Manypedia: Comparing Language Points of View of Wikipedia Communities

Paolo Massa  
Bruno Kessler Foundation  
Via Sommarive, 14  
38123 Trento - Italy  
massa@fbk.eu

Federico Scrinzi  
Bruno Kessler Foundation  
Via Sommarive, 14  
38123 Trento - Italy  
fscrinzi@fbk.eu

ABSTRACT
There are more than 270 Wikipedias in different languages. On every article, each Wikipedia community strive to reach consensus on representing all significant views fairly, proportionately, and without bias. But are Linguistic Point Of View (LPOV) of each Wikipedia diverse? Manypedia is a mashup combining and translating Wikipedia pages that helps to answer this question. Manypedia can be accessed at http://sonetlab.fbk.eu/manypedia/

Keywords
Wikipedia, mashup, API, javascript, jQuery, natural language processing, linguistic point of view (LPOV)

1. MOTIVATION FOR MANYPEDIA
Possibly Wikipedia is becoming the largely most accessed Web resource for information needs. 53% of American Internet users look for information on Wikipedia as of May 2010 and this number increased from 36% in February 20071. Every month 78 million visitors access Wikipedia, as of January 20102. It is undeniable that a large share of people rely on Wikipedia for forming their representations of facts, of what is true or not.

This fact is even more interesting considering that every single word on which so many people rely could have been added by anyone. In fact Wikipedia’s slogan is “the free encyclopedia anyone can edit” and indeed the more than 3 million pages of the English Wikipedia, started in 2001, has received more than 453 million edits by more than 14 million registered users3. It is even possible to edit Wikipedia without performing the login with the personal username and hence to edit the encyclopedia anonymously.

Given the importance in shaping the wisdom and world view of so many people, we believe it is important to raise awareness on who are the people who edit Wikipedia. This concern is fully shared by the Wikipedia community itself: for instance on Wikipedia there is a page in the “Wikipedia:” namespace, the one devoted to policies and rules titled “Wikipedia:Systemic bias”4

which states “the Wikipedia project suffers from systemic bias that naturally grows from its contributors' demographic groups, manifesting as imbalanced coverage of a subject, thereby discriminating against the less represented demographic groups.”

The page clearly lists the main biases: “The average Wikipedian on the English Wikipedia is a male, technically inclined, formally educated, an English speaker (native or non-native), European-descent, aged 15–49, from a majority-Christian country, from a developed nation, from the Northern Hemisphere, and likely employed as a white-collar worker or enrolled as a student rather than employed as a labourer”.

However here we are not interested in the biases internal to a specific wiki such as the English Wikipedia but, on the other hand, we focus on the existence (or absence) of different biases in different language communities of Wikipedia. In fact, while the largest and oldest Wikipedia is in English language, there are currently more than 270 Wikipedias in as many different languages, ranging from many with more than 700,000 articles such as the German, French, Polish, Italian, Japanese and Spanish ones, up to smaller ones such as Wolof, Catalan, Latin, Tibetan, Haitian and more.

So the motivating question for this contribution is “do people who self-elect for editing the page about Osama Bin Laden in the English Wikipedia have and represent the same points of view of people who self-elect to edit the counterpart article on Arabic Wikipedia?” We call this Linguistic Point of View (LPOV).

One of the three core content policies of Wikipedia is “Neutral Point of View” (NPOV) which states that “Editors must write articles from a neutral point of view, representing all significant views fairly, proportionately, and without bias.” And in fact, inside a specific Wikipedia, the community of millions editors strive to find a balance among the many different biases and points of view: sometimes articles are the battlefield of so-called edit wars. “An edit war occurs when two or more users who disagree about the content of a page repeatedly override (revert) each other's contributions, rather than trying to resolve the disagreement by discussion.” And often on talk pages associated with articles it is possible to observe heated discussions, usually about the fact the article is not neutral. In cases of edit wars and heated discussions, admins can block pages in editing until consensus is reached. In fact most of the time editors of a Wikipedia are able to discuss and reach a consensus about representing the different points of view in a neutral and balanced way.

However the platform does not “force” editors to discuss and negotiate their points of view across different language Wikipedias. There isn’t too much relationship and

---

3 This data are taken from http://en.wikipedia.org/wiki/Special:Statistics . All data reported in this paper are taken from the Wikipedia site as they appeared on March 29, 2011.
4 Since this article deals with Wikipedia pages, it cites many of them. In order not to clutter the paper with too many footnotes or citations, we simply report the title of the Wikipedia page between brackets such as “Page title”. Wikipedia pages were accessed on March 29, 2011. to get a version of “Page title” at that date using the history feature of Wikipedia, visiting the URL http://en.wikipedia.org/w/index.php?action=history&offset=20110329000000&title=Page_title
possibility/forcing of engaging in discussions between, for example, people who strive to reach consensus about the page "Jerusalem" on Arabic Wikipedia and people who strive to reach consensus about the page "Jerusalem" on Hebrew Wikipedia.

The possibility is that the each different language Wikipedia community is able to develop and reach its own intra-Wikipedia consensual (neutral) point of view. However it might be the case that the points of view of the different language Wikipedias are indeed very diverse. We call this concept Linguistic Point Of View (LPOV) or Language Point Of View.

The goal of Manypedia is to provide a web mashup by which it is possible to easily and quickly check for different Linguistic Points Of View between different language Wikipedias.

2. TECHNICAL DETAILS OF MANYPEDIA

Manypedia is a Web mashup that offers new functionalities for comparing the Linguistic Points of View across different language Wikipedia by combining, aggregating and transforming resources and services available on the Web.

Figure 1 shows the interface of Manypedia. A description of the functionalities offered can be accessed by clicking on the “About” link on the Web mashup page.

On top of page, there is a drop-down element which becomes fully visible when you move the mouse over it. On this HTML element, you can search for a specific page in one of the 56 available Wikipedias and get a translation of this page into one of the 56 languages in which this page is available. It is also possible to see the popular, latest and featured comparisons. On the bottom of the interface there are buttons for sharing comparisons using social bookmarking services APIs of Facebook, Twitter, Digg, StumbleUpon, Delicious and GoogleBuzz.

In Figure 1, there is a screenshot of Manypedia comparing the page "Osama Bin Laden" from English Wikipedia (left) with the page "أسامه بن لادن" from the Arabic Wikipedia (translated into English).

So, when the user starts typing “Osama”, Manypedia uses the Wikipedia API\(^5\) in order to get suggestions about titles of Wikipedia pages that starts with this string.

Once the user has chosen a page, this page is loaded dynamically using AJAX (Asynchronous JavaScript and XML) calls tunneled through a PHP proxy and placed inside a DIV HTML element (left part of the interface). The content of every requested page is retrieved from Wikipedia's index.php that provides a pure HTML version of every page\(^6\). At this point the Wikipedia API is invoked in order to know in which other language Wikipedias there is an equivalent version of this page. Connections between Wikipedias are created by Wikipedians by inserting so called interwiki links: for example the Wiki Markup code [[ar: أسامه بن لادن]] at “Osama Bin Laden” page in English Wikipedia specifies that there is an equivalent page titled "أسامه بن لادن" in the Arabic Wikipedia.

\(^5\) Wikipedia API is at http://en.wikipedia.org/w/api.php
The languages in which the left-side page is available are inserted in the drop-down menu so that the user can choose the language Wikipedia on which to compare the left-side page. The page is retrieved with the ajax call and the content is channeled through the Google translate API.

In fact Manyepedia, as a mashup, uses natural language processing, specifically machine translation, for translating one Web page into a language into another language. The mashup uses the Google translate API7 by means of the jquery-translate plugin8. This means that most of the work is done on the client, i.e. the web browser. The translation is automatic and hence it is not perfect, but we believe most of the times and for most language pairs, it is enough to get a precise idea of what the translated page is speaking about. Being able to “understand” the result of hundreds of edits by Wikipedians who edited a certain page in Arabic or Chinese (without knowing Arabic or Chinese!) is something that was impossible up to 20 years ago and we still marvel at how powerful this is for cross-cultural studies. The languages currently available for translations are 56, the ones provided by Google translate, ranging from English, Spanish, German to Yiddish, Tagalog, Catalan and Swahili.

Manyepedia is a Web mashup, combining information collected from different Web sources. The main source is Wikipedia. Each Wikipedia article is released under Creative Commons Attribution-Share Alike License, so this means that anyone, including us with Manyepedia, can copy, distribute, transmit and also remix the content as long as he or she attributes it to the authors and copyright holders: we do so by giving credit to the specific Wikipedia articles incorporated in each comparison on Manyepedia specifying that the source is Wikipedia and linking to the specific article. As a consequence, the content of Manyepedia is released under a Creative Commons Attribution-Share Alike License.

Manyepedia also exploits information extraction and automatic text summarization. In fact, as Figure 1 shows, on top of each compared Wikipedia page, we show:

(1) images embedded in the Wikipedia page. If you move the mouse over them and you'll see a larger version of the image. This is useful so you can get a one-second visual grasp of the differences between the 2 compared Wikipedia pages.

(2) word cloud of most frequent words (the larger the font, the most frequent the word) by DynaCloud jQurey plugin9. Stop words and words in page title are excluded. This functions allows to quickly spot the main textual differences of the 2 pages.

(3) number of total edits received. This is useful to get an idea of the relative attention each page is receiving from each language community (take into account that the English Wikipedia community is much larger than, for example, the Japanese Wikipedia community that is much larger than the Swahili Wikipedia community). Moreover, if a page has received only few edits by only one or two editors, it is possible that it does not reflect an at-least-partially shared vision but only the point of view of the one or two editors. On the other hand, if it has received a large number of edits by a large number of editors, you can assume a little bit more that the current page is the up-to-date neutral result of the negotiation of the all significant viewpoints about the issue shared by the specific language Wikipedia community.

(4) creation date and creator. This is useful in order to ponder if the page has existed since enough time to get enough attention.

(5) date of last edit. If the last edit is very recent, it might indicate recent attention received by the page. Moreover, if you find any sign of vandalism or very biased point of view, it might be because the community didn't have enough time to react and fix the vandalism yet.

(6) number of editors. Similar consideration of number of edits hold: many Wikipedians involved mean the page is likely to represent most relevant points of view. On the other hand, if the number of editors is small, it might be that the page is biased towards the points of view of few Wikipedians.

(7) the 5 Wikipedians who edited this page the most, with a link opening statistics about them10. This is useful in order to understand if there is one single user "owning" the page (Wikipedia clearly states that "you do not own articles"). Moreover you can get an idea of the relative influence exercised by top editors of this page by comparing their edits and the total number of edits: again, a small percentage might indicate a more shared and neutral point of view. Even more interestingly, there might be cases in which the same Wikipedian is one of the most active editors in both the 2 compared language Wikipedia articles. All this information (images, most frequent words cloud, ...) are computed live and on the client side, i.e. via javascript (jquery) on the visitor's browser. The last pieces of information (3 to 7) comes from Ajax calls retrieving information from Articleinfo X! Tool11.

In order to provide a fully functional navigation inside Manyepedia every link of any Wikipedia page is modified to open it in the current browsing window, so clicking on any Wikipedia page link in Manyepedia will launch a new request inside the mashup.

The URL structure is clean as typical of Web mashups, for example Figure 1 is the result of a HTTP call to http://sonetlab.fbk.eu/manyepedia/#/en/Osama_Bin_Laden/jar

When the user visits Manyepedia without specifying 2 languages and a page title, a random comparison is extracted from the database of featured comparisons inserted by the authors and it is visualized.

Manyepedia is a Web mashup and can be accessed at http://sonetlab.fbk.eu/manyepedia/

7 Google translate API is at http://code.google.com/apis/language/translate/overview.html
9 http://johannburkard.de/blog/programming/javascript/dynacloud-a-dynamic-javascript-tag-keyword-cloud-with-jquery.html
10 “X! edit counter” at http://toolserver.org/~soxred93/ec
11 “X! article info” at http://toolserver.org/~soxred93/articleinfo/