Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2016) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

13th International Conference on Current Research Information Systems, CRIS2016, 9-11 June 2016, Scotland, UK

Crowdsourcing opportunities for research information systems

Ivan Nevolin[a,b]*

[a]*Central Economics and Mathematics Institute RAS (CEMI RAS), Nakhimovskiy prospect 47-909, Moscow 117418, Russia*
[b]*Moscow Institute for Physics and Technology, Institutskij pereulok 7B-406, Dolgoprudnij, 141700, Russia*

Abstract

Research information systems provide data for scientific work valuation. An example of Russian system called Elibrary demonstrate that a number of errors could distort the data. An existing mechanism for data correction relies on the manual validation by the moderator of the system. Research organisations are allowed to reveal and report the data errors under a paid service. Manual moderation, however, increases the time of applications processing. One could speed up the moderation while enabling users to report the errors and to decide about the correction by means of voting. Converting only one of the variety of paid function is hardly to harm the business interests of the operator of the system. Meanwhile the simulation modelling demonstrate the weakness of internal motivation to restore missing citations. One should suggest an external motivation. As an example of external factor, the article suggests scoring system that prevents money transactions. The scores collected could be exchanged for paid service access. Nevertheless, the operator of the system benefits while choosing the parameters of the scoring system in order to ensure that crowdsourcing costs beat the fulltime moderator.

Peer-review under responsibility of the Organizing Committee of CRIS2016.

Keywords: crowdsourcing; digital library; data quality; research information system

* Corresponding author. Tel.: +7-499-724-2476; fax: +7-499-129-1011.
E-mail address: i.nevolin@cemi.rssi.ru

Peer-review under responsibility of the Organizing Committee of CRIS2016.

1. Problem statement

Data quality in research information systems are of big importance since they are widely used for the impact demonstration[2]. Variety of indicators – for individuals and organisations – are calculated in order to track research productivity and to illustrate the impact of scientific results. Operators of CRISs (Current Research Information Systems) are concerned about the full and correct representation of data, so they employ methods to reveal and eliminate errors in data[4,5]. In case of Elibrary – CRIS, widely adopted in Russia – organisations have an opportunity to manage their publications for a fee. Collections of publications, researchers, organisations and some indicators are freely available. However, if an organisation is willing to correct metadata of stored publications or add missing ones, it has to pay for access to the additional functionality. While correcting or adding data, organisation creates an application, that holds the changes suggested. All the applications must pass through the manual moderation before the changes come into force. The employees of the Elibrary review every single request, and this is what produces problems. Moderation takes not even weeks, but months. Examples of waiting for a half of a year are common. As a result, one could not expect indicators values to be adequate, because all the changes simply have no time to come into force before reporting dates. This is the moderation technique what is the issue of particular research.

2. Crowdsourcing as a tool for citation linkage

Some applications in the Elibrary system deal with rather simple issues. One can think about mistakes in citations: misprint in the title, wrong order of the authors and etc. Machine could give only probabilistic decision on whether two citations with one of them holding misprint refer to the same publication or not. People solve this task intuitively, and no special qualification or algorithm is involved. Having this idea one could suggest a free access to the now paid functionality of citation linkage. Every registered author would get an opportunity to distinguish the particular citation in the system as a reference to her publication. As a result users ensure correct linkage and the operator of the system could use available resources to reinforce moderation of other applications. Additional activity from the users is a way to improve the data quality and also known as crowdsourcing.

The effect of the crowdsourcing could be demonstrated based on the Elibrary usage statistics. During one year researchers from the Central Economics and Mathematics Institute RAS (CEMI RAS) established 927 citation links for their 93 publications. The institute has a paid subscription for advanced functionality that allows representatives of CEMI RAS to edit data in the Elibrary system. Representatives of other organizations also contributed to the citation linkage and uploaded the missing publications. These entire jobs yielded 7 575 new citation links and 327 new publications for CEMI RAS. The Elibrary system holds metadata for 4 251 publications, associated with CEMI RAS. These publications were cited 27 471 times in total, and data are valid on 28 December 2015. If the proportion of errors is valid for all Elibrary data, one could say, that 7,7% of publications and 27,6% of citations should be corrected. Absolute values on 4 February 2016 are 1 704 611 for publications and 57 782 968 for citations. Big numbers and they influence strongly on the values of bibliometric indicators. For example, according to the Elabrary on 22 January 2015 publications of CEMI RAS were cited 2 207 times in 2012, 1 753 times in 2011 and 1 277 times in 2010. Within a year users submitted applications to add the papers already published, but not recorded in the system, as well as applications to establish citations links. Note that these applications were made by means of paid service. The entire work resulted in new citation values for CEMI RAS. The citation values on 28 December 2015 were 3 540 in 2012 (60,4% increase), 3 045 in 2011 (73,7% increase) and 2 301 in 2010 (80,2% increase) respectively. The correction of data for CEMI RAS allowed the organization to move from the 120-th place to the 73-th place in the list of all Russian research bodies, ranged by h-index. This is the contribution of only small part of the community, and one could think about the scaling this effect due to the engagement of all users.

3. The crowdsourcing model for CRIS

Crowdsourcing, however, rises at least two questions: motivation and quality control. These issues are addressed to after the description of the crowdsourcing model. Modern CRISs establish citation links automatically. If the database has no publication to match the reference, but stores publication with the similar title and author list, let this pair – reference and publication – to be considered as a suspected citation link. Let all the suspected citation links to

be placed in special list called List of Tasks. Therefore, the first two elements of the model suggested are the reference analyzer and the List of Tasks. Not only the analyzer adds the suspected links to the list, but also users do the same job. Managing her profile user could specify a publication citing her paper. This new pair is considered as suspected citation link. But before the pair from the user would be placed in the List of Tasks, she has to examine some number of pairs from the list. To be clear, suggest this number equals three. While examining the pair user must conclude whether the reference corresponds with the publication from the pair of not. The answer on each pair is also stored in the List of Task as an attribute of the pair. As a result, one new pair would increase the List of Task and the decision on three pars would get an additional score. Three users should examine each pair from the list before the decision is made by the simple majority rule. However the number of votes for the decision to be made is a value defined by the operator of particular CRIS and it may be greater than three. When a pair from the List of Task collects two affirmative votes, the citation link is established and the pair is removed from the List of Tasks. When two negative votes are collected, the link is disproved. Note that newly created links could be stored in the separate database or in general one, but with a special label assigned. The recognition of citation links established by the users do not harm the original database.

4. Hurdles for the model adoption

One should keep in mind two issues when developing crowdsourcing methods. One is the system operator's reluctance because of the possible revenue reduction. The second is insufficient users' activity as crowdsourcing workers.

In the Elibrary system the full management of the personal profile is available under the paid subscription. Thus, the free function of linkage a reference to the publication rises a natural question, whether free citation linkage damages the number of paid subscriptions. While answering this question one should keep in mind that the advanced features are available for paid subscription not to the authors, but to the research organisations. Having paid for subscription, an organisation gets an access to the full profile of its every single employee. And if the particular author is not recognized as a single entity in the system, the orgsnisation could create her profile and to link all her publications stored in the Elibrary. Metadata of every publication associated with the organisation could be edited under the paid subscription: misprints in the title or abstract could be corrected, the classifier could be specified, full text could be uploaded and the list of references could be edited manually. The last function is very important since imperfect algorithms could recognize two citations following each other in the reference list as a single citation. All the above mentioned functionality would remain as a paid service for organisations. Authors could link unrecognized citations without an access to the advanced features. In this case, business interests of the administrator of the system are affected very slightly.

Another hurdle – users' involvement – should be discussed in more details. Shown above are the estimates of the errors numbers in the profile of the organisation as well as in the entire system. The users of the system could face a huge workload and this rises a reasonable question on how fast would be moderated all the suggested links by the crowdsourcing technique. In order to investigate this issue one should compare two values. The first is the rate of moderation by the users, stimulated solely by the intrinsic motivation to improve their personal profiles. The second is the growth rate of the List of Tasks. If the intrinsic motivation could push the users to moderate more tasks than the algorithm puts in the List of Tasks, the administrator of the system could save money for the reward. The intrinsic motivation in moderation involves the analysis on how do authors distribute their attention and efforts.

5. Researchers as crowdsourcing workers

The time spent on the personal profile moderation depends strongly on how do authors distribute their attention – an important human resource in the intangible economy[1]. Some authors are passionate about the bibliometric indicators while the other are concerned about the writing an article of good quality. Indeed, more cited researchers tend to hold senior positions so the bibliometric indicators compete with different obligations for their attention. The time of such authors is too valuable for the society and it is spent on the issues that are more important. However, these authors receive a big number of citation and this number could be associated with the most reference errors. Suspected citation links could be revealed automatically and are placed in the List of Tasks – researchers on senior

positions are unlikely to apply for citation link manually. On the contrary, manually are likely to be placed the pairs from the authors, who begin their career and seek to improve values of indicators. Crowdsourcing workers driven by intrinsic motivation would work as long as they submit suspected citation links. When they stop to apply for new citation links, they will not moderate the List of Tasks any more. Consequently, answering the question about the moderation rate one should compare the number of the examinations versus the number of automatically revealed pairs. There are a lot of works no the optimal scheduling in crowdsourcing, however, this article presents quantitative estimation using the empirical data on the activity of the authors associated with CEMI RAS.

Consider the citation statistics of CEMI RAS as a data source for estimation the relationship between inflow and outflow in the List of Tasks. Elibrary demonstrate the citation numbers for researchers. While managing the profile of the institute in the Elibrary under the paid subscription, one can distinguish employees –active users. They are the best candidates for active crowdsourcing workers. About 10% of the institute's employees examined their profiles and reported mistakes to the representative of CEMI RAS. The representative edited the profiles under the contract between CEMI RAS and Elibrary. Variation of citation numbers allow to build probability distribution for active employees and the rest researchers. Median for citations in 2015 among employees-active users is 14 with the maximum value of 274. Suppose this distribution describes the number of citations the crowdsourcing worker gets every year.

Researchers with the high citation rates are rare interested in managing their profile. But they are cited more frequently, and the number of references to their works represents a significant value. Incorrect citations would be included in the List of Tasks automatically. Active users will later examine these suspected citation links while submitting references to their publications. Median for citations in 2015 among non-active users is 25 with the maximum value of 1 443. Note the difference in numbers for active and non-active users.

6. Evaluation of crowdsourcing activity

Empirical distributions help in answering the question whether the active users examine more pairs than input to the List of Tasks. Simple model is used to estimate the number of tasks that could be done by the crowdsourcing workers with intrinsic motivation and the probable number of errors. The issue of crowdsourcing works distribution is investigated in modern literature as optimal pricing subjected to the budget constraint and the fixed level of quality. The pricing of work is beyond the scope of this article. The information system under discussion already has active users, who are ready to spend some time in order to improve the quality of the data. Indeed, the number of crowdsourcing workers could be increased in case of the reward. And thorough investigation of workflow requires a lot of data that are not available for this research. Instead, data at hand make it possible to employ simple Monte-Carlo simulation in order to answer the question on the amount of work that could be accomplished by the interested users.

6.1. Simulation

Suppose, that after launching automatic algorithm produces N suggested reference links for the List of Tasks. Users can increase this list on n pairs of suspected citations. In this case a single user j submits n_j pairs while examining k pairs from the List of Tasks. Suggest that citation confirmation (or rejection) requires m votes with the majority supporting the corresponding decision. Adequate algorithm for the tasks ranking and the majority decision before the collection of m votes contribute to the decrease of the number of comparisons to be made. However, at worst each of $(N + n)$ pairs should be examined m times. Along with that the users can ensure only $k \times n$ examinations. Thus, all the works from the List of Tasks would be accomplished only if $m \times (N + n) \leq k \times n$, or $N/n \leq (k - m)/m$.

Simulation model calculates the probability distribution of the (N/n) ratio and relies strongly on the data of CEMI RAS activity in the Elibrary system. Suggest that the number of citation errors n_j for a single crowdsourcing worker j is a random variable with the same distribution as the annual citation number for active users from CEMI RAS. Assuming the portion of errors to be constant among different groups of users one could ground on the distribution of citation numbers because the relative value (N/n) is calculated. Thus, simulation exploits absolute values for citations under the assumption of constant portion of errors for a single author and for the electronic library as a whole. During

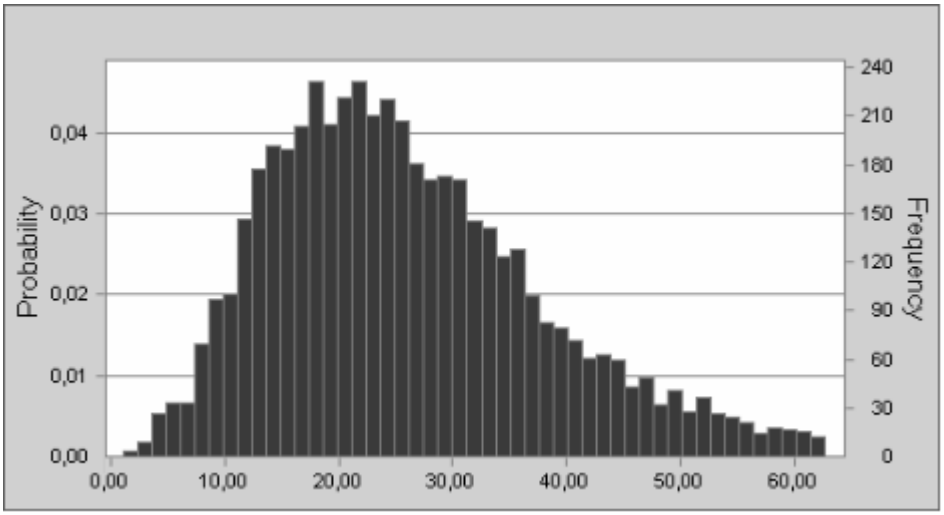

Fig. 1. Probability distribution for *(N/n)* ratio.

the one simulation the computer program calculates n_j for 4 000 active users ($j = 1..4000$) according to the empirical distribution. Four thousand active users is rather reasonable assumption. Elibrary reports 240 000 registered users, 93 000 of them manage their profile. When the set of data is created, the value of $N/(n_1 + \cdots + n_{4000})$ is calculated for a particular simulation. Every simulation creates unique set of n_j and results in a new value for $N/(n_1 + \cdots + n_{4000})$. N equals to the total number of citations in Elibrary, that could result in 57 782 968 errors mentioned earlier. Five thousand simulations in total were performed and this number results in probability distribution for (N/n), that is sufficient to estimate the efforts needed for the citation linkage correction. The simulated result of the (N/n) probability distribution is shown on Fig.1.

6.2. Simulation results discussion

Analyzing the percentiles of the distribution one could state (N/n) exceeds the value of 35,75 in the 20% of cases. That is with the probability of 80% the value $(k - m)/m = 35{,}75$ is sufficient in order to exhaust the List of Tasks by the 4 000 of crowdsourcing workers driven solely by the intrinsic motivation. However, one application to the List of Tasks from crowdsourcing worker would require from her to make about $k = 37 \times 3 = 111$ examinations, if the decision rule requires $m = 3$ votes in order to establish or reject citation link. The value of 111 examinations at a time is extremely high and is likely to undermine the intrinsic motivation. Thus, one should reject the hypothesis about the ability of intrinsic motivation as an only mechanism in crowdsourcing. Without an external motivation, users fail to accomplish all the available tasks within a reasonable time. But, the operator of the system should not deny the workers with the intrinsic motivation. Instead, two types of crowdsourcing workers – internally interested ones and paid workers – should be properly managed. Those who want to improve their indicators values, examine the pairs from the List of Tasks before the application is placed in the list. The rest users examine the pairs for a reward. An external motivation involves those who have no missing citations but is willing to get some prize. Worker differentiation of such kind is reasonable, and as an additional value for the active users, one could think about moving their applications to the top of the List of Tasks. Further, in order to implement the particular pricing procedures one should address the parameters of the particular information system and to the methods, described in the literature[3].

6.3. Optimistic estimates in case of high users' activity

The number of active users who manage their profiles properly depends strongly on the size of the system and the quality of the data. If the number is large enough, users accomplish the tasks driven solely by intrinsic motivation. If not, the List of Tasks grows faster than users examine its content, and additional incentives are needed. In order to

motivate users the operator of the system could assign some score per task. A user examines a pair from the List of Tasks and receives one score as a reward. When a threshold score level is reached, the user obtains an invitation to access some paid service in exchange for the scores collected. In Elibrary system an opportunity to register 10 publications could be an example of such service. The usage of scores instead of money reward prevents money transactions and all difficulties involved: taxes, payment processing, security issues, etc. Therefore, duplication – the same job, assigned to different users – ensures quality control and the reward ensures motivation and involvement.

In case of external motivation, the moderation promises optimistic results. Suppose one active user could correct 10 publications per day, spending 3 minutes to make one application. Given 4 000 active users all publications would be examined in 137 days (four and a half months). As for CEMI RAS experience 10% of employees volunteered to correct errors in the Elibrary system. While 4 000 active users among 93 000 form only 4,5% of the audience. Of course, 137 days is a rough estimation because the content is updated and more than 2 million publications are uploaded each year. One could suggest that 7,7% of them have errors in metadata or about 155 000 in absolute values. Therefore, active users would spend about 11 days per year to examine the content in the current mode.

7. Conclusion

The crowdsourcing model for research information systems is very clear. Reasons against the implementation of the procedure discussed could ground on the business interests of the operator of the system and technical issues. Second ones are well developed – algorithms for task distribution and dynamic pricing in crowdsourcing are published. What is new in the procedure is the method of data management by the community in the CRIS. But coding a new functionality is still challenging and the difficulty depends strongly on the platform of the particular CRIS.

Business interests suffer less than one could think. As the paid subscription for organisations is the main source of revenue, only one free feature among many do not harm: institutions would retain their subscriptions to access full functionality.

Simulation modelling demonstrate that the efforts of active users with intrinsic motivation could be insufficient to accomplish all the work. If this is the case, the operator of the information system should introduce a reward. Expressed as a set of scores such a reward could ensure access to the paid services for crowdsourcing workers without any money transactions. Estimates demonstrate that only small portion of users with strong motivation would examine a database with tens of millions items within a few months.

Acknowledgements

The financial support of this study through the Russian Science Foundation grant № 14-18-01999.

References

1. Falkinger J. Limited Attention as a Scarce Resource in Information - Rich Economies. *The Economic Journal*, 2008. 118(532), 1596-1620.
2. Mahieu B. et al. *Measuring scientific performance for improved policy making.* 2014. DOI 10.2861/57414
3. Minder P. et al. Crowdmanager-combinatorial allocation and pricing of crowdsourcing tasks with time constraints. *Workshop on Social Computing and User Generated Content in conjunction with ACM Conference on Electronic Commerce (ACM-EC 2012).* 2012
4. Pasula H. et al. Identity uncertainty and citation matching. *Advances in Neural Information Processing Systems 15 (NIPS 2002).*2003
5. Stempfhuber M. Information quality in the context of CRIS and CERIF. *Proceedings of the 9th International Conference on Current Research Information Systems*. 2008.