Authoritative Authors Mining within Web Discussion Forums

Kristína Machová, Michal Sendek
Dept. of Cybernetics and Artificial Intelligence
Technical University of Košice
Košice, Slovakia
kristina.machova@tuke.sk, michal.sendek@student.tuke.sk

***Abstract*—The paper is focused on authoritative users of some web discussion searching and their authority measure estimation. The paper describes design of a method for authority calculation for all discussants of some selected web discussion forum. The designed method can be used in the process of web authority mining. This method involves both the conversational content mining and the conversational structure mining. The resulting implementation can be used for simulation of a dynamic change of the authority measure of social web users. The implementation of the presented method can be used by some firm or other organization for searching authorities (experts) in a special field, in the case, when the organization needs some experts – employees from this field and so such implementation can be used for preselecting on some position in the organization.**

Keywords-authority identification; web mining; conversational content; conversational structure; discussion forums

I. INTRODUCTION

Nowadays, the Web has become the phenomenon of our age. Mainly the social web and its platforms (chats, discussion forums, blogs and so on) enable interactions between actors. The actor, within this paper, is a web user, who has added one or more contributions to a given web discussion and so participated in the conversational content creation. Such web user is also called a contributor within this paper. Within these interactions, web users (contributors) communicate and influence each other. This communication creates so called conversational content, which is an important source of large-scale databases of information about knowledge, opinions, and attitudes of particular users. These data offer many possibilities for web mining from *conversational content*, *conversional structure* and from *conversional web usage*. The social web mining can be focused on different analysis tasks:

- Social networks analysis
- Opinion analysis
- Safety issues analysis
- Authority analysis

The social net analysis is usually focused on some social network structure analysis from the point of view of searching suitable metrics for dynamic analysis, prediction of changes in social networks, developing algorithms for social networks monitoring, dynamic visualization of social networks, research of social networks and their time characteristics. The social net analysis represents mining from conversional structure.

The opinion analysis is concerned with the classification of some web discussion to positive or negative opinion in an automatic way. The summarizing information about positivity or negativity of major part of users opinions about some object (a product, person, event, organization, etc.) can be obtained without necessity to read all discussions and can be useful in the course of a decision making process. The opinion classification represents mining in conversational content.

The safety issues analysis comprises two subtasks: theme modelling and authorship identification. Conversational content can be a source of some information connected with safety issues, for example suspicious activities and identification of authors of the conversation about these subtle themes.

The authority analysis represents the identification of authorities of some web discussions. The contribution of the paper is the design of a method for the authority identification using web mining from a conversational content and also from a conversional structure. There are no known approaches to authority identification from conversational content, which would be compared with our solution. The most known approaches determine the authority degree only from the conversation structure [1][3][9]. Our approach is not based on Natural Language processing (NLP) and it is not based on some known information retrieval algorithm as well.

The contributions of the paper are the following: at first the list of authorities of a given web discussion, ordered according to their estimated measure of authority, at second simulation of dynamic changes of the measure of these authoritative web users. Authority identification can be used in various real situations. For example some web user searches for some authority, which is able to give him/her some advice or information for decision making support. Another example – some Information Technology (IT) organization needs specialists - authorities in the given fields and is able to offer them an interesting position. The paper is related to the conference topics, especially to the track "Complex systems – Simulation and data mining" because: at first, the resulting implementation can be used for simulation of a dynamic change of the authority measure of social web users and at second, the designed method can be used in the process of data mining, especially for web authority mining.

Section II discusses various types of web authorities. The analysis of various kinds of discussion forum contributors is introduced in Section III. Our design of the approach to authority estimation is described within Section IV and next Section V presents the refined model of authority estimation.

Section VI is concerned with dynamic changes of the estimated authority of related social web users.

II. Authority Analysis Types

An authority is reputable competence of some person, society or organization to affect somebody. Authoritative opinions and attitudes have been approved in real situations. The authority can be informal and formal.

The *informal (natural) authority* is a person with naturally authoritative behaviour. Other persons are willing to respect informal authority. Such authority is the result of a personal profile, capability, adequate self-confidence and social activities, which ensure the status for a person – authority. In the field of authorities, a mechanism of associating dynamics and psychology can be found. The people, who let an authority to lead them, enforce the weight of this authority.

The *formal (functional) authority* is a person, which other persons do not want to respect, but they have to. Such authority is the result of a position, title or function of some person within an organization (an arbiter, teacher, politician and so on). The formal authority can be at the same time the informal one as well. The formal authority can sometimes change his/her status. A leader could require submission, although his authority is missing honesty, braveness, predictability and ability of quick decision making.

Different problems can be formulated in the field of authority searching within the Internet. We can search for friend authority, for influence authority or for authority within a given web discussion or a given social network. The *friend authority* is a user with great number of relationships with other users of the Web. The *influencer authority* is a person, who impresses others because of his/her opinions and knowledge on some subject. Our attention was after all concentrated on the problem of searching for *authorities within web discussions*. We took into account the number of relationships - communications between web discussion contributors (mining in the structure of web) as well as the strength of influence in the form of estimation of opinion impression – influencer authority (mining in the content of web).

A. Searching for Authorities within Given Field of Research

The problem of authorities searching was solved by a method for searching scientific papers of a given research field within the Association for Computing Machinery Digital Library (ACM-DL) and Institute of Electrical and Electronics Engineers (IEEE) database. Consequently, we have found authorities between citations presented in the reference part of the searched papers. The Tag Cloud method was used for our results visualization. This approach can be used also for authoritative sources searching. This approach has been implemented and the application was named "Tag Cloud Authority" (TCA). The expected input of the system is a key word or a key phrase characterizing the given research field. The system searches for relevant documents within the ACM-DL and IEEE databases. All authors who have been mentioned in the reference part of each selected paper are considered. The authority degree of an author is increased if he/she is the first author mentioned in the processed reference. The complication is variability of citation standards. Many institutions create their own standards or let authors to choose the form of references.

The orientation on the first author simplifies the processing of selected documents and it has the following interpretation. The first author has usually the greatest share (portion) on the paper creation so he/she is the highest authority from the co-authors. The authority degree of a particular author represents the number of his/her publications citations related to the given research field. More information about this approach can be found in [4].

B. Searching for Authorities within Web Discussions

But, the main attention of this paper is on the analysis of authorities within web discussion forums. Each social web user can establish the discussion on some theme which is interesting for him/her. Other actors can add their contributions to this theme. Many people can adjoin this discussion but not all of them are experts in the discussed field. It is very important to let us be influenced only by authoritative contributors of the discussion forum. Thus, the recognition of authorities within the contributors is a matter of principle.

III. Web Discussion Forums

Users of the Internet can play roles of producers of web content within various platforms of the social web. The attention of this paper is focused on discussion forums. Discussion forum represents the area on a web page, which is created by the given page owner interactively. In the case the discussion on some theme is established (see Figure 1), other users can express their opinions within their contributions to the given theme.

These contributions ("Cont 1", … , "Cont n" in Figure 1) create the discussion forum, which can be represented by a graph – an acyclic tree (in the right part of Figure 1).

People have various reasons for contributing to some discussion forum. Great majority of contributors are people, who want to find answers on their questions or want to obtain informed advices from more experienced people for decision making. They expect truthful information. These contributors create a core of discussion but they are not very authoritative ones, and so not very interesting for us.

A smaller group of contributors are actors, for whom the discussion is the opportunity to express their knowledge, to ensure about truth of their ideas or to revise their opinions. These users access to the discussion seriously, add only truthful information, and join discussion only when they are acquainted with the topic. They are really authoritative contributors and they are interested to be distinguished from the other actors. Therefore, an approach for these contributors identification was designed. For these identified authorities, a measure of their authority should be estimated. This authority value should be represented by a numerical value. Design of this numerical value - estimation of the authority is presented in the Section IV.

The last group of contributors is the group of troublemaking actors. They are provocateurs, who are not reputable and they only seek for an opportunity to present their opinions on web discussions. They often contribute not truthful information, invoke conflicts and they want to present their significance.

They usually degrade all web discussions. These actors are not authoritative. They should be eliminated from the discussions.

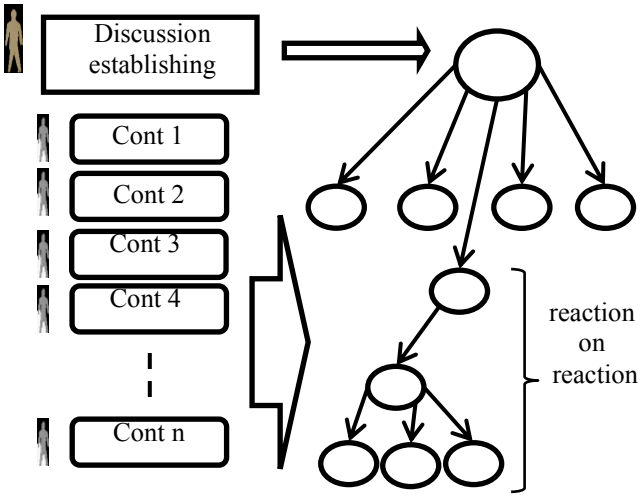

Figure 1. Tree representation of a discussion forum. ('Cont' represents a contribution. Arows lead from a contribution to all reactions on this contribution.)

The elimination of unsuitable contributions can be provided in the web discussion control. There are two ways for discussion control: manual and semiautomatic. *Manual* discussion control is made by a moderator. The moderator checks each new contribution from the point of reliability for the given discussion, law-breaking, whether it is not abusive and so on. Only suitable contributions are consequently added to the web discussion. This kind of discussion control is time consuming and difficult mainly in the case of discussions with great amount of contributions. Thus, semiautomatic discussion control was designed. It uses programs, which filter unsuitable contributions with the aid of words recognition. All contributions, which are denoted by such program as faux or abusive, are redirected to the moderator to decide about deletion of the contribution from the given discussion. There is another way for this problem solving. The moderator can enable actors to denote improper and abusive contributions by a special mark and this marking causes an automatic redirection of the contribution to the moderator for evaluation. It enables to integrate all actors of the web discussion into the process of discussion control. The combination of these three approaches creates so called the three level discussion control. It can ensure nearly zero occurrences of improper contributions. Faux and abusive contributions are forbidden because they can damage good reputation of some firm or organization. But, negative and serious contributions are accepted.

IV. Design of the Authority Estimation

The controlled web discussions can be used for opinion mining and authorities mining. The results of opinion mining can be a part of information, which is necessary for authority estimation.

A. Opinion Mining

The Opinion mining was used for acquisition of opinion polarity of all contributions of the web discussion. These opinions polarities were used for one parameter of a function of authority estimation named "polarity matching".

A web discussion carries a lot of information, for example various themes, opinions and attitudes concerning to various objects (a product, political situation, book, film, physician and so on). Actors who have created some discussion but also other actors who simply have the similar problem as the given discussion is about, want to know the whole opinion of all contributors to the given theme. If the discussion consists from a great number of contributions, reading the whole discussion is time consuming process. There arises the need of automatic classification of the web discussion to positive or negative opinion. This classification has to be based on classification of each particular contribution to positive or negative opinion. More about opinion mining can be seen in [2][7][8]. Opinion classification can be used in those fields where the aggregation of a large amount of opinions into integrated information is needed. The input to opinion classification can be represented by a large amount of conversational content (e.g., content of a discussion forum, blog, chat and so on) and the output of the classification is summary information about opinion polarity. It was solved also within our previous works [5]. This aspect of conversational mining – opinion mining of particular contributions – was used in our design and application of authority estimation (authority mining – Subsection IV.B) in the form of "polarity matching" parameter (Subsection IV.C).

B. Authority Mining

The search for authoritative web users within some web discussion represents mining within data about the web discussion as a whole. We are interested not only in contributions' content but also in the structure of the whole discussion.

The authority is related to contributors not to contributions. Thus, from the beginning it is necessary to collect all data about one contributor together. This process is illustrated in Figure 2. It can be seen in this figure, that all information (selected values) about the "Author A" is completed within first item of all selected values repository.

The authority value estimation was designed as a function of particular selected parameters. These parameters were chosen, because they have influence to authority identification. This function was designed to be composed from two parts: primary part and secondary part.

The result of the process of authority estimation of all discussion actors is the rating of authorities ordered in a descending way, which should indicate:

- contributors who showed the best knowledge concerning to the given theme,
- contributors who invoked most reactions,
- contributors who initialized diversion from the given theme most often.

Testing of this approach was oriented on web discussions, which contain more than 30 contributions, their contributions are not too short, which have more than one level of reactions, their discussion is controlled and they support creation of discussion trees. The implementation of this approach has been designed in such way, which enables using it for any theme. Thus, two different themes in two different web discussions were chosen for testing. The first one was a technically

oriented discussion about Windows7 and the second one was a discussion about TeleVision (TV) serial "Neighbours".

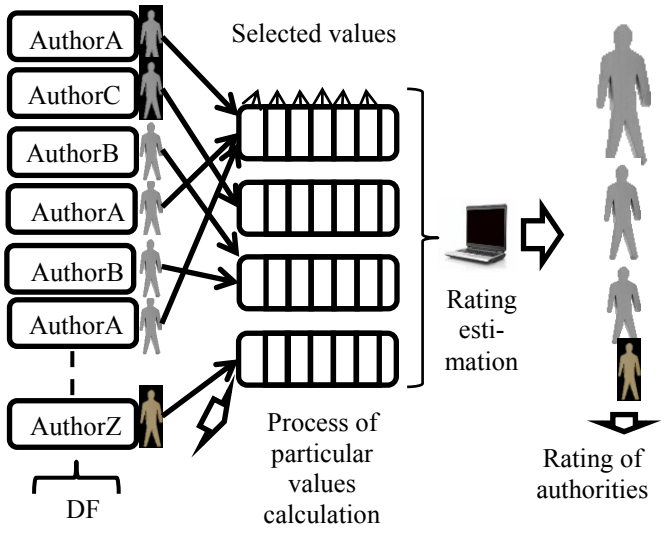

Figure 2. The process of authority estimation from the discussion forum (DF). The human-shape figures represent rated authorities.

C. Design of the Function for Authority Estimation

Our design of the function for authority calculation has the following parameters:

- *Number of discussion contributions (NC)* of the given contributor. It seems, that somebody who understands the theme (authority) will contribute to the discussion more often than other actors. A specific group of contributors are users, who are not so knowledgeable but they use to join discussions to put questions and to learn from answers. They are valuable contributors, because they shift the discussion on higher levels.
- *Number of reactions* (NR) on the contribution(s) of the given discussant. This parameter represents the number of reactions which support or negate a statement of the user, whose authority is examined. It is assumed, that more authoritative contributor could evoke higher number of responses and more reactions.
- *Number of occurrences on the bottom level (NBL)* of the discussion tree. The contributor, whose contributions are located on the bottom level of the discussion tree, usually has added the exhaustive commentary which answered all questions. This kind of contributors can be authoritative.
- *Polarity matching (PM)*. Within this parameter, the whole polarity of all contributions of the given actor is compared with the polarity of the whole web discussion. The polarity of particular contributions was determined using the Opinion Classification Application (OCA) [6]. This application uses the highest degree of positive polarity equal to 3 and the most negative contribution is marked by -3 degree. The greatest difference between the polarity of all users contributions and the polarity of the whole discussion can be 6. The polarity matching in the form of value from the interval <0, 3> is interpreted as an agreement between polarity of discussant's opinions and opinion of the whole discussion. Similarly, polarity matching in the form of value from the interval (3, 6> is interpreted as a disagreement. Greater authority should be assigned to the users, polarity of whose contributions agrees with polarity of the whole discussion. The less authority should be assigned to actors with significant opinion disagreement between their opinions and overall discussion opinion.
- *Position in the discussion tree (PT)* expresses the number of all levels of the discussion tree the contributions of the user are situated in. Each level is considered only once regardless the number of contributions on this level. Exactly, PT is the ratio of this number and the number of maximal possible occurrences of contributions in the discussion tree.
- *Words number (WN)* represents ratio of all words within all discussant's contributions to all words of the whole discussion.

All these parameters, taking separately, indicate rather chatty contributors than authoritative ones. But, taking them together as one entity, the emergency phenomenon arises. This phenomenon can indicate the authoritative contributors.

The first three parameters (number of contributions, number of reactions and number of occurrences on the bottom level) create the primary part of actor's authority value, which influences final authority value in a significant way. The primary part has greater weight in the process of authority estimation. The last three parameters (polarity matching, position in the tree and words number) create the secondary part of actor's authority value. The secondary part serves on precise tuning of the authority value of very similar authorities or is used for refining this value to prevent the case, when the primary parts of the authority value of different actors are the same.

We have experimented with a function for authority estimation in the form of a simple sum of all selected parameters. But, the final precision was very low (from 10% to 20%). The precision was counted as a ratio of the number of all correctly stated authorities (by our application in comparison with an expert opinion) to the number of all recognized authorities by our application (including these authors, which the expert do not consider as authorities). Equation (1) represents the design of the function for estimation of the Authority of the Contributor (AC).

$$AC = 4NC + 2NR + 4NBL + PM + PT + WN \quad (1)$$

The greater weight was connected with those parameters, which appeared more significant during the testing phase. In this case, the testing precision was higher (54.8% for technical domain and 52.6% for real life domain) but not overly satisfying. So this function was refined to the following one (2):

$$AC = 4NC^2 + 2NR^3 + 4NBL + PM + PT + WN \quad (2)$$

In this case, the testing precision was more satisfying (77.4 % in technical domain and 80.7 % in real life domain). Testing was provided in technical and life domains. Technical domain was represented by web discussions on the theme Windows 7 (http://www.verejnadiskusia.sk, http://www.kulman.sk and http://www.warxtreme.eu). Life domain was represented by web discussion on the TV serial story "Neighbours" (http://www.warxtreme.eu).

This form has been implemented. The implementation provides final rating of authoritative users, what is illustrated in Figure 3. This rating is situated in the left dialog window. This window obtains names of web discussion actors. These names are followed by numbers, which represent their authority values.

V. Refined Model of the Authority Estimation

The refined version of the function for authority calculation has been developed. Within this new version logarithmic functions were used. In addition, two new parameters were involved into our method of authority estimation:

- *Number of reactions of the contributor* (NRC) represents the number of reactions of a given contributor on contribution(s) of other discussants.
- *Frequency (F)* represents the number of the given contributor reactions within a time period.

The result was a modified model of the contributor authority estimation (3):

$$AC = 5(1+\log_{10}(NC))+13(1+\log_{10}(NR))+15(1+\log_{10}(NRC)) +(1+\log_{10}(NBL))+3(F+PT+WN) \quad (3)$$

This modified model of the authority estimation has been tested on three various discussion forums related to the following themes:

- The TV "PLUS" discusses about moderators' authorities according to the number of their "likes" on the Facebook.
- The presentation of a well known Slovak politician has been accepted inconsistently.
- Rockets, air attacks and sirens. Near east region drifts toward conflicts.

The results of these tests are presented in table 1.

VI. Dynamic Change of the Authority

Within our research, the dynamic change of the authority value was considered as well. It is an important aspect, related to authority identification. Somebody, who is searching for authorities in some given field, wants to know a dynamic change of the estimated authority value of candidates on authority in recent time.

A natural tendency to increase or decrease the strength of the estimated authority according to activities of the given user – discussant within a web discussion forum has to be considered. Any user of our application would like to have actual information about contributors to the given discussion forum.

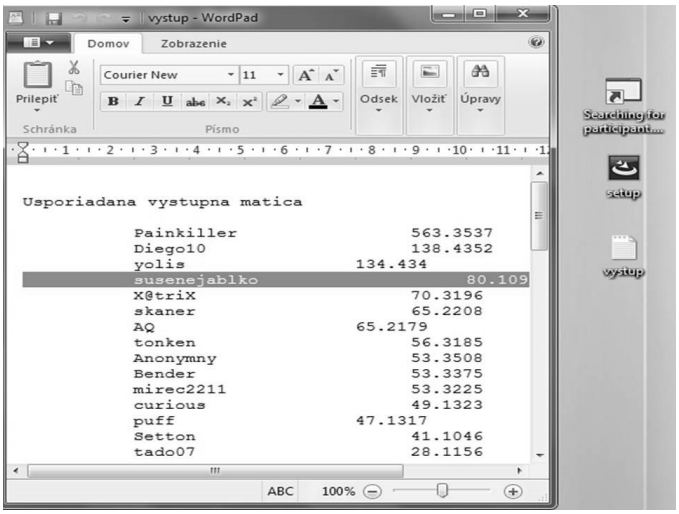

Figure 3. The implementation of the designed method of authority estimation with the resulting rating of authoritative actors.

TABLE I. Results of the Authority Estimation Testing

Theme of the Discussion Forum	***Precision***
Authority and the number of "likes"	0.94
Slovak politician	0.96
Rockets, air attacts and sirens	0.93

The approach to dynamic authority (DA) determination can be modelled by (4):

$$DA = (AC - D)T*P \quad (4)$$

Where:

AC is authority of the contributor according to (3)
D is dynamic change according to (5)
T is time characteristics, which represents the value of the percentage of authority change. It represents increasing or decreasing of the authority on the basis of his/her activity within particular day.
P is penalty (P = 1 for the number of banned contributions from interval $(-\infty;0>$, P = 0.5 for the number of such contributions from interval $(0;2>$ and P = 0 for the number of deleted contributions from interval $<2;\infty)$).

$$D = 1{,}7\sqrt{1 + 2 * \log AND} \quad (5)$$

Where: AND is the average number of days of web discussion continuance.

The visualization of the dynamic authority for 5 the most authoritative contributors is illustrated in Figure 4. You can see, that the authority value of some contributors is rising, but authority of some is degreasing during a period of seven days.

VII. Conclusion and Future Work

The paper introduced the novelty approach to authority estimation of actors of web discussions and implementation of

this approach, which provides the resulting rating of authoritative actors. The resulting implementation is able to simulate the dynamic change of the authority measure of social web users. This approach has been implemented in the programming language Java and in the development environment NetBeans IDE. The results of this implementation were valuable, but they were connected with selected domains from a real life and from a technical world. The novelty of this approach is in estimation of an authority on the base of various parameters obtained not only from the structure but also from the content of the conversational content.

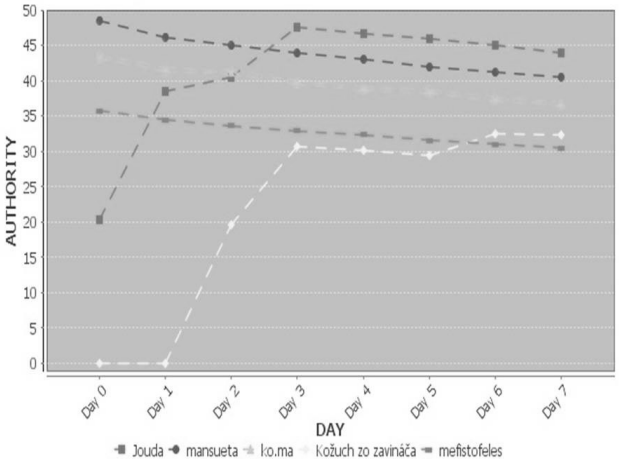

Figure 4. The dynamic authority of the five most significant contributors. Each of them has its own color.

The presented approach can be used for refining the process of opinion classification of some web discussions to positive or negative opinion. Within known approaches to opinion classification of the whole web discussion, the resulting classification to positive (negative) opinion is made when there are more positive (negative) contributions in this discussion and each contribution has the same weight within creation of the resulting polarity. The novelty approach could multiply the positivity value "1" (negativity value "-1") of the given contribution with the weight represented by the estimated authority value of the contributor, who is the author of the given contribution.

The authority identification can be used also by common people, who are interested in some web discussions because of decision making about some purchase, a holiday destination choosing and so on. They can obtain information about authorities and consequently put a question directly to the most authoritative actors of the web discussion. The presented application of authority identification can be used also by organization searching for skilled and valuable employees. The organization can establish some professional discussion and consequently appeal on persons interested in this work position to join the established web discussion concerning to key problems and tasks, this organization has to face up, or to key technologies, this organization uses. Final rating of authoritative actors of this established discussion can serve as the result of preselecting. Or simply, a responsible person of this organization can search for authoritative actors on various web forums, which are focused on technologies and tasks, which are concerned to this organization. Thus, the research in the field of authority identification has big importance for the future.

In future, we would like to refine the model of the authority identification to achieve higher precision. We tend to test our implementation in an extended web space. Also, we would like enrich possibilities of dynamic analysis of the web actor authority change.

ACKNOWLEDGMENT

The work presented in this paper was supported by the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic within the 1/1147/12 project and the 1/0663/14 project.

REFERENCES

[1] M. Bouguessa, B. Dumoulin, and S. Wang, "Identifying authoritative actors in question-answering forums – the case of Yahoo! answers," Department of Computer Science, University of Sherbrooke, Quebec, CA, 2008, 1-9.

[2] Y. Choi and C. Cardie, "Learning with compositional semantics as structural inference for subsentential sentiment analysis," Proc. of the EMNLP 2008, Conference on Empirical Methods in Natural Language Processing, 2008, pp. 793-801.

[3] B. Dom, I. Eiron, A. Cozzi, and Y. Zhang, "Graph-based ranking algorithms for E-mail expertise," Proceedings of 8th ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'03), 2003, pp. 42-48.

[4] S. Kmet and K. Machova, "Web authorities estimation within the given domain," Department of Cybernetics and Artificial Intelligence, Technical University, Košice, 2012, pp. 0-71.

[5] K. Machová, "Opinion analysis from the social web contributions," Computational Collective Intelligence – Technologies and Applications. Lecture Notes in Artificial Intelligence Vol. 6922, No.1, Springer – Verlag Berlin Heidelberg, 2011, pp. 356-365, ISSN 0302-9743.

[6] K. Machová and M. Krajč, „Opinion classification within thread discussions on the web," Proc. of the 10th annual international conference Znalosti 2011, Stará lesná, 31.1.2011 – 2.2.2011, Publisher: FEI Technická univerzita Ostrava, Czech Republic, 2011, pp. 136-147, ISBN 978-80-248-2369-0 (in Slovak).

[7] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," Computational Linguistics, Vol. 37, No. 2, 2011, pp. 267-307.

[8] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," Journal of the American Society for Information Science and Technology, Vol. 61, No. 12, 2010, pp. 2544-2558.

[9] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in on-line communities: structure and algorithms," Proceedings of the 16thACM International World Wide Web Conference (WWW'07), 2007, pp. 221-230.