WikiReviz: An Edit History Visualization for Wiki Systems

Jianmin Wu and Mizuho Iwaihara

Graduate School of Information, Production and Systems, Waseda University,
Fukuoka 808-0135, Japan
jianmin.wu@moegi.waseda.jp, iwaihara@waseda.jp

Abstract. Wikipedia maintains a linear record of edit history with article content and meta-information for each article, which conceals precious information on how each article has evolved. This demo describes the motivation and features of WikiReviz, a visualization system for analyzing edit history in Wikipedia and other Wiki systems. From the official exported edit history of a single Wikipedia article, WikiReviz reconstructs the derivation relationships among revisions precisely and efficiently by revision graph extraction and indicate meaningful article evolution progress by edit summarization.

Keywords: Wikipedia, Mass Collaboration, Visualization.

1 Introduction

As a collaborative project, online encyclopedia Wikipedia receives contribution from all over the world [13] and its content is well accepted by those who want reliable social news and knowledge. Guided by the fundamental principle of "Neutral Point of View", Wikipedia articles need plenty of extra editorial efforts other than simply content expanding and fact updating. Users can choose to edit on an existing revision and override the current one or revert to a previous revision. However, there is no explicit mechanism in Wikipedia to trace such derivation relationship among revisions, while the trajectories how such collaboration appears in Wikipedia articles in terms of revisions are valuable for group dynamics and social media research [4]. Also, research exploiting revision history for term weighting [1] requires clean history without astray, which can be accomplished by such trajectories.

Wikipedia and other Wiki systems generally keep all revisions' texts for each article and make the edit history publicly available. The meta-data of the edit history, such as timestamps, contributors, and edit comments are also recorded. We propose WikiReviz to model the article evolution process as *revision graph*. Here ReViz stands for "Revision Visualization". A *revision graph* is a DAG (directed acyclic graph) where each node represents one revision and each directed edge represents a derivation relationship from the origin node to the destination node [6]. Users create a new revision by editing either the current revision, or one of past revisions. Also, a completely new input may replace the current revision. Most existing research

L. Chen et al. (Eds.): APWeb 2014, LNCS 8709, pp. 670–673, 2014.

modeling Wikipedia's revision history choose tree [4][5] or graphs [2] to represent the relationship, but few of them concern about the accuracy.

2 WikiReviz Overview

WikiReviz takes the original XML file of edit history dump as an input, and automatically generates the revision graph with edit summaries in the GraphML format [9]. It is intended for English Wikipedia and other languages that have inter-word separation. WikiReviz is implemented in Java and currently consists of two functional parts: the Revision Graph Extraction [6] Unit, which reconstructs the graph structure from original edit history; and Edit Summarization [7] Unit, where we generate supergram summaries on revision graph.

2.1 Revision Graph Extraction (RGE)

For a given revision r, at least one parent revision r_p should be identified from r's previous revisions, which involves comparison between those revisions. The best candidate is decided by a certain similarity measure as well as the characteristics of Wikipedia editing. In WikiReviz, revisions are split into supergrams, which are maximal-length phrases and retrieved by word transition graph and path contraction. After comparing with others' supergrams, supergram diff score can be computed for each revision pair in the comparison scope. The whole process is shown in Fig. 1.

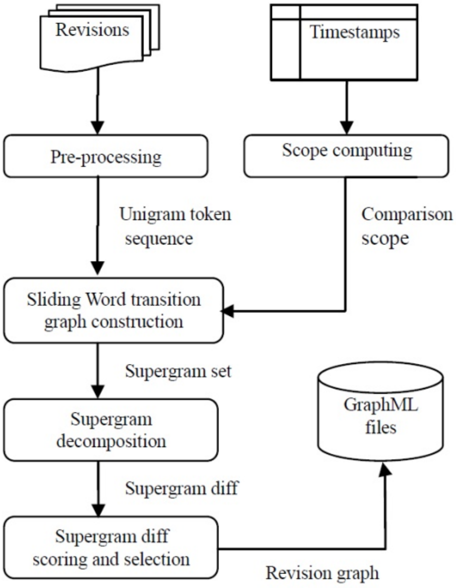

Fig. 1. RGE Process

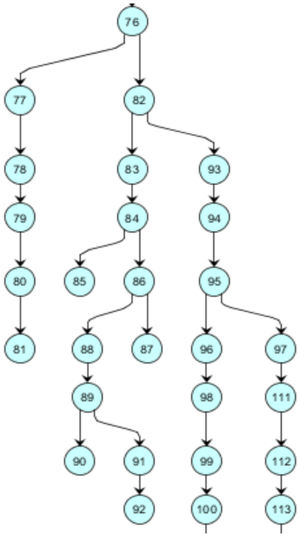

Fig. 2. Part of result graph "PhpBB", from revision #76 to #113

The output of the RGE part is a GraphML file illustrating the revision graph, which can be viewed by existing software such as yEd graph editor. Fig. 2 shows a part of the graph of article "PhpBB" about an Internet forum package written in the PHP scripting language [8], where each node's id represents a revision number. Since the whole revision graph is too large, we only capture a close look at a small portion of it.

2.2 Edit Summarization

In the process of edit summarization, WikiReviz automatically summarizes contributed contents during a specified edit period in the revision graph of a Wikipedia article, into a group of maximal-length phrases, i.e. supergrams, and attaches to the original revision graph. From the supergram generated in previous RGE part, two supergram selection algorithms, TF-IDF and Extended LDA ranking are developed to pick up representative supergrams.

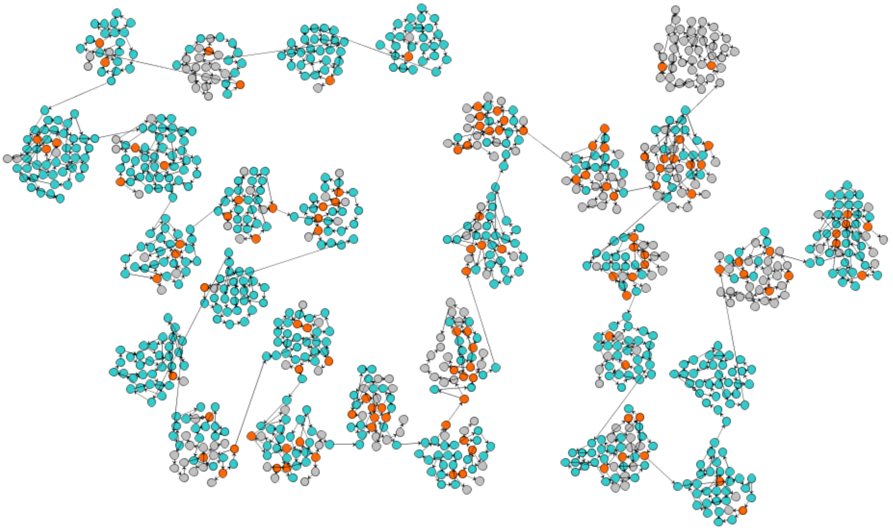

Fig. 3. Revision graph of article "Nazi Germany", first 1000 revisions, compact layout

The scopes upon which summaries would be applied are determined based on the type of topics. In a revision graph, one revision usually contains multiple topics, and the amount of topics will increase over time. We classify these topics based on which context in the revision graph the topics represent:

1. **Popular topic** is a topic that is most prominent among all the revisions in a view. We can discover such topics by LDA from the revision graph.
2. **Surviving topic** is a topic that appears at a revision, and continues to appear until the latest (current) revision. It can also be described as a surviving topic in the mainstream. After a period of edits, certain topics become stable and survive to the latest.

3. **Extinct topic** is a topic that is not surviving to the current revision. The definition of surviving topics is relative to the current revision, so if there are large amount of deletes after the current revision, several topics may be lost and surviving topics can be changed to extinct.

Fig. 3 shows an example, the whole revision graph of the first 1000 revisions of article "Nazi Germany", where scopes are divided by orange nodes. The blue nodes represent revisions that contain the surviving topics and the grey nodes are for extinct topics.

3 Conclusion

In this demo, we describe the motivation and features of our edit history visualization system WikiReviz, with visualization of revision graphs representing evolution processes of Wikipedia articles.

References

1. Aji, A., Wang, Y., Agichtein, E., Gabrilovich, E.: Using the past to score the present: extending term weighting models through revision history analysis. In: CIKM 2010, pp. 629–638. ACM, New York (2010)
2. Keegan, B., Gergle, D., Contractor, N.: Staying in the Loop: Structure and Dynamics of Wikipedia's Breaking News Collaborations. In: WikiSym 2012. ACM, New York (2012)
3. Lih, A.: Wikipedia as participatory journalism: Reliable sources: Metrics for evaluating collaborative media as a news resource. In: Proc. Int. Symp. Online Journalism (2004)
4. Ekstrand, M., Riedl, J.T.: rv you're dumb: identifying discarded work in Wiki article history. In: WikiSym 2009. ACM, New York (2009)
5. Flöck, F., Vrandečić, D., Simperl, E.: Revisiting reverts: accurate revert detection in Wikipedia. In: Proc. Hypertext and Social Media, pp. 3–12. ACM, New York (2012)
6. Wu, J., Iwaihara, M.: Revision graph extraction in Wikipedia based on supergram decomposition. In: Proc. WikiSym 2013(OpenSym 2013), Hongkong (August 2013)
7. Li, B., Wu, J., Iwaihara, M.: Tracking Topics on Revision Graphs of Wikipedia Edit History. In: Li, F., Li, G., Hwang, S.-w., Yao, B., Zhang, Z. (eds.) WAIM 2014. LNCS, vol. 8485, pp. 204–207. Springer, Heidelberg (2014)
8. `http://en.wikipedia.org/wiki/PhpBB`
9. `http://graphml.graphdrawing.org/specification.html`