Chapter 5
Visualizing Wikipedia Article and User Networks: Extracting Knowledge Structures using NodeXL

Shalin Hai-Jew
Kansas State University, USA

ABSTRACT

As a public and open-source resource, Wikipedia is used by many in the public as a quick reference; academic researchers have tapped Wikipedia for human- and machine-based insights. The MediaWiki understructure enables a wide range of transparency, enabling users to very easily search and access template-structured text and image contents, source citations, contributors, page histories, and others. Transparency has been hard-wired into the platform technology. Developers have been building tools to extend the transparency of data built on a MediaWiki understructure. Network Overview, Discovery and Exploration for Excel (NodeXL) features a third-party graph data importer that enables the extraction of MediaWiki article (and user) networks, which include all of the languages of Wikipedias. This chapter highlights the uses of graphs from two main types of Wikipedia pages for increased knowledge transparency: (1) topical article pages and (2) contributor user pages (whether human or robot).

INTRODUCTION

One of the main questions of the Web 2.0 or Social Web age is a core one related to human collective action: Is it possible for technology to change-up the interests of (generally selfish) human participants so that they can collaborate productively and pro-socially over time? The basic understanding is that there is a difficult value proposition at play: contribute effort and resources to a common effort (and submit to various risks) with little promise of direct benefit. If Wikipedia were the only consideration, the answer to the initial question would be: “Of course!”

Wikipedia, the crowd-sourced wiki-based encyclopedia run by the Wikimedia Foundation, regularly ranks in the top ten visited sites on the Web. As of the moment of this chapter’s creation, the English

DOI: 10.4018/978-1-4666-8844-5.ch005
version contains 4,672,164 articles. Wikipedia is “the largest existing repository of encyclopedic knowledge, freely available in 282 languages. It combines free-form natural language content with structural information, represented by intra- and inter-language links” (Tonelli, Giuliano, & Tymoshenko, 2013, p. 204). Its articles have been improved with human editorial oversight and robot-based endeavors (to update data, clean up the spelling, and deny human vandals access to permanent changes), and there are a number of metrics applied to understand the quality of its contents. One approach, for example, suggests that article quality may be effectively indicated by the “number of edits, number of editors and intensity of cooperative behavior” (Wilkinson & Huberman, 2007, p. 157).

The underlying MediaWiki software enables basic page templating, topic categorizing, alphabetical structuring of pages (within categories), full electronic memory of page edits (the historical evolution of each article), and roll-back functions; it enables accommodation of virtually all comers—those humans who want to create accounts vs. those who want to make changes anonymously (albeit with the capturing of their Internet Protocol or “IP” addresses). Contributing is simple with pages already templated and set up for markdown syntax, which is a highly simplified way to format data for the page (http://www.mediawiki.org/wiki/Extension:MarkdownSyntax). The editing process is incremental and often born-digital. Figure 2 shows “The MediaWiki Landing Page.”

Those who contribute to Wikipedia have been studied fairly extensively, from their motivations to their “survival” as contributors over time to the power contributors (known as the 10,000-contributors strong “Wikipedians” as a subset of the 22 million registered accounts on the English Wikipedia but with only 73,000 active editors as of May 2014) (“Wikipedia Statistics,” May 2014). Multiple researchers have observed wide disparities between those who contribute much over time and those in the long tail who contribute little, with high inequality distributions based on both the Lorenz curve (which shows how far off a distribution is from perfect equality) and Gini coefficient (which shows inequality in a frequency distribution). Applying these statistical techniques, the authors observed initial oscillations during the early lives of various language editions of Wikipedias before settling in to a certain level of inequality of contribution; they found high levels of inequality “for every language edition processed,

Figure 1. An annotated screenshot of a Wikipedia page from English Wikipedia