Regression Methods in the Authority Identification within Web Discussions

Kristína Machová[1], Jaroslav Štefaník[1]

[1] Department of Cybernetics and Artificial Intelligence, Technical University, Letná 9,042 00, Košice, Slovakia
kristina.machova@tuke.sk, jaroslav.stefanik@student.tuke.sk

Abstract. The paper describes the problem of authority identification within web discussions solving using linear and nonlinear regression methods. The goal is to find an approximation of dependency of the authority value on variables representing parameters of the structure and particularly the content of selected web discussions. The approximation function can be used at first for computation of the authority value of a given discussant, at second, for discrimination of an authoritative discussant from non-authoritative contributors to the web discussion. This information is important for web users, who search for truthful and reliable information in the process of decision making about important things. The web users would like to be influenced by some credible professionals. The various regression methods were tested. The best solution was implemented in the Application for the Machine Authority Identification.

Keywords: Authority identification · social web mining · linear regression · non-linear regression · web forums

1 Introduction

We live in the information era. A volume of information, which is discovered each day, is too large and too time consuming to be processed by a human. Everybody from us needs sometimes an access to the relevant supporting information for our decision making. To know the relevance of information we have found, we need information about sources of the obtained information and their credibility. In other words it is important to know the sources, which are authoritative ones. A web forum discussion can be a repository of various kinds of useful information: facts, opinions, ideas, attitudes, and so on. However, useful information is mixed with non-useful or misleading information. Every web user can join the web discussion but many of them have not sufficient experiences or theoretical knowledge about the discussed themes. The web discussion often contains an opinion spam and an information trash. So, it is the matter of principal to search for authoritative discussants to let them influence our important decisions. And just the searching for an authority and its machine identification among all discussants of web forum is our challenge.

To achieve our main goal – machine authority identification, we had to do the following three steps:

1. To find such variables - parameters of the structure and content of the web discussion, which are the most related to the authoritative contributing.
2. To define a dependency of the variable "Authority" of a web discussion on the independent variables selected in the first step. We tried to find an approximation of this dependency using the Linear and Nonlinear Regression [1] based on the method of the Ordinary Least Squares (OLS) [2].
3. To use this approximation function for the discrimination of the authoritative from non-authoritative contributors to the web discussion.

Before starting the machine authority identification, we had to solve a number of technical problems. The first one was the automatic extraction of the conversation content and structure from the web page with the web discussion. The second one was to extract the values of selected independent variables from previously obtained information about the discussion. Another problem was how to obtain the values of dependent variable "Authority" for regression function training. We decided for two alternative ways – to obtain values of "Authority" from human "expert" and to extract them directly from the web discussion as so called "wisdom of a crowd".

2 Authority and Web Discussion

2.1 Web Discussion Group

Our attention was on an authority of a web discussion forum. The discussion group was developed in the society Usenet from the beginning of 80th years of 20thcentury [3]. Two computer specialists Jim Ellis and Truscott have come with a new idea to create a system of rules for the contributions creation. Nowadays, WWW society becomes the main organization, which supports and spreads various platforms for Internet discussion groups using various settings up of different web servers. The internet discussion is represented by a web page, where users insert their contributions (opinions and reactions). Within this paper, the web users joining a web discussion will be called the contributors or discussants. They add their opinions, ideas and attitudes to the web discussion and in this way they create so called "conversational content". The authority identification represents the mining of this conversational content and its internal structure. There are different types of Internet discussion forums according to their scope [4]: a discussion to web article, guestbook, discussion forum etc. The paper focuses on the web discussion dedicated to some given theme.

2.2 Authority Identification in General

The concept "authority" comes from the Latin word "augere". It denotes a person, whose opinions, attitudes or decisions are respected by other members of the group and whose decisions and advices are expected by other members of the group. The authority is derived from the relations between people (web users), positions and hier-

archies [5]. There are many kinds of authorities. For example according to prestige, authority can be:

- *Formal (functional) authority*–coming from his formal position regardless of his personal properties. It is a leadership of a person who is mandated to make decisions. It is obviously the result of a position of a person within an organization.
- *Informal (natural) authority*–is based on someone's personal properties and professional assumptions. Such person has a spontaneous influence on others, because of his persuasiveness and good experiences with his advices/decisions. The people, who let an authority to lead them, reinforce the weight of the authority.

The formal authority can be at the same time the informal one. The formal authority can sometimes change his status to informal and vice versa.

2.3 Authority of a Web Discussion

The virtual web authority has different characteristics as the authority in real life. It is related to the structure of the web, which is based on hyperlinks among web pages. The Google has discovered very complicated relations among web pages and references. Well known tool for the web page authority calculation is PageRank [6]. Other known approaches to the web page authority calculating are HITS algorithm [7] and SALSA [8]. These approaches are also based on an input and output hyperlinks of the evaluated web page. There are also tools of the respected portal "Seomoz", for example MozTrust [9] and Open Site Explorer [10]. All these tools cannot be easily used for calculating of an authority of the web discussion forum.

The authority identification from web discussion forums is a similar problem as web page authority calculation, because authority identification from web discussion is concentrated on web page, the discussion runs on. On the other hand, it is also a different problem, because no input or output references between this page and other pages are taken into account. Only references inside this page between various discussants are considered. These references are represented by reactions on contributions. All mentioned methods (PageRunk, HITS, SALSA, MozTrust and Open Site Explorer) calculate authority of each web page separately. One page leads to one measure of authority. Within the authority mining from the web conversation, not only one but all contributions of the given discussant are evaluated. All information about all contributions related to one discussant has to be concentrated and used for the authority estimation. Nevertheless, we can inspire ourselves by these techniques and take into account the number of references as reactions on an actual contribution.

In our previous work [11], we have taken into account mentioned number of reactions on all contributions of evaluated discussant, but also the number of all contributions of this discussant, the number of reaction of the discussant on the bottom level of the conversation tree (Fig. 3), the polarity matching between opinion of the discussant and opinion of all discussion, the positions of contributions in the discussion tree and the length of his/her contributions. Some of these variables have appeared to be not so important for the precise estimation of the authority. Another problem of this approach was in way of the estimation function generation. For these reasons, we

decided to modify the set of variables - arguments of the conversational structure and to use the regression methods for training the authority estimation function.

3 Used Methods

We tried to solve the problem of the authority estimation within the web discussion forum using a machine learning method based on regression analysis. The regression analysis can be a simple Y = f(x) or a multiple regression, when we are searching the dependency of one dependent variable(*Y*) on more other independent variables (x_1, x_2, …x_N) -see equation (1). These variables are called "regresses" or "predictors".

$$Y = f(x_1, x_2, \ldots, x_N). \quad (1)$$

Within the regression analysis, it is very important to realize, which one of variables is dependent and which are independent. The goal is to describe this relation by a suitable mathematic model, for example by linear or nonlinear function. The result will be a regression curve, which should optimally match the empirical polygon [12].

3.1 Linear and Non-linear Regression

Within the two dimensional space, the linear regression can be described by the equation (2) and is illustrated in the Fig.1 for two dimensional space.

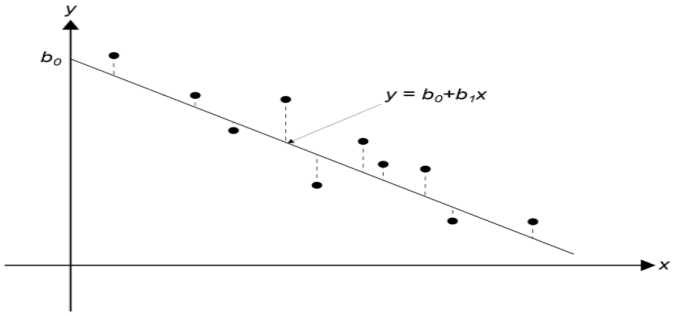

Fig.1. Linear regression in two dimensional space

The goal is to find such values of constants b_0, b_1, ..., b_n (in two dimensional space b_0 and $b_1$of the linear line, see Fig.1) to achieve the optimal matching with the point graph consist of *m* points (observations). These constants can be dedicated from the point estimation using the Ordinary Least Squaresmethod (OLS) [2].

$$y_i = b_0 + b_1 x_{i1} + \ldots + b_n x_{in} + \varepsilon_i \quad (2)$$

Sometimes, it is not possible to find a satisfactory precise linear relation. In this case, the relation can be modeled by somenon-linear function, the most frequently exponential function ($y = ae^{bx}$) or logarithmic function ($y = a + b\,\ell n\, x$) [1].

3.2 Specification of the Variables of a Discussion Structure

We have selected 120 discussants from the portal "www.sme.sk". Consequently, the following variables for each discussant were extracted from all his contributions:

- *AE*– Average Evaluation of the contribution
- *K*– value of the Karma of the user, which is the contribution author
- *NCH*– Number of Characters within his/her contributions
- *AL*–Average Layer in the conversation tree (see Fig.3)
- *ANR* – Average Number of Reactions on his/her contributions
- *NC* – Number of Contributions of given discussant

These variables were used to form the training set (is illustrated in the Fig.2) for selected regression method.

Nickname	AE	K	NCH	AL	ANR	NC
Peter Z	60	108	26	0	1	1
V12	80	182	220	2	0,5	2
fer	80	171	548,5	3	2,5	2
sandokan555	80	162	57,5	4	0,5	2
Peter_S	50	99	112,5	6	0	2
Darkman	80	167	117	3	0	1
Jesse Pinkman	40	74	210,5	1,5	1,5	2
mmm	60	108	22	1	1	1

Fig. 2. Each line of the training set represents one discussant and contains the values of variables *AE, K, NCH, AL, ANR and NC.*

Average evaluation of the contribution (*AE*) is represented by the ratio of the sum of all reactions (agree (+) and disagree (-)) on the contributions of given discussant to the number of all his contributions. This average evaluation is available on the web discussion page. The range of the *AE* is the number from 0 to 80.

Value of *karma* (*K*) of the discussant is also available on the discussion web page. The karma is a number from 0 to 200, which represents activity of the discussant from last 3 months (within the portal "www.sme.sk").

Number of characters (*NCH*) represents the length of discussant contributions. It penalized authors with too short and so less informative contributions.

Average Layer (*AL*) in the conversation tree (see Fig. 3.) is the average number of all layers, which the contributions of the discussant are situated in. The conversation tree is a graphical representation of the web discussion. The *AL* represents the information, when the discussant joined the discussion, from the beginning or at the end.

Average number of reactions (ANR) on the all contributions of the given discussant is the number of reactions per one his contribution.

Number of contributions (NC) is simply the whole number of contributions of the given discussant.

It may happen that a good contribution of already well-known authority finishes the discussion on the Web. It is truth that in such a case there is no reaction on this contribution. It does not disturb the measure of the authority, because of high probability that there were more previous contributions of this contributor with many reactions within the given discussion. These reactions can balance the lack of reactions on the finishing contribution.

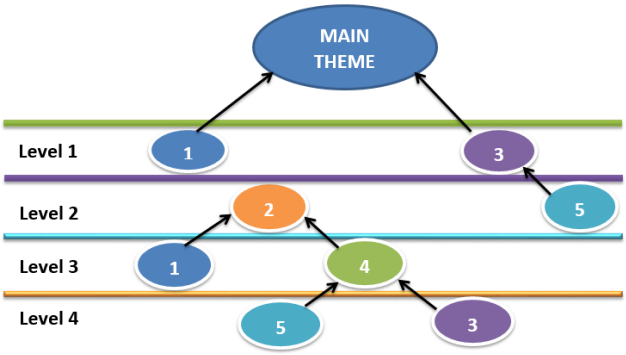

Fig. 3. The conversation tree has 4 levels. The main theme is in the root and reactions are situated on levels 1 – 4. All reactions of the same discussant have the same color.

All these variables were considered to be independent variables. The dependent variable of the regression function *Y* was dedicated from:

1. evaluation of each discussant by "human expert",
2. evaluation of each discussant by other discussants and it represents "wisdoms of the crowd".

4 Implementation and Testing

The authority value A≡Y was estimated by a linear and non-linear function of selected variables (AE, K, NCH, AL, ANR and NC). The four regression functions for authority estimation were generated in the process of machine learning:

1. Linear function learned from the "human expert" (L-EXPERT) is represented by formula (3):

$$A = 0{,}4383AE + 0{,}0746K + 0{,}0281NCH - 2{,}1932AL - 3{,}4386ANR + 8{,}0102NC \quad (3)$$

2. Linear function learned from the "wisdoms of the crowd" (L-CROWD) is represented by formula (4):

$$A = 0{,}4385AE + 0{,}325K + 0{,}002NCH - 0{,}2928AL - 0{,}0853ANR + 1{,}0728NC \quad (4)$$

3. Non-linear function learned from the "human expert" (NL-EXPERT) is represented by formula (5):

$$A = 0{,}0382AE^{1{,}7192} - 0{,}3295K^{0{,}959} + 0{,}4470NCH^{0{,}681} + 0{,}1825AL^{0{,}0001} - 0{,}6269ANR^{3{,}2394} + 20{,}2509NC^{0{,}2977} \quad (5)$$

4. Non-linear function learned from the "wisdoms of the crowd" (NL-CROWD)is represented by formula (6):

$$A = 0{,}0185AE^{1{,}8135} + 141{,}5704K^{-78{,}39} + 0{,}0018NCH^{1{,}0457} - 0{,}0011AL^{3{,}7717} - 0{,}5562ANR^{0{,}0001} + 37{,}6642NC^{0{,}0038} \quad (6)$$

All these functions were created using standard MATLAB functions: "regress" in the case of linear and "lsqnonlin" in the case of non-linear relations. No auxiliary regularization method was used, because the input data matrix was regular. The input data can hardly be considered as noise-data obtained for example from a device. These used input data map the structure of the given web discussion using defined variables. In the case of nonlinear regression, also exponential parameters were elicited from the training data using the function "lsqnonlin". It solves nonlinear least-squares (nonlinear data-fitting) problems and uses numerical optimization method "Trust-Region-Reflective Least Squares Algorithm". The default settings were used, only the number of iterations was extended.

We had considered also polynomial functions to be used for solving the problem of authority identification, but we decided to use a more general form of the function with parameters in its exponents, where exponents need not to be integer values.

All the four functions (from (3) to (6)) were tested. The concise results of these tests are illustrated in Tab.1 and Tab.2.

Table 1. Average deviation of four versions of authority estimation function

Version	Average deviation
L-EXPERT	17,3489
L-CROWD	3,2998
NL-EXPERT	18,1131
NL-CROWD	6,5618

At first, the average deviations were calculated. According to the results in Tab.1, the better functions were obtained by learning from the "crowd" than by learning from the "expert". The deviations for some of tested discussants for the best version L-CROWD are illustrated in Fig.4.

At second, these four versions of regression function were tested using obvious measures of a machine learning efficiency: precision and recall. The regression problem, when the value of A (authority) attribute should be estimated from the interval <0, 100> using formulas (3-6), was adopted to classification problem in the following

way. A threshold T has been stated experimentally (T=70) and discussants were classified into categories: “authority” and “non-authority”. The discussants were classified to the class “authority” when their value of A was equal to or greater than T and they were classified to the class “non-authority” when their value of A was smaller than T. The precision π and recall ρ were computed according to formulas (7) and (8):

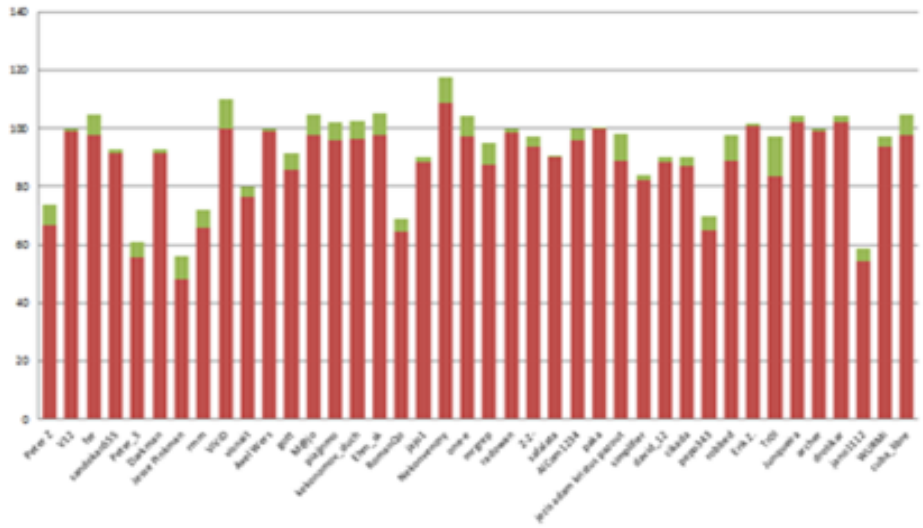

Fig. 4. The values of Authority (red colour) and deviations (green colour) for some of tested discussants for the best version L-CROWD

$$\pi_j = \frac{TP_j}{TP_j + FP_j} \tag{7}$$

$$\rho_j = \frac{TP_j}{TP_j + FN_j} \tag{8}$$

Where:

TP is the number of True Positives (the method classifies these examples as positive (authority) and they are truly positive according to the expert’s (crowd’s) opinion).

FP is the number of False Positives (the method classifies these examples as positive (authority) but they are not positive according to the expert’s or crowd’s opinion).

FN is the number of False Negatives (the method classifies the examples as negative (non-authority) but they are positive according to the expert’s (crowd’s) opinion).

Some key and the most important achieved results of tests are presented in Tab.2.

The linear regression learned from the “crowd”, with best results of testing, was implemented in the Application for the Machine Authority Identification (AMAI). This application provides the list of all discussants with the actual value of their Au-

thority. The AMAI also displays the value of the authority of the discussant, which was selected by a user. This value is from the interval <0, 100>. The application provides not only the binary decision whether the discussant is or is not the authority, but also it provides a precise numeric value of its authority.

Table2. Values of precision and recall of four versions of regression functions were obtained in the three-time cross validation.

		PRECISION		**RECALL**	
Test	Version	**EXPERT**	**CROWD**	**EXPERT**	**CROWD**
Cross val. 12_3	Linear regression	0.78	0.99	0.69	0.99
	Non-linear regression	0.72	0.99	0.66	0.88
Cross val. 13_2	Linear regression	0.65	0.98	0.65	0.93
	Non-linear regression	0.67	0.97	0.67	0.86
Cross val. 23_1	Linear regression	0.68	0.97	0.67	0.67
	Non-linear regression	0.69	0.97	0.69	0.67
Average	Linear regression	**0.70**	**0.98**	**0.67**	**0.80**
	Non-linear regression	**0.67**	**0.97**	**0.67**	**0.80**

5 Conclusions

The design of solving the problem of the authority identification from conversational content using the linear and nonlinear regression was presented. The measure of the authority A was estimated as dependency on variables (AE, K, NCH, AL, ANR and NC) - parameters of the structure and content of given web discussions. The four generated estimation functions were tested. According to the values of average deviations (see Tab.1) the best solution is the linear function learned from crowd (L-CROWD). The second one is the nonlinear function learned for crowd (NL-CROWD). Linear and non-linear functions learned from a single human evaluator – expert – seem to be worse. The same conclusions can be deduced from the resulting average values of precision and recall in Tab.2. It can be hardly said who is the expert on the authority identification. Also an opinion of a psychologist may be subjective. On the other hand, combined opinion of many discussants can be objective.

There are other existing authority identification methods, as Klout, TwentyFeet, My Web Carrer [13] and our previous work [11]. All these methods use formulas for authority estimation, but these formulas were generated more experimentally without considering a theoretically based way. For this reason, we tried to generate the relation between the authority and the structure of web discussion using the classic mathematical approach based on the linear and nonlinear regression.

For the future we plan to elicit the constants of linear and nonlinear equations using evolutionary algorithms [14, 15] in order to calculate not only constant values but the form of a non-linear regression function as well.

Acknowledgements. The work presented in this paper was supported by the VEGA project 1/1147/12 “Methods for analysis of collaborative processes mediated by information systems”.

References

1. Pazman, A., Lacko V.: Lectures from Regression Models (in Slovak). University of Comenius Bratislava, Bratislava, Slovakia, 132, ISBN 978-80-223-3070-1 (2012)
2. Pohlman, J.T., Leitner, D.W.: A Comparison of Ordinary Least Squares and Logic Regression. The Ohio Journal of Science. vol.103, no.5, 118-125 (2003)
3. What is Usenet?, www.usenet.org
4. Internet forum, http://en.wikipedia.org/wiki/Internet_forum
5. Chavalkova, K.: Authority of a teacher (in Czech). Philosophic faculty of the University of Pardubice, Pardubice, Czech republic, (2011)
6. Fiala, D.: Time-aware PageRank for bibliographic networks. Journal of Infometrics.vol.6, no.3, 370-388 (2012)
7. Li, L., Shang, Y., Zhang, W.: Improvement of HITS-based algorithms on web documents. In: 11th International Conference on the WWW, pp.527-535, ACM, Hawaii, USA (2002)
8. Lempel, R., Moran, S.: The stochastic approach for link structure analysis (SALSA) and the TKC effect. Computer Networks: The International Journal of Computer and Telecommunication Networking. vol.33, no.1-6, 387-401 (2000)
9. Hallur, A.: MozRank and MozTrust: Everything you Should Know, http://www.gobloggingtips.com/mozrank-and-moztrust/
10. Fishkin, R.: Open Site Explorer News Link Building Opportunity Section, http://moz.com/blog/open-site-explorers-new-link-building-opportunities-section
11. Machová, K., Sendek, M.:Authoritative Authors Mining within Web Discussion Forums.In:9th International Conference on Systems, pp.154-159, International Academy, Research and Industry Association, Nice, France (2014)
12. Introduction to regress analysis [in Czech], http://www.statsoft.cz/file1/PDF/newsletter/2014_26_03_StatSoft_Uvod_do_regresni_analyzy.pdf
13. Štefaník, J.: Aproximation of the relation of an authority on the parameters of the structure of web discussion (in Slovak).Technical University of Košice, Košice, Slovakia (2015)
14. Mach, M.: Evolution algorithms – problems solving (in Slovak). FEI Technical University, Košice, 135 ps., ISBN 978-80-553-1445-7 (2013)
15. Ćádrik, T., Mach, M.: Evolution classifier systems (in Slovak). Electrical Engineering and Informatics IV. –Proc. of the FEI Technical University of Košice, Košice, 168-172, ISBN 978-80-553-1440-2 (2013)