BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 14 2013, pages 1837–1839
doi:10.1093/bioinformatics/btt284

Databases and ontologies

Advance Access publication June 3, 2013

AuthorReward: increasing community curation in biological knowledge wikis through automated authorship quantification

Lin Dai[1,2], Ming Tian[2], Jiayan Wu[1], Jingfa Xiao[1], Xumin Wang[1], Jeffrey P. Townsend[3,4] and Zhang Zhang[1,*]

[1]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, [2]School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China, [3]Program in Computational Biology and Bioinformatics and [4]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520, USA

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Community curation—harnessing community intelligence in knowledge curation, bears great promise in dealing with the flood of biological knowledge. To exploit the full potential of the scientific community for knowledge curation, multiple biological wikis (bio-wikis) have been built to date. However, none of them have achieved a substantial impact on knowledge curation. One of the major limitations in bio-wikis is insufficient community participation, which is intrinsically because of lack of explicit authorship and thus no credit for community curation. To increase community curation in bio-wikis, here we develop *AuthorReward*, an extension to MediaWiki, to reward community-curated efforts in knowledge curation. *AuthorReward* quantifies researchers' contributions by properly factoring both edit quantity and quality and yields automated explicit authorship according to their quantitative contributions. *AuthorReward* provides bio-wikis with an authorship metric, helpful to increase community participation in bio-wikis and to achieve community curation of massive biological knowledge.

Availability: http://cbb.big.ac.cn/software.

Contact: zhangzhang@big.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 1, 2013; revised on May 8, 2013; accepted on May 13, 2013

1 INTRODUCTION

Biological knowledge is generated at ever-faster rates and dispersed among researchers and across literatures. As each new biological study has become increasingly dependent on the availability of existing knowledge, comprehensive and up-to-date collection of biological knowledge across a wide variety of research fields is of critical significance in life sciences (Clark, 2007).

Traditionally, biological knowledge has been aggregated through expert curation, conducted manually by dedicated experts. However, with the burgeoning volume of biological data and increasingly diverse densely informative published literatures, expert curation becomes more and more laborious and time consuming, increasingly lagging behind knowledge creation. Accordingly, community curation—harnessing community intelligence for knowledge curation—has gained significant attention as a solution to this issue (Salzberg, 2007; Waldrop, 2008; Zhang *et al.*, 2011). A successful example that engages community intelligence in knowledge aggregation is Wikipedia that features up-to-date content, huge coverage and low cost for maintenance. Spirited by the extraordinary success of Wikipedia, multiple biological wikis (bio-wikis) have been built to date (Supplementary Table S1).

However, bio-wikis have not achieved a substantial impact on community curation of biological knowledge (Finn *et al.*, 2012). One of the major limitations in bio-wikis is insufficient participation from the scientific community, which is intrinsically because of lack of explicit authorship and thus no credit for community-curated contributions (Finn *et al.*, 2012; Howe *et al.*, 2008). A valuable attempt has been made to motivate community contributions in wikis by means of social rewarding techniques (Hoisl *et al.*, 2007), but it does not provide explicit authorship for any wiki page. Although authorship has been introduced in a non-MediaWiki–based system (Hoffmann, 2008), it only links every sentence to its author but does not provide a quantitative measure of authorship, and most important, it is inapplicable to extant bio-wikis that are largely built on MediaWiki (a free, open source and widely used wiki engine, which is adopted by Wikipedia). Several initiatives based on semantic web technologies have already emerged for biological knowledge management (Antezana *et al.*, 2009). However, they do not promise to manage or quantify authorship of the free text in bio-wikis. To increase community curation in bio-wikis, here we develop *AuthorReward*, an extension to MediaWiki, to reward community-curated efforts in bio-wikis by contribution quantification and explicit authorship.

2 ALGORITHMS

MediaWiki allows anyone to develop customized functionalities by packaging a bunch of codes as MediaWiki extensions. Thus, *AuthorReward* is implemented as an extension to MediaWiki. Although MediaWiki itself includes an infrastructure for individual contributions to be recognized, it only records the revision history and provides no explicit authorship.

*To whom correspondence should be addressed.

A wiki page contains a collection of knowledge on a specific subject, where multiple researchers are most likely to collaboratively provide edits. *AuthorReward* aims to provide a viable quantification for researchers' contributions in bio-wikis. A major concern to automated authorship has been ensuring that authorship cannot be 'manipulated' by spurious, short-lived edits (Supplementary Text S1). For any wiki page p, we assume there are a series of edit versions $v_0, v_1, v_2, \ldots, v_n$, where version v_0 is empty and $n>0$. *AuthorReward* counts multiple successive versions edited by a researcher as one version. Thus, any neighboring versions, v_{i-1} and v_i (where $1 \le i \le n$), are edited by different researchers. The edit distance between v_i and v_j, termed as $d(v_i, v_j)$ (where $i<j$), is computed by the Levenshtein distance (LD) (Levenshtein, 1966) that measures the minimum number of edit operations (insertions, deletions and substitutions) required to transform one string into the other. In *AuthorReward*, the contribution score of version v_i, $CS(v_i)$, is formulated straightforwardly as

$$CS(v_i) = c[d(v_{i-1}, v_n) - d(v_i, v_n)], \quad (1)$$

where c is the scale factor, $d(v_{i-1}, v_n)$ is the edit distance between v_{i-1} and v_n and $d(v_i, v_n)$ is the edit distance between v_i and v_n.

In Equation (1), $CS(v_i)$ factors edit quality as well as edit quantity in an implicit manner; the edit quantity of version v_i, $QTY(v_i)$, amounts to the edit distance between v_i and its previous version v_{i-1}, viz., $d(v_{i-1}, v_i)$ [Equation (2)], and the edit quality of version v_i, $QAL(v_i)$, corresponds to whether the edit persists in comparison with the last version v_n [Equation (3)].

$$QTY(v_i) = d(v_{i-1}, v_i) \quad (2)$$

$$QAL(v_i) = \frac{d(v_{i-1}, v_n) - d(v_i, v_n)}{d(v_{i-1}, v_i)} \quad (3)$$

According to the triangle inequality, $QAL(v_i)$ ranges from -1, when the edits were entirely reverted, to $+1$, indicating that the edits were totally preserved in the last version. Therefore, $QAL(v_i)$, in other words, measures how long the edit lasts in the latest version; a high (or low) quality score is given for version v_i, if it is long-lived (or short-lived). Consequently, $CS(v_i)$ can be expressed by $QTY(v_i)$ multiplied by $QAL(v_i)$, namely, $CS(v_i) = QTY(v_i) \times QAL(v_i)$. Thus, $CS(v_i)$ is not easily gamed, providing a viable quantification for researchers' contributions.

Considering that one researcher may provide many discontinuous edits across the evolution of a wiki page, and thereby contribute multiple versions in one wiki page, the contribution score of researcher r in page p, $S(r, p)$, is quantified as the sum over all contributed versions,

$$S(r,p) = \sum_{vi \in E(r,p)} CS(v_i), \quad (4)$$

where $E(r, p)$ is a set of versions contributed by researcher r in page p. As a consequence, the total contribution of researcher r in a bio-wiki is termed as the sum of multiple contribution scores in all participated pages,

$$S(r) = \sum_{p \in P} S(r,p), \quad (5)$$

where P is a set of pages in which researcher r provides edits.

3 APPLICATION AND FEATURES

To test the functionality of *AuthorReward*, we installed it in RiceWiki (http://ricewiki.big.ac.cn). For testing purposes, we chose the semi-dwarfing gene (*sd1*), which is one of the most important genes deployed in modern rice breeding and is also known as the 'green revolution gene' affecting plant height of rice. There were nine researchers collaboratively annotating the *sd1* gene, providing 87 versions as of August 23, 2012 (Supplementary Table S2; http://ricewiki.big.ac.cn/index.php/Os01g0883800).

As testified on the *sd1* gene (Supplementary Fig. S1), *AuthorReward* is capable of yielding sensible quantitative contributions and providing automated explicit authorship, consistent well with perceptions of all participated contributors. Moreover, *AuthorReward* features good compatibility with any MediaWiki-based system and simple installation, consequently possessing a broad scope for its application and providing a consistent appearance and functionality as Wikipedia.

4 CONCLUSION

AuthorReward provides bio-wikis with an authorship metric, featuring robust contribution quantification and automated explicit authorship. When contribution is appropriately quantified and authorship is duly rewarded, it is possible to exploit the full potential of the scientific community in knowledge curation.

Although *AuthorReward* does not contribute directly to the integration of biological knowledge, it provides a standard practice to reward community-curated efforts, which in return can increase community participation in bio-wikis for knowledge curation. Thus, our intention here is to produce an automated, simple and robust authorship metric and no automated measure will be able to gauge scientific content. *AuthorReward* can be used in combination with semantic web technologies, potentially promising a significant advance for harnessing community intelligence for knowledge curation. In addition, social rewarding techniques (e.g. peer rating) can be used together with *AuthorReward* for contribution evaluation. Moreover, it is likely in the long term to integrate community-curated efforts across multiple bio-wikis for each researcher, which accordingly requires close collaborations among bio-wikis and standardized mechanisms for individual identity recognition (e.g. OpenID at http://www.openid.net).

AuthorReward provides a standard practice to reward community-curated efforts in bio-wikis, and it is of interest to the scientific community intending to perform knowledge curation collectively and collaboratively in bio-wikis and also other domain wikis.

ACKNOWLEDGEMENTS

The authors thank Jun Yu, Lina Ma, Gang Wu, Hao Wu, Chao Xu, Jian Sang and Ang Li for their valuable comments on this work.

Funding: National Programs for High Technology Research and Development (863 Program; 2012AA020409); the '100-Talent' Program of Chinese Academy of Sciences (Y1SLXb1365);

National Natural Science Foundation of China (60803050, 61132009); USA National Institutes of Health P01 (GM068067).

Conflict of Interest: none declared.

REFERENCES

Antezana,E. *et al.* (2009) Biological knowledge management: the emerging role of the Semantic Web technologies. *Brief. Bioinform.*, **10**, 392–407.

Clark,T. (2007) Knowledge integration in biomedicine: technology and community. *Brief. Bioinform.*, **8**, E1–E3.

Finn,R.D. *et al.* (2012) Making your database available through Wikipedia: the pros and cons. *Nucleic Acids Res.*, **40**, D9–D12.

Hoffmann,R. (2008) A wiki for the life sciences where authorship matters. *Nat. Genet.*, **40**, 1047–1051.

Hoisl,B. *et al.* (2007) Social rewarding in wiki systems–motivating the community. In: Schuler,D. (ed.) *Online Communities and Social Computing*. Springer, Berlin Heidelberg, pp. 362–371.

Howe,D. *et al.* (2008) Big data: the future of biocuration. *Nature*, **455**, 47–50.

Levenshtein,V. (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, **10**, 707–710.

Salzberg,S.L. (2007) Genome re-annotation: a wiki solution? *Genome Biol.*, **8**, 102.

Waldrop,M. (2008) Big data: Wikiomics. *Nature*, **455**, 22–25.

Zhang,Z. *et al.* (2011) Data integration in bioinformatics: current efforts and challenges. In: Mahdavi,M.A. (ed.) *Bioinformatics–Trends and Methodologies*. InTech, Rijeka, Croatia.