MISTA 2013

Guiding wiki crowds through resource scheduling

Ioanna Lykourentzou • Wassila Aggoune-Mtalaa • Dimitrios J. Vergados • Yannick Naudet

1 Motivation

Wikis are among the most popular technologies for collaborative knowledge production. In brief, a wiki is a collaborative content creation system, where users contribute knowledge content in the form of articles, while they can also edit and even delete the contributions of others [1]. Wikis have received significant interest in the past few years and they are increasingly being used to support knowledge development in many domains, from education, to scientific research, and from activities of the public sector to enterprise environments. Last, one of the most well-known and studied wikis, the popular Wikipedia, is an ever-growing source of information with millions active users and articles.

The rapid expansion and success of wikis is based on the open form of user collaboration that they are based on. That is, wiki users are free to edit any article they wish, with almost no restrictions on their access and edit rights. This open collaboration enables the massive production of wiki articles, which cover a broad spectrum of topics and expertise backgrounds. However, this same self-coordination poses significant limitations in terms of content quality. Take as an example Wikipedia: although it features a number of qualitative articles, it has also a very long tail of low-quality ones [2]. This inability to guarantee quality lowers the reliability of wikis and hinders their adoption.

To guide the wiki crowd, systematize contributions and help them utilize their knowledge competencies more efficiently [3], we propose a coordination scheme that can be viewed as a scheduling problem [4]. The wiki is seen as a system with resources, which are the users and their expertise, and tasks, which are the wiki articles that need quality improvement. The objective is to match users to articles, in such a way as to maximize the average quality of the articles inside the wiki, regarding specific constraints such as user workload.

The work is presented as follows: Section 2 formulates the problem and gives the methodology adopted to solve it. Section 3 shows some promising results. Last, Section 3 concludes the work and gives perspectives for the future.

Ioanna Lykourentzou, Wassila Aggoune-Mtalaa and, Yannick Naudet
Public Research Centre Henri Tudor, Luxembourg
E-mail: {ioanna.lykourentzou, wassila.mtalaa, yannick.naudet}@tudor.lu
Dimitrios J. Vergados
NTNU, Norway
E-mail: dimitrios.vergados@item.ntnu.no

2 Problem formulation and Methodology

The wiki scheduling problem can be formulated as follows. Given a set of:

- wiki articles $A = \{i_1, i_2, \dots, i_{|A|}\}$, with quality below the threshold. Each article i has two characteristics: a current article quality q_i which changes after a new user contribution and a knowledge topic $D_i \in D = \{D_1, D_2, \dots, D_{|D|}\}$. Each article is considered to belong to exactly one knowledge topic.
- wiki users $U = \{u_1, u_2, \dots, u_{|U|}\}$. Each user u_j has an *expertise vector* $\boldsymbol{e_j}$, with length equal to the knowledge domains $|D|$. Expertise indicates the quality that the system estimates that user u_j can bring to an article belonging to a specific domain. The user expertise can be computed using past user data (e.g. ratings of past user contributions, with the help of Feed Forward Neural Networks, as detailed in [5]),

the problem is to find which article should each user be requested to contribute to, so that the average quality of articles inside the wiki C is maximized:

$$C = \frac{\sum_{i=1}^{|A|} q_i}{|A|} \quad (1)$$

The constraints considered in this problem formulation are:

- all articles need to surpass a given quality threshold T, $0 < T < 10$. This threshold is considered to be fixed for all the domains, but in a more general problem setting it could be set individually for each domain, depending on the community requirements.
- a user can be recommended only one article at a time. The maximum number of articles that can be given to him is therefore a binary workload $w = \{0,1\}$.
- Non-preemptive process. Once a user has been assigned with an article, he cannot be interrupted, to be recommended with another one.

The wiki problem has an additional complexity compared to classical scheduling problems. First there is *uncertainty* regarding *resource availability*: users enter a wiki when they want, remain connected for as long as they like and they may or may not accept to make a contribution. Secondly there is *uncertainty* regarding *resource capacity* (user expertise): the mechanism cannot know a priori if an expert user will enter the system, or whether the users that will enter are not knowledgeable enough to improve a certain article. Thirdly, a wiki article might require the contributions of multiple users before reaching the quality threshold, which brings upon the *need for sequenced, chain scheduling*, with hands-off resource dependencies (the input of one user starts when another user has finished contributing to the same article). Due to the above constraints, the problem complexity is high. Therefore, in the scope of this paper it was decided to opt for a heuristic greedy algorithm: once a user enters the wiki, the mechanism should suggest him with an article that the system estimates, at that moment, that the user is best fit for. This solution is not expected to necessarily lead to optimality but, in case it can improve the current quality state of wiki systems, it can be judged as sufficient. The overall wiki coordination mechanism, featuring the scheduling algorithm, functions as follows. For every article that is inserted into the wiki, either as a new article or as a contribution to a previous article version, the mechanism evaluates the article's quality, as a single numerical value. The mechanism then compares this quality to a pre-defined quality threshold. If the article does not surpass it, then the article needs to be enhanced by an additional user. The selection of which user will be asked to contribute to which article is handled by the greedy algorithm. The process of successive article contributions, quality evaluations and user selections continues until the article surpasses the quality threshold.

3 Evaluation

To evaluate the proposed algorithm we model two systems:

- *Benchmark*. In this system, users enter the wiki, view and contribute to articles as they would in a typical wiki, without any recommendations. The system is calibrated to function similarly to the English Wikipedia[6],on various statistical parameters

including the arrival rate of users, the user expertise distribution, the user probability of viewing each specific article and the user probability of contributing to an article. This calibration was validated using two indicative factors: workload and quality.

- *Smart*. The smart system extends the benchmark system, by applying on it the greedy algorithm that suggests articles to users. Apart from that, the other characteristics are identical to those of the benchmark. We refer to the smart system as a CI (Collective Intelligence) system, in line with the definition given in [7], and to highlight the fact that the system improves the emergence of the collective intelligence of the involved community, by combining user and algorithmic intelligence.

Comparing the performance of the benchmark and the CI system (Fig. 2), we may observe that the scheduling algorithm results in a significant improvement in the objective function, i.e. on the average article quality. We may also observe that this improvement is quicker as compared with wiki articles which often reach adequate quality levels in a slow manner.

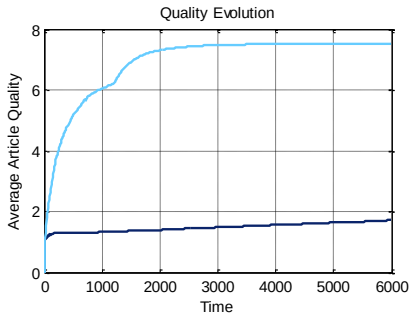

Fig.2. Evolution of the article quality achieved through the use of CI and benchmark systems

Another interesting feature to examine (Fig. 3a) is that the CI system leads to slightly more edits performed by participating users per article, compared to the benchmark system, however with a significant shift of the article quality distribution (Fig. 3b). Thus, the community manages to produce more qualitative articles through the use of the CI system, compared to the respective result achieved through the use of the benchmark system.

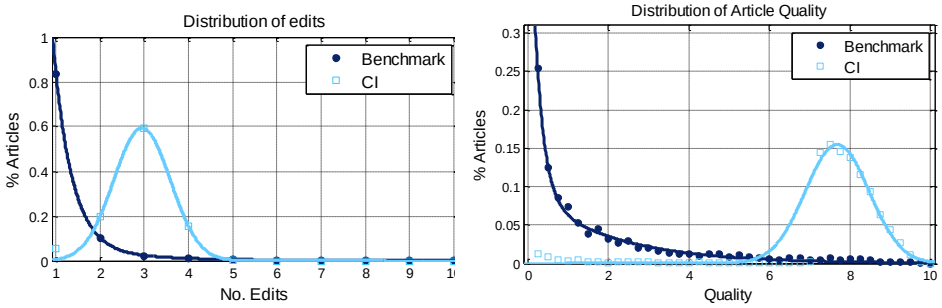

Fig.3. Comparison of distribution of a) edits and b) quality produced by the benchmark and CI system

The above results indicate that the scheduling-enabled version of the wiki system can help increase the produced article quality, better use expertise and reduce the time needed for the articles to reach satisfactory levels of quality.

4 Conclusions

In this work, we address the problem of low quality in wikis. A heuristic greedy algorithm is used to allocate users to the wiki articles that they can improve the most.

Experimental results show that this approach can indeed increase average article quality inside the wiki. Finally, viewing user coordination in modern web 2.0 systems as a resource scheduling problem can be extrapolated to other crowd-involving domains, such as crowdsourcing and this is a very interesting future research direction.

Acknowledgements: The work of I. Lykourentzou is supported by the National Research Fund, Luxembourg and co-funded under the Marie Curie Actions of the European Commission (FP7COFUND).The work of D. J. Vergados is supported by the ERCIM "A. Bensoussan" Fellowship.

References

1. P. Louridas, Using Wikis in Software Development, IEEE Software, 23, 88-91 (2006).
2. S.K. Lam and J. Riedl, Is Wikipedia growing a longer tail?, Proc. Proceedings of the ACM 2009 international conference on Supporting group work, ACM (2009).
3. A. Kittur, et al., Coordination in collective intelligence: the role of team structure and task interdependence, Proc. Proceedings of the 27th international conference on Human factors in computing systems, ACM (2009).
4. R. Lewis, A survey of metaheuristic-based techniques for University Timetabling problems, OR Spectrum, 30, 167-190 (2008).
5. I. Lykourentzou, et al., CorpWiki: A self-regulating wiki to promote corporate collective intelligence through expert peer matching, Inf. Sci., 180, 18-38 (2010).
6. F. Ortega, "Wikipedia: A quantitative analysis," Doctoral Thesis, Departamento de Sistemas Telemáticos y Computación (GSyC), Universidad Rey Juan Carlos, Móstoles, Spain, 2009.
7. T.W. Malone, et al., Harnessing Crowds: Mapping the Genome of Collective Intelligence, MIT Center for Collective Intelligence, 2009.