PRACTICE OF WEB DATA MINING METHODS APPLICATION

WEB DATA MINING METOŽU LIETOŠANAS PRAKSE

P. Osipovs, A. Borisov

Keywords: web data mining, duplicate document detection, internet users behavior patterns

Abstract - Recent growth of information on the Internet imposes high demands on the effectiveness of processing algorithms. This paper discusses some algorithms from the field of Web Data Mining which have proved effective in many existing applications. The paper is divided into two logical parts; the first part provides a theoretical description of the algorithms, but the second one contains examples of their successful use to solve real problems. Search algorithms of vague duplicates of documents are currently actively used by all the leading search engines in the world. The paper describes the following algorithms: shingles, signature methods and image-based algorithms. Such methods of classification as a method of fuzzy clustering to-medium (Fuzzy c-means/FCM clustering) and clustering by ant colony (Standard Ant Clustering Algorithm SACA) are considered. In conclusion, the experience of the successful application of fuzzy clustering in conjunction with the software toolkit DataEngine to improve the efficiency of the bank «BCI Bank» is described as well as the sharing of the ant colony clustering method in conjunction with linear genetic programming to meet the increasing efficiency of predicting the load on the servers of high load Internet portal Monash Institut.

Introduction

The meaning of Web Data Mining (WDM) is growing together with Internet meaning for current society, and this is an almost linear dependence. New WDM methods and algorithms are created and old ones are improved. This paper considers some of most effective methods and reviews their usage in real projects. The paper is based on some publications in the area.

The considered algorithms are successfully used in banking services, business intelligence, medical researches and all biggest Internet search engines.

Vague duplicate document detection algorithms include*:*

- shingle-based algorithms;
- signature methods;
- image-based algorithms.

While analyzing article [1] about practical application of methods for finding behavioral templates of bank Internet site users, some most frequently used methods will be discussed, e.g.,

- Fuzzy c-means;
- Data Mining tools software: "*DataEngine*".

Also an example is considered of successful application of clusterization using Artificial Ant Colony algorithms together with Linear Genetic Programming in real project [2]. In addition, the basics of artificial ant algorithms and software package Discipulus™ [3] are described.

Vague Duplicate Document Detection (DDD)

The main tasks of DDD [4] are listed below:

- Duplicate document detection; it is used in all main Internet search engines, such as Google, Bing, Yandex to disallow excessive indexing of similar documents;
- Detection of plagiarism;
- Document archiving (to optimize space required to store documents by refusing from storing duplicate documents);
- Document clustering by the values of their similarity (used for more relevant search results, when from each of allocated information clusters most relevant documents will be returned);
- Spam search and filtering.

Shingles Algorithm

Shingles [5,6] algorithm is employed to get the numerical value of two documents similarity. In the simplest case to detect documents content identity, checksum calculation can be used. But this technique is 100% sensitive to even minimal document content changes; even after changing one character the final checksum will be totally different.

One possible way to avoid this is to use Shingles algorithm, which allows one to determine percentage value of two documents similarity.

Shingles Algorithm Description

The main idea of Shingles algorithm is to divide the whole document text into some equal parts, for example by 10 words. Before this separation, the text is

reduced to canonical form by removing prepositions, conjunctions, possible HTML tags; also some researches recommend removing adjectives, too, because they almost do not bear the semantic load but can be used in some automated software to add semblance of uniqueness.

The second most important feature of Shingles algorithms is that all word sequences are taken jointly but not in the overlap, which disallows possible information loss at block borders. For every shingle the value of some hash function (CRC32, MD5 etc.) is calculated and after that hashes (signatures) of two randomly selected documents are compared. Even a single coincidence will be the sign of large probability of documents similarity.

The third advantage of shingles algorithm is time complexity. A lot of common algorithms require to compare all analyzed documents with each other, which gives $O(N^2)$ complexity (where N – count of analyzed documents), instead, using shingles can give complexity close to $O(N * \log(N))$ [5]; however with large values of N all information cannot be processed by one computer; requirements of distributed computing appear, which increases the value of complexity approximately back to $O(N^2)$.

Currently, the algorithms based on shingles are actively developing. One of promising directions is super-shingles [6]; in this case the values of all document shingles are first clusterised, and whole clusters signatures are then compared. This approach has a much better execution speed, but the author himself pointed to unsatisfactory results when comparing small documents.

Also the idea of shingles union has evolved in mega-shingles algorithm, which is a pairwise combination of six super-shingles. In this case documents are declared similar, if even one mega-shingle is the same in both documents.

Term-Based Algorithms

Unlike Shingles, in Term-Based Algorithms [7] the main measure unit is a separate word. This class of algorithms is much more focused on semantic, but not syntax document similarity, refusing from document structure and clearance analysis. The creation of dictionary of most common words in all analyzed documents is employed; as a measure of closeness the value of intersection $\cup$ between analyzed document keywords set and common keywords set is used.

The numerical value of similarity can be obtained using the metric of semantic similarity:

$$r(A,B) = \frac{|S(A) \cap S(b)|}{|S(A) \cup S(B)|} \quad (1)$$

where

$r(A,B)$ - value of A and B documents similarity;

$S(A); S(B)$ - sets of A and B documents keywords (k-grams).

In the analysis, each document should be compared with others, which gives $O(N^2)$ algorithm complexity. It is believed that this algorithm is more complicated than Shingles, and therefore it is rarely used.

Alternatively, instead of constructing a dictionary using the method *Tf-IDF* [8] (*TF* — term frequency, *IDF* — inverse document frequency) is suggested when each document is represented as a numeric vector reflecting the importance of each word in the document. The dimension of the vector depends on the number of words in a common set. This approach is called a vector model (VSM) and allows you to compare documents using cluster analysis.

More Information about the *Tf-IDF* Method

The *Tf-IDF* method is used for statistical evaluation of the importance of words in the context of the individual, but being part of the collection instrument. The weight of each word is proportional to its use in the document and inversely proportional to the frequency of its use in other documents.

This measure is the product of two relations:

Term Frequency(TF) – is the ratio of the number of occurrences of a word to the total number of words in the document

$$TF = \frac{n_i}{\sum_k n_k} \quad (2)$$

where

n_i is the number of occurrences of words in the document;

$\sum_k n_k$ the total number of words in the document.

And *inverse document frequency (IDF)* - inversion of the frequency with which a word occurs in the documents collection (feature of the IDF in the fact that this measure reduces the weight of frequently used words)

$$IDF = \log \frac{|D|}{|d_i \supset t_i|} \quad (3)$$

where
$|D|$ - number of documents in the collection;
$|d_i \supset t_i|$ - number of documents where t_i is present.

As a result of using that measure, the bigger weight have the words with high frequency of usage in one document and low frequency in others.

Image-Based Algorithms

In this group of algorithms [9,10] the documents are submitted and compared in the form of images. One of the simplest algorithms of this area is *GQView* [11]. Its idea is simple: the image pixel by pixel is divided into squares measuring 32 by 32 pixels, and then takes the average color value of pixels in each block. In this case, the difference is the sum of the values of differences for identical blocks of two images, and the value of difference for each block is normalized to the range $0 \div 1$.

$$dif(img1, img2) = \sum_{i=1}^{n} norm(img1_i - img2_i) \quad (4)$$

where
$dif(img1, img2)$ - value of difference between image 1 and image 2;
n – number of pixel blocks in each image;
$img1_i$ - average value of color in the i-th block;
norm – normalization procedure.

Classification Systems of Users Based on Behavioral Patterns

One of the three main components of Web Mining is the Web Usage Mining. The main purpose of this direction is analysing and modelling patterns of user behavior on the Internet resource for adaptive tuning of information issued for the most relevant data to individual users, or prediction of the behavior of users in the future.

The major directions in which Web Usage Mining is used are as follows:

- Internet portals of banks;
- business Intelligence;
- forecasting of attendance at high-load projects;
- Internet shops personification systems.

Web Usage Mining for Banks Internet Portals

Before we proceed directly to the description of the study, the methods used in it should be described.

Fuzzy c-means/FCM Clustering

When using the algorithms of fuzzy clustering, each object can belong to more than one cluster.

As a result of method execution, for each object under consideration, x, there is returned not cluster number to which it belongs, but the probability of belonging to the k-th cluster $u_i(k)$. Usually the sum of these probabilities is 1.

$$\forall x(\sum_{k=1}^{n} u_k(x) = 1) \quad (5)$$

where
n – number of clusters, for which probability of belonging to cluster is larger than 0.

Using this algorithm, the center of each cluster is determined as the average value of all its points, with the weights of the probability of belonging to the analyzed point cluster:

$$center_k = \frac{\sum_x u_k(x)^m x}{\sum_x u_k(x)^m} \quad (6)$$

where
m – exponential weight (*m* > 1).

Usually *m* is chosen equal to 2, but this choice is not theoretically justified, so some of the methods offer other values. There are studies [12], recommending to choose the values of *m* in the range $3.2 \div 3.7$.

The measure of point membership in a cluster is inversely proportional to its proximity to the center of the cluster:

$$u_k(x) = \frac{1}{d(center_k, x)}. \quad (7)$$

The fuzzy c-means algorithm is very similar to the classic k-means, with the exception that the point can belong to multiple clusters simultaneously.

Features of DataEngine Package

Multi-functional suite of software tools DataEngine [13] is a complete solution for working with various Data Mining algorithms. The package uses a variety of approaches to solving Data Mining tasks.

- Client - server architecture to provide scalability;
- Algorithms of fuzzy calculations;
- Automated creation of neural networks of various topologies;
- Kohonen networks;
- Algorithms for normalization of different scales;
- Different strategies for handling missing and abnormal values;
- Rich visualization availabilities;
- Standard interfaces for creation add-on modules and API for automation tasks.

Description of the Study

In the "BCI Bank" in Santiago, Chile a study was conducted that clearly showed the benefits derived from the introduction of Web Mining techniques in the banking sector.

Since for that bank the cost of processing customers' transactions made through the Internet is approximately 90% lower than in conventional offices, there is a real benefit of attracting the largest possible number of customers to use exactly this type of payment.

Despite more favorable prices, most of customers prefer to use classical methods of use of banking services, thus the main task was to classify customers by their opinions on their Internet services on the basis of existing customer profiles and in the future to conduct explanatory work with them to demonstrate all the advantages of using Internet services.

In the beginning, segmentation of user profiles for different attributes was held. Also, the procedure of profiles anonymization was carried out, mainly for privacy purposes, but also because the attribute «name» in itself is not important information.

The following summary table was obtained:

Table 1

Sample table with selected parameters

Attr. 1	Attr. 2	Attr. 3	Attr. 4	Attr. 5
38	1	1702	62	234
26	0	833	21	34
41	0	500	58	123
40	1	1240	46	314

Then, using the fuzzy clustering algorithm (fuzzy c-means) customer segmenting by 5 classes was performed.

- **L1** – «Young», rarely used Internet banking services, operate mostly in small amounts.
- **L2** – «Very young» still less use of Internet transactions.
- **M1** – «Old», use the Internet fairly frequently, operate with greater amounts than the «young».
- **M2** - «The average age», use the Internet fairly frequently, operate with greater amounts than the «young».
- **H** - «Modern», actively use all the possibilities of the Internet.

Then the whole set of profiles was divided into two parts, training sample 20% and primary part 80%, respectively. Using the software DataEngine MLP (multilayer perceptron) neural network was established, which was trained to distribute profiles of the classes at the 20% sample.

After training, the whole base of profiles of bank clients was forwarded at the input of the network and classification was made.

As a result, data were obtained that 21% of profiles are potentially active users of the Internet portal of the bank and they were sent booklets detailing the advantages of using this type of payment.

The results of the study showed an increase in the number of users of Internet transactions by 1.4%, which means a good increase in profits for the bank.

Prediction of Attendance for High Load Projects

Another important area of Web Mining is the prediction of attendance Internet resources. The most important and main source of data for this are the server requests logs and tools for their processing and analysis. Also, their advantage consists in the prevalence, because servers generate them automatically; whereas specialized tools for collecting statistics though allow all required information, but need to be installed and configured.

From the logs of server queries the following data can be extracted:

- user navigation paths;
- browsing time;
- the structure of hyperlinks and content pages.

All this can increase efficiency in areas such as e-business, e-services, e-learning, getting new members, retain existing ones, increase the effectiveness and usefulness of marketing and advertising companies.

Let us first consider two important technologies used for solving problems of that type.

Standard Ant Clustering Algorithm (SACA)

Clustering using ant algorithms, along with other «biological» methods, began to develop rapidly in the early 21 century. It was then suggested and confirmed that the synergistic effect of the use of algorithms similar to the natural, together with the ever increasing computing power of computers can give excellent results in various fields of science.

Consider a classical clustering method of ant colony in more detail. This method is based on a model of behavior when searching for food, when first the ants are moving in random directions, but as soon as food is found, the most profitable paths are marked by ever growing amount of pheromone, which, however, is eroded over time, if this path is no longer used.
Movement at each iteration is determined by the following formula:

$$P_i = \frac{\left(l_i^q * f_i^p\right)}{\sum_{n=0}^{N} \left(l_n^q * f_n^p\right)} \qquad (8)$$

where

P_i - probability of moving using the *i*-th path;

l_i - length of the *i*-th path;

f_i - count of pheromone on the *i*-th path;

q – algorithm's «greedy» factor;
p – algorithm's «herd» factor.

Despite the fact that the algorithms of this type issue an approximate result, but because of their statistical properties an increase in the number of iterations increases the accuracy of the result.

In continuation of this ideology, a lot of different options for increasing the classical algorithm were offered, e.g., ACLUSTER [14], an adaptive ant clustering method [15] and the method ATTA [16].

Capabilities of the Software Discipulus™

This package is based on genetic algorithms and is designed to solve a variety of regression and classification problems. Additionally it allows the use of algorithms of neural networks, classification trees, Support Vector Machines and some others.

It is alleged that the main advantage of this software system is its unprecedented high speed (thanks to the proprietary technology AIMLearning™), which allows one to make huge gains in speed compared with other modeling tools.

Linear Genetic Programming (LGP)

The concept of LGP is the development of ideology of genetic programming (GP), but the main difference is that if in a simple GP chromosomes (sites under construction tree program) can only be basic operators or variables programming language, in more complex cases - the whole logic blocks are already used, such as functions or their sequence (subprograms). Also the most important feature is that there are used imperative programming languages (such as C) in LGP, in contrast to the functional languages (e.g., LISP) in GP.

Commonly some array of registers r is used, which can contain variables and constants. Also one of the variables (usually r[0]) is used to display the values of the chromosome. This classic version is also called Single Solution Genetic Programming (*SS-LGP*).

In the course of execution, the algorithm uses 2 types of operations: crossover and mutation. Crossover exchanges the instructions from the parents. The mutation is possible in two versions:

- Macromutation - can add or delete a random instruction;
- micromutation – can delete or modify an operator or instructions constant.

There are no other important differences from the genetic algorithms. Iterations occur, each of which calculates the value of fitness function, on whose basis the natural selection and testing stops is carried out.

To illustrate the use of Web Mining techniques in this area, let us consider the research conducted at Monash University [17]

The average load of servers of the University is estimated at 7 million hits a day, but depending on the season, month, day, week or even hour, the load may vary very considerably, and their effective forecasting enables adjusting effective policies to configure load on the servers. The main purpose of this study was to compare the effectiveness of clustering by ant colony and the subsequent building of classifier program using the methods of linear genetic programming, as compared to other possible methods.

The subject of the study was pattern identification in the use of Internet portal Monash University to predict the distribution of loads on the servers.

Initially a training set was allocated, the data from server log for a certain period of time have been

cleaned up and clustering was made using ACLUSTER method according to parameters such as the number of bytes requested from the domain, the number of requests per hour and per day.

The result was the clustering of the number of requests to all domains with regard to the time of their commission. Of course, depending on the parameters of the algorithm the results somehow varied, but it is a property of this kind of approximate algorithms, moreover, due to its probabilistic nature, the results are improving as the number of iterations increases.

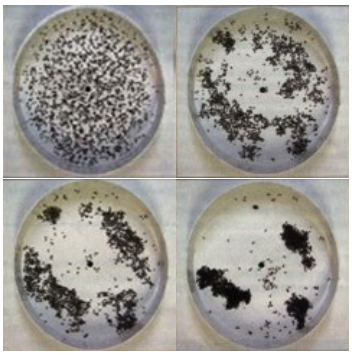

Fig. 1. Improving the quality of clustering using the algorithm ACLUSTER with an increase in the number of iterations (from the authors' algorithm [18])

Furthermore, using genetic methods for linear programming, it was necessary to build a classifier to predict loads.

To create a classifier program, the Discipulus™ package has been used which employs the ideology of linear genetic programming.

As a result of different settings of that package, several classifier programs have been created , of which the one having the best results was chosen.

The correlation coefficient between actual and predicted results of the program was 0.9921 for hourly predictions and 0.9963 for the daily prediction.

The main purpose of the study was to compare the effectiveness of using this group of algorithms as compared to other algorithms. The results of the study are provided below.

The following abbreviations were used:

ANT-LGP considered algorithm, ant colony + linear genetic programming;

i-Miner – hybrid system of fuzzy clustering + fuzzy output (same authors);

SOM-LGP – self-organizing maps + linear genetic programming;

SOM-ANN – self-organizing map + neural network.

Table 2

Comparison results of different methods of forecasting

Method	Values		Correlation coefficient
	Real	Test	
ANT – LGP	0.2561	0.035	0.9921
i-Miner	0.0012	0.0051	0.9981
SOM-LGP	0.0546	0.0639	0.9493
SOM-ANN	0.0654	0.0516	0.9446

Table 2 demonstrates that in solving the problem this group of algorithms has showed results comparable with the best.

Conclusions

The use of modern methods of Data Mining provides a real economic effect. Also, based on analysis of actual implementation experience, it can be stated that the combination of several methods for solving subproblems is reasonable and gives the expected effect.

References

1. Araya S., Silva M., Weber R. Identifying web usage behavior of bank customers // Berlin: Springer, October 2003. – P 951-958.
2. Ajith A., Vitorino R. Web Usage Mining Artificial Ant Colony Clustering and Linear Genetic Programming // CEC´03 - Congress on Evolutionary Computation / IEEE Press, ISBN 078-0378-04-0, 8-12 December 2003. - Canberra, Australia, P. 1384-1391.
3. Software Discipulus™ Web Site URL: http://www.rmltech.com/ - Visit date September 2009.
4. Ye S., Wen J., Ma W. A systematic study on parameter correlations in large scale duplicate document detection // Knowledge and Information Systems. 14(2), (2008), - University of California Postprints P. 217-232.
5. Manber U. Finding similar files in a large file system // Usenix Winter 1994 Technical Conference, January 1994.
6. Broder A., Glassman S., Manasse M. Syntactic Clustering of the Web // Sixth World Wide Web Conference, September 1997.

7. Chowdhury A., Frieder O., Grossman D., McCabe M.C. Collection statistics for fast duplicate document detection // ACM Transactions on Information System 20(2), (2002) P. 171-191.
8. Salton, G. and McGill, M. J. Introduction to modern information retrieval // New York McGraw-Hill, ISBN 0-07-054484-0 Chapter 3, (1983) P. 52-117.
9. Daniel P. L. Models and Algorithms for Duplicate Document Detection // Information Retrieval, Kluwer Academic Publishers Hingham, MA, USA Volume 4 , Issue 2, (2001) P. 153-173.
10. Bharat K., Broder A. A systematic study on parameter correlations in large-scale duplicate document detection // London: Springer ISSN 0219-1377, March 2007. - P. 217-232.
11. Description of GQView algorithm URL: http://www.elliotglaysher.org/2006/03/19/duplicate-image-algorithms/ - Visit date September 2009.
12. Киселева Е.М., Блюсс О.Б. Особенности некоторых алгоритмов многокритериальной нечеткой кластеризации // Questions of Applied Mathematics and Mathematical modeling, Dnepropetrovsk National University of Oles Gonchar KB № 5713 (2008) (In Russian).
13. Software DataEngine Web Site URL: http://www.dataengine.de/ - Visit date September 2009.
14. Ramos V., Muge F., Pina P. Self-Organized Data and Image Retrieval as a Consequence of Inter-Dynamic Synergistic Relationships in Artificial Ant Colonies // Soft Computing Systems Design, Management and Applications, 2nd Int. Conf. on Hybrid Intelligent Systems, IOS Press, 2002.- P. 500-509.
15. Vizine A. L., de Castro L. N. et al. Towards Improving Clustering Ants: An Adaptive Ant Clustering Algorithm // Informatica 29 (2005) ISSN 0350-5596, P. 143-154.
16. Handl J., Knowles J., Dorigo M. Ant-based clustering and topographic mapping // Artificial Life Volume 12 , Issue 1, Cambridge, MA, USA: MIT Press ISSN:1064-5462, January 2006. – P. 35-61.
17. Monash University Web site URL: http://www.monash.edu.au/ - Visit date September 2009.
18. Artificial Ant Colonies alghoritm description URL: http://www.chemoton.org/ref39.html - Visit date September 2009.

Pavel Osipov is a Ph.D. student in the Institute of Information Technology at Riga Technical University. He received his master's diploma in Computer Science from Transport and Telecommunication Institute. His research interests include web data mining, machine learning and knowledge extraction.

Arkady Borisov is Professor of Computer Science in the Faculty of Computer Science and Information Technology at Riga Technical University. He holds a Doctor of Technical Sciences degree in Control in Technical Systems and the Dr.habil.sci.comp. degree. His research interests include fuzzy sets, fuzzy logic and computational intelligence. He has 205 publications in the area.

Pāvels Osipovs, Arkadijs Borisovs. Web Data Mining metožu lietošanas prakse
Interneta informācijas pieauguma tempi mūsdienās izvirza augstas prasības informācijas apstrādes algoritmu efektivitātei. Šajā rakstā ir apskatīti daži Web Data Mining jomas algoritmi, kuri ir pierādījuši savu efektivitāti vairākās eksistējošās lietojumprogrammās. Raksts ir sadalīts divās daļas, no kurām pirmajā ir dots teorētisks algoritmu apraksts, bet otrajā piemēri to veiksmīgam pielietojumam uzdevumu atrisināšanai. Dokumentu neprecīzo dublikātu meklēšanas algoritmi tiek aktīvi izmantoti visās galvenajās meklēšanas sistēmās. Ir doti sekojošu algoritmu apraksti: jostas roze, parakstu ‚metodes, uz attēliem balstītie algoritmi, kā arī tādas klasifikācijas metodes kā klasterizācija pēc izplūduša k-vidēja (Fuzzy c-means/FCM clustering) un skudru kolonijas klasterizācijas algoritms (Standard Ant Clustering Algorithm SACA). Noslēgumā aprakstīts veiksmīgs pielietojums izplūdušajai klasterizācijai, pielietojot programmatūras rīku komplektu Data Engine, bankas «BCI Bank» efektivitātes paaugstināšanas uzdevumā. Tāpat aprakstīta vienlaicīga skudru kolonijas klasterizācijas algoritma izmantošana kopā ar lineāro ģenētisko programmēšanu tam, lai piedāvātu ļoti noslogota Interneta portāla „Monash" slodzes prognozes efektivitātes palielināšanos.

Павел Осипов, Аркадий Борисов. Практика использования методов Web Data Mining
Современные темпы роста количества информации в Internet предъявляют высокие требования к эффективности алгоритмов её обработки. В данной работе рассмотрены некоторые алгоритмы из области Web Data Mining, доказавшие свою эффективность во многих существующих приложениях. Статья разбита на две логические части: в первой части рассматривается теоретическое описание алгоритмов, а во второй - примеры их успешного использования для решения реальных задач. Алгоритмы поиска нечётких дубликатов документов активно используются всеми ведущими поисковыми системами в мире. В статье приведены описания следующих алгоритмов: шинглы, сигнатурные методы, алгоритмы, базирующиеся на изображениях. Рассмотрены такие методы классификации, как кластеризация методом нечётких к-средних (Fuzzy c-means/FCM clustering) и кластеризация методом муравьиной колонии (Standard Ant Clustering Algorithm SACA). В заключение описан опыт успешного применения методов нечёткой кластеризации совместно с программным инструментальным пакетом DataEngine для повышения эффективности работы банка «BCI Bank», а также - совместное использование кластеризации методом муравьиной колонии и линейного генетического программирования для решения задачи увеличения эффективности прогнозирования нагрузки на серверы высоконагруженного Internet портала института Монаш.