In Search of the Ur-Wikipedia: Universality, Similarity, and Translation in the Wikipedia Inter-language Link Network

Morten Warncke-Wang[1], Anuradha Uduwage[1], Zhenhua Dong[2], John Riedl[1]

[1]GroupLens Research
Dept. of Computer Science and Engineering
University of Minnesota
Minneapolis, Minnesota
{morten,uduwage,riedl}@cs.umn.edu

[2]Dept. of Information Technical Science
Nankai University
Tianjin, China
dongzhh@mail.nankai.edu.cn

ABSTRACT

Wikipedia has become one of the primary encyclopaedic information repositories on the World Wide Web. It started in 2001 with a single edition in the English language and has since expanded to more than 20 million articles in 283 languages. Criss-crossing between the Wikipedias is an inter-language link network, connecting the articles of one edition of Wikipedia to another. We describe characteristics of articles covered by nearly all Wikipedias and those covered by only a single language edition, we use the network to understand how we can judge the similarity between Wikipedias based on concept coverage, and we investigate the flow of translation between a selection of the larger Wikipedias. Our findings indicate that the relationships between Wikipedia editions follow Tobler's first law of geography: similarity decreases with increasing distance. The number of articles in a Wikipedia edition is found to be the strongest predictor of similarity, while language similarity also appears to have an influence. The English Wikipedia edition is by far the primary source of translations. We discuss the impact of these results for Wikipedia as well as user-generated content communities in general.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Systems and software—*Information networks*; H.5.3 [**Information Systems**]: Group and Organization Interfaces—*computer-supported collaborative work*

General Terms

Human Factors, Measurement

Keywords

Wikipedia, Tobler's Law, First Law of Geography, Multilingual

WikiSym '12, Aug 27-29, Linz, Austria

1. INTRODUCTION

The world: seven seas separating seven continents, seven billion people in 193 nations. The world's knowledge: 283 Wikipedias totalling more than 20 million articles. Some of the content that is contained within these Wikipedias is probably shared between them; for instance it is likely that they will all have an article about Wikipedia itself. This leads us to ask whether there exists some ur-Wikipedia, a set of universal knowledge that any human encyclopaedia will contain, regardless of language, culture, etc? With such a large number of Wikipedia editions, what can we learn about the knowledge in the ur-Wikipedia?

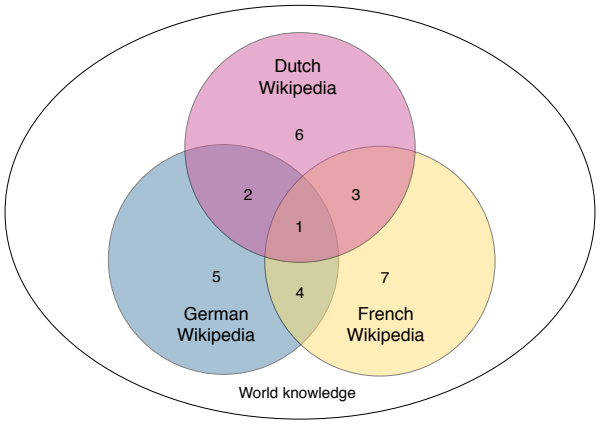

Figure 1: Illustration of shared and non-shared concepts in the world of knowledge between three of the largest Wikipedia editions.

Shared concepts can be illustrated as shown in Figure 1, where we have simplified the case by using only three Wikipedia editions of approximately the same size. In this case there are seven numbered classes of concepts: the ones that are shared by all (1), those shared by two (2, 3, 4), and those unique to a specific edition (5, 6, 7). There is also a part of world knowledge that is not covered by any Wikipedia, as Wikipedia content usually needs to meet a notability threshold and have verifiable sources. We are interested in understanding what forces influence the number of concepts in each of the numbered classes, expanded to encompass all 283 Wikipedia editions. For instance it has been shown that

location bias exists in some Wikipedias [6], and Dahinden [4] found a connection between where a language is used and the location of articles in the same Wikipedia language edition. If it is possible to locate certain Wikipedia editions to specific geographic regions we might therefore be able to decide what influence geographic distance between Wikipedias has on the similarity between them.

Articles about the same concept have been linked across the different language Wikipedias from the early days. The inter-language link (ILL) network contains as of early January 2012 more than 224 million links between the more than 20 million articles. It is a decentralised system where each article is responsible for linking to all other languages where the same article exists[1]. Originally this network was maintained manually, now Wikipedia users have a collection of software robots called *interwiki bots* to help them. The bots follow the ILLs to decide which languages have an article in order to add, update, or remove links as necessary. Because they use only the ILLs they are unable to discover isolated articles that *should* be linked, and the number of missing links varies between editions [7]. Research on the existing network has identified errors from incorrect links [5], from differences when deciding how to cover certain topics [7], and because the boundaries of concepts vary [2].

It is not difficult to find anecdotal evidence to show that there are differences in *content* coverage of the same concept. For example the English edition's article about the Norwegian politician Erik Solheim[2] contains several paragraphs about his work in the Sri Lankan peace negotiation process, including a section on controversies about alleged bias towards one side of the conflict, while the Norwegian edition's article[3] contains only a single three-sentence paragraph and completely omits the Sri Lankan point of view. There are two official languages in Sri Lanka: Sinhala and Tamil. The Sinhala Wikipedia edition contains as of March 2012 about 6,000 articles, but does not have an article about Erik Solheim. Tamil Wikipedia has more than 43,000 articles, and the one about Erik Solheim[4] lists only the basic facts about him and that he was involved in the peace talks from 2000 to 2006.

There are also similarities and differences among Wikipedia editions as a whole. In the book "The Wikipedia Revolution" [16], Lih mentions the story of how contributors to the Spanish edition branched out and created Enciclopedia Libre in 2002 after disputes about financing Wikipedia with advertisements, leaving Spanish Wikipedia nearly inactive for a year and a half. The book also mentions that Japanese Wikipedia has a much larger proportion of users who do not create an account (so called "anonymous users") compared to other Wikipedias of similar size.

Research has also emerged describing similarities and differences between Wikipedias. Pfeil et al. [17] looked at the coverage of the article "Game" in four Wikipedias and found correlation with some of Hofstede's dimensions of cultural influence. Callahan and Herring [3] analysed coverage of famous persons in the English and Polish Wikipedias and discovered systematic differences but no indication of intentional bias. Yasseri et al. [32] used regularity in the edit patterns of the larger Wikipedia editions to estimate the geographical distribution of their editors. Adar et al. [1] studied tables with summary information about a topic, called *infoboxes*, mining data from the English, German, French and Spanish Wikipedias to show that the information could to a large extent be discovered and consolidated automatically through machine learning. The featured article approval process of Arabic, English, and Korean Wikipedias was explored by Stvilia et al. [23], finding that both their quality models and understanding of quality differed. Hecht and Gergle [7] studied 25 of the largest Wikipedia editions and found strong evidence against there being a global consensus of world knowledge. In this paper we extend the existing literature by exploring metrics to quantify the similarity between Wikipedia editions, and by using these metrics to identify what factors affect similarity.

Much of the existing research has used a single language edition as a data source, English Wikipedia in particular. Some of the issues that have been explored are: how users work together [29, 30, 13, 25], how article quality develops [31, 24, 15], and how different users contribute vastly different amounts of content to the encyclopaedia [18, 11]. There are also examples of research using other Wikipedias as a source, such as Stein and Hess [22] who looked into how certain contributors greatly affected article quality in the German edition.

Researchers have also looked at other networks. Roth et al. [20] mined data from about 360 different wikis to understand how macroscopic indicators, structural features, and governance policies affect growth patterns, while Kittur and Kraut [12] sampled 6,118 wiki production groups to understand how a selection of coordination mechanisms affect task quality and conflict in systems that are similar to Wikipedia. On the popular social networking site Twitter, Hong et al. [10] discovered that across languages users differ in their inclusion of URLs in tweets, as well as Twitter-specific features like hashtags, mentions, and retweets. A paper on the blogging site LiveJournal has also shown that there existed at least four distinct networks of users writing blog entries in languages other than English: Russian, Portuguese, Finnish, and Japanese [9].

2. RESEARCH QUESTIONS

The existing literature raises a number of fascinating questions about information flow across languages and cultures, with the ILLs in Wikipedia providing a rich source of data about those questions. This paper explores three perspectives of the ILL network by answering the following three research questions:

RQ1-Universality: **What are the universal concepts that nearly every Wikipedia writes about?**

The concepts that appear in nearly all Wikipedia editions should provide us with a strong indicator of what the global network of Wikipedias regard to be *universal knowledge*, which we will call the ur-Wikipedia. Understanding what these concepts are can affect what user-generated content (UGC) communities should focus on, e.g., what topics to

[1]A system where all ILLs are stored in a central repository has been discussed since at least late 2002: http://lists.wikimedia.org/pipermail/wikitech-l/2002-December/001686.html

[2]http://en.wikipedia.org/w/index.php?title=Erik_Solheim&oldid=463794085

[3]http://no.wikipedia.org/w/index.php?title=Erik_Solheim&oldid=9669080

[4]Due to Tamil characters in the URI, we chose to omit referencing the article, a link to ta.wikipedia.org is found in the articles of both the Norwegian and English editions

cover first, or whether there is some type of bias they need to control for.

RQ2-Similarity: **How can we use the ILL network to measure the similarities and differences between Wikipedias?**

We want to understand the underlying factors that affect similarity between Wikipedias by applying existing methods of calculating similarity. Previous research [8] has found that the internal link network between articles in a Wikipedia follows Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things." [27] Can we locate Wikipedias to specific geographic areas, and will we find that those close to each other will have a higher degree of similarity? We also want to investigate other properties of culture. For instance does language similarity also have a measurable influence on similarity? The driving factors of similarity across language editions of Wikipedia might transfer to other communities where users generate content, e.g., the Twitter and LiveJournal language networks mentioned earlier.

RQ3-Translation: **How much of the information in a Wikipedia comes from translations from other languages?**

This research question looks at the network of Wikipedias from a different perspective; whereas the ILL network itself is about concept coverage, translation is about *information flow* and how that affects relationships between Wikipedias. The English Wikipedia has nearly three times the number of articles of the second largest Wikipedia, is this dominance also found in translations? In addition to showing us how information flows in Wikipedia the results can give suggestions on how user effort should be prioritised in interlingual UGC communities.

The rest of this paper will cover each of the research questions in turn before summarising results and discussing implications for Wikipedia as well as user-generated content communities in general. We begin by examining articles with a universal appeal.

3. UNIVERSALITY AND UNIQUENESS

RQ1-Universality asks "What are the universal concepts that nearly every Wikipedia writes about?" To find these universal concepts we gathered all articles with ILLs from each of the 283 Wikipedia editions in mid-March 2012 and ranked them by number of Wikipedias that have that article. The collections of articles were combined by matching titles, using the English language title if an article linked to it. We then limited the selection to the articles covered by a majority of languages (>142). This approach has limitations in that we will not discover articles that are missing links into the ILL network and we might not capture the exact number of languages that have a specific article. Solving the first of those issues would require extensive content analysis, which is beyond the scope of this paper. The probability that the English Wikipedia does not have the authoritative list of ILLs should be inverse correlated with the number of ILLs, and because we investigate articles with a high number of ILLs the influence of the second issue should be minimal.

We also calculated the total amount of content in bytes across all Wikipedias for articles that were not about time-related topics. This content measure is not a measurement of the amount of information found about a specific topic due to differences in word length and information density between languages. Because the articles we measure are mostly found in a very large number of languages (>175) these kind of differences should distribute evenly, making this approach a suitable estimate for the amount of global content on a topic.

Category	Number of articles	%
Airport	2	0.24
Animal	6	0.73
Chemical	6	0.73
Celestial object	13	1.58
City	42	5.10
Continent	8	0.97
Country	183	22.24
Currency	1	0.12
Festival	1	0.12
Food & drink	3	0.36
General subject	30	3.65
History	2	0.24
Language	16	1.94
List	1	0.12
Ocean	6	0.73
Organisation	2	0.24
People	18	2.19
Religion	6	0.73
US States	8	0.97
Technology	3	0.36
Time	465	56.50
Wikipedia	1	0.12

Table 1: Categorisation of the 823 inter-language linked articles found in more than half of all 283 Wikipedia editions.

From a candidate list of 877 articles we removed 17 non-articles, e.g., "Category:Geography", 17 articles with non-English titles that were all duplicates of English ones, 19 errors from incorrect links, and one misspelling, resulting in a list of 823 articles. Those 823 articles were then categorised into general categories as shown in Table 1.

The majority (56.5%) are time-related subjects, particularly years and specific days of the Gregorian calendar year (e.g. January 4). Two other subjects are also large: countries (22.2%), and cities (5.1%). Wikipedia articles about years and specific dates are used to list events that happened at a specific time, therefore it is unsurprising to see "Time" being a large category. More surprising is the fact that the two airports are both in Vietnam, that there are eight US states on the list, and that the only currency is the Euro. This suggests that the ur-Wikipedia consists mainly of general topics and listings of events, with less than 1.5% being an eclectic mix of articles.

The top 20 articles by number of languages and by total content size are shown in Table 2. The two columns on the left show articles ranked by number of languages, and we see prominent countries like the United States, Russia, Germany and France rank high, as well as the article on Wikipedia itself. Again we also find rather odd topics like the cities of Uetersen, Germany with about 18,000 citizens (#19) and Kurów, Poland (#13, pop. ≈2,800). The article found in the most languages is "True Jesus Church" with 254. Since there are 283 Wikipedias this means that none of the articles are found in all Wikipedias, a result that questions the existence of an ur-Wikipedia. Future research could look

Title	Lang	Title	Size
True Jesus Church	254	United States	10.08
United States	251	World War II	8.64
Russia	250	Russia	7.03
Wikipedia	249	United Kingdom	6.74
Europe	244	Germany	6.64
Germany	242	World War I	6.39
France	235	Adolf Hitler	6.15
Africa	235	France	5.87
Asia	235	Japan	5.81
Italy	233	India	5.77
Spain	233	China	5.64
English language	232	List of sov. states	5.63
Kurów	230	Israel	5.57
Poland	228	Spain	5.51
India	227	Canada	5.40
Lithuania	225	Earth	5.31
Canada	225	European Union	5.27
Vietnam	224	Islam	5.18
Uetersen	224	Africa	5.15
United Kingdom	220	Switzerland	5.08

Table 2: Top 20 articles by number of languages (left), and by article size across languages (right, in millions of bytes), based on data from all 283 Wikipedia editions.

into this in more detail, for instance examine which articles a Wikipedia edition creates first.

Investigating the history of the True Jesus Church articles reveals that 97 of them were created by users likely from New Zealand. Another 79 were started by one single user who then requested help with translation from other Wikipedia users, often using the English language article as the source. This indicates that an organised effort by determined users can affect content in Wikipedia on a global scale.

Ranking the articles by total amount of content, shown in the two columns on the right in Table 2, results in a list where more than one half is countries and one entry is the "List of sovereign states". World history is also present in articles about both World Wars and Adolf Hitler. The remaining four entries are Earth, European Union, Islam, and Africa. Unlike when ranked by number of languages there are no odd topics on the list. Instead we see that they are all important encyclopaedic topics with a large amount of content.

In addition to investigating the articles with a universal appeal we are also interested in the articles that are found in only a single Wikipedia edition in order to understand the properties of knowledge that appears to have a more limited audience. The amount of these unique articles varies greatly between Wikipedia editions. As of October 2011 the Waray-Waray Wikipedia had only 81 unique articles, or 0.078%. On the opposite end was Hindi Wikipedia with 70.29% unique articles. The edition with the largest number of such articles was English; its nearly 1.8 million unique articles was about half a million more than the total number of articles in German Wikipedia.

Many Wikipedias have a category system to categorise articles under a common theme. These categories also link between each other, making it possible to follow a path from more specific categories (e.g., "Turing Award laureates") to more general categories (e.g., "Computer science"). The larger Wikipedias tend to have a well-developed category system which should make it possible to do such a walk between categories and use the general categories to describe unique articles.

We limited ourselves to the English, Malay, Swedish, Norwegian, and Danish Wikipedias due to their size and our authors' understanding of language and culture for those. For each of these five we identified the set of top-most general categories, e.g., "Category:Main topic classifications" in the English Wikipedia. A software tool was developed that gathers unique articles, and for each article walks the category graph until it reaches either a general category, a loop, or a dead-end. We gathered the number of articles and amount of content (in bytes) for each top-level category and its immediate sub-categories.

After walking the category graph for these five Wikipedias we analysed the distribution of number of articles per general category and inspected a random sample of articles. We found that unique articles are generally about people, places, organisations, historic events, and cultural artifacts like music, artists, and TV/radio. The articles appear to be clearly notable in the region of the more localised Wikipedias, i.e. the four non-English ones. This suggests that the articles do not have ILLs due to a limited scope of interest.

The category walk also reveals some interesting aspects of these Wikipedias. Malay Wikipedia appears to have a fair amount of unique articles related to Indonesia, which one would expect to be non-unique since Indonesian Wikipedia is fairly large (more than 185,000 articles as of early 2012). One example is the category "Geografi mengikut tempat" (Eng: Geography by place) which has a large number of articles about villages in Indonesia.

Some articles were not linked to other Wikipedias due to differences between how different language versions cover topics, or simply because links are missing. Norwegian Wikipedia has an article on the New York County District Attorney's Office that has no ILLs, perhaps because of naming differences; English Wikipedia has articles on the person holding the office, and on what a district attorney does. Malay Wikipedia's article "Nick Drake (penyair)" appears to be identical to English Wikipedia's article "Nick Drake (Poet)", but no ILL exists. The interwiki bots that help maintain the inter-language link network do not look for content similarities. Instead they require at least one link into the network to function, and thus fail to discover unique articles that might be similar.

4. MEASURING SIMILARITY BETWEEN WIKIPEDIA EDITIONS

As we have seen, some concepts are covered by many Wikipedias and some by only few. To investigate further the forces that lead to these article-level differences we next look at the macro-level similarities and differences between Wikipedias. ***RQ2-Similarity*** investigates how we can measure the degree of similarity and what factors influence it. Geographic distance and language similarity are particularly interesting; previous research has found the former drives relatedness between articles within a Wikipedia edition [8].

4.1 Similarity based on shared concepts

Calculating similarity based on Wikipedia editions is complicated by the fact that they do not share concepts. Instead

they share articles and the ILLs connect shared articles to each other. Many concepts are likely covered by a single topic and Wikipedias only have one article per topic. Our approach is therefore to use shared articles as our best estimate of shared concepts.

The ILL network is in a constant state of change as articles and links get created and deleted. We used a snapshot of the ILL network from July 2011 so all our analyses are internally consistent. The disadvantage of this method is that links might be incorrect due to human error or differences in how certain topics are covered [5, 7, 2]. Previous research has shown that the error rate is low [7, 21], meaning a quantitative approach to measuring similarity between the larger Wikipedia editions should give useful insight.

The next challenge is how to measure this similarity. One straightforward approach is the Jaccard coefficient, which has for instance been used in ecology to compare diversity in species. Adapted to Wikipedia it measures the ratio of the number of shared articles between two Wikipedias A and B to the union of all articles in the same two Wikipedias, as defined in formula 1. A potential issue with applying the Jaccard coefficient is how it will handle the large differences in number of articles between Wikipedias, an issue which we will examine in detail in section 4.2.

$$S(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

In order to understand if geographic distance is a driving factor of similarity it is necessary to locate a Wikipedia edition to a specific geographic region or country. The Wikimedia Foundation (WMF) makes statistics available for every Wikipedia[5]. Since our similarity measure is based on analysis of content we chose to use statistics from March 2011[6] on the proportion of all edits (i.e. by both anonymous and registered users) from specific countries for each of the 283 language editions.

The list of Wikipedias was narrowed down from all 283 to the 38 that had more than 100,000 articles as of July 2011. There is some precedence from earlier research for selecting these Wikipedias [8, 32]. Hindi Wikipedia was barely below the threshold and was tens of thousands of articles ahead of the next one, so we included it to make the total 39.

The statistics from the WMF shows that many of these 39 Wikipedias have the vast majority of their contributions coming from a single country; German Wikipedia has 82.2% of its edits made from Germany; French Wikipedia has 78.6% of its edits made from France. Some editions are less localised. Users from the United States are the primary contributors to the English Wikipedia, but they only account for 44% of the total. Another large Wikipedia without a majority country is Spanish where there are more contributors from Mexico, Argentina, Chile, Colombia, and Peru (45.2% combined) than from Spain (38.1%). The Chinese and Arabic Wikipedias could also not be placed in a single country; most edits to Chinese Wikipedia come from outside China; contributions to the Arabic Wikipedia are mainly from countries across the northern coast of Africa as well as the Middle East.

We chose to place a Wikipedia edition in a specific country if a clear majority of the contributions came from said country. Instead of choosing a simple majority (>50%), we made our cut-off ⅔ because it clearly distinguished two groups of Wikipedias while only removing 11 of the 39 editions larger than 100,000 articles. There were two editions close to the borderline, Romanian (62%) and Serbian (61.3%). In those two cases we looked at edits from their neighbouring countries and found that they amounted to 5-11% of the total. If we chose to add this traffic it would make both countries eligible without appreciably changing the location. After adding these two countries the total number of geolocated Wikipedias ends up being 30.

To measure distance between geolocated Wikipedias we calculated the great circle distance (GCD) in kilometres between the geographic centres of their corresponding countries. Another approach is to calculate the distance between capital cities. We found that the latter only resulted in minuscule changes to our models and our results are therefore based on the GCD between geographic centres.

Our goal is to discover the driving factors behind similarity between Wikipedias by building linear regression models. The models based on the Jaccard coefficient are found in Table 3. Model I only uses distance as a predictor and we find it to be statistically significant ($P \ll 0.001$), explaining approximately 4.2% of the variation in our data. We also see that the coefficient has a negative sign, indicating that the relationship between Wikipedia editions follows Tobler's first law of geography; Wikipedias that can be located to specific countries share fewer concepts as the distance between them grows.

Model II tests whether size (in number of articles) is a significant predictor, controlling for interaction between the two size variables. Because size of Wikipedia editions has a non-linear distribution the variables are log-transformed (in base 10) in order to make them normally distributed. From the results we see that size by itself is also statistically significant and a very good predictor of similarity since it explains just over half the variation in our data. This result suggests that the availability of topics to write about might be a limiting factor on number of articles in a Wikipedia as described by Suh et al. [26]

Model III is a linear combination of the previous two models. In this model all variables are statistically significant and we see this model explains slightly more of the variance in the data than the individual models (55.5% combined). Our selection of Wikipedias is limited to ones with more than 100,000 articles and English Wikipedia is not one of them because it could not be geolocated, thus one would not expect size to be such a dominant factor.

	Model I	**Model II**	**Model III**
Intercept	0.1058	6.6075	6.571
Distance	-1.954e-6		-2.053e-6
Size of Wiki A		-1.2159	-1.209
Size of Wiki B		-1.2169	-1.210
Size A × Size B		0.2272	0.2262
Adj. r^2	0.0417	0.5087	0.5553

All P-values for all predictors in all models: $\ll 0.001$

Table 3: Linear regression models for Jaccard coefficient similarity. All size variables are $\log_{10}$-transformed.

[5]See http://stats.wikimedia.org/

[6]http://stats.wikimedia.org/archive/squid_reports/2011-03/SquidReportPageEditsPerLanguageBreakdown.htm

In addition to distance and size we also examined some other potential predictors of similarity. Language similarity will be covered in section 4.3; here we will focus on shared contributors. Similar to how Wikipedias can share concepts they might also share contributors. The Wikimedia Foundation conducted a survey of Wikipedia contributors in April 2011[7], with the survey translated into 21 different languages. 51% of the respondents said they contributed to two or more languages. Thus it is likely that many Wikipedias, particularly the larger ones, share many contributors.

There are two ways of identifying shared accounts on Wikipedia: either use their centralised account system or match usernames. The system of centralised accounts has been available since 2008[8], which should reduce the potential for accounts with the same username belonging to different individuals. This led us to use matching usernames as a proxy. We gathered usernames from several Wikipedias to identify the number of matches with software robots removed. We found a strong correlation between the number of shared contributors and the number of shared articles ($r \approx 0.4$). It is therefore possible that shared articles are created by contributors to both languages, or that shared articles lead contributors to move among Wikipedias – or that some hidden cause, such as shared culture, leads to both shared contributors and shared articles. Future research could aim to discover the direction of causality.

4.2 Improving the similarity calculations

One of the issues that arose from our modelling of similarity is that size was by far the dominant predictor. Take the Danish and German Wikipedias as an example: the former has about 160,000 articles while the latter has about 1.3 million. With the Jaccard coefficient this leads to a very low similarity score because the denominator is nearly a factor of ten larger than the numerator. Even if a disproportionally large percentage of Danish Wikipedia articles are also in German Wikipedia, the Jaccard coefficient is not sensitive enough to capture that.

A potential problem when reducing the impact of size on the models is that it can reduce the explanatory power. As we saw previously size by itself explains about half of the variation in our data. As we will see shortly some alternative approaches will result in a reduction in adjusted r^2. We find this to be a reasonable trade-off as it might also allow future research to discover other important predictors of similarity.

One alternative to the Jaccard coefficient is to use cosine similarity, which measures similarity using the angle between two vectors and normalises them to unit length in the process. This approach is used in information retrieval to handle term vectors for documents of differing length, a situation very similar to ours, and one where the Jaccard coefficient is not ideal.

In order to adapt cosine similarity for sets of Wikipedia articles we first observe that given an article set A, we can represent it by a vector $\vec{a}$ where a component $a_i = 1$ if an article is present, and $a_i = 0$ if it is not. The standard formula for cosine similarity between vectors then turns into a formula for cosine similarity between sets of articles in two Wikipedias A and B as defined in formula 2.

$$cos(A, B) = \frac{|A \cap B|}{\sqrt{|A|}\sqrt{|B|}} \quad (2)$$

We calculated the cosine similarity for the same 30 geolocated Wikipedias used previously and built multiple regression models to understand how distance and size affect similarity. Again we found distance to be a significant predictor with a negative sign, explaining slightly more of the data with adjusted $r^2 = 0.077$. Size becomes a less influential predictor with $r^2 = 0.3541$.

Aiming to improve our results we chose to build upon the work of Amos Tversky[28], where the calculation of similarity between two objects is altered in such a way as to allow for one of the objects to have a larger influence on the similarity. Formula 3 is the Tversky Index. When $\alpha = \beta = 1$ the Tversky Index is equivalent to the Jaccard coefficient. If β is kept constant while α is reduced, the result of $A \setminus B$[9] will have less of an influence on the similarity, and β can be similarly manipulated to alter the influence of $B \setminus A$.

$$T(A, B, \alpha, \beta) = \frac{|A \cap B|}{|A \cap B| + \alpha * |A \setminus B| + \beta * |B \setminus A|} \quad (3)$$

The Tversky index has the same issue with size as the Jaccard coefficient. In the case of Danish (A) and German (B) we know $B \setminus A$ will dominate the calculation because there are about 1.15 million articles in German Wikipedia that cannot possibly be in Danish Wikipedia. We want to mitigate this situation so that the shared articles and those articles Danish Wikipedia does not share with German will have an increased influence on the similarity calculation.

We systematically tested functions for calculating α and β based on the number of articles in each Wikipedia edition, for example functions based on the ratio of the two sizes (e.g., $\alpha = \frac{|A|}{|B|}$) and log-transformations ($\alpha = log_{10}(|A|) - log_{10}(|B|)$). We also tested set-based log-transformations and whether keeping either α or β constant while altering the other provides better results than altering both. Our chosen measurement was the adjusted r^2 of the size variable in regression models to confirm if we had successfully reduced its explanatory power.

The log-based modifier for α and β described in formula 4 produced the best results. It is based on the idea that if the sizes of the two Wikipedias are roughly the same then α and β should be close to 1. This results in the measure behaving much like the Jaccard coefficient under those conditions. As the size difference grows the influence of the larger edition should be decreased. The multiplication factor of 0.5 was determined through iterative testing of values in the range $[0.1, 2]$. This modified metric which we will call the Size-Normalised Tversky Index, is strongly correlated with the Jaccard coefficient ($r = 0.8$).

$$\begin{aligned} &\textit{if } |A| = |B| \textit{ then } \alpha = 1, \beta = 1 \\ &\textit{if } |A| > |B| \textit{ then } \alpha = \frac{1}{1 + 0.5 * log_2(\frac{|A|}{|B|})}, \beta = 1 \\ &\textit{if } |A| < |B| \textit{ then } \alpha = 1, \beta = \frac{1}{1 + 0.5 * log_2(\frac{|B|}{|A|})} \end{aligned} \quad (4)$$

[7] http://blog.wikimedia.org/2011/04/18/launching-our-semi-annual-wikipedia-editors-survey/

[8] http://meta.wikimedia.org/wiki/Help:Unified_login

[9] $A \setminus B$ is the set of objects in A that are not also in B.

	Model I	Model II	Model III
Intercept	0.1407	2.0436	2.025
Distance	-2.632e-6		-2.919e-6
Size of Wiki A		-0.3950	-0.3906
Size of Wiki B		-0.3950	-0.3906
Size A × B		0.0805	0.0799
Adj. r^2	0.0744	0.3369	0.4291

All P-values for all predictors in all models: $\ll 0.001$

Table 4: Linear regression models for the Size-Normalised Tversky Index metric. Size variables are $\log_{10}$-transformed.

Modelling results using this function are shown in Table 4 and as we see distance explains slightly more of the variation in the data (Model I: $r^2 = 0.074$). Our goal was to reduce the influence of size and Model II shows a significant decrease ($r^2 = 0.3369$) compared to the same model in Table 3 ($r^2 = 0.5087$). We also find that Model III, the linear combination of the first two models, shows a larger increase in explanatory power compared to the the Jaccard coefficient models. Unfortunately we also see a decrease in the overall explanatory power of Model III. As previously mentioned we find this to be a reasonable trade-off in order to allow for potentially discovering other predictors.

We inspected the results of similarity calculations for some of the Wikipedias used in our data sets and found that the calculations behave as expected. If two Wikipedias are of roughly the same size the Size-Normalised Tversky Index (T) is nearly identical to the Jaccard coefficient (S), e.g., for the Danish and Korean Wikipedias $S = 0.1137$ and $T = 0.1187$. When the size difference is large there is a substantial increase in similarity, for instance comparing Danish and German we have $S = 0.0635$ while $T = 0.1372$. This leads us to conclude that the size-normalised Tversky Index is a clear improvement and should be the preferred approach.

4.3 Language as a predictor of similarity

So far we have examined distance, number of articles, and shared editors as potential explanations of similarity. Only the first two of these were found to be significant predictors. Language is also an important aspect of culture, suggesting that language similarity could impact similarity between Wikipedias. French and Italian are both romance-based languages, thus we might expect that their respective Wikipedia editions will have high similarity because contributors can fairly easily understand content in both languages.

How can we quantify the similarity between languages? One approach is to use the language tree as a search tree. In the language tree there are connections based on research into ancestor or prototype languages that existed previously, thus relating many of the Western European languages to each other since they are all members of the Indo-European branch of the tree. The idea was to give each link in the tree a uniform distance and then measure the distance between two languages, but this turned out to be difficult.

We chose to use Ethnologue [14] as a source for our language tree and discovered that because the language tree is based upon languages' historic development it might not reflect current understanding of how similar languages are, e.g., in Ethnologue's tree there are 14 steps between English and French, but only 8 between English and Italian.

Unable to find a suitable solution to this problem we instead chose to look into whether lexical similarity could help us calculate similarity. Ethnologue has a small data set of lexical similarity between pairs of languages available. Lexical similarity is calculated by comparing controlled vocabularies for words that are similar in form and meaning. The Ethnologue data contains 56 of 110 possible language pairs for eleven languages, mostly romance-based languages from Europe. Constrained to the 39 large Wikipedias the resulting data set has 34 pairs for which we calculated the size-normalised Tversky Index. One concern with this data set is a potential auto-correlation between distance and lexical similarity given that the languages are mostly European. Spanish is one of them and as we saw earlier it could not be geolocated because nearly half of its edits are from countries in the Americas. Another language in the data set is Portuguese, which we successfully geolocated to Brazil (82.2% of edits). Lastly, because we have lexical similarity for languages we could not geolocate, for instance English and Spanish, distance is not a variable in these models.

	Model I	Model II	Model III
Intercept	0.1921 ***	-0.4010 *	-0.8925 ***
Langsim	0.0160		0.1479 ***
Size of Wiki A		0.0515 *	0.0848 ***
Size of Wiki B		0.0515 *	0.0848 ***
Adj. r^2	-0.0268	0.2629	0.5

P-values: * < 0.05, *** < 0.001

Table 5: Modelling results for language similarity as a predictor. Wikipedia similarity is calculated using the Size-Normalised Tversky Index as described in section 4.2. Sizes are $\log_{10}$-transformed.

Regardless of the method used to calculate the similarity the results are similar. Due to size constraints we choose to report only the specific results for one of the measures, the Size-Normalised Tversky Index. The results are found in Table 5 and as we can see language similarity is not a significant predictor on its own, while size continues to be. When combined they are all significant and the $r^2 = 0.5$. The results indicate that language similarity correlates positively with similarity between two Wikipedias, though given our limited data set we find that future work is needed to tease out whether there is direct causation.

5. THE FLOW OF TRANSLATIONS

RQ3-Translation investigates translations between different languages. We are interested in these translations because they identify movement of content, while the creation of ILLs connects articles that may have been written independently. Translating articles might be easier than writing them from scratch; if the source article is of good quality it can be worthwhile to re-use parts or all of it. At least three projects have been created for translation efforts in Wikipedia: Google's Translator Toolkit-based project[10], Microsoft Research's WikiBhasha[11], and Duolingo[12].

Wikipedia articles are licensed using the Creative Commons CC-BY-SA license[13]. It defines a translation as an

[10]http://googletranslate.blogspot.com/2010/07/translating-wikipedia.html

[11]http://www.wikibhasha.org/

[12]http://www.duolingo.com

[13]http://en.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License

adaptation and requires that the user "takes reasonable steps to clearly label, demarcate or otherwise identify that changes were made to the original Work." Many Wikipedias use the MediaWiki template system to make it easy to add a note that an article is translated, with a reference to one or more source articles. We identified how the ten largest Wikipedias (as of July 2011) used the template system to attribute translations. Most put the template on an article's talk page, where discussions about the article's content goes, while some instead put the template in the article itself. In addition to those ten editions we also mined data from the Chinese, Norwegian, Swedish, and Danish Wikipedias.

This approach is limited in that it does not capture translations that are not attributed correctly. Users who are unfamiliar with the CC-BY-SA license might not attribute at all. It is also possible that they mention the translations in MediaWiki's edit comment field, a field where a contributor can provide a short description of the changes made in a given edit. Mining the template usage is therefore the strictest approach, which should provide us with a lower bound on translations.

The results from our data gathering of translations are shown in Table 6, and we can see that English Wikipedia is the primary source of translated content. Closer inspection of the numbers show that it often is by far the most used source, e.g., German Wikipedia has 3,834 articles labelled as translations, of which 3,162 articles are from English, while the second language is Italian with only 205 articles. There is a coordinated translation project called Translation of the Week[14] (TotW) which primarily translates English Wikipedia articles. For all but four of our investigated Wikipedias there are thousands of translated articles from English, thus TotW's 400-something total articles does not explain the English dominance.

As described earlier, mining template usage is a very strict method of measuring translations. The results give clear indications that Wikipedia users choose other approaches to attributing translations. For instance the Danish, Norwegian, and Swedish Wikipedias are very similar and all have templates for translated pages. While both Norwegian and Swedish have attribution in more than 1% of their articles, Danish Wikipedia's template is never used. Japanese Wikipedia is also an outlier with only 30 translated articles, of which 26 were from English. Japanese Wikipedia has one of the largest proportions of unique articles (52.05%), thus it is not surprising that they also appear to have fewer translated articles than others, but it is surprising that the proportion is so low.

We also see two Wikipedias that have a very large number of translated articles: French with 40,280, and Italian with 34,689. In both cases the vast majority of these articles are translated from English: 29,517 articles into French, and 29,271 articles into Italian. Given these large numbers we were interested in understanding what these translated articles were about. Earlier work has discovered that French Wikipedia has a bias towards articles that are located in France or the French-speaking area of Canada[6]. Are these translated articles about subjects that are distinguishably different? To answer this question we mined the translated articles for geolocation data and removed those that were not located on Earth, e.g., craters on the Moon and interstellar objects. This method found 5,187 geolocated translations in French Wikipedia, and 4,480 in Italian Wikipedia.

[14]http://meta.wikimedia.org/wiki/Translation_of_the_week

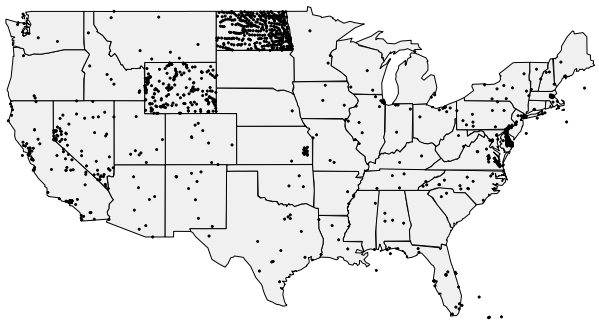

Figure 2: Plot of locations in the continental United States of geolocated articles in Italian Wikipedia that were translated from English Wikipedia

Plotting the locations on a world map revealed interesting trends. For instance we can easily spot some translations from former colony areas, e.g., Eastern Africa in Italian Wikipedia. Another trend was interesting area coverage, exemplified with the excerpt of the Italian Wikipedia translation map shown in Figure 2, where we see the US states of North Dakota and Wyoming having many translations.

We examined the edit history of a small random selection of the translated articles located within the borders of North Dakota and Wyoming. They were all contributed by the same user who was a part of a project working on Italian Wikipedia's coverage of government and counties of the United States[15]. A similar trend was found in the map of translations in French Wikipedia, with good coverage of locations in New Zealand. These translations were contributed by a single user who had joined a project on French Wikipedia for users interested in improving content about New Zealand[16]. Future research could look into the causes of these peculiar patterns.

6. DISCUSSION

In our first research question, ***RQ1-Universality***, we searched for the ur-Wikipedia by asking "What are the universal concepts that nearly every Wikipedia writes about?" We found that they are broad topics with a general appeal, mainly countries, cities, and lists of events. On the opposite end were articles without inter-language links (ILLs) which are instead about narrower topics that appear to have a limited scope of interest.

There is some cause for concern about a bias towards Europe and the United States in the universal articles. If we look at the 100 most linked to articles, the list only contains two African countries, few Asian countries, and no cities in neither Africa nor Asia. In "The Reach and Richness of Wikinomics" [19] Rask shows that Wikipedia activity at a national level is correlated with score on the United Nations Human Development Index (HDI). The Wikimedia Foundation has a focus on increasing activity in countries where it

[15]http://it.wikipedia.org/wiki/Progetto:Amministrazioni/Comuni_degli_Stati_Uniti

[16]http://fr.wikipedia.org/wiki/Projet:Nouvelle-Zélande

Language	No. of articles	No. translated	%	Top languages
English (en)	3,672,210	9,794	0.267	de, fr, es, it, ru, nl, ja
German (de)	1,253,523	3,834	0.306	en, it, sv, fr, nl, hu
French (fr)	1,120,897	40,280	3.594	en, de, it, es
Italian (it)	815,734	34,689	4.253	en, fr, de, es, pt
Polish (pl)	811,717	419	0.052	en, de, cs
Spanish (es)	784,017	7,275	0.928	en, ca, fr, it, de, pt
Japanese (ja)	756,331	30	0.004	en, fr, ru, ko
Russian (ru)	730,190	10,314	1.416	en, uk, de, fr, pl
Dutch (nl)	712,615	3,529	0.495	en, de, fr
Swedish (sv)	401,309	6,161	1.535	en, de, fi, no, da, fr, es
Chinese (zh)	362,519	2,182	0.602	en, ja
Norwegian (no)	306,389	3,500	1.142	en, da, nn, sv, de
Danish (da)	151,680	0	0	

Additional language codes: hu: Hungarian, ca: Catalan, cs: Czech, ko: Korean, fi: Finnish, nn: Norwegian Nynorsk

Table 6: Translated articles in the ten largest Wikipedias as well as Swedish, Chinese, Norwegian, and Danish, ordered by number of articles per July 2011.

is lacking[17]. As that activity grows the local editions should be careful to adjust for this bias.

For user-generated content (UGC) communities in general our results suggest that they should be aware of potentially introducing bias. It can be bias on a language level, e.g., when Italian Wikipedia translated articles about former colonies in Africa; or bias on a user level, e.g., when French Wikipedia translated articles about New Zealand. If the community's goal is unbiased coverage it can for instance help visualise existing bias by identifying areas of high/low coverage for geolocated articles [6]. Another approach could be to design so users are aware of whether they are covering specific or general topics, or to help them easily discover topics that have broad impact so they can create versions in their own language.

Three out of the top twenty inter-language linked articles were surprising and it would be interesting to know more about why. Are there other examples of individuals or small groups having global impact on Wikipedia? How did these articles come to be universal, who did it, and why did they do it? While our research has been quantitative this is likely a strong opportunity for qualitative investigations, including interviewing some of these users.

Our second research question was ***RQ2-Similarity:*** How can we measure the similarities and differences between the Wikipedias using the ILL network? We found that Tobler's first law of geography holds across language editions: similarity decreases as distance increases. When it comes to predictors of similarity number of articles was the dominating factor, but its influence can be greatly reduced as demonstrated by our Size-Normalised Tversky Index. Language similarity was also found to be positively correlated with similarity between Wikipedias.

These results make it perhaps a bit counter-intuitive who your nearest neighbour is. Cultural or geographic proximity might be what people generally use to explain similarity, whereas we find size to have a stronger influence. This suggests that Wikipedias of similar *size* should consider collaborating, for instance by organising competitions or having specific cross-wiki projects.

In future research it would be interesting to investigate the effects of secondary language proficiency, perhaps through case studies. Lastly one can also look into whether the similarity properties we discovered hold with other UGC communities across cultures and geographies.

RQ3-Translation was our last research question, looking into how much of the information on a Wikipedia comes from translations from other languages. We found that English is the *lingua franca* of the Wikipedia world with English Wikipedia being the primary source of translations, often an order of magnitude above the rest.

For Wikipedia this result suggests that one model for broader distribution of content is to push local content to English so it can spread to the rest of the network from there. We see at least three potential issues, the first being the language of sources cited in the article. English Wikipedia does allow non-English sources[18], but it might be difficult to verify that non-English sources support the claims an article makes. The reverse is likely less of a problem due to wide-spread English proficiency. Future research could study both how non-English sources fare in English Wikipedia as well as vice versa. Issue number two is the notability threshold; an article will likely be deleted if the threshold is not met. Other Wikipedias should therefore start content locally and make sure it is of high quality with reliable sources before translating it to English. Issue number three is that a potential centralisation towards the English Wikipedia conflicts with our previous description of the Wikipedias as decentralised. If the goal of Wikipedia is to share the world's knowledge it may be worthwhile to forsake some decentralisation in favour of the potential increase in sharing of said knowledge.

For other UGC communities that are multilingual our results suggest designing for translations, specifically for having English as a hub language and making it easy for users to push high-quality content towards English. A value model that allows us to reason about how value changes depending on language could help decide which language content should be written in. For Wikipedia it could take into consideration factors like readership, existing demand, probability of translations, availability of sources, and whether one should prioritise preservation of local culture over number of read-

[17]http://blog.wikimedia.org/2011/07/22/year-in-review-and-the-road-ahead-for-global-development/

[18]http://en.wikipedia.org/wiki/Wikipedia:NONENG

ers. The model can then be altered to fit properties of other UGC communities.

We also see possibilities for studying how content spreads across the language editions. It could contrast work done in projects like Translation of the Week or WikiProjects like the ones we found in Italian and French Wikipedia with work done by individuals. Interviewing users to understand their motivation and practises could provide useful insight for designing multilingual UGC communities where sharing of work is made easy.

Lastly we think it would be worthwhile to look into what the role of a Wikipedia edition is in the global network of Wikipedias. The wiki way is decentralised and allows for massive parallelisation of small tasks. The many language editions are decentralised in a similar way. At the time of writing the larger Wikipedias are now more than eleven years old. With the Wikimedia Foundation's push on global access it is time to ask: what should Wikipedia do in its teenage years to make sure that it continues to be a freely accessible global repository of human knowledge?

7. ACKNOWLEDGEMENTS

We would like to thank Nivashini Harikrishnan for help with Malay translations, Brent Hecht for valuable input during the writing stage, the other members of GroupLens Research for their support, and the anonymous reviewers for their helpful suggestions. We also thank the Wikimedia Foundation and Wikimedia Deutschland for facilitating access to Wikipedia data, and all Wikipedia contributors for creating this great encyclopaedia. This work is funded by the National Science Foundation (grants IIS 08-08692, 09-68483) and by the China Scholarship Council.

8. REFERENCES

[1] E. Adar, M. Skinner, and D. S. Weld. Information arbitrage across multi-lingual wikipedia. In *Proc. WSDM*, pages 94–103, 2009.

[2] P. Bao, B. Hecht, S. Carton, M. Quaderi, M. Horn, and D. Gergle. Omnipedia: Bridging the wikipedia language gap. In *Proc. CHI*, 2012.

[3] E. S. Callahan and S. C. Herring. Cultural bias in wikipedia content on famous persons. *Jour. ASIST*, 62(10):1899–1915, 2011.

[4] T. Dahinden. Estimation of the locations of the language-versions of Wikipedia – A case study on geographic data mining. In *Advances in Cartography and GIScience*, volume 6, pages 471–487. 2011.

[5] G. de Melo and G. Weikum. Untangling the cross-lingual link structure of wikipedia. In *Proc. ACL*, 2010.

[6] B. Hecht and D. Gergle. Measuring self-focus bias in community-maintained knowledge repositories. In *Proc. C&T*, pages 11–20, 2009.

[7] B. Hecht and D. Gergle. The tower of Babel meets web 2.0: User-generated content and its applications in a multilingual context. In *Proc. CHI*, pages 291–300, 2010.

[8] B. Hecht and E. Moxley. Terabytes of tobler: evaluating the first law in a massive, domain-neutral representation of world knowledge. In *Proc. COSIT*, pages 88–105, 2009.

[9] S. C. Herring, J. C. Paolillo, I. Ramos-Vielba, I. Kouper, E. Wright, S. Stoerger, L. A. Scheidt, and B. Clark. Language networks on LiveJournal. In *Proc. HICSS*, pages 79–89, 2007.

[10] L. Hong, G. Convertino, and E. Chi. Language matters in Twitter: A large scale study. In *ICWSM*, July 2011.

[11] A. Kittur, E. Chi, B. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World Wide Web*, 1(2):19, 2007.

[12] A. Kittur and R. E. Kraut. Beyond wikipedia: coordination and conflict in online production groups. In *Proc. CSCW*, pages 215–224, 2010.

[13] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi. He says, she says: conflict and coordination in wikipedia. In *Proc. CHI*, pages 453–462, 2007.

[14] M. P. Lewis, editor. *Ethnologue: Languages of the World*. SIL International, sixteenth edition, 2009. Online version: http://www.ethnologue.com/.

[15] A. Lih. Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource. In *Proc. ISOJ*, 2004.

[16] A. Lih. *The Wikipedia revolution: How a bunch of nobodies created the world's greatest encyclopedia*. Hyperion Books, 2009.

[17] U. Pfeil, P. Zaphiris, and C. S. Ang. Cultural differences in collaborative authoring of wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113, 2006.

[18] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in wikipedia. In *Proc. GROUP*, pages 259–268, 2007.

[19] M. Rask. The Richness and Reach of Wikinomics: Is the Free Web-Based Encyclopedia Wikipedia Only for the Rich Countries? *Proc. of the Joint Conference of ISMD and the Macromarketing Society*, 2007.

[20] C. Roth, D. Taraborelli, and N. Gilbert. Measuring wiki viability: an empirical assessment of the social dynamics of a large sample of wikis. In *Proc. WikiSym*, 2008.

[21] P. Sorg and P. Cimiano. Enriching the crosslingual link structure of wikipedia - a classification-based approach. In *Proc. of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*, 2008.

[22] K. Stein and C. Hess. Does it matter who contributes: A study on featured articles in the German Wikipedia. In *Proc. HT*, pages 171–174, 2007.

[23] B. Stvilia, A. Al-Faraj, and Y. J. Yi. Issues of cross-contextual information quality evaluation–the case of arabic, english, and korean wikipedias. *Library & Information Science Research*, 31(4):232 – 239, 2009.

[24] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Information quality work organization in wikipedia. *J. Am. Soc. Inf. Sci. Technol.*, 59:983–1001, April 2008.

[25] B. Suh, E. H. Chi, B. A. Pendleton, and A. Kittur. Us vs. them: Understanding social dynamics in wikipedia with revert graph visualizations. In *Proc. VAST*, 2007.

[26] B. Suh, G. Convertino, E. H. Chi, and P. Pirolli. The singularity is not near: Slowing growth of wikipedia. In *Proc. WikiSym*, pages 8:1–8:10, 2009.

[27] W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Econ. Geography*, 46:234–240, 1970.

[28] A. Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

[29] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proc. CHI*, pages 575–582, 2004.

[30] F. B. Viégas, M. Wattenberg, J. Kriss, and F. van Ham. Talk before you type: Coordination in wikipedia. In *Proc. HICSS*, 2007.

[31] F. B. Viégas, M. Wattenberg, and M. M. McKeon. The hidden order of wikipedia. In *Proc, OCSC*, 2007.

[32] T. Yasseri, R. Sumi, and J. Kertész. Circadian patterns of wikipedia editorial activity: A demographic analysis. *PLoS ONE*, 7(1):e30091, Jan 2012.