# Examining Wikipedia across Linguistic and Temporal Borders

Ramine Tinati
University of Southampton
Web and Internet Science
r.tinati@soton.ac.uk

Paul Gaskell
University of Southampton
Web and Internet Science
pvg1g10@soton.ac.uk

Thanassis Tiropanis
University of Southampton
Web and Internet Science
tt2@ecs.soton.ac.uk

Olivier Phillipe
University of Southampton
Web and Internet Science
op1e10@ecs.soton.ac.uk

Wendy Hall
University of Southampton
Web and Internet Science
wh@soton.ac.uk

## ABSTRACT

The Web has grown to be an integral part of modern society offering novel ways for humans to communicate, interact, and share information. New collaborative platforms are forming which are providing individuals with new communities and knowledge bases and, at the same time, offering insights into human activity for researchers, policy-makers and engineers. On a global scale, the role of cultural and language barriers when studying such phenomena becomes particularly relevant and presents significant challenges: due to insufficient information, it is often hard to establish the cultural or language groups in which individuals belong, while there are technical difficulties in establishing the relevance and in analysing resources in different languages. This paper presents a framework to the end of addressing those issues by leveraging data on the use of Wikipedia. Resources available in different languages are explicitly correlated in Wikipedia along with time-stamped logs of access to its articles. This paper provides a framework to enable temporal page views in Wikipedia to be associated with specific geographic profiles. This framework is then used to examine the exchange of information between the English speaking and Chinese speaking localities and reports initial findings on the role of language and culture in diffusion in this context.

## Categories and Subject Descriptors

H.4.m [**Information Systems Applications**]: Miscellaneous

## Keywords

Wikipedia; Dynamic Content Analysis; Social Machines

## 1. INTRODUCTION

Over the last two decades, a co-constructive relationship between humans and technologies has given rise to the development of the World Wide Web [1]. This has been a socio-technical evolution from its original read-only state – Web 1.0 – to a much richer, interactive, and collaborative system, often labelled as Web 2.0 [2], the Social Web [3], and now the Web of data [4]. This evolution, which is a product of both the reflective nature of new technologies and social practices, has been a driving force for the Web's success; becoming a platform for a number of previously offline-only activities: networking, knowledge sharing, news, and entertainment. The Web has become an integral part of society, offering the potential to overcome the linguistic, spatial, and temporal barriers that are faced within the offline world. Embedded in the Web's development and evolution is the influence of different cultures, languages, politics, economic positions; such factors ultimately affects the way the Web is developing and accessed [5].

It is no longer sufficient to examine the Web based on the technical dimension; one has to also ask questions about the underlying social and cultural phenomena. As a result of the growing interest in these types of questions, Web Science aims to observe this phenomenon [6] using interdisciplinary approaches, theories and methodologies [7] [8].

Often, the Web is perceived as a homogenizing, border less technology, offered as a solution to the "global village" of modern society, yet there exists evidence to suggest that this is not the case [9]; the Web is a network of networks [10], and consequently, introduces issues concerning language [11], gender and culture [12], and both physical and perceived geographic boundaries [13] [9]. The cultural and language barriers are hard to capture when performing analysis on the Web since relevant information is not available or consistent. Wikipedia, which a large-scale social-collaborative Web platform represents the world's largest online free encyclopedia, growing rapidly, and currently contains over 30 million collaboratively created and edited pages, which are offered in 287 different languages [14]. Unlike traditional social networks, Wikipedia is not restricted to barriers such as language or global reach, and its open structure does not impose limitations for users to access, edit or create content.

This paper presents a methodological and technical approach to identify linguistic cross-over based on Wikipedia page view logs. Building upon previous research dedicated to understanding the phenomena of Wikipedia [15] [16], we present a methodology to compare user page view cross-over between Wikipedia languages and demonstrate how to associate page views with user localities based on time zones. The methodology developed in this paper situates itself in a growing area of Wikipedia research, analysing the access rather than editing of Wikipedia articles, as well as un-


*WWW'14 Companion,* April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2576931.

derstanding the cross-over of access between translated Wikipedia pages languages.

To evaluate the proposed methodology, We compares Wikipedia page views between two different languages, English and Chinese. We performed a comparison between the entire corpus of Wikipedia pages and found that pages could be categorised by their 'page-view' timeseries profile, and that where there was cross-over between languages, pages tended to relate to specific topics or subjects.

## 2. RELATED WORK

Wikipedia has grown at an astonishing rate, not only in the number of articles written in multiple languages, but also in the number of editors and views [17] [18], with over of 30 million articles in 287 different languages, which are supported by 20 million registered users [14]. The availability of Wikipedia research data [19] has led to a vast body of literature examining a number of key topics including, the analysis of Wikipedia's structure and growth [20] [21] [19], using network science and graph theory to analysing the structure Wikipedia, in terms of article linking structure, and its growth overtime; often showing similar scale-free properties synonymous to the World Wide Web [22].

Increasingly, there is a growing area of interest with regards to the multi-lingual support that Wikipedia offers [23] [24] [25] [16]. Within this area, research offers insights to understand the communities involved with the translation process of articles, identifying barriers to adoption, and the social processes of an articles lifecycle [16]. Research has also examined the effects of culture within the collaborative environment, identifying how Wikipedia is far from culturally neutral, which directly influence the collaborative efforts in article creation [15], and how external factors such as political, regional, or linguistic differences affect the policy and governance of Wikipedia in different countries [26]. Studies have shown the relationship between the position of an editor and their contribution as an editor [27], and also the structural similarities between via semantic linking [28] [29]. In addition to this, there has been the development of technical solutions such as WikiTranslate [30] or CLWE [24] to aid the translation process, supporting individuals to cross-collaborate and discuss their translation work flow [24]. Moreover, in response to the call for improved content and consistency within translated Wikipedia articles, automated techniques have also been developed for improving content of translated Wikipedia articles [31].

Underpinning the aforementioned studies is the desire to find better metrics to measure the factors involved in Wikipedia in terms of its collaborative environment, and the eventual development of techniques to help it develop in the future. As a result of this, there is a significant body of research which examines the collaborative nature of Wikipedia and the technologies that can support it. Nevertheless, there is very little insight in regards to the access of Wikipedia, what is being viewed and how this differs between languages and cultures. Apart from a small number of studies which have examined the impact that Wikipedia has on current and future pedagogy approaches [32] [33] or the most viewed pages within the English corpus of articles [34], the development of an in-depth understanding of the use of Wikipedia and also the flow of information or diffusion between cultural and linguistic barriers is yet to be achieved.

## 3. EXAMINING WIKIPEDIA PAGE VIEWS BETWEEN LANGUAGES

To understand how viewing patterns differ in Wikipedia pages requires a method to determine the common viewing patterns of a page (or collection of pages) of a specific language. In this paper we focus on two languages, English and Chinese. This choice was made primarily due to the suspected cultural differences between English and Chinese speaking populations combined with the large number of global Chinese and English speakers. We employed four structural and statistical variables to explore patterns of page viewing behaviour:

1. The language of the page - Wikipedia provide an hourly log of all the page-views that occur on the site. Within this log is the URL of the page with a count of views for that hour. A Wikipedia page URL is prefixed with a two character country code for the different language versions, which was used as a translation mechanism in this paper. We do note however that 'translated' pages do not necessarily contain like-for-like content [31], however, [27] has shown that between the Chinese and English articles there tends to be a high percentage of similarity.

2. The language that the reader of the page would be expected to speak - building upon Honeycutt and Herring [35] study of using a time delta to identify language of a Web user, the language of a Wikipedia user has been defined by a time-series viewing profile for a page of a given language. We use the time delta between CET and CST as well as the daytime/diurnal page-view profile as a way to distinguish between English (en) and Chinese (cn) Web users. The amount of interest that the page has stimulated outside of its own language can then be thought of as the relative difference of the page-viewing profile from typical viewing behaviour in that language.

3. The similarity of viewing patterns for a page in either language - The relation of a page to a particular time-series viewing profile is relevant to the similarity of the English and Chinese versions of a given page. However, we acknowledge that there are also different usage patterns to consider; as described in [36] [37] pages are sometimes accessed as a constant source of professional or scholarly reference material, or alternatively, for a one off-event viewing [38].

4. The category of content to which the page belongs - OpenCalais has been used to tag and categorise the pages by content. Previous studies have used openCalais to successfully tag semantic meaning, including tweets [39] [40] [41], news articles [42] and Wikipedia pages [43] [44]. However, the service does not provide a service for Chinese language documents, thus tagging semantic information is purely from an English language perspective, although the content of pages can be assumed to be in the same topic area [43].

## 4. COMPARING WIKIPEDIA PAGE VIEWS: EXPERIMENT SETUP

Wikipedia page-views log files were downloaded for the time period running from the 1st June 2012 to the 14th October 2012. These files were then processed and every example of a URL with both an English and Chinese language version and their corresponding time-series of page-view counts stored in a separate file. The resulting data can be described as a series of labels each attached to

two time-series sets, one for English language page-views and on for Chinese language page-views.

$$\begin{aligned} &page_1[english(day_1, day_2 day_n)] \\ &page_1[chinese(day_1, day_2 day_n)] \\ &page_2[english(day_1, day_2 day_n)] \\ &page_2[chinese(day_1, day_2 day_n)] \end{aligned} \tag{1}$$

Each day was assigned to a category of usage behaviour. Reference or browsing usage as described in [36] [37] has no formal definition in the literature. What is required is some way of identifying pages that are consistently viewed at a steady rate. After reviewing the distribution of page views on a number of days a value of double the median for either language was selected as the minimum activity in a day in order to be considered as a browsing event. This value was found by selecting a third of page view events in either language as browsing events.

A metric for trending behaviour was then selected. Studies such as [38] [45] define trending as events with the highest standard deviation change from on time period to the next. This definition is slightly problematic in our case because, although we define discrete days as 24-hour time periods, effectively one typical language profile will always lead or lag the other. The ratio change from one day to the next would then be mismatched across the two series. Instead, trending events were described as those greater than 2 standard deviations from the mean for the page over the whole series. This has the advantage of allowing the time-series to keep the memory of the last days page-views and so smooths over the issue of mismatched time-series. A lower bound for trending behaviour was set at twice the median so that for a day to be considered a trending event it must first be considered a browsing event. The final result is a set of pages and time-series of the form:

$$\begin{aligned} &page_1[english(br_{t1}, ne_{t2}, ne_{t3}, br_{t4}, br_{t5}, br_{t6} br_{tr})] \\ &page_1[chinese(br_{t1}, ne_{t2}, ne_{t3}, brt4, br_{t5}, br_{t6} br_{tr})] \end{aligned} \tag{2}$$

Where *br* is a browsing event, *tr* represents a trending event and *ne* represents days with no events recorded.

These metrics were then defined to represent browsing and event similarity across languages for each page. With each metric taking a value of 1 when all days overlap and a 0 if there was no overlapping occurrences. Note that for trending events this condition was applied even if there was a trend in one language and not the other, in this case the metric equation would give a value of 1 if this condition were not applied in advance.

$$\begin{aligned} browsingsimilarity = \\ (overlapingbrdays)/(englishbrdays + chinesebrdays) \\ trendingsimilarity = \\ (overlapingtrdays)/(englishtrdays + chinesetrdays) \end{aligned} \tag{3}$$

The next stage was to then assign a value to day (24 hour period) in each series representing the likelihood the page viewing was concentrated in an English speaking or Chinese speaking time-zone. There are many distance metrics that have been applied to time-series data in the literature. Clustering techniques often use Euclidean distance or dynamic time warping [46] [47]. In this case a distance metric relying on raw page view numbers is problematic because total English page-view numbers are typically much larger than their Chinese counterparts. Instead, the decision was taken to construct a metric originally suggested in the bio-informatics literature and use the Pearsons correlation function. Although this function is not typically used for data mining applications as it does not define a metric space [48], by ensuring $1 - p(x, y)$ where p is the Pearsons correlation applied to the variables x and y, it is possible to satisfy a generalized form of the triangle inequality where $1 - p(x, y)2((1 - p(x, z)) + (1 - p(y, z)))$. It is then possible to treat the distance $2(1 - p(x, z)) - 2(1 - p(y, z))$ as an unbiased estimator of the relative proximity of z to x and y providing each of these values have been normalized against the scale running 1-p(x, y) to 4 or:

$$\begin{aligned} (2(1 - p(x, z)))/(4 - (1 - p(x, y))) - 2(1 - p(y, z)) \\ /(4 - (1 - p(x, y))) \end{aligned} \tag{4}$$

In terms of the problem presented here x is a 24-hour time-series of total page-views of English Wikipedia and y is the corresponding 24 hour time-series of total page-views of Chinese Wikipedia. For any given page z, the relative influence of a English-speaking time profile to Chinese-speaking time profile will be given by the equation above. Further, this distance will be comparable for any pair of pages on different days, irrespective of changes in the distance between English or Chinese-speaking time profiles. The decision made with regards to the normalization approach is that there are likely to be variations in the typically Chinese and English viewing profiles over time. If typically Chinese viewing becomes more different on a particular day this will make it less likely the different language versions of a page will be highly correlated. Note also that the English language page is always on the left hand side of the equation so a positive value indicates closer proximity to the Chinese time zone and a negative value indicates closer proximity to the English time zone.

Using this procedure for each page, every day with an overlapping browsing or trending event occurrence was rated for the Chinese/English of the viewing behaviour over the 24 hour period. For each Wikipedia page with both an English and Chinese language version the whole page was downloaded, all HTML mark-up and images removed and the full text was submitted to the OpenCalais API for tagging. There are various options for tagging available from the API. The option selected was a mixture of social and topic based tagging. Social tags are derived from documents tagged by the OpenCalais user community, while topic tags are based on a set ontology developed for OpenCalais. The decision was taken to use a combination of social and pre-defined topic tags. This creates a combined ontology of both structured pre-determined classification but also provides an insight into the opinions of individual user communities who may be more familiar with particular topic areas. Each page was tagged with every category that could be associated with it. As a result each page was associated with general categories and also with some more specific to the page content. After the pages were tagged a random sample of 200 were checked for tagging errors. The 200 pages had a total of 745 category tags, of these there were 19 errors giving an error rate of 2.6%. Each category contained a varying number of different pages; therefore, counting just the number of sharing events per category would be biased towards very general categories. Instead, the rate of sharing per category was used:

$$\begin{aligned} Rate\ of\ sharing = \\ ((number\ of\ pages)/(number\ of\ shared\ events))/136 \end{aligned} \tag{5}$$

136 represents number of days of the sample data, thus the rate of sharing is normalised to the number of data. This provides a metric of information flow; a large metric indicates a category that has many instances of Chinese and English speaking individuals accessing the same type of content at the same time. Finally the distance metric for comparative Chinese to English of a viewing event was calculated for every event and every page in each cate-

gory, the metric was the averaged over all of the pages to give a score for the tagged page.

# 5. RESULTS

In total there were 7603 pages viewed over the period analysed with both an English and Chinese version available. The median level of page views for the Chinese language pages was 5 per day and 46 per day for the English language page views. This gave a minimum level of page views to be considered a browsing event of 10 for Chinese pages and 92 for English language pages. There were 6848 pages that met the criteria of having browsing or trending days. This leaves a complete sample set of 931328 days which could potentially have a browsing or trending event:

There were subsequently 392167 browsing event days and 239345 overlapping browsing events giving a browsing similarity of 0.61. There were 26774 total trending days with 2053 of these overlapping giving a trending similarity of 0.077. These results suggest that the major driver of sharing between the two languages is because of Wikipedia's use as a reference material. It is, however, difficult to interpret these statistics without another language to compare with.

## 5.1 Comparison of English and Chinese Page Viewing

Figure 1.a and 1.b show scatter plots of the English-Chinese scale plotted for all categories for browsing and trending days respectively. The first point of note is that no Chinese language pages show trending or browsing behaviour related to the English time zone. This is not entirely surprising but does confirm that none of the categories have stimulated significant interest in Chinese language Wikipedia in the English speaking world. Given that there are a set of pages with positive values for both English and Chinese language pages, we assume that there are a number of categories where viewing of English language pages happens most often in Chinese speaking time zones.

## 5.2 Browsing Events

15 categories were identified by OpenCalais for pages that were predominantly browsed in the Chinese time zone. Examining these we noticed a strong emphasis towards Asian culture and popular music residing outside of China. *SM Town* and *S.M. Entertainment* refer to a Korean production company and record label respectively. *J-pop* refers to Japanese pop music. A final category refers to 'International Relations'. Closer inspection of the actual pages in these categories revealed that the technology categories tended to be related to technical computing documentation rather than consumer products.

In comparison to this, we identified 16 categories where pages have been viewed most often in separate time zones. Similar to before, technology was a predominant category, with pages relating to 'Orthography' (the study of writing systems), and 'Multi-touch' (relating to touch screen technology). However unlike before, there are no categories referring to protocols or networking standards, and instead there was a greater emphasis on browsers, operating systems and Web (rather than networking) standards. Another difference is that 'Orthography' and 'Language-Linguistics' create a large number of overlapping browsing events (0.72-0.87 per page per day) but particularly in the case of 'Orthography' these events occur firmly in separate time zones.

We also identified 15 categories irrespective of their average English-Chinese rating. Noticeably 6 categories were primarily viewed in the Chinese time-zone viewing profile. Two of these categories again refer to Japanese and Korean pop culture, one to pop music in general and the further three to mobile communications technology, with 'Software-defined radio' referring to a particular technology used in cell phones (amongst other things). These results indicate that the highest rates of information exchange between English and Chinese language Wikipedia articles are not between Chinese and English speaking time-zones. Instead, there is a set of English language pages that Chinese speaking people use as a source of reference within typically Chinese-speaking time zones. These types of pages refer quite specifically to Asian pop culture, and mobile communications and networking technologies.

## 5.3 Trending Events

We also examined how the profiles of pages could provide a method to examine trending events. There were 45 categories referring to trending events where both pages are accessed primarily in the Chinese time zone, Pop music and Asian pop culture are present in the top trending events, and there are also several categories referring to similar types of computing and mobile technologies. Out of the 45 categories identified, all but one was viewed in the English speaking time zones and the Chinese language page being viewed in the Chinese speaking time zones, which related to 'films and cinema'.

## 5.4 Summary of Key Findings

1. Shared browsing events are much more common than shared trending events. This is difficult to interpret without the context of a third language for comparison, however, there is strong evidence that a major driver of shared events across languages actually comes from the access of English language Wikipedia in Chinese-speaking time zones. This being the case a plausible hypothesis would be that trending events are made unlikely by the fact Chinese and English speaking people are interested in different categories of English language pages on Wikipedia.

2. The types of categories people in Chinese-speaking time zones view most frequently tend to be largely related to Asian pop culture and technical computing products (as opposed to consumer information about computer products).

3. Where there is evidence of sharing across languages and time zones, these categories tend to be related to less technical, more consumer-related computing products with a significant shared interest also in linguistics.

4. Although there is significant interest in Asian pop culture within Chinese time zones there is little evidence of consistent shared viewing of pages relating to English speaking pop culture. There is some evidence of shared viewing of categories of pages related to film and cinema but these only form trending and not consistent browsing events.

# 6. CONCLUDING REMARKS

In this paper we have explored the page views of English and Chinese translated pages in Wikipedia and have developed an approach to identify different categories of Wikipedia pages based on their language and time zone page-view profile.

The key finding has been that, although Wikipedia provides a platform for global information sharing, there is strong evidence that English and Chinese speaking people browse quite different categories of pages. Further, viewing of English language pages from within typically Chinese time zones is responsible for higher rates of cross-language viewing than shared viewing between time

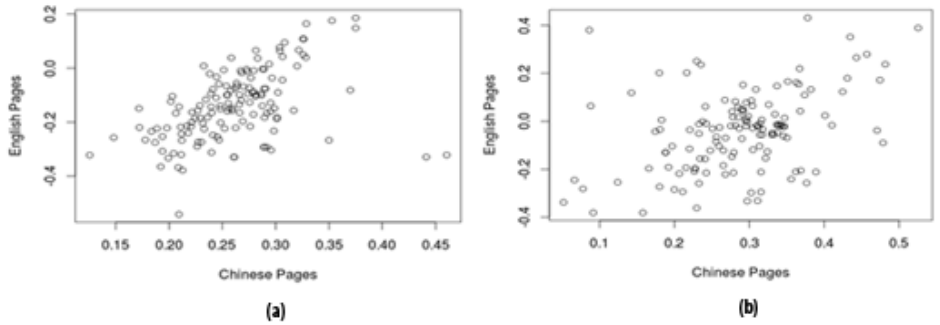


**Figure 1: (a) Scatter plot of the Chinese-English metrics for Chinese and English language pages on browsing event days. Negative values on either axis indicate the page is being viewed more frequently in an English time zone and positive values indicate the page is being viewed more frequently in a Chinese time zone. The pages at the top of the plot have positive values for both the Chinese and English language version of the page, indicating both are most commonly viewed in the Chinese time zone (b) Same plot as (a) but with the values for trending event days. There are a significant number of categories where both language versions of the page trend in the Chinese time zone and no categories where both language versions of a page trend in the English time zone.**

zones. It would appear then that Wikipedia is acting to increase interest in the English language from within Chinese time zones but this does not necessarily generate as strong an interest in the same types of content. The fact that Wikipedia has stimulated interest in English language content within China, but not particularly in Chinese content within the English speaking world is to some extent intuitive. It does to some extent confirm that our methodological approach is appropriate for identifying cross language and time-zone browsing behaviour. What is more interesting is that we have been able to identify particular types of content that are most responsible for creating this interest. Particularly, the fact that English language Wikipedia is being used as a lingua franca by Chinese people wishing to read about wider Asian culture is an interesting finding that merits further investigation.

Considering the wider societal reasons of the findings, specifically crossover of Asian culture, this might be due to the large population of Korean nationals living in China (the largest ethnic population living outside Korean), and indeed, the cultural similarities shared between nations. We also question the reason for the low number of Chinese page views (in comparison to the English subset of pages); this again may be due to cultural differences in Web use and knowledge discovery, but it also may be a result of Government control and restrictions to certain Web services. Future research will aim to generalize the distance metric for use over several languages at once, and explore ways of tagging categories of pages that will work for different languages, without relying on the English interpretation of the page only.

## 7. ACKNOWLEDGMENTS

This work is supported under SOCIAM: The Theory and Practice of Social Machines. The SOCIAM Project is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J017728/1 and comprises the Universities of Southampton, Oxford and Edinburgh

## 8. REFERENCES

[1] S. Halford, C. Pope, and L. Carr, "A Manifesto for Web Science?," *Proceedings of the Web Science 2009 Conference*, pp. 1–6, 2009.

[2] T. O'Reilly, "What Is Web 2.0," 2005.

[3] E. H. Chi, "The Social Web: Research and Opportunities," 2008.

[4] W. Hall and T. Tiropanis, "Web evolution and Web Science," *Computer Networks*, vol. 56, pp. 3859–3865, Dec. 2012.

[5] S. Goel and J. M. Hofman, "Who Does What on the Web : A Large-scale Study of Browsing Behavior," in *International Conference on Weblogs and Social Media*, 2012.

[6] T. Tiropanis, W. Hall, N. Shadbolt, D. De Roure, N. Contractor, and J. Hendler, "The web science observatory," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 100–104, 2013.

[7] T. Berners-Lee, D. J. Weitzner, W. Hall, K. O'Hara, N. Shadbolt, and J. a. Hendler, "A Framework for Web Science," *Foundations and TrendsÂő in Web Science*, vol. 1, no. 1, pp. 1–130, 2006.

[8] J. Hendler, N. Shadbolt, W. Hall, T. Berners-Lee, and D. J. Weitzner, "Web science : an interdisciplinary approach to understanding," *Communications of the ACM - Web Science*, vol. 51, no. 7, 2008.

[9] A. Halavais, "National borders on the world wide web," *New Media & Society*, vol. 2, no. 1, pp. 7–28, 2000.

[10] W. Hall, "The Ever Evolving Web: The Power of Networks," *Journal of Communication*, vol. 5, pp. 651–664, 2011.

[11] A. Kralisch and T. Mandl, "Barriers to information access across languages on the internet: Network and language effects," vol. 3, pp. 54b–54b, 2006.

[12] N. Li and G. Kirkup, "Gender and cultural differences in Internet use: A study of China and the UK," *Computers & Education*, vol. 48, pp. 301–317, Feb. 2007.

[13] K. Bharat and B. Chang, "Who links to whom: Mining linkage between web sites," in *Proceedings of the 2001 IEEE International Conference on Data Mining*, pp. 51–58, 2001.

[14] Wikipedia, "Wikipedia Statistics," 2014.

[15] U. Pfeil, P. Zaphiris, and C. S. Ang, "Cultural Differences in Collaborative Authoring of Wikipedia," *Journal of Computer-Mediated Communication*, vol. 12, no. 1, pp. 88–113, 2006.

[16] A. Hautasaari and T. Ishida, "Analysis of discussion contributions in translated wikipedia articles," pp. 57–66, 2012.

[17] R. Glott, P. Schmidt, and R. Ghosh, "Wikipedia survey–overview of results," *United Nations University: Collaborative Creativity Group*, 2010.

[18] A. Kittur, E. Chi, B. A. Pendleton, and T. Mytkowicz, "Power of the Few vs . Wisdom of the Crowd : Wikipedia and the Rise of the Bourgeoisie," *Algorithmica*, pp. 1–9, 2007.

[19] J. Stuckman and J. Purtilo, "Measuring the wikisphere," *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, 2009.

[20] L. S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi, "Temporal Evolution of the Wikigraph," in *Proceedings of Web Intelligence*, pp. 45–51, IEEE CS Press., 2006.

[21] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, "Preferential attachment in the growth of social networks: the case of Wikipedia," *Physical Review E*, vol. 74, no. 3, p. 4, 2006.

[22] A.-L. Barabasi and R. Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, no. October, pp. 509–512, 1999.

[23] S. P. Ponzetto and M. Strube, "Extracting world and linguistic knowledge from wikipedia," pp. 7–8, 2009.

[24] L.-P. Huberdeau, S. Paquet, and A. Désilets, "The cross-lingual wiki engine: enabling collaboration across language barriers," in *Proceedings of the 4th International Symposium on Wikis*, p. 13, ACM, 2008.

[25] A. Hautasaari and T. Ishida, "Discussion about Translation in Wikipedia," *2011 Second International Conference on Culture and Computing*, pp. 127–128, Oct. 2011.

[26] H.-T. Liao, "Conflictual consensus in the chinese version of wikipedia," pp. 1–10, 2008.

[27] C. A. Wang and X. M. Zhang, "Network Centrality and Contributions to Online Public Good - The Case of Chinese Wikipedia," *2012 45th Hawaii International Conference on System Sciences*, pp. 4515–4524, Jan. 2012.

[28] B. Hecht and D. Gergle, "The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context," *Proceedings of the SIGCHI Conference on Computer Human Interaction*, pp. 291–300, 2010.

[29] P. Bao, B. Hecht, S. Carton, M. Quaderi, M. Horn, and D. Gergle, "Omnipedia: Bridging the wikipedia language gap," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, (New York, NY, USA), pp. 1075–1084, ACM, 2012.

[30] D. Nguyen, A. Overwijk, C. Hauff, D. R. B. Trieschnigg, D. Hiemstra, and F. D. Jong, "WikiTranslate : Query Translation for Cross-Lingual Information Retrieval Using Only Wikipedia," pp. 58–65, 2009.

[31] E. Adar, M. Skinner, and D. S. Weld, "Information arbitrage across multi-lingual Wikipedia," *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09*, p. 94, 2009.

[32] A. Forte and A. Bruckman, "From wikipedia to the classroom: exploring online publication and learning," in *Proceedings of the 7th international conference on Learning sciences*, pp. 182–188, International Society of the Learning Sciences, 2006.

[33] N. Augar, R. Raitman, and W. Zhou, "Teaching and learning online with wikis," in *Proceedings of the 21st ASCILITE Conference*, no. December, pp. 95–104, 2004.

[34] A. Spoerri, "What is popular on Wikipedia and why?," *First Monday*, vol. 12, no. 4, pp. 1–22, 2007.

[35] C. Honeycutt and S. C. Herring, "Beyond Microblogging: Conversation and Collaboration via Twitter," in *42nd Hawaii International Conference on System Sciences*, vol. 0 of *42nd Hawaii International Conference on System Sciences*, pp. 1–10, IEEE Computer Society, 2009.

[36] H.-l. Chen, "The use and sharing of information from Wikipedia by high-tech professionals for work purposes," *The Electronic Library*, vol. 27, no. 6, pp. 893–905, 2009.

[37] S. Lim, "How and why do college students use wikipedia?," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 11, pp. 2189–2202, 2009.

[38] B. Ahn, B. V. Durme, and C. Callison-Burch, "WikiTopics: what is popular on Wikipedia and why," in *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, pp. 33–40, 2011.

[39] S. Ardon, A. Bagchi, A. Mahanti, and A. Ruhela, "Spatio-temporal analysis of topic popularity in twitter," in *arXiv preprint arXiv:1111.2904*, 2011.

[40] D. Quercia, L. Capra, and J. Crowcroft, "The social world of twitter: Topics, geography, and emotions," in *Sixth International AAAI Conference on Weblogs and Social Media*, no. Hansen 1999, 2012.

[41] F. Abel, Q. Gao, G. Houben, and K. Tao, "Semantic enrichment of twitter posts for user profile construction on the social web," in *The Semantic Web: Research and Applications*, pp. 1–15, 2011.

[42] S. Amer-Yahia, S. Anjum, and A. Ghenai, "MAQSA: a system for social analytics on news," in *Proceedings of the 2012 international conference on Management of Data*, pp. 653–656, 2012.

[43] T. D. Nies and S. Coppens, "Automatic discovery of high-level provenance using semantic similarity," in *Proceedings of the 4th International Provenance and Annotation Workshop IPAW*, 2012.

[44] F. Iacobelli, N. Nichols, L. Birnbaum, and K. Hammond, "Finding new information via robust entity detection," in *Proactive Assistant Agents (PAA2010) AAAI 2010 Fall Symposium*, 2010.

[45] M. Osborne, S. Petrovic, and R. McCreadie, "Bieber no more: First Story Detection using Twitter and Wikipedia," in *Proceedings of the SIGIR Workshop in Time-aware Information Access*, 2012.

[46] H. Ding, G. Trajcevski, and P. Scheuermann, "Querying and mining of time series data: experimental comparison of representations and distance measures," in *Proceedings of the VLDB Endowment*, vol. 1, 2008.

[47] T.-c. Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, vol. 24, pp. 164–181, Feb. 2011.

[48] J. Ernst, G. J. Nau, and Z. Bar-Joseph, "Clustering short time series gene expression data.," *Bioinformatics (Oxford, England)*, vol. 21 Suppl 1, pp. i159–68, July 2005.