Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 106 (2017) 19 – 24

Procedia Computer Science

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
doi:10.1016/j.procs.2017.03.004

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