WikiSim: simulating knowledge collection and curation in structured wikis

Luna De Ferrari  
School of Informatics  
University of Edinburgh  
Edinburgh, United Kingdom  
luna.deferrari@ed.ac.uk

Stuart Aitken  
AIAI  
School of Informatics  
University of Edinburgh  
Edinburgh, United Kingdom

Jano van Hemert  
National e-Science Centre  
School of Informatics  
University of Edinburgh  
Edinburgh, United Kingdom

Igor Goryanin  
School of Informatics  
University of Edinburgh  
Edinburgh, United Kingdom

ABSTRACT
The aim of this work is to model quantitatively one of the main properties of wikis: how high quality knowledge can emerge from the individual work of independent volunteers. The approach chosen is to simulate knowledge collection and curation in wikis.

The basic model represents the wiki as a set of of true/false values, added and edited at each simulation round by software agents (users) following a fixed set of rules. The resulting WikiSim simulations already manage to reach distributions of edits and user contributions very close to those reported for Wikipedia. WikiSim can also span conditions not easily measurable in real-life wikis, such as the impact of various amounts of user mistakes.

WikiSim could be extended to model wiki software features, such as discussion pages and watch lists, while monitoring the impact they have on user actions and consensus, and their effect on knowledge quality. The method could also be used to compare wikis with other curation scenarios based on centralised editing by experts. The future challenges for WikiSim will be to find appropriate ways to evaluate and validate the models and to keep them simple while still capturing relevant properties of wiki systems.

Categories and Subject Descriptors
H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces—Computer-supported cooperative work; I.6.7 [Simulation and Modelling]: Simulation Support Systems; K.4.3 [Computers and Society]: Organizational Impacts—Computer-supported collaborative work

General Terms
Simulation

Keywords
Wikis, Wiki simulation

1. INTRODUCTION
This work is based on agent modelling and builds on recent literature analysing quantitative measures of existing wikis, especially Wikipedia. Recent work has extensively measured the growth, authorship and edit distributions of various language editions of Wikipedia [14, 10]. Other work has concentrated on finding information quality metrics and on capturing the social processes who regulate Wikipedia activity [11, 12].

Our idea is to develop the simplest model of collaborative data curation which still shows some of the emergent properties of wikis. Our boolean model of wikis (an article is represented as a true/false value) is akin to structured wikis, such as OntoWiki [5] which support finely grained data collection and editing, down to individual numbers and strings.

Having complete control over the simulation allows us to make all parameters explicit and to explore which aspects affect the completeness and correctness of the knowledge collected in the wiki. The model also allows to simulate a variety of conditions not easily measurable or reproducible in real-life wikis (for example, to vary the amount of user mistakes).

2. THE MODEL
The basic model of wiki curation described in this work contains a wiki and some users guided by a fixed action plan (Figure 1). For the purpose of our experiment, a wiki is a set of knowledge elements with either a correct (true) or incorrect (false) value. The wiki is initially empty. Each user knows a subset of all the knowledge elements in the system. The users’ knowledge is partly overlapping and, since human knowledge is imperfect, some elements of each user’s knowledge are incorrect (false instead of true). At each simulation step a user can add a knowledge item to the wiki or change...
an item which is already in the wiki. Users have a given energy and each action (search, add and edit) consumes a fixed amount of it. When the user energy falls to zero, the user stops acting (in real life, users only have a limited time to spend on the wiki).

![A simple wiki model](image)

Figure 1: The wiki model: adding and editing elements.

The simulation parameters, such as the number of users, the user energy, the percent of mistakes or the action costs can be varied at will. With this model we can already quantify how external and internal parameters affect the basic knowledge quality, for example, measuring the number of correct items present in the wiki at the end or during the simulation. External parameters (which depends on the user community but not on the particular wiki software used) include: the number of users, how much the users knowledge overlaps and the distribution of correct knowledge in the population of wiki users. Internal parameters (which depend on the wiki software) include the absolute and relative cost of each ‘search’, ‘add’ or ‘edit’ action. For example, editing an existing wiki entry generally takes less time than adding a completely new one, so in the model an ‘edit’ action will cost less than an ‘add’ action.

The probability distributions play a particularly important role in randomising repeated simulations. Probability distributions (uniform or power-law) are used to: 1. Extract knowledge elements from a pool to build the internal knowledge of each user (power-law). 2. Introduce random mistakes in the users knowledge (uniform). 3. Randomise the order in which the users act on the wiki (uniform). 4. Extract which knowledge element (from its internal knowledge) the user acts on (uniform). 5. Assign the user energy (power-law).

The WikiSim software tracks the system states and all user actions during the simulation, allowing from general measurements to potentially very detailed analysis of user actions and edit wars in the style of graphical history flows [13].

The model has been implemented in Java 1.5 (using a test driven methodology [7]). A MySQL database provides data storage. All the code, tests and documentation is available under the University of Edinburgh Gnu Public License on our wiki.

3. THE RESULTS

People knowledge and favourite topics tends to be power-law (long-tail) distributed in many contexts [3]. Qualitatively, this corresponds to saying there are things everybody tends to know (or is interested in) and then a long tail of specialistic knowledge. Contributions to Wikipedia tend to be power-law distributed too, with less than 3% of the users responsible for more than 50% of the edits [14, 8].

To simulate the power-law effect, we have used a power-law distribution (power-law shape factor=2.0 and scale factor=1.0) both to build the individual user knowledge and to assign the user energy. We have then compared the distributions of user edits in our simulation with those of Wikipedia [14]. Figure 2 shows the number of authors versus the number of distinct articles they have authored. The similarity with the Wikipedia plot is remarkable (shape factor 1.3 in the simulation versus 1.5 in Wikipedia) and could be used to further refine the model parameters.

Another example of simulation results shows how an increasing percentage of user mistakes affects the number of correct entries accumulated in the wiki over time (Figure 3). In the figure, the total correct system knowledge (100% mark) is the union of all correct (true) values known by the users (the wiki cannot become better than the sum of the correct knowledge of its users). The data on individual user actions allows to explain why the wiki never reaches more than 60% of correct elements. The reason is that the users spend most of their energy in edit wars over the most popular topics. The model does not provide yet a way of reaching consensus (see “Future work”).

4. MODEL APPLICATIONS AND FUTURE WORK

WikiSim offers a framework which could be applied to several areas. Four could be of particular interest for our group: 1. modelling the user internal knowledge, 2. modelling the wiki software (and the structured semantic wiki features), 3. comparing models of centralised curation and wiki curation and 4. measuring the impact of machine learning on the curation process.

1http://www.bioinformatics.ed.ac.uk/wiki/PublicCSB/WikiSim

2Model parameters Number of simulation steps: 1000; number of users: 10; knowledge elements per user: 20% of the 100 system elements (power-law distributed); percent of mistakes in the user knowledge: 20% for Figure 2, 5% to 80% for Figure 3; user energy: power-law distributed (truncated at 1000); action costs: add=3 energy units, edit=2 units, search=1 unit.
4.1 The user competence model

Our current model is based on an implicit premise: a user acts more on the knowledge items it is more interested in. If we go a step further and say that wiki users are reasonably good judges of their own knowledge and competence, we can introduce a second principle: the more the user is interested in a topic, the more likely it is to have some correct knowledge about that topic - on average -.

Obviously, this does not mean that the user knowledge will be perfect, it just introduces a potential correlation between the user awareness of its expertise, the quality of the user knowledge and the probability the user acts on it. In the model, this could be represented as a user having less mistakes in its 'favourite' knowledge (the interest in that topic correlates with a higher probability of knowing something correct about the topic) while acting on it more often (the desire to contribute to a favourite topic overcomes the effort of adding or editing).

This principle is likely to have a powerful effect on wiki knowledge quality, however, it is difficult to prove: in real life there is no easy way to measure 1. how good a person’s knowledge of a topic is, 2. how good the person is at judging his own expertise on the topic and 3. how that relates to the person’s actions, such as the willingness to edit the corresponding Wikipedia article.

4.2 Modelling wiki software affordances

The WikiSim model could be enriched to represent typical wiki software features such as discussion spaces and watch lists and measure their impact on user actions and knowledge quality. Discussion could be represented in the model by allowing the user to consider, before acting, both its own knowledge and a count of other users actions (true/false edits). Automatic notification of watched topics could be modelled by pushing items of interest - recently edited by other users - to the top of a user’s action list. The simulations could then quantify how these social affordances, such as visibility of other authors choices, can help to build consensus and better allocate user energy.

Semantic structured wikis (such as OntoWiki [5]) could be modelled by introducing relations between knowledge elements. Before changing a knowledge element in a correlated group, the user could be influenced by the value of the other connected elements. This would also create another measure of knowledge quality in the model: whether the element has been connected to an appropriate semantic group/class.

Structured wikis are particularly suited to create editable knowledge bases or to reannotate database data. The results of the WikiSim simulations could be interesting for biological sciences scenarios, where curation efforts using wikis are currently under way in many areas, from microorganism [4, 1] and molecule resources [2] to more ambitious initiatives like WikiProteins [9], pre-populated with 1,000,000 concepts from PubMed articles, EMBL, SwissProt and Gene Ontology.

4.3 Centralised versus wiki curation

In Biology the wiki approach to data curation is relatively recent. The current curation scenarios (Figure 4) are either centralised (with edit reserved to official curators) or mixed, with users allowed to report mistakes but not to perform or discuss the edits. It would be very interesting to simulate and compare the three scenarios. Expert curators could be modelled by giving them more correct knowledge and more energy than basic users. Edits could be either exclusively...
initiated by expert curators or channelled through them.

Figure 4: Curation scenarios: wiki curation versus centralised curation.

4.4 Integrating machine learning
Considering the scenario of biological structured wikis, it has been proved that the amount of raw data to be curated largely exceeds the availability of human curation [6] even if a wiki approach is used. This limitation has already brought into Biology extensive application of machine learning techniques to predict biological function, profiting from the fact that evolutionary mechanisms make inference of biological function possible across species. However, machine learning needs training on known data: structured semantic wikis could provide the machine readable examples for automatic learning. The next step in WikiSim could be to model the integration of machine learning into the curation model. For instance, data mining could be used to prioritise user edits. Simulations on real data could highlight the circumstances under which machine learning can or cannot help in making the wiki knowledge more coherent and more correct over time.

5. CHALLENGES
The first challenge will be to keep the models simple while still representing some essential aspects of individual author behaviour, wiki software features and social collaboration. The second parallel challenge will be to evaluate the model results appropriately. This will require rigour and attention in applying a computer science approach to insights collected for very different purposes, such as analysing data quality, neutrality of news or collective organisation of work in open knowledge/source projects. An additional challenge will be to represent and evaluate a model of semantic wikis, as not many have been extensively used or measured so far.

The simulation approach though, however simplistic and inherently biased, still tries to make the parameters under study explicit, and through this effort it highlights many additional questions on what is a wiki and what makes it a functioning collaboration. WikiSim has the potential to add a dynamic and measurable component to many insights and theories that will be developed in wiki research in the next few years.

6. REFERENCES