An Approach for Using Wikipedia to Measure the Flow of Trends Across Countries

Ramine Tinati, Thanassis Tiropanis, Leslie Carr
University of Southampton
Southampton
United Kingdom
{rt506, tt2, lac}@ecs.soton.ac.uk

ABSTRACT

Wikipedia has grown to become the most successful online encyclopedia on the Web, containing over 24 million articles, offered in over 240 languages. In just over 10 years Wikipedia has transformed from being *just* an encyclopedia of knowledge, to a wealth of facts and information, from articles discussing trivia, political issues, geographies and demographics, to popular culture, news articles, and social events. In this paper we explore the use of Wikipedia for identifying the flow of information and trends across the world. We start with the hypothesis that, given that Wikipedia is a resource that is globally available in different languages across countries, access to its articles could be a reflection human activity. To explore this hypothesis we try to establish metrics on the use of Wikipedia in order to identify potential trends and to establish whether or how those trends flow from one county to another. We subsequently compare the outcome of this analysis to that of more established methods that are based on online social media or traditional media. We explore this hypothesis by applying our approach to a subset of Wikipedia articles and also a specific worldwide social phenomenon that occurred during 2012; we investigate whether access to relevant Wikipedia articles correlates to the viral success of the South Korean pop song, "Gangnam Style" and the associated artist "PSY" as evidenced by traditional and online social media. Our analysis demonstrates that Wikipedia can indeed provide a useful measure for detecting social trends and events, and in the case that we studied; it could have been possible to identify the specific trend quicker in comparison to other established trend identification services such as Google Trends.

Categories & Subject Descripters

H.1.2 [Information Systems]: User/Machine Systems

Keywords

Web Observatories, Wikipedia, Social Machines, Web Science

1. INTRODUCTION

The Web is a 'Social Technology' [6], evolving as a network of networks [15] representing a collection of different human activities, sharing characteristics – both socially and technically – with practices reflective of society, or activities which are novel and unique to the online environment. Undoubtedly an integral part of society, the Web offers the potential to overcome the linguistic, spatial, and temporal barriers that are faced within the offline world. Whilst communication and information exchange at a global scale was possible before the Web, the associated costs and practicalities are now less intense and demanding [33].

As a consequence of the Web's global network, humans have the ability to obtain and share information at an astonishing rate; the emergence of large-scale social networking platforms enhanced further the potential for sharing information on a large scale on the Web [22,29]. However, as a result of the socio-technical relationship between the Web and Society [14], the implications are not only an online phenomena, but offline as well. Structures in society such as the film and music industry [4], and even traditional establishments such as government [24] have all been affected, requiring to adapt in response to the Web's ability to enable information to go viral or turn into 'Memes'.

Examining the Web from a data-centric perspective, the flow and exchange of information and human communication provides to some extent a pulse of human activities, an indicator which is embedded with different cultures and societies reflecting the worldwide usage of the Web. Examining this *pulse* can be done at scale, both at the micro – examining specific Web platforms [20] – or at the macro level, examining the Web as a holistic platform [10]. Both approaches provide their own level of granularity to examining the current pulse of the Web, and serve different purposes accordingly.

Working on the micro level, we are already beginning to develop methods to establish trending information [22], and filtering approaches to distill valuable content [18,32]. Research at the micro level tends to favor relatively well-established online social networking platforms such as Twitter and Facebook, as they offer a firsthand insight into the communications interactions of humans, providing a lens to explore the network structures that form through these communications and how information diffuses. Web Search engines also now provide services to examine the flow of information and trends online, with examples such as Google Trends offering new ways to uncover human search activity, including the detection and monitoring of worldwide flu epidemics [11]. These services are complementing our understanding of human activity, providing a platform to explore the current 'popular', both at scale and by geographic and potentially demographic granularity.

These services offer a good starting point to measure the pulse of the Web; however, restrictions within the platforms – in terms of linguistic barriers, restrictions to data access, and also lack of global reach– can limit the analytical capabilities. Wikipedia, one kind of Web activity provides another perspective on human activity and the exchange of information, not only at global scale, but also across linguistic borders.

Wikipedia is a large-scale, extensive collaborative Web platform which provides a vast collection of knowledge similar to that of a

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

traditional encyclopedia [19]. Although sharing many similarities with its offline counterpart, unlike a traditional knowledge system, Wikipedia enables any[1] individual to freely create, contribute and edit articles (Wikipedia articles) without the need for authorization from administrators or moderators. 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 Website traffic [12,19], with current figures estimating a number of 18 million articles in 285 different languages, which are supported by 29 million registered users [17]. Many Wikipedia articles contain several if not hundreds of translated versions, providing a truly global resource of knowledge. Unlike traditional social networks, Wikipedia is not restricted to barriers such as language or network structures that can impose limitations to the ability for users to access, edit or create content; we can assume that almost everybody knows that (i) Wikipedia is available and that it (ii) it will have an article on almost any topic. Could it be that data on the use of Wikipedia can provide a valuable reflection of human activity in addition to online social networks and Web search engine usage data?

In this paper we explore the potential of using Wikipedia as a way to measure human activity on the Web. With the ever growing number of Wikipedia articles in multiple languages, we examine if it is possible to use the service to reveal trends on the Web; not only in terms of information online, but offline as well. Building upon current research that explores traditional social networking platforms as an indicator for information diffusion, we develop a methodology to examine the article view access in Wikipedia of specific and categories of articles, and aim to understand whether it is a suitable candidate for providing a measure for human activity on the Web, and its application within a Web Observatory.

2. BACKGROUND

Current Wikipedia research can be broadly distilled into a number of categories. Firstly, the analysis of Wikipedia's structure and growth [7,31], using graph theory and approaches developed in network science. Such research typically examines Wikipedia article linking structures and its growth over time revealing similar scale-free properties to those observed in the evolution of the World Wide Web [5].

Another line of research concerns identifying and classifying common motivations for the communities of contributors and editors responsible for the articles produced in Wikipedia [13,30]. Such literature not only examines the actual practices of the communities, but also aims to develop an understanding of the social and cultural factors that underpin their actions [19]. Certain studies also examine the role of the reader as a participant within the Wikipedia ecosystem, investigating whether the reader has an impact on the contribution process that drives the development of new Wikipedia articles [3,35]. Finally, studies also identify different motivations for individuals to view Wikipedia articles, which has been suggested as a source of quick reference material related to professional [8] or scholarly tasks [23], or as a method to instantly learn about a one-off news event [2].

Increasingly, there is a growing area of interest in the multi-lingual support that Wikipedia offers [9,16,28]. Within this area, research extends to investigating topics such as the communities of users that are involved with the translation process of articles, providing an analysis of discussion logs and edit history of translated articles to identify barriers to adoption, and the social and technical processes that exist during and after an articles reaches a state of translation [16]. Additional research has investigated the effects of culture within collaborative environments, maintaining that Wikipedia is far from culturally neutral, which directly influences the collaborative efforts in article creation [27].

Finally, various technical solutions have been developed such as WikiTranslate [25] and CLWE [9] to aid the translation of articles, supporting individuals to cross-collaborate and discuss their translation workflow [9]. 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 [1].

Despite this work, there is a lack of methodological or technical approaches to examine the human activity surrounding the use of Wikipedia, with more attention towards examining the phenomena of collaborative editing and authoring rather than Wikipedia readership. Nevertheless, there are research methodologies for examining the role of traditional social platforms in the spread of information in terms of how they become viral [21,29], and also for performing event detection based on human communication [34], or detecting where viral information first originates from [26].

Applying these concepts to Wikipedia, the following sections provide an initial attempt at a methodological and technical approach to event and trend detection based upon the views within Wikipedia.

3. APPROACH FOR DATA COLLECTION AND ANALYSIS

This section outlines the approach that was developed in order to examine activity occurring on Wikipedia. The premise of this investigation is to examine whether Wikipedia can provide a method to explore human Web activity based upon volume of views that an individual, or set of articles receive over a given time period; with a specific focus towards its functionality for event and trend detection. We then compare the outcome of this approach to that of more traditional Web trend detection platforms in order to assess and discuss its performance.

3.1 Data Collection and Analysis Architecture

The initial step is dependent on the focus of the study. As Wikipedia contains over 28 million articles, available in 285 languages, the breadth of knowledge that it contains is vast. However, due to the large corpus of articles, the scope of the study needs to be limited to specific topics. Tackling this, our approach either examines specific Wikipedia articles, or a subset of articles based upon the Wikipedia mapping ontology[2] used to categorize and define articles based upon their content, i.e. the article contains knowledge about a *Person*, a *Location*, an *Event* etc. If a subset of articles is required, it is possible to query the DBpedia SPARQL endpoint[3] to return the list of corresponding article names that are associated with the Wikipedia category required.

Based on the initial step, it is then possible to examine the unique views for a given article – or set of articles – on an hourly basis. Using the Wikipedia article view logs[4], every article within the

[1] Exceptions to this do exist, where some articles are 'locked-down' as a result of edit wars or controversial articles

[2] http://wiki.dbpedia.org/Ontology

[3] http://dbpedia.org/sparql

[4] http://dumps.wikimedia.org/

Wikipedia corpus is listed (with its corresponding translated version semantically linked by a shared URL), and contains the number of unique views in a given hour. By determining a time period of analysis and extracting the required articles (and their translated versions as necessary) within each hourly log, the final step required is to transform the data into a matrix of articles with number of views and timestamps.

Based on this organized dataset, we are now able to either calculate a baseline value for the number of hourly views within a category of Wikipedia articles, or examine the dataset at a more granular level, investigating a specific article for unexpected spikes in article views.

Finally, in order to establish whether Wikipedia is useful as an indicator for events or trends detection, we cross-examine the findings using Google Trends, a trend analysis services that provided a normalized measure of human search traffic related to specific search terms or keywords. Using this data we see whether the activity within Wikipedia correlates to Google Trends metrics to reveals any major trends or events that have occurred.

4. ANALYSIS

In order to examine whether Wikipedia is a suitable source for identifying tends and events on the Web, the following analysis has chosen a worldwide trend that happened during 2012. For this study we are examining the South Korean pop song 'Gangnam Style', which became a global phenomenon since its release on July 15th 2012; breaking previous YouTube records with the video reaching over 1 billion views in less than 5 months[5]. Although originating in South Korea, the song became a global success, reaching music top chart worldwide.

Using this topic as a basis for the study, the following analysis uses the methodology described in Section 3 and applies it to two studies; (1) an analysis of the English and South Korean Wikipedia articles 'Gangnam Style' and songs artist 'PSY'; (2) an analysis of a subset of English Wikipedia articles categorized as 'artist'[6] within the DBpedia. Based upon the different levels of granularity offered between these two studies, we discuss the suitability of Wikipedia for trend and event identification.

4.1 Article View Analysis: 'Gangnam Style' and 'PSY'

Applying the methodology in Section 3, we examined two Wikipedia articles, 'Gangnam_Style' and 'PSY' in both the English and South Korean versions, during the time period of 1st June – 31st December 2012. During this time period, the English Wikipedia article for 'Gangnam_Style' and 'PSY' had 139,202,55 and 1,325,493 unique views respectively, and the South Korean articles received 95,278 and 145,503 unique views.

As Figure 1 shows, the number of article views for the English Wikipedia article 'Gangnam_Style' was 2 orders of magnitude greater than the South Korean article, and in addition to this, The English article was created (29th July 2012) 19 days before the South Korean article. Also noted was that for the English article the highest rate of article views was between September 19th and October 8th 2012, where the South Korean article spanned for a much shorter time, albeit around the same time period of 22nd – 27th September 2012.

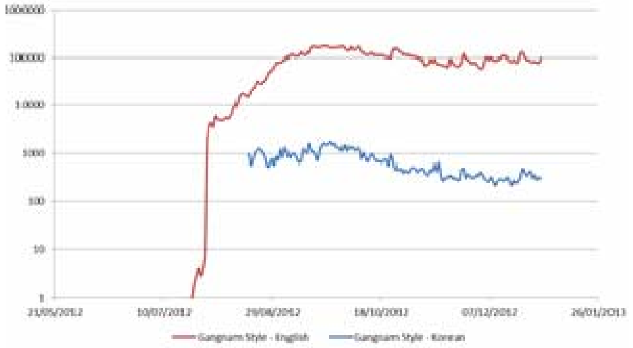

Figure 1. Number of article views (log-scale) for the English (Red) and South Korean (Blue) Wikipedia Article 'Gangnam_Style'

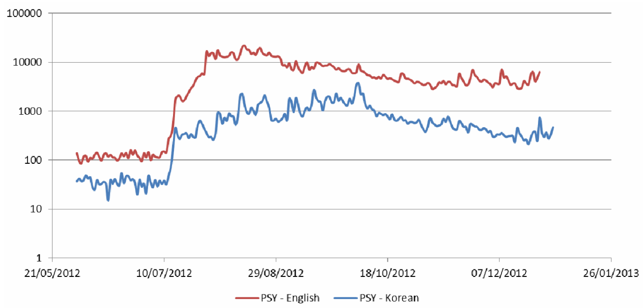

Figure 2. Number of article views (log-scale) for the English (Red) and South Korean (Blue) Wikipedia Article 'PSY'

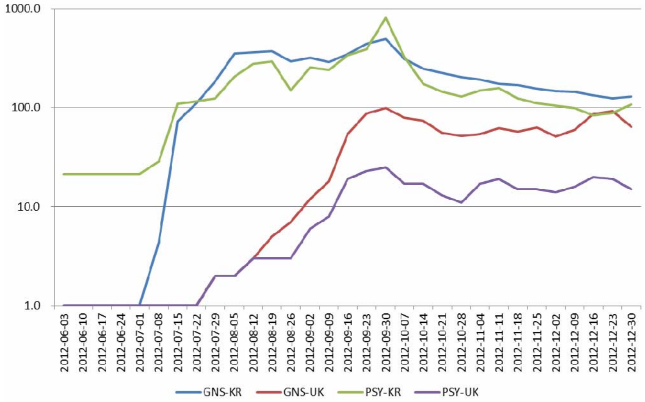

Figure 3. Google Trends search statistics for 'Gangnam Style' and 'PSY' for United Kingdom (Red and Purple) and South Korea (Blue and Green). Values represent peak search volume (log scale)

As Figure 2 illustrates, the article views for 'PSY' were different to that of 'Gangnam_Style' as both articles existed before the release of the song. Although the English article received a greater article views during the time period analyzed, what was noted was the initial spike in views occurred first on the South Korean article 2 days before the English article, and the highest peak in article views took longer to reach for the South Korean article (4th October 2012) compared with the English article (15th August 2012).

Comparing these findings with the search trends harvested from Google Trends in Figure 3, we find that in South Korean the volume of searches for both 'Gangnam Style' and 'PSY' (using the Korean Character set for the search criteria) are greater than the United Kingdom, and that the article views on the Wikipedia article for 'Gangnam Style' for both the South Korean and English reached its highest volume before the Google trends statistics; this was also true for the 'PSY' article. What was also noted was the

[5] http://youtube-global.blogspot.co.uk/2012/12/ytvev.html (Accessed on 1st February 2012)

[6] http://dbpedia.org/ontology/artist

correlation between spikes in search activity reported by Google Trends, and number of article views in Wikipedia. Examining Figure 3, the peak in search volume for both 'Gangnam Style and 'PSY' occurred on the 9th October 2012, which corresponded with a concert in Seoul, South Korea; during this time period both the English and South Korean articles for 'Gangnam_Style' and 'PSY' incurred a spike in article views.

4.2 Article View Analysis: DBpedia 'Artist'

The previous section shows that analysis of Wikipedia views for the chosen trending band and song seems to be in line with the analysis of search engine use. However, the hierarchy of topics on Wikipedia, as reflected in DBpedia, can give us further insights. In this case, it can show us how access to the articles on the specific artist or song relates to access to articles of other artists.

Using the same methodology as before, we now take a subset of Wikipedia articles listed within the Wikipedia category of 'artist' and examine their unique article views. The DBpedia SPARQL query returned a subset of 7,752 articles, and using this dataset we gather article view statistics between the January 1st 2010 and December 31st 2012. Finally, we discarded articles that had 0 views, as these were deemed articles that were no longer existent or active. The initial analysis of the total number of views per article within the extracted subset of Wikipedia articles exhibits the familiar properties of preferential attachment often characterized by the link structure of Web, producing the 'long tail' effect. Effectively, this means that the majority of articles have a low volume of traffic.

Using the total number of views of all articles per day as a simple measure to form a baseline value, we are presented with a '*typical viewing*' count to examine the subset of articles for event detection. As Figure 4 shows, over the 1096 days of collected data, it is possible to plot a time series of page views (normalized) to help identify different levels of activity around a specific Wikipedia category. As the dataset extracted is concerned with the category of 'artist' we are therefore presented with a metric of daily views within this category of Wikipedia articles. Interrogating the data, if we examine the spikes in activity shown in Figure 5, the rise in the number of article views during the timestamp of 14th February 2011 and 13th February 2012 correspond to the 'Grammy Awards', which is a US based award ceremony for achievements in the music industry.

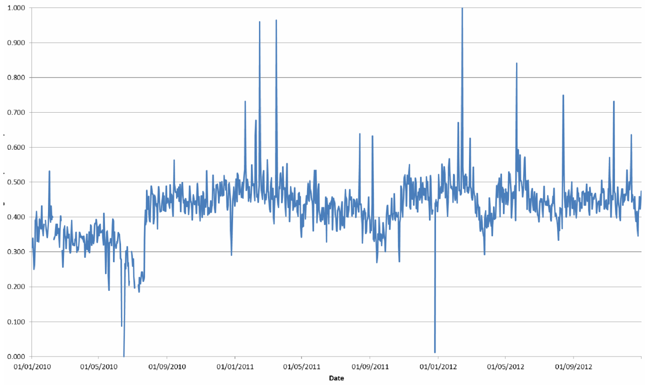

Figure 4. Total number of daily article views (Normalized) for the subset of 7,752 articles categorized as 'artist'

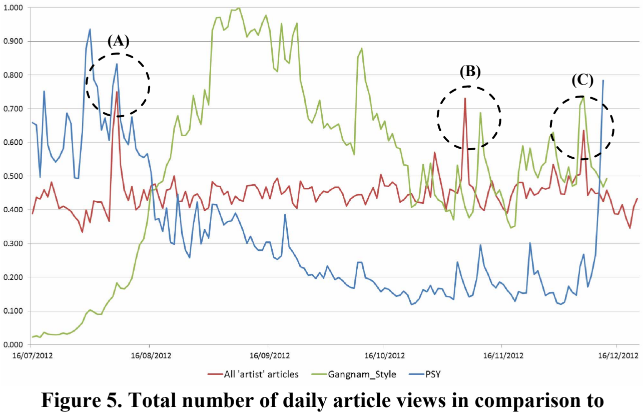

Figure 5. Total number of daily article views in comparison to views for 'Gangnam_Style' and 'PSY' (Normalized). Jul – Dec 12

Drawing upon the individual articles analyzed in 4.1, by comparing the total number of article views against the views for 'Gangnam_Style' and 'PSY' (normalized), it is also possible to identify – as shown by the two dashed circles on Figure 5 – how the popularity of an individual article can correlate to an entire collection of articles (A)(C), but also in the case of spike identified by (B), how monitoring a subset of articles may provide an indication of articles '*soon-to-be*' popular.

Whilst these baseline values provide a method of identify specific events and trends, they also provide a way to examine the use of Wikipedia and as a lens to examine current levels of human interests and activity. The baseline values shown in Figures 5 not only show the rise in activity with a subset of articles, but also provide an signal to time periods of dormant activity, indicated by the troughs in activity during the months of December and March, which correspond to worldwide holidays of cultural and religious significance, including Christmas and Easter Vacation.

5. CONCLUDING REMARKS

In this paper we have explored the potential of using Wikipedia as a way to measure social trends and events happening worldwide. By developing a methodology that enables Wikipedia human activity to be measured, we have demonstrated that it is possible to use such a metric to identify trends or events happening both online and offline.

Despite being an initial attempt at this kind of study, we have shown that Wikipedia provides some level of analysis for exploring the flow of information across different languages, which potentially may transpose to examining different country specific trends. By using the music phenomena of 'Gangnam Style' as a reference point, we have discovered the different patterns of Wikipedia usage for the English and South Korean Wikipedia article. Although we cannot be sure that those viewing the South Korean article are located in South Korean (nor can we assume that all English article views are within the United Kingdom), what we do notice is the South Korean article is viewed much less than the English article, and also the English article tends to peak far earlier than the South Korean equivalent. Speculating at this, it may be that Wikipedia is as popular in South Korea as it is English speaking countries.

We have also explored the potential of using a subset of Wikipedia articles selected by using the Wikipedia categories of articles to create a baseline value for the number of views expected. We have shown that this also helps to provide an indicator for events and trends on the Web. Developing this further, if we were to take another subset of articles, say for instance politics, we could then provide a measure of the current level of interest in comparison to that of artists or music.

The initial analysis of Wikipedia has shown promising results for offering an observatory lens to monitor and identify worldwide trends and events, we now wish to take the methodology further and explore the potential of cross-comparison event detection between different Web trend services, including Google Trends and Twitter.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Adar, E., Skinner, M., and Weld, D.S. Information arbitrage across multi-lingual Wikipedia. *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09*, (2009), 94.
2. Ahn, B., Durme, B. Van, and Callison-Burch, C. WikiTopics: what is popular on Wikipedia and why. *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, (2011), 33–40.
3. Antin, J. and Cheshire, C. Readers are not free-riders: reading as a form of participation on wikipedia. *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, (2010), 127–130.
4. Arewa B., O. YouTube, UGC, and Digital Music: Competing Business and Cultural Models in the Internet Age. *Northwestern University Law Review 104*, January (2010).
5. Barabasi, A.-L. and Albert, R. Emergence of Scaling in Random Networks. *Science 286*, October (1999), 509–512.
6. Berners-Lee, T., Weitzner, D.J., Hall, W., O'Hara, K., Shadbolt, N., and Hendler, J. a. A Framework for Web Science. *Foundations and Trendsin Web Science 1*, 1 (2006).
7. Capocci, A., Servedio, V.D.P., Colaiori, F., et al. Preferential attachment in the growth of social networks: the case of Wikipedia. *Physical Review E 74*, 3 (2006), 4.
8. Chen, H. The use and sharing of information from Wikipedia by high-tech professionals for work purposes. *The Electronic Library 27*, 6 (2009), 893–905.
9. Désilets, A. The Cross-Lingual Wiki Engine□: Enabling Collaboration Across Language Barriers Categories and Subject Descriptors. *Proceedings of the 4th International Symposium on Wikis and Open Collaboration.* (2008).
10. Dourisboure, Y., Geraci, F., and Pellegrini, M. Extraction and classification of dense implicit communities in the Web graph. *ACM Transactions on the Web 3*, 2 (2009), 1–36.
11. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature 457*, 7232 (2009), 1012–4.
12. Glott, R., Schmidt, P., and Ghosh, R. Wikipedia survey–overview of results. *United Nations University.* (2010), 1–11.
13. Gokhman, S. and McDonald, D. Wiki architectures as social translucence enablers. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration* (2011),
14. Halford, S., Pope, C., and Carr, L. A Manifesto for Web Science? *Proceedings of the Web Science 2009*, (2009), 1–6.
15. Hall, W. The Ever Evolving Web□: The Power of Networks. *Journal of Communication 5*, (2011), 651–664.
16. Hautasaari, A. and Ishida, T. Discussion about Translation in Wikipedia. *2011 Second International Conference on Culture and Computing*, (2011), 127–128.
17. Hautasaari, A. and Ishida, T. Analysis of discussion contributions in translated Wikipedia articles. *Proceedings of the 4th international conference on Intercultural Collaboration - ICIC '12*, (2012), 57.
18. Hurlock, J. and Wilson, M.L. Searching Twitter: Separating the Tweet from the Chaff. *ICWSM2011*, 161–168.
19. Kittur, A., Chi, E., Pendleton, B.A., and Mytkowicz, T. Power of the Few vs . Wisdom of the Crowd□: Wikipedia and the Rise of the Bourgeoisie. *Algorithmica*, (2007), 1–9.
20. Kwak, H., Lee, C., Park, H., and Moon, S. What is Twitter, a social network or a news media? *19th international conference on the World Wide Web*, (2010), 591–600.
21. Leskovec, J., Adamic, L.A., and Huberman, B.A. The Dynamics of Viral Marketing. *Machine Learning 1*, May (2007), 1–46.
22. Leskovec, J., Backstrom, L., and Kleinberg, J. Meme-tracking and the Dynamics of the News Cycle. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining.* (2009), 497-506
23. Lim, S. How and why do college students use Wikipedia? *Journal of the American Society for Information Science ... 60*, 11 (2009), 2189–2202.
24. Nahon, K., Hemsley, J., Walker, S., and Hussain, M. Fifteen Minutes of Fame: The Power of Blogs in the Lifecycle of Viral Political Information. *Policy Internet 3*, 1 (2011), 1–30.
25. Nguyen, D., Overwijk, A., Hauff, C., Trieschnigg, D.R.B., Hiemstra, D., and Jong, F. De. WikiTranslate□: Query Translation for Cross-Lingual Information Retrieval Using Only Wikipedia. *Workshop of the Cross-Language Evaluation Forum.* (2009), 58–65.
26. Osborne, M., Petrovic, S., and McCreadie, R. Bieber no more: First Story Detection using Twitter and Wikipedia. *Workshop in Time-aware Information Access*, (2012).
27. Pfeil, U., Zaphiris, P., and Ang, C.S. Cultural Differences in Collaborative Authoring of Wikipedia. *Journal of Computer-Mediated Communication 12*, 1 (2006), 88–113.
28. Ponzetto, S. Extracting world and linguistic knowledge from Wikipedia. *Proceedings of Human Language Technologies*, June (2009), 7–8.
29. Simmons, M., Adamic, L.A., and Adar, E. Memes Online: Extracted, Subtracted, Injected, and Recollected. *ICWSM 2011*, (2011).
30. Slattery, S. and Ave, W.B. "Edit This Page": The Socio-technological Infrastructure of a Wikipedia Article. *Proceedings of the 27th ACM international conference on Design of communication* (2009), 289–295.
31. Stuckman, J. and Purtilo, J. Measuring the wikisphere. *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, (2009).
32. Tinati, R., Carr, L., and Hall, W. Identifying communicator roles in twitter. *Proceedings of the 21st international conference companion on World Wide Web*, (2012).
33. Welzer, T. Cultural awareness in social media. *Proceedings of the international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, (2011), 3–7.
34. Weng, J. and Lee, B. Event Detection in Twitter. *ICWSM2011*. (2011)
35. Zhang, X. Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *The American Economic Review 101*, June (2011), 1601–1615.