Computer Science • 16 (3) 2015 http://dx.doi.org/10.7494/csci.2015.16.3.265

Grzegorz Kowalik

COMPETITION-BASED RATING SYSTEM FOR MEDICAL WEBSITE CREDIBILITY

Abstract *In this paper, we propose a new approach to the aggregation of monadic ratings (5-step scale) done by crowdsourcing users for the evaluation of medical websites. We compare them pairwise with other evaluations done by the same users for other websites (whether they are higher or lower), and we will use an Elo rating algorithm to calculate website "credibility" values. Results show that this method of crowdsourcing evaluation is highly correlated with expert evaluations. As proposed, a competition-based model uses a 5-step scale as ordinal and only compares which website is rated higher or lower by the same user. This approach can solve many problems associated with a 5-point scale, such as different understanding by users, user bias, and distribution skewness that can be clearly observed in results.*

Keywords rating systems, Elo rating, credibility of online content, web credibility, rating aggregation

Citation Computer Science 16 (3) 2015: 265–280

1. Introduction

Researching the difficult matter of website "credibility", we (as a research group) used crowdsourcing to collect opinions (evaluations) about different websites from many users. This paper focuses on the problem of combining all evaluations into one overall evaluation for each website, which is then used as a recommendation of their credibilities for other users. In other words, we want to create a rating system for websites using multiple user evaluations/votes.

The goal of this research is to show that simple methods of aggregation, such as mean or majority vote, are exposed to many errors resulting from the characteristics of crowdsourcing data, as well as the necessity of using more-advanced methods and models that can handle the characteristics of crowdsourcing.

In this paper, we propose modeling our problem as a competitive game between websites, using competition-based rating systems that are commonly used in games like chess or football. This rating stands for the "skill" of players or teams. In our case, it will stand for "credibility" – how likely each website will convince users that they are more or less credible than others.

Our project about website credibility uses multiple evaluations made by users on a 5-point scale. We faced many problems using such ratings from crowdsourcing, such as problems with scale, user bias, and distribution skewness [12]. To learn more about credibility evaluation, two use cases were performed (regarding programming in Java [11] and medicine). Both use cases give us the opportunity to compare results from users to expert evaluations that we have collected. In our previous research, we already proposed a game approach in [14] and [19], where content consumers and producers competed. In this research, we propose a much simpler model – where websites are competing for better ratings from users.

In this paper, we introduce our dataset and how to model 5-step scale ratings into a series of "games" between pairs of websites, with three possible outcomes (win, draw, or lose) and calculate a competition-based rating, such as the Elo rating system from chess. After that, we present the results from a medical use-case dataset.

2. Related work

In our study, we have an issue about aggregating multiple opinions into some kind of decision (about website credibility recommendations). This situation is very similar to issues known in group decision-making, economy, and market research with methods from mathematical psychology, where we ask multiple people about their opinions or preferences about some objects to create some sort of ranking.

There are different approaches to this issue that we can group by which task a respondent is completing:

- **Monadic ratings**, where people are asked to rate objects on some scale (like the Likert scale, or a 1 to 10 point scale, etc.). These kinds of ratings were used in Reconcile use cases.

- **Comparisons**, where people have to rate a presented option relative to other options; for example, saying that one option is better than another. The most well-known method is **pairwise** or **paired comparison**, where there is a choice between two objects.

As proven in [18], paired comparisons give better results than monadic ratings and have many other pros (especially that they are much easier for respondents, which is very important in the CS topic).

Regardless of the selected method, all of them can be described using terms like in [4, 5, 6, 7, 20, 21]

- **Respondents** or **decision-makers** who are performing comparison or rating task. In our Reconcile use cases, we call them users.
- **Objects** or **acts** which will be compared.
- **Criteria** or **attributes** that describe objects and are used by respondents to compare.

Focusing on comparison methods, there are many different approaches on how a comparison might be done, with the most known examples of such a task: respondents might be asked to place multiple objects in order from best to worst or vice-versa. In another approach from methods such as MaxDiff [16], respondents are asked to select some objects; for example, the best and worst option (like in MaxDiff). Of course, these tasks can be mixed; for example, by asking respondents to select the three best objects and order them by preference. Paired/pairwise comparisons are a special case, as it consists of order and selection. In such a system, respondents have only two objects and must compare them and decide which is better. Sometimes, indifference/tie is also possible. Some methods are using paired comparison without directly asking respondents to compare two elements – for example, as in MaxDiff, pairs and paired comparisons are computed from a larger comparison where respondents selected the best option (which is better in pair with each other option) and worst option (which is worst in pair with each other option). The same approach can be used in many similar situations; for example, in [15], the answer marked as "best" in a question and answer portal is better than each other answer. A very important issue is that our respondents might give different, even contradictory comparisons with others. This way, we need methods that assume that we have many respondents (as opposed to having one decision-maker), like using dominance in [7]. As the main goal is to find the best objects from a whole set or to create some kind of ranking, some methods focus on criteria that are used by respondents. For example, conjoint analysis (based on pairwise comparison) uses multiple criteria to compare even artificial objects (described by criteria values) to calculate the importance of each criterion, and, for example, to create the best-possible object or to research how many (even not compared by respondents) combinations of criteria values would be compared. The problem is, to use methods like conjoint or Dominance-based Rough Set Approach (DRSA) [4, 5, 6, 7, 20, 21], we require a list of criteria that are used and their values for each object – we need a rough set. In our case, we do not have it. If we

can describe each website in some clear criteria, we could use such methods; but the problem is when the criteria are unknown and might require new, dedicated methods of how to rate tasks using comparison tasks. We can also try to adapt and evaluate existing methods, even from different fields (like [15], using competition-based pairwise ratings for Q&A portal). A competition-based system that we propose is the Elo rating system, which is designed to measure the skill of chess players [2]. Currently, it is a commonly-used solution for many competitions, both "real" like football [10], and virtual, like online games [9]. There are, of course, many other ratings, like one considered for future research – the Glicko rating system [3], or systems for different game conditions; for example, to apply them to games with more than two players, like the TrueSkill rating system [8]. The idea is to use such ratings outside of conventional games and apply them to Internet activities like web 2.0 portals, which were recently researched for Yahoo Answers question and answer portal [15]. Liu's work shows that we can apply competitive-based rating systems to activities that are not directly a "game" between users. His model compares pairs of answers in topics where the answer with "best answer" status from the original poster wins, calculating a rating for each author (user). Our problem is also similar to reputation systems, sharing some common issues. For example, a good solution (both for reputation and rating system) should be aware of the issue about number of evaluations (or sales volume to rate, like in [17]). This is a well-known problem for rating systems, when highly-ranked players could avoid playing to maintain their status. In our case, we must be aware that some websites might have a higher number of evaluations and shouldn't be considered as more credible simply due to to the fact that they are evaluated more often. Using such ratings for websites as competitors, with user votes as "games", is a new approach. We propose to interpret monadic-scale results as pairwise comparisons, and pairwise comparisons as a game.

3. Dataset

In this research, we have used data from Reconcile medical use case. In this experiment, our research group selected 180 medical websites, trying to have both credible and non-credible websites (some websites were selected intentionally as to be non-credible). Using Amazon mTurk crowdsourcing, we collected evaluations for each website. The most important task given to users was to evaluate a presented website's credibility using a 5-step scale (where 1 is completely not credible and 5 is completely credible). We collected 1,134 evaluations. Each user could evaluate multiple websites, and the average number of websites evaluated was 6.3 per user.

The same websites were evaluated by experts – students of Medical University of Warsaw and doctors recruited by the team. These ratings we treat as the "ground truth". We use them to evaluate how crowdsourcing can estimate website credibility and how good our methods of aggregation are in this paper.

As we can see by comparing Figure 1 and Figure 2, users tend to give higher ratings as compared to experts. This essentially shows the problems with crowdsour-

cing ratings – they are skewed, biased toward positive values, exactly as described in previous analyses from our team (described in [12]).

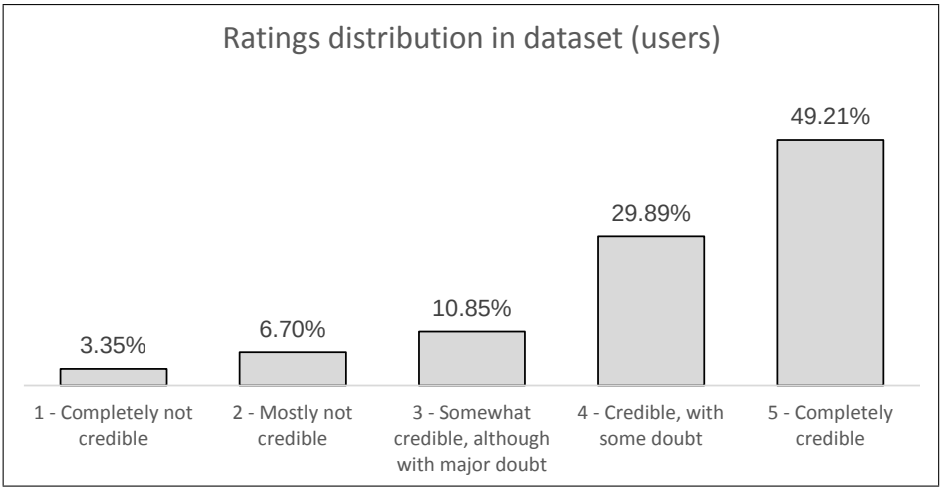

Figure 1. Frequency of ratings in population of all evaluations ($n = 1134$).

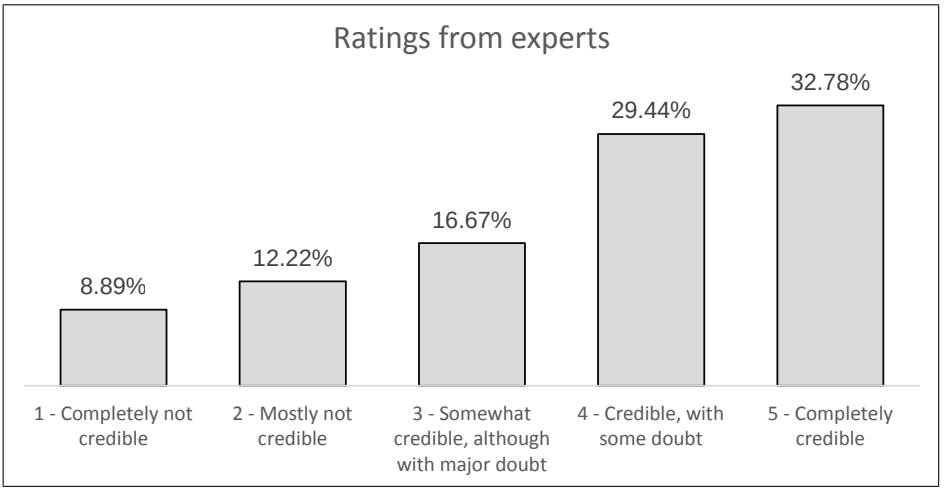

Figure 2. Distribution of websites with specified expert rating in population of websites ($n = 180$).

The provided ratings are on a monadic, ordinal scale. However, as it is similar to the Likert scale, there is a common assumption that we can treat these ratings as on an interval scale, usually to calculate means. The method proposed in this paper

avoids this. Results from using means will be presented as comparison for evaluation sake, along with majority vote (modal value).

4. Competitive game model

To sum up: in the dataset, we have multiple ratings on a 1–5 scale for each website. Each user could evaluate multiple websites, and each website could been evaluated by multiple users. Our idea is that each user can understand each scale differently. We know nothing about his or her single value. For example, "4" given to one website means nothing alone. But, as a user can evaluate more than one website, we can clearly say about the relationships between the evaluated websites – preferences and indifferences. However, we do not compare ratings from different users, as their scale understanding might be different – some users can give mostly "good" ratings, and for them, "3" is reserved for non-credible websites, etc. This approach is resistant to different-scale understanding and any skewness toward good or bad ratings. We want to model it as there is some hidden value for each website, "credibility". This value affects users and their evaluations. We want to create a rating to find this value for each website. As mentioned before, our model uses a well-known algorithm from chess – the Elo rating – to calculate website credibility. To do this, first we built a theoretical model of our "game" to compare it to a chess tournament. To build such a model, we had to deal with some restrictions that come from the original application – chess:

- It is a two-player game.
- It is a game with three possible results for each player: win, lose, or draw.
- Only one player can win and only one player can lose. If so, the second player loses. A draw concerns both players.

Following this, to build the model, we must define:

1. **Players** – Who plays the game, who competes, and for whom do we calculate rating.
2. **Game match** – What kind of games are considered. For example, in chess, we can take into account only tournament games.
3. **Game rules** – who wins, who loses, and what defines a draw. In chess, it is described by rules – the outcome can be decided by checkmate, surrender, timeout, disqualification, stalemate, threefold repetition, etc.
4. **Rating interpretation** – What the calculated rating from results means. In chess – it is assumed to be the "skill level" of the player.

In our problem, users are part of a crowd, and we do not need to make ratings for them, so they are not "players." However, we can also try to do this, as we will mention later in this paper. Our main issue is website credibility, so we model websites as competing players. Users give them rankings – so we can compare rankingss from one user pairwise, saying that site A is more / less / as credible as B, comparing ranks. This way, we are very strict to ordinal scale limitations and can build a well-described rating. To sum up, we can model Reconcile ratings as a game with:

1. Players – websites
2. Game match – every user with more than one evaluated website. Each user's ratings are compared pairwise (but not between players, only between the same user ratings).
3. Game rules – the website with a higher rating on a 1–5 scale wins. If ratings are the same, there is a draw.
4. Rating interpretation – website credibility

This way, we will receive a rating for each website, using ratings from users strictly as an ordinal scale. Moreover, as we do not compare ratings from different users, we allow them to understand each scale differently – as long as they understand which value is higher and lower. In this approach, giving ratings of 5 and 4 to two different websites have the same result as giving 4 and 3 ratings instead. In our opinion, this can help with many problems in an ordinal scale, leading to results with good interpretation and correct methodology. Here is an example of how we can use crowdsourcing ratings to create such a credibility ranking:

Table 1

Example of Reconcile ratings.

User	Website 1	Website 2	Website 3
User 1	5	3	3
User 2	2	1	–
User 3	–	5	5

In Table 1, we can see three users that could evaluate three websites. They can give one ranking from 1 to 5 for each website or not rank them (for example, if we randomize website samples for them). As long as they evaluate more than one website (to do pairwise comparison), it does not matter how many website they evaluated. Each row represents one user. As we can see, for example, User 1 used only "5" and "3" ratings. We only need information that rating "5" is greater than rating "3". "–" means no evaluation for website by particular user. This way, we translate these results into:

1. From User 1:
   (a) A vs B, A wins
   (b) A vs C, A wins
   (c) B vs C, draw
2. From User 2:
   (a) A vs B, A wins
3. From User 3:
   (a) B vs C, draw

Ratings systems give points to "winners" and takes away points from "losers", taking into account their current ratings. Players with higher ratings can get fewer

points for a victory over players with lower ratings, but can lose more points if they lose, giving more points to players with lower rankings. In case of a draw, players with higher rankings lose a smaller number of points and players with lower rankings receive a larger number. This way, they are "moving closer" in ranking. If we look back at the example, we can see that our approach will say that website A is the best (as it beats all other websites), website B is the worst (as it loses 2 times and has 1 draw), placing C in the middle. Compare these results to using mean – where A has 3,5, B has 3 and C has 4, the order is different. Of course, we should have a lot more users and/or ranks to compare; but if we think about "biased" users like user 3 here (who gives only 5's), or if we, for example, know that user 2 is generally skeptical about websites and gives lower rankings, we can see that our results can improve the aggregation of ratings.

4.1. User memory limit

Additionally, the assumption that users are actually comparing (in their minds) each website to give ratings might be too unrealistic with big numbers. To avoid this, we propose an additional, optional, modification to the model – instead of taking into account each possible pair when we create pairs as games with results, we use a fixed value of previous ratings for each rating. For example, if we set memory for 3 and users have evaluated 10 websites, 6th rating is compared with 3rd, 4th, and 5th, and 10th rating is compared with 7th, 8th, and 9th, etc (three previous ratings).

We intentionally lose some pairwise comparison, as we think that users can "learn" more about their task and can change their scale usage in time.

5. Rating calculation

5.1. Elo rating calculation

For calculations, we have used the Elo rating system [2]. The Elo system is based on expectations about results and differences between them and the actual results. Expectation about results is calculated using the current ratings of players and assume how their difference in skill (measured by rating) affects results (currently, by observing a large number of results, a logistic curve is considered as most fit for results and used by the United States Chess Federation). Each website starts from the same rating value: 1500 (the same as in chess). Next, for each user with more than one website evaluation, we compare website ratings in every possible pair created from his or her ratings. In memory variant from chapter 4.1, we use only fixed value of previous evaluations. Each comparison has one of three possible results (score, S): win (1), draw (0.5), or lose (0). Win is when a vote on a 5-point scale is higher than its opponent's. If these values are the same, both websites score a draw. A lower value means lose.

For each website in each pair, the rating is updated using the procedure below:

1. Firstly, expected score for website (E_W) is calculated using logistic curve:

$$E_W = \frac{1}{1 + 10^{\frac{R_O - R_W}{400}}}$$

   Where R_w is current Elo rating of website and R_O is rating value of "opponent" (compared) website.
2. Then, expected value is compared with actual score and used to update ranking (updated website rating R^*_W), using formula below:

$$R^*_W = R_w + K_U(S_W - E_W)$$

Where S_W is the actual score for player/website (1, 0.5 or 0). K_U is our equivalent of K-value from Elo rating that can be used for weights. In Elo rating, for example, $K = 16$ is used for experienced chess players and $K = 32$ for newcomers. As it affects how much ratings can change due to one "game", our idea is to use it not for competitors but for users. Crowdsourcing users can have lower K than experts, if we have them. If we have any additional information about expertise of a user, we can use it for K-value for each of his or her rating. In this research, as all users are from an anonymous crowd, we use $K = 16$. What is important, we do not calculate mean or other values, nor tests that should not be used for ordinal scales on which user votes are (unlike many other methods of aggregation used for these datasets). Also, one of the main disadvantages of using Elo rating is the possibility that "top player" activity would drop due to strategy (as their estimated result is closer to 1, they have a lot to lose but almost nothing to gain) does not exist here, because websites cannot avoid evaluations (as chess player can avoid playing). Moreover, we can recommend some websites for users, update ratings, and avoid the "rating inflation" problem known to Elo rating.

5.2. Final rating calculation

Our result will be an Elo rating for each website. It will usually be a large value that has good theoretical interpretation (can be used in pairs to estimate "game outcome"), but might not be clear for the final user (the goal of our project is to inform users about website credibility) and hard to compare with our expert ratings that are given in a 5-step scale. To do so, we normalize Elo ratings into 5-point scale.

5.3. Elo rating values

As all users were from an unknown crowd, we use the same K-value for them. We chose $K = 16$. Different levels of K-value could be used when we have any knowledge about users. In standard Elo rating in chess, it is used to calibrate newcomers faster (higher K-value results with bigger changes in ratings from each game). In our model, the K-value could be defined not to players (websites) but for users (games). If we know

that some users have better knowledge about a topic, we could give them a higher K-value to increase their impact on the final ratings. However, at this stage, we chose to set K-value to 16 for each website and user. Each website starts with the 1500 Elo rating. After applying user evaluations, we achieve following results according to our model (see Fig. 3).

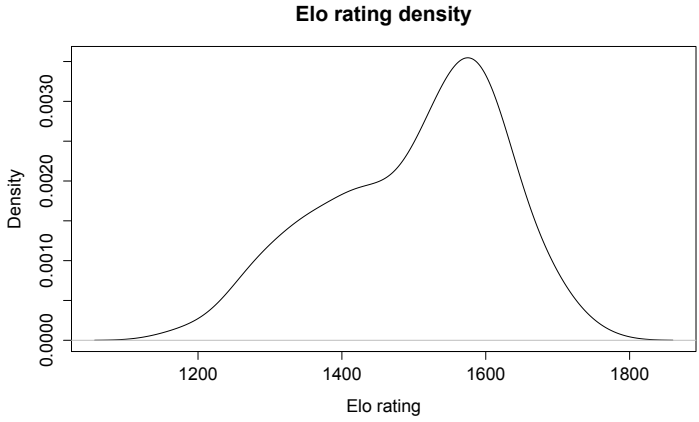

Figure 3. Elo rating density.

To evaluate our results, we compare them with expert ratings (see Fig. 4).

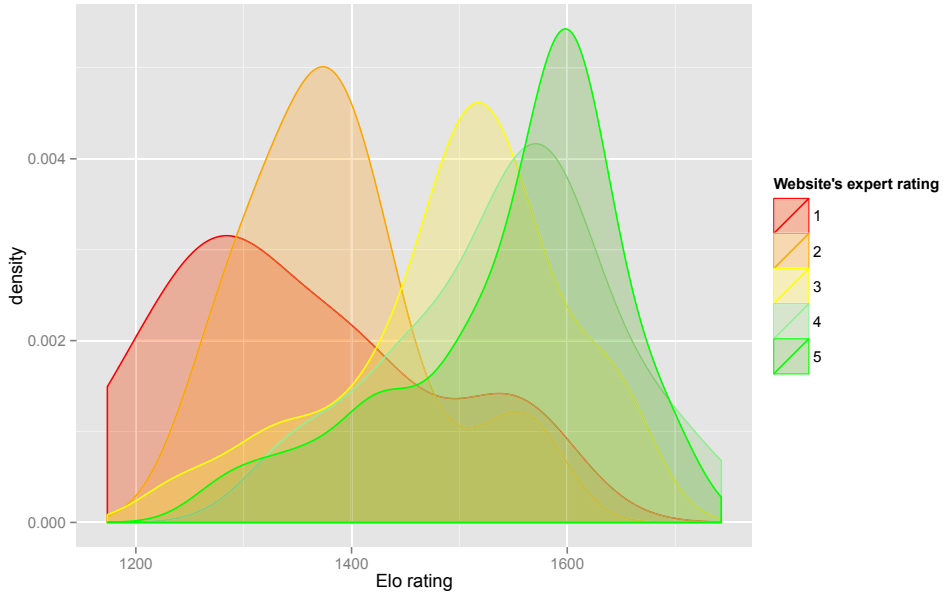

Figure 4. Elo ratings of websites (x-axis) for websites rated by specific value by experts.

As we can clearly see in Figure 4, Elo ratings of websites based on user evaluations are highly correlated with expert ratings. This, of course, can be verified using a correlation coefficient – as expert ratings are on a 1–5 ordinal scale, we have used Spearman's rank correlation coefficient.

Table 2

Example of Reconcile ratings.

Method	Spearmann's correlation with experts ratings
Elo rating	0.49
Mean from users*	0.48

As we can see in Table 2, correlation is high. It is very similar to values calculated using means from user ratings. At this point, we prove that our method is not worse than calculating means. But there are advantages – our methods do not make any additional assumptions on scale, so it is methodologically more correct than calculating means from an ordinal scale. Moreover, we achieve more-detailed values than 1–5 with good interpretation – we can compare ratings to estimate how many times one website "wins" over another. As we can measure the degree of difference between values, we have an interval scale – this allows for more calculations than with an ordinal scale.

5.4. Memory variant

As mentioned before, to make a more-detailed evaluation, we use normalization (from Elo rating values, using min and max values, normalizing into a 1-5 scale). To compare, as previously, we will use expert ratings and results from using means from user ratings (rounded) and majority (lowest modal).

Firstly, to select the best variant of our model, we tested multiple memory values. As criteria, we check how many times our rating match expert rating and, if does not, how big is the error (distance between ratings).

Results are:

As we can see in Table 3, the memory = 8 option gave us the best results in both highest accuracy ratings (41.11%) and lowest mean of error (0.78). This is 10 percentage points better than user mean. Our memory variant clearly improves the results with the exception of low values, like 3 or 1. In low values, we probably drop too much information. With higher values (best at 8), we can observe that our idea about memory can be right.

As this variant gives us the best results, we proceed to compare it with user mean and majority to show the advantages of our method.

As we can see in Figure 5, our Elo-based rating gives more varied distribution than mean and majority, which are dominated by ratings 4 and 5. What is important here, mean distribution is very skewed toward positive values, with 0% websites rated as "1".

Table 3
Example of Reconcile ratings.

Error (distance from expert's rating)	Elo (no memory)	Elo (memory = 10)	Elo (memory = 8)	Elo (memory =5)	Elo (memory =3)	Elo (memory =1)	Users mean	Majority
0	37.22%	40.00%	41.11%	38.89%	32.78%	33.89%	31.11%	36.11%
1	43.33%	42.78%	42.78%	43.89%	49.44%	50.56%	58.33%	42.22%
2	14.44%	13.33%	12.78%	13.89%	13.89%	12.22%	8.89%	15.56%
3	4.44%	3.89%	3.33%	3.33%	3.89%	2.78%	1.67%	3.89%
4	0.56%	0.00%	0.00%	0.00%	0.00%	0.56%	0.00%	2.22%
Mean of error	0.88	0.81	0.78	0.82	0.89	0.86	0.81	0.94

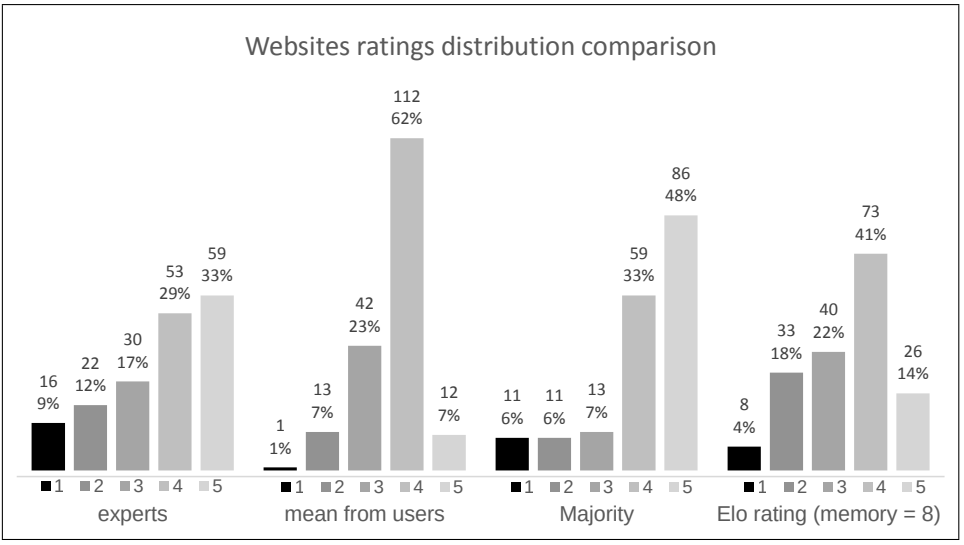

Figure 5. Comparison between expert ratings, means and majority from users, and proposed solution distributions on website population ($n = 180$).

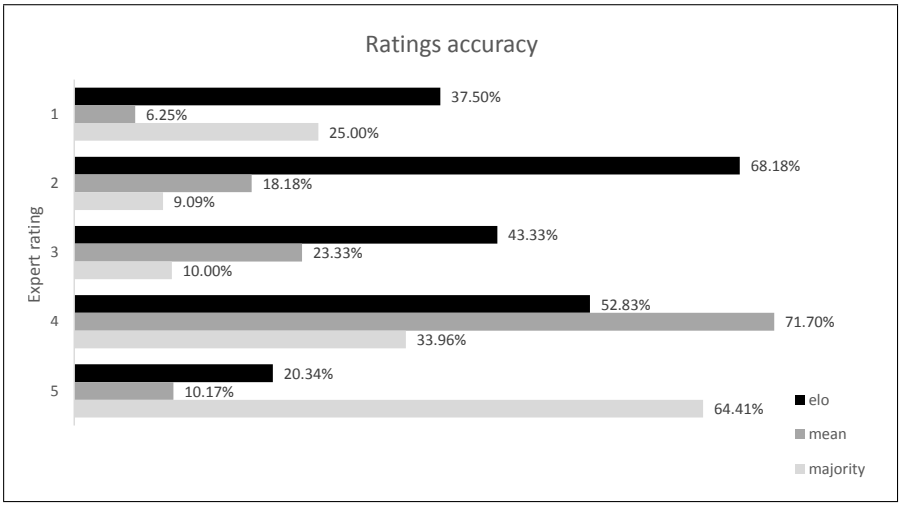

Figure 6. Rating accuracy – percentage of websites rated exactly as experts for each expert rating.

Table 4

Crosstable between ratings from users **(Elo rating method with memory 8)** and experts rating.

Expert evaluation	Elo (memory = 8) [%]				
	1	2	3	4	5
1	**3.33**	3.89	0.56	1.11	0.00
2	0.56	**8.33**	2.22	1.11	0.00
3	0.56	1.67	**7.22**	5.00	2.22
4	0.00	2.22	6.11	**15.56**	5.56
5	0.00	2.22	6.11	17.78	**6.67**

Total correct: 41.11

Table 5

Crosstable between ratings from users **(Majority)** and expert rating.

Expert evaluation	Majority [%]				
	1	2	3	4	5
1	**2.22**	1.67	1.67	1.11	2.22
2	3.33	**1.11**	1.67	5.00	1.11
3	0.00	1.11	**1.67**	6.67	7.22
4	0.56	1.11	1.67	**10.00**	16.11
5	0.00	1.11	0.56	10.00	**21.11**

Total correct: 36.11

If we use this distribution, we could reach a false conclusion that we have no non-credible websites in the sample (and only 4% rated as "2"). This is, of course, not enough to evaluate our method – more important is how accurate the rating is (how close to expert ratings).

Table 6

Crosstable between ratings from users **(Mean)** and expert rating

Expert evaluation	Mean [%]				
	1	2	3	4	5
1	**0.56**	3.33	3.33	1.67	0.00
2	0.00	**2.22**	8.33	1.67	0.00
3	0.00	1.11	**3.89**	11.67	0.00
4	0.00	0.56	4.44	**21.11**	3.33
5	0.00	0.00	3.33	26.11	**3.33**

Total correct: 31.11

As we can see in Figure 6, our method is much better for recognizing websites with low credibility ("1", "2") and "neutral" ("3"). On the other hand, results for highly-credible websites ("5") are worse than using majority. As we can see in Table 4, most of them are classified as "4"; so, in our opinion, this error can be accepted. If we assume that the most important issue (especially with medical websites) is to warn users about non-credible websites, the proposed method gives much better results, if we look into the first two rows of Tables 4, 5 and 6 (non-credible websites).

6. Conclusions and future work

In conclusion, this research shows that modeling crowdsourcing data with ratings as a game is possible and can give better results than simple methods of aggregation. Known issues due to characteristic of crowdsourcing data clearly lead to errors when using simple methods without thought. Conscious about it, when we resign from some part information (like here, comparison between users) that we think are the most exposed to errors and build a competition-based model using each user evaluation separately, we achieve better results. Moreover, more modifications to the algorithm (like introduced "memory") lead us to even better results. This is due to, as we think, a more realistic model of human behavior. In the future, we will consider adapting different completive-based systems instead of only Elo, as the examples mentioned in the related work section. Optionally, team-based systems can be used for group websites in domains and calculate ratings for them. Solutions with clustering or rounding into a 5-step scale can also be improved by testing methods other than simple normalization. In the near future, we plan to run another medical use case, this time using our Reconcile website with more features, like reputation, gamification, and sentence evaluation. We can try to use values like the reputation of a user to set different k-values for users, making some evaluations more or less significant (k-value

affects rating change). We are also aware of any manipulations that can come from users. The model proposed in this version can be fragile to users that give a lot of ratings. In future research, we want to introduce modifications to reduce such effects and test the model with some dishonest users and to compare our method with other, more-advanced methods from data analysis. We might consider similar issues and solutions, like in reputation systems [1, 13, 22], where we can use some information about users to modify their impact. Finally, we can try to test this solution for other systems that can be similar to evaluating website credibility – such as reviews and crowdsourcing tasks.

Acknowledgements

This work is supported by Polish National Science Center grant 2012/05/B/ST6/03364.

References

[1] Borzymek P., Sydow M., Wierzbicki A.: Enriching trust prediction model in social network with user rating similarity. In: *Computational Aspects of Social Networks, 2009. CASON'09. International Conference on*, pp. 40–47, IEEE, 2009.

[2] Elo A. E.: *The rating of chessplayers, past and present*, vol. 3, Batsford, London, 1978.

[3] Glickman M. E.: *The glicko system*. Boston University, 1995.

[4] Greco S., Matarazzo B., Słowiński R.: Decision rule approach. In: *Multiple criteria decision analysis: state of the art surveys*, pp. 507–555, Springer, 2005.

[5] Greco S., Matarazzo B., Słowiński R.: Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research*, vol. 129(1), pp. 1–47, 2001.

[6] Greco S., Matarazzo B., Słowiński R.: Dominance-based rough set approach to case-based reasoning. In: *Modeling Decisions for Artificial Intelligence*, pp. 7–18, Springer, 2006.

[7] Greco S., Matarazzo B., Słowiński R.: Dominance-based rough set approach on pairwise comparison tables to decision involving multiple decision makers. In: *Rough Sets and Knowledge Technology*, pp. 126–135, Springer, 2011.

[8] Herbrich R., Minka T., Graepel T.: Trueskill$^{\text{TM}}$: A Bayesian skill rating system. In: *Advances in Neural Information Processing Systems*, pp. 569–576, Cambridge, MA, USA, 2006.

[9] Hinnant N. C.: Practicing Work, Perfecting Play: League of Legends and the Sentimental Education of E-Sports, 2013.

[10] Hvattum L.M., Arntzen H.: Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, vol. 26(3), pp. 460–470, 2010.

[11] Juźwin M., Adamska P., Rafalak M., Balcerzak B., Kąkol M., Wierzbicki A.: Threats of Using Gamification for Motivating Web Page Quality Evaluation.

In: *Proceedings of the 2014 Mulitmedia, Interaction, Design and Innovation International Conference on Multimedia, Interaction, Design and Innovation*, pp. 1–8, ACM, 2014.

[12] Kąkol M., Jankowski-Lorek M., Abramczuk K., Wierzbicki A., Catasta M.: On the subjectivity and bias of web content credibility evaluations. In: *Proceedings of the 22nd international conference on World Wide Web companion*, pp. 1131–1136, International World Wide Web Conferences Steering Committee, 2013.

[13] Kaszuba T., Hupa A., Wierzbicki A.: Advanced feedback management for internet auction reputation systems. *Internet Computing, IEEE*, vol. 14(5), pp. 31–37, 2010.

[14] Kowalik G., Adamska P., Nielek R., Wierzbicki A.: Simulations of Credibility Evaluation and Learning in a Web 2.0 Community. In: *Artificial Intelligence and Soft Computing*, pp. 373–384, Springer, 2014.

[15] Liu J., Song Y. I., Lin C. Y.: Competition-based user expertise score estimation. In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 425–434, ACM, 2011.

[16] Louviere J. J., Islam T.: A comparison of importance weights and willingness-to-pay measures derived from choice-based conjoint, constant sum scales and best–worst scaling. *Journal of Business Research*, vol. 61(9), pp. 903–911, 2008.

[17] Morzy M., Wierzbicki A.: The sound of silence: Mining implicit feedbacks to compute reputation. In: *Internet and Network Economics*, pp. 365–376. Springer, 2006.

[18] Orme B.: Scaling multiple items: monadic ratings vs. paired comparisons. In: *Sawtooth software conference proceedings, Sequim*, pp. 43–59, Sequim, WA, USA, 2003.

[19] Papaioannou T. G., Aberer K., Abramczuk K., Adamska P., Wierzbicki A.: Game-theoretic models of web credibility. In: *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pp. 27–34, ACM, 2012.

[20] Słowiński R., Greco S., Matarazzo B.: Rough sets in decision making. In: *Encyclopedia of complexity and systems science*, pp. 7753–7787, Springer, 2009.

[21] Słowiński R., Greco S., Matarazzo B.: Rough-set-based decision support. In: *Search Methodologies*, pp. 557–609, Springer, 2014.

[22] Wierzbicki A.: The case for fairness of trust management. *Electronic Notes in Theoretical Computer Science*, vol. 197(2), pp. 73–89, 2008.

Affiliations

Grzegorz Kowalik
Polish-Japanese Academy of Information Technology, Warsaw, Poland,
grzegorz.kowalik@pjwstk.edu.pl

Received: 19.01.2015
Revised: 05.03.2015
Accepted: 05.03.2015