subject field [1, 2].

Bill Inmon, author of DW concept (1992 y.), defined them as "subject oriented, integrated, unchanged, supporting data collection, organized to support management", designed to act as "unique and only source of truth", providing managers and analysts with trustworthy information necessary for operative analysis and decision making [3, 4].

Richard Hackathorn, another founder of this concept, wrote that the objective of DW – is to provide "unique image of existing reality" for organizations [5]. Ralph Kimball, one of DW authors, described the warehouse as "a place, where people can access their data". He also formulated in [6, 7] main requirements to data warehouses.

For correct comprehension of data concept, it is necessary to understand following principal moments: DW concept – is not a data analyzing concept; rather it is a concept of preparation for data analyzing; DW concept does not predetermine architecture of purposive analytical system.

II. Relevance

Apparently, at modern development level of information technologies, simple relocation of data from one DW to another does not cause difficulties. Filling of warehouses, as a rule, is carried out by information from several data sources. Human factor and partial absence of control at submittal lead to occurrence of distortions in data. Misprints and omissions are present almost in all details of saved objects, as well as in identification sets.

Exactly presence of point sets of information collection makes the clearing process especially relevant. Generally speaking, errors are always committed, and it is impossible to completely dispose of them.

There are different kinds of errors. There are errors characteristics to a certain subject field or task. Errors, which do not depend of task: contrariety of information; data omissions, anomalous values; noise; data entry errors etc. There are different kinds of solutions for each of these problems. Data omissions are a very serious problem for majority of DW.

III. Objective

Clearing of domain name registration data is carried out in the works and domain names registration data identification method is developed based on decision tree apparatus application.

Domain is a domain name space field and is characterized with independence of data allocation, inclusion of information system in domain contents, presence of special information systems (DNS servers) containing data on domain names allocated in domain and carries out the function of domain name space organization [8].

Registration data of domain include: domain name (domain), registry identifier (registrar), full name of the physical person (person), contact address of the physical person (address), domain administrator identifier (admin-o), organization identifier for administrative communication (admin-c), title of organization (organization), domain registration time (created), domain free date (free-date), telephone numbers with international codes (phone), e-mail address (e-mail), list of DNS servers supporting domain (nserver), domain type (type), information source (source), domain registration payment time (paid-till) [9].

Following tasks are formulated for research purposes:

1) Processing of domain name registration data, data clearing using Damerau-Levenstain algorithm;

2) identificattion of domain name registration data using deicison tree construction

IV. Solution

A. Algorithm choice

Generalized string matching task which includes detection of substrings of text strings is also called fuzzy string matching task. One of the most popular fuzzy string matching methods is editing distance calculation methods [10-12]. Generally, editing distance means metrics, numerically calculating the value of transformation of one string to another. There are different several operations, each of which can have a value of its own: character stuffing, deleting, replacement and transposition of proximate symbols.

There are different fuzzy string matching algorithms, which are based on different editing distances. Hamming distance – is a number of positions, in which corresponding symbols of two words of the equal length are different [13].

If matching of two strings of different lengths is allowed, then as a rule, insertion and deletion are also required. If they are given the same weight as replacement, minimal general value of transformation will be equal to one of the metrics proposed by Levenstein [14].

Upon entry of domain name registration data in DW, abbreviations, misprints, omissions, double recordings and other distortions are encountered. In order to increase the quality of input registration data such as "registrar", person", "address", "organization", "admin-o" etc, prevention of errors and inconsistencies of duplications in records is required. For example, in names of countries (cities) misprints can be as Kanata (Canada), Russian (Russia), Frankfrut Am Main (Frankfurt Am Main) etc; organization identifier registration data (registrar) "MONIKER ONLINE SERVICES, INC" can be identified as "MONIKER, INC", "MONIKER", "MONIKERS ONLINE SERVICES". In this case, words included in phrases must be processed separately. Despite difference of these strings, it is clear that, all of these titles stand for the same registrar. But let's also note that, upon comparing strings "MONIKER ONLINE SERVICES, INC" and "MONIKER, INC" using Levenstein metrics, we receive a larger value for editing distance.

Let's use Damerau-Levenstein distance – difference measure of two strings of symbols, defined as minimal quantity of insertion, deletion, replacement and transposition (rearrangement of two proximate symbols) operations necessary for transfer of one string to another. It is the modification of Levenstein distance, and differs from it by addition of transposition operation.

B. Symbol fields for registration data recording

As a rule, symbol fields consist of a string which contains one or several words, divided in gaps and punctuation marks. Data is entered manually by an operator, often in distorted condition. In this regard, punctuation marks such as ".", ",", ":", """, "-" etc that do not carry a functional significance, are replaced by "blank" sign.

Name of physical person (person), contact address of physical person (address), domain administrator identifier (admin-o), organization identifier for administrative communication (admin-c), organization title (organization) are the key fields upon identification of registration data.

Components of these fields can be present in random order, which is a difficult task for automatic processing of such information. For content analysis, we enter so-called templates. For example, address template is a combination of address components:

[apartment], [block], [building], [street], [city], [district], [index], [country]

For example, for domain – azerbaijan.info, address template is important:

, , 4676, Admiralty Way, Marina del Rey, California, 90292-660, US.

C. Decision tree

Decision trees is the most comfortable decision making method for record (object) identification, for their demonstrativeness while use, minimal calculation resources and simplicity of realization.

Decision trees – is a method of rule presentation in hierarchic, consecutive structure, where each object corresponds to single knot that gives decision. Under rule, we understand a logical construction, presented in "if…then" form. Main advantages of decision trees are generation of rules in fields where experts formalize their knowledge with difficulty; extraction of rules in natural language; intuitively understandable classification model; high forecast precision, comparable to other methods (statistics, neural networks); construction of non-parametric models [15].

Currently, there is a significant number of algorithms, realizing decision trees CART, C4.5, Newld, ITrule, CHAID, CN2 etc [16, 17]. But following two are the most widespread and popular: CART (Classification and Regression Tree) and C4.5.

We will be generating the decision tree from available decision tables. For precise object identification, we will be using fields unequivocally identifying the object. Domain registration data will be identified by following attributes: registrar, person, address, admin-o, admin-c, organization, created, updated, free-date, phone, e-mail, nserver, type, source. Intermediate nodes of the tree correspond to these attributes, and arches – to possible alternative comparison values of these attributes "+" (same), "±" (similar), and "–" (different). Tree leaves are indicated as one of three classes as "=" (compared objects are identical), "≠" (objects are different), "?" (unknown).

Following are the most significant among attributes: registration identifier (registrar), full name of the physical person (person) contact address of the physical person (address), domain administrator identifier (admin-o), organization identifier for administrative communication (admin-c), title of organization (organization), list of DNS servers supporting domain (nserver).

Less significant, but as informative attributes are: domain registration time (created), domain update time (updated), domain free date (free-date), telephone numbers with international codes (phone), e-mail addresses (e-mail), domain type (type), domain registration payment time (paid-till), information source (source) etc.

Decision tree is formed based on the knowledge of experts of the subject field. Depth of the tree is selected based upon objects necessary for precise identification. Conducted experiments with above-described decision tree gave positive results.

V. Conclusion

Domain names registration data category identification was reviewed in this work. Abbreviations, misprints, omissions, conscious data corruption, record duplicates etc which were allowed on information collection stage are considered as errors.

Currently, variety of alternative proximity functions were proposed, but from our point of view, in order to conduct clearing and consistency checking, Damerau-Levenstein distance most accurately corresponds to intuitive similarity concept. Also, domain name registration data records identity method based on decision tree was proposed.

Developed identification method was implemented for creation of a DW on information from several DNS Servers based on the example of domain name registration data serving the interests of the Republic of Azerbaijan.

References

[1] V.V. Przhiakovski. Complex analysis of large volume data: new perspectives of computerization // SUBD, 1996, № 4, pp. 71-83.

[2] A.A. Sakharov. Construction concept and realization of information systems oriented on data analysis // SUBD, 1996, № 4, pp. 55-70.

[3] W. H. Inmon. Building the Data Warehouse. Second Edition. New York: John Wiley & Sons, Inc.1993, p. 298.

[4] W. H. Inmon. Building the Data Warehouse. Third Edition. New York: John Wiley & Sons, Inc. 2002, p. 412.

[5] W. H. Inmon and R.D. Hackathorn. Using the Data Warehouse. New York: John Wiley & Sons, 1994, p. 285.

[6] R. Kimball. The Data Warehouse Toolkit, New York: John Wiley & Sons, 1996, 388 p.

[7] R. Kimball, M. Ross. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, Second Edition, John Wiley & Sons, 2002, p. 464.

[8] A.G. Sergo. Domain Names. M.: Bestseller», 2006, p. 368.

[9] A.A. Venedrukhin. Domain Wars SPb.: Piter, 2009, p. 224.

[10] D.M. Sunday. A very fast substring search algorithm // Communications of the ACM, vol. 33, No. 8, 1990, pp. 132-142.

[11] A. Hume, D. Sunday. Fast string searching // Software – Practice and Experience, Vol. 21, No. 11, 1991, pp. 1221-1248.

[12] A.V. Aho. Algorithms for finding patterns in strings. Handbook of Theoretical Computer Science, Elsevier Science Publishers, Amsterdam. Vol. A, Algorithms and complexity, J. van Leeuwen ed., Chapter 5, 1990, pp. 255-300.

[13] R.W. Hamming. Coding and Information Theory. Englewood Cliffs, N.J.: Prentice-Hall, 1980, 239 p.

[14] V.I. Levenstein. Binary codes capable of correction deletions, insertions and reversals. (Reports of AS of USSR) Soviet Physics Doklady1965, 163.4, pp. 845-848.

[15] O.G. Berestneva, E.A. Muratova. Construction of logical models using decision tree // Tomsk Polytechnic University News, v. 207, issue 2, 2004, pp. 55-61.

[16] J.R. Quinlan. C4.5 Programs for Machine Learning. Morgan Kaufmann pub., San Mateo, CA, vol. 240, 1993, 302 p.

[17] J.R. Quinlan. Improved Use of Continuous Attributes in C 4.5. //Journal of Artificial Intelligence Research, vol. 4, 1996, pp.77-90.