Indexing and Analyzing Wikipedia's Current Events Portal, the Daily News Summaries by the Crowd

Giang Binh Tran
L3S Research Center, Hanover, Germany
gtran@L3S.de

Mohammad Alrifai
L3S Research Center, Hanover, Germany
alrifai@L3S.de

ABSTRACT

Wikipedia's Current Events Portal (WCEP) is a special part of Wikipedia that focuses on daily summaries of news events. The WikiTimes project provides structured access to WCEP by extracting and indexing all its daily news events. In this paper we study this part of Wikipedia and take a closer look into its content and the community behind it. First, we provide descriptive analysis of the collected news events. Second, we compare between the news summaries created by the WCEP crowd and the ones created by professional journalists on the same topics. Finally, we analyze the revision logs of news events over the past 7 years in order to characterize the WCEP crowd and their activities. The results show that WCEP has reached a stable state in terms of the volume of contributions as well as the size of its crowd, which makes it an important source of news summaries for the public and the research community.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]; H.3.1 [**Information Storage and Retrieval**]: Content Analysis

Keywords

News Events; Wikipedia; Analysis

1. INTRODUCTION

Wikipedia's Current Events portal [1] (WCEP) provides a platform for creating and archiving daily summaries of relevant news events by the crowd, for the crowd. Users insert short summaries of news events on a daily basis to the main page of the portal. One page is created for each day and incorporated into the portal home page (there is one container for each day of the current month on the home page). Events are then simply bullet listed in the right container. At the end of each month, all events of the ending month are automatically archived into one Wikipedia page, which has a unique identifier of the form Month_Year (e.g. http://en.wikipedia.org/wiki/October_2013).

WCEP can be seen as a platform for collaboratively creating and archiving daily summaries of relevant news events by the crowd, for the crowd. Unlike Wikinews [2], WCEP includes only short summaries of daily news, thus providing a compact overview of relevant news. A key advantage of WCEP is the manual annotations of event summaries as can be shown in Figure 1. Each event is typically assigned to a specific category (e.g. *Armed conflicts, Business and economy* etc.), and (when applicable) is linked to a news storyline (e.g. *2013 United States embassy bombing in Ankara*), which is described in a separate Wikipedia article. Entity mentions (e.g. *Pakistan, Antony Jenkins* etc.) are linked to Wikipedia articles. In addition, links to external resources about the event (typically online news articles) are provided. For these reasons, WCEP appears as an invaluable resource of human created knowledge about news events that can be harvested and exploited in many applications.

Figure 1: A screenshot from Wikipedia's Current Events portal

While there have been lots of studies on Wikipedia in general (e.g., [6], [5], [8]), in this paper we are focusing on WCEP given its special characteristics and time-centered structure. Our main goal in this study is to take a closer look into WCEP in order to better estimate its quality, reliability and stability. More precisely, we try to answer the following questions:

1. **Descriptive Analysis**: what type and how many news events can be found in WCEP and how is it evolving over time? (see Section 3)
2. **Comparative Analysis**: how much is the coverage of this portal in comparison with summaries created by professional journalists? (see Section 4)
3. **Crowd Activity Analysis**: how big and dynamic is the community behind this portal and how is it evolving over time? (see Section 5)

[1] Wikipedia's Homepage → left Sidebar → Current Events or http://en.wikipedia.org/wiki/Portal:Current_events

[2] http://www.wikinews.org/

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2576942 .

In addition, we introduce the WikiTimes system (Section 2), which extracts, indexes and provides a structured access to the WCEP collection of news events. It is worth mentioning that in this study we focus on the English part of WCEP and leave the analysis of other languages for future work.

2. WIKITIMES

In this section we briefly describe the WikiTimes[3] system for extracting and indexing news events (and their attributes) from WCEP. We have written scripts for extracting the daily events from monthly pages starting from January 2000 and up to date. The extracted events are then indexed in a database, as well as full-text indexed using Apache Lucene in a regualr basis. Currently, we are also working on building an RDF knowledge base of the events. The system provides a GUI for structured and keyword based search as well as a RESTful web service interface (a SPARQL endpoint will be added soon) for downloading the datasets.

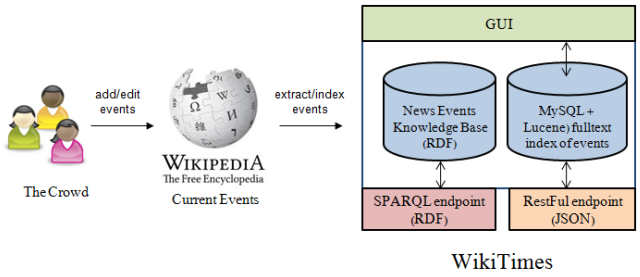

Figure 2: High level overview of WikiTimes architecture

Table 1 shows the key attributes of a news event that are extracted from WCEP and being indexed in WikiTimes indexes. Using WikiTimes interfaces, users can query for events that are connected to a certain entity or new story. Users can also query for the timeline of a certain news story. More details on the WikiTimes interfaces can be found on the project website[3].

Concept	Description
Description	a brief summary of the news event
Date	the date at which the event occurred
Entities	a list of entities mentioned in the description (e.g. person, location, or simply anything that is described by a Wikipedia article)
Category	the high level (topic) category the event belongs to (e.g. Armed conflicts, Disasters, Business etc.)
StoryLine	a news story that spans a period longer than one day, this is typically a story that is described in Wikipedia in a separate article
References	links to external online news articles

Table 1: Event information indexed by WikiTimes

3. DESCRIPTIVE ANALYSIS

In this section, we present the results of the descriptive analysis on the news events extracted from WCEP.

3.1 Statistics on events and event topics

In total, there has been about 50K news events found in WCEP between 1 January 2000 and 30 November 2013. Figure 3 shows that the number of events per month has been growing over the past years, which indicates that this portal is still active and gaining more content. By looking into the categories of the events (as manually given by the editors) we see (in Figure 4) that there is a wide coverage of topics although most of the reported news events belong to the armed conflicts, politics, crimes and natural disasters categories, while less events have been reported in the science and arts categories. It is also worth mentioning that about 44% of the events have not been classified into any category. This is due to the fact that news events categories have been more recently introduced in WCEP starting in 2010 and, hence, all events reported before 2010 were unclassified (Figure 5 also confirms this finding).

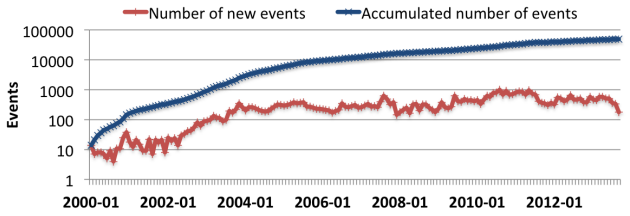

Figure 3: Number of events per month

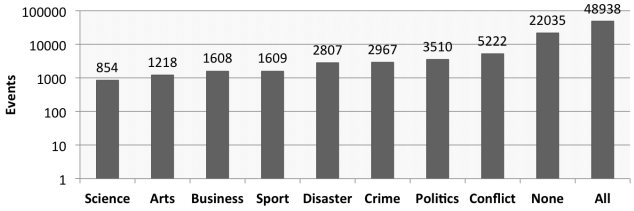

Figure 4: Distribution of events over major categories

3.2 Top cited news agencies

Table 2 shows the most cited news agencies based on the references to external news articles. Interestingly, we observe that BBC is the most cited news agencies in all categories. By looking at the location from editors IP addresses (13K editors in WCEP do not have Wikipedia accounts and their edits are logged by their IP address), we found that most of them are from US (53%), and following by UK (10%), Canada (9%) and India (4%). Nevertheless, we observe that most contributions come from the group of Wikipedia admins (as will be discussed in more details in Section 5). We don't have information on the origin of those admins, but it turns out that this is a relatively small group of about 60-70 admins. Following discussions about how to cite news in Wikipedia [4], we found that BBC appears as a preferred source for citation because of its neutrality, urls reliability and freshness, among others. In addition, we spot that WCEP editors also often cite domain specific news agencies such as *espn* and *goal* in Sport, *bloomberg* and *WSJ* in Business topics.

All	Politics	Business	Sport
bbc 14656	bbc 1123	bbc 397	bbc 478
reuters 4021	aljazeera 331	reuters 295	espn 180
cnn 3342	reuters 314	aljazeera 85	guardian 122
google news 2185	google news 184	cnn 73	usatoday 58
aljazeera 2132	cnn 182	guardian 73	telegraph 49
guardian 1847	guardian 167	*bloomberg 60*	cnn 46
nytimes 1359	wp 124	nytimes 56	goal 44
xinhuanet 981	nytimes 90	*wsj 50*	nytimes 43
yahoo news 908	smh 64	smh 43	reuters 42
wash. post(wp) 866	rte 62	google news 37	aljazeera 34

Table 2: Top 10 news agencies from citations

[3]http://wikitimes.l3s.de

[4]http://en.wikipedia.org/wiki/Template_talk:Cite_news/Archive_5#Re-opening_discussion.2C_June_2010

3.3 Annotations in events

Manual annotations of news events are one of the key advantages of WCEP over automatic news summarization tools. Editors typically assign events to categories and news stories and annotate entity mentions with links to their Wikipedia pages. These annotations enrich the news events collection and allow for more structuring and linking of the data. In this study, we take a look into the evolution of this annotation behavior in WCEP and report the results in Figure 5. The figure shows the ratio of annotated events every month. We observe that categories were introduced in 2010 and since then, almost all events are assigned to a category (see Figure 4 for all categories). We also observe that assigning events to story lines (thus forming news timelines) started in 2004 and has been growing since then. For entity annotations, we plot the number of events that have at least 2 entity annotations and observe that there has been an increase of annotated events over the first 9 years, with a slightly dropping trend since 2010.

By tracking the edit changes from WCEP revision logs, we found that 62% of entities were annotated by Wikipedia admins, following by 24% done by regular Wikipedia users and 14% by anonymous editors. Our hypothesis for explaining this slight drop in the number of entity annotations is that this is affected by the slight drop in number of admins in the same period, as can be shown later in Figure 14 in Section 5.

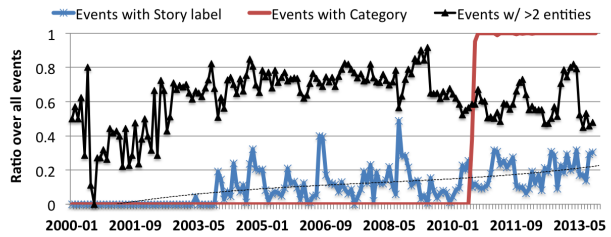

Figure 5: Ratio of events with story label, category and more than 2 entities over all events per month

3.4 Number and duration of news stories

The daily news events in WCEP are often associated to (longer) news stories that span more than one day. By indexing the story lables in WikiTimes, we are able to construct news story timelines, i.e. chronologically ordered lists of (daily) events that belong to the same story. A feature that is currently not available in WCEP. In Figure 6 we show the statistics on the number and duration of news stories (i.e. length in days, months or years). It is worth mentioning that in WCEP sometimes different labels are used for the same story. We detect this by matching the Wikipedia URL of the story page. In addition, we use redirection information of news story pages in Wikipedia to detect "simialr" stories, when applicable. The events of similar stories are then merged in WikiTimes and linked to one story timeline.

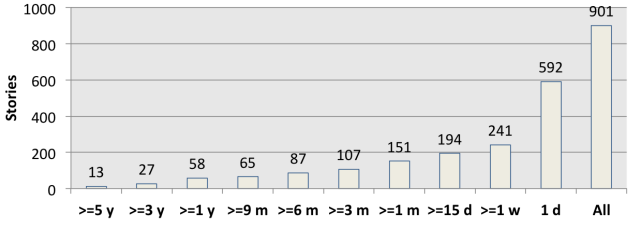

Figure 6: Distribution of storylines by duration

4. COMPARATIVE ANALYSIS

4.1 Comparison with Expert timelines

In this section, we compare the summaries generated by the crowd with those created by professional journalists. In particular, we collected 21 timeline summaries on four example news stories (Egypt Revolution, Syria Civil War, Yemen Crisis and Libyan Civil War) that were published by 25 popular news agencies (such as CNN, BBC, Reuters, New York Times, etc.) between January 2011 and July 2013. Each timeline includes a list of short summaries of relevant dates. We refer to those timelines as the '*Expert*' timelines. We chose those stories because they lasted over a relatively long period and had many news updates. We then extracted all news events from WCEP that belong to matching storylines.

In total we collected a number of 936 events from news agencies and 1940 events from WCEP. We measured the overlap between WCEP timelines and Expert timelines in terms of: dates coverage, text overlap and entity coverage. For each news story, we first measure the mutual overlap between Expert timelines and report the average (red bars in Figure 7). Next, we measure the overlap between the WCEP timeline and each of the Expert timelines and report the average (blue bars in Figure 7).

Date coverage: We observe some variation in the number of dates included in Expert timelines of the same story. Therefore, we selected for each story a set of dates that are mentioned by at least 2 Expert timelines and considered those as the most important time points of the story. We call this set the *relevant* set of dates. The date coverage of a timeline was computed as the ratio of common dates it has to that *relevant* set of dates.

Text overlap: We measure the text overlap using ROUGE score [7]. The ROUGE score between a timeline and another timeline is the average recall of word overlap between the description of the common events in both timelines.

Entity coverage: The entity coverage of a timeline was computed as follows: first, we extracted the list of named entities from all events of an Expert timeline using Illinois Named Entity Tagger [10]. For WCEP timelines, we consider the entity labels given by the event editors. After that, we measured the average overlap between the set of entity tokens of a timeline to another timeline whenever they have a date in common.

The results in Figure 7 show scores with standard deviations of our mentioned metrics. First, we observe that event coverage (i.e. dates) of WCEP is higher than individual Expert timelines. This is probably due to the fact that WCEP editors add events to stories day by day without any space constraints, while journalists might be more selective to keep the timeline summary as compact as possible (e.g. within one page). Second, we observe that event descriptions and entity coverage of WCEP summaries are comparable with Expert summaries, which suggests that the quality of WCEP summaries are comparable to those of professional journalists.

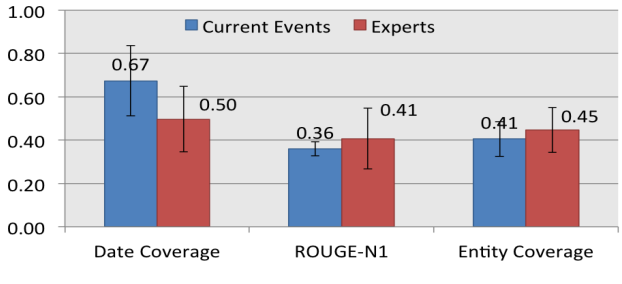

Figure 7: Comparison of WCEP news summaries with summaries provided by news agencies

4.2 Comparison with entity events

Revision logs of Wikipedia articles have been recently shown to be a potential resource for detecting events (e.g., [2]). We compare the list of events found in WCEP with the events that can be extracted by this method. More specifically, we considered a sample of news stories and extracted (from WCEP) the list of events that belong to each story. We then extracted the list of most frequently mentioned entities in (the events list of) each story. Finally, we retrieved the revision logs of the Wikipedia pages of those frequent entities and aggregated their revision statistics.

Figure 8 shows the analysis results of the example story: Libya War 2011. The most frequent entities of this story include: Libya, Gaddafi, Benghazi etc. We collected revision logs of the frequent entities for over 350 days, each day is represented by a point on the x-axis of Figure 8. On the y-axis we plot a normalized value of both: the number of edits in entity pages, and the number of events found in WCEP in each of the those days. We observed a correlation between the two plots indicating peeks on salient periods. That suggests Wikipedia users described more events and edited more often on Wikipedia pages of the top frequent entities when important events happened.

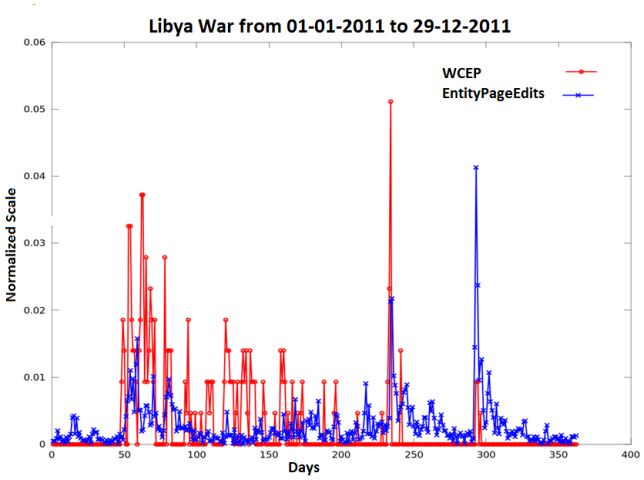

Figure 8: Number of WCEP events in comparison with number of edits from Wikipedia pages of top frequent entities in the example story: Libya War 2011

5. CROWD ACTIVITY ANALYSIS

In this section, we analyze the revision logs of the monthly and daily pages of WCEP in order to understand how big and active is the community behind this portal. We collected the revision data of daily pages using Wikipedia API. However, there was no daily pages in WCEP before June 2006, instead only monthly pages were used. We limit our consideration in this part of the analysis to the period between June 2006 and November 2013.

5.1 Type and duration of revisions

Figure 9 shows the number and type of edits (i.e. adding new content vs. revising existing content) per month. Interestingly we observe that the overall number of edits is stable over the past 7 years, unlike the findings of [6]. Our hypothesis to explain this observation follows. Unlike typical Wikipedia pages, which can be arbitrarily created, the WCEP pages are created in a periodic and fixed basis (i.e. one page per day). Consequently, the growth of the WCEP part in terms of pages is rather stable, while the growth of the rest of Wikipedia is not restricted by similar constraints. Moreover, we observe that most of the edits to a daily page occur in the first few days after creation. Figure 10 shows the total number of edits observed on the page creation day (day 0) and the following days (day 1, 2 and so on). Hence, it is unlikely that a WCEP page is edited far after its creation; a constraint that typical Wikipedia articles (e.g. on persons, organizations etc) do not have.

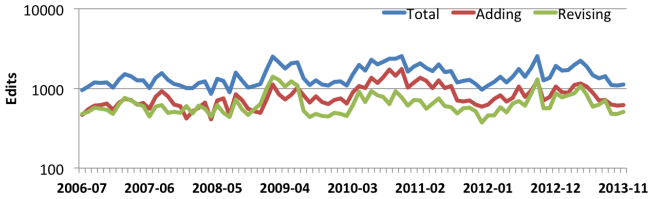

Figure 9: Number and type of edits per month

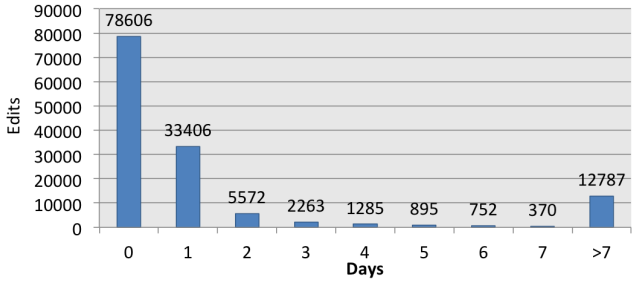

Figure 10: Number of edits since the creation day

5.2 Active vs. inactive editors

We further classify the editors of the WCEP news events by terms of activity level into two categories: active editors and inactive editors. An editor is considered active if her number of contributions within a year is above a certain threshold (*50 edits/year* in our experiments), otherwise, she is classified as inactive.

Figure 11 shows that in total there are about 600-700 editors every month with about %10 of them being active editors and 90% inactive editors, still, Figure 12 shows that most of the edits come from the active editors.

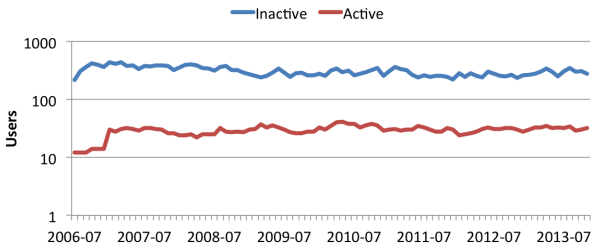

Figure 11: # editors per month by activity level

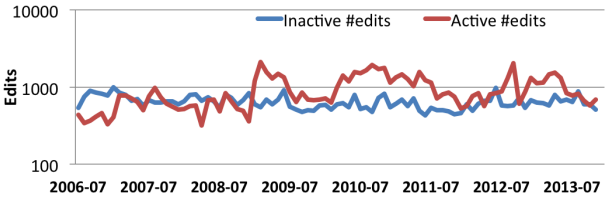

Figure 12: # edits per month by active and inactive editors

Furthermore, by looking at the frequency and duration of editing activities as shown in Figures 13, we observe that there is a core group of "senior" or "loyal" editors, who contribute more often and for longer periods than average editors. The *frequency* dimension on Figure 13 shows the number of editors the contributed at least once per 1, 2, 3, 6, 9 or 12 months, while the *continuity* dimension shows the number of editors that contributed for more than one month, three months, 6 months up to more than 6 years.

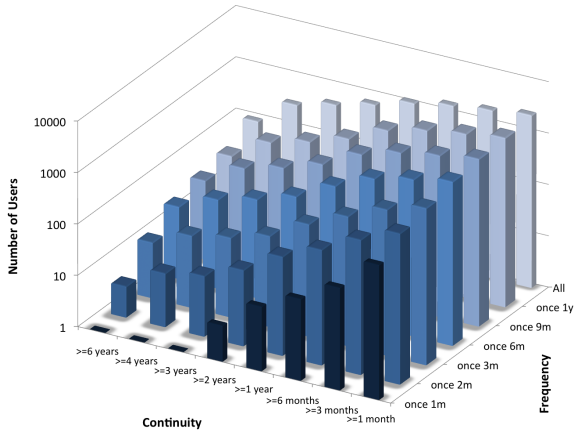

Figure 13: # of editors by frequency and continuity

5.3 Grouping editors by account type

Apart from activity level, we also distinguish between three groups of editors based on their account type. An editor can be either an *admin* (with special access rights), a *wikipedian* (with normal Wikipedia account) or an *anonymous* editor (i.e. without an account). Edits made by the latter group is typically associated with the IP address of the editor.

Figure 14 shows the size of each group over time. We observe that the size of the wikipedian and anonymous groups is almost stable in the past 7 years, while the admins group is shrinking very slowly. On the other hand, the results on Figure 15 show that the the anonymous, wikipedian and admin groups are very dynamic with around 80%, 50% and 20% members leaving or joining every month, respectively (with slightly more leaving than joining editors in the admin group). In this computation, we considered one editor is leaving if after his last contribution, there was no more edit from him, while one is joining if we noticed there was no contribution from him before his first edit.

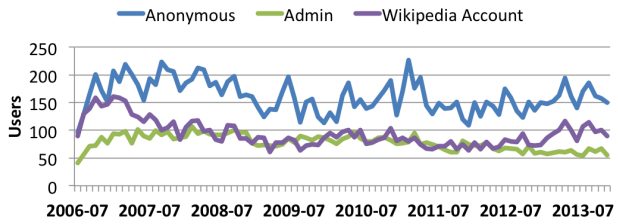

Figure 14: Number of editors by account type

Different than the findings of [6] on general Wikipedia, we found that the average number of edits (Figure 16) made by admins in WCEP is significantly higher than the number of edits made by wikipedians and anonymous editors (although the admins group remains the smallest group in terms of size, and number of new

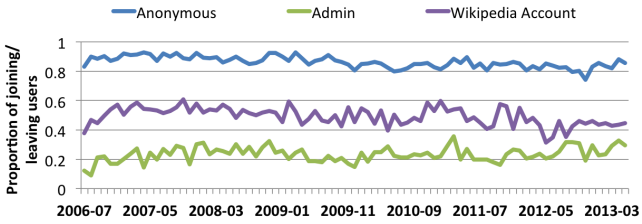

Figure 15: Number of joining/leaving editors by account type

comers). By having a closer look into the type of edits, we observed that admins actually not just revising existing content but also significantly adding more new content (i.e. news events) than other groups as shown in Figures 17 and 18 by number of edits. We spotted the same trend when plotting the number of words added by these groups.

This observation is also confirmed in Figure 19, which shows that the admins have contributed most of the new content in almost all categories (except the conflicts category, where most of the content come form anonymous editors).

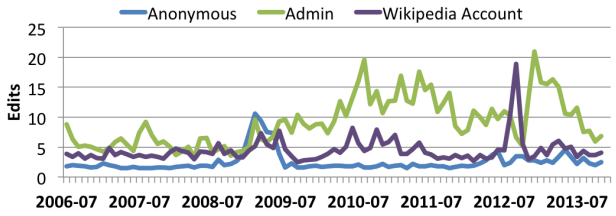

Figure 16: Average number of edits by account type

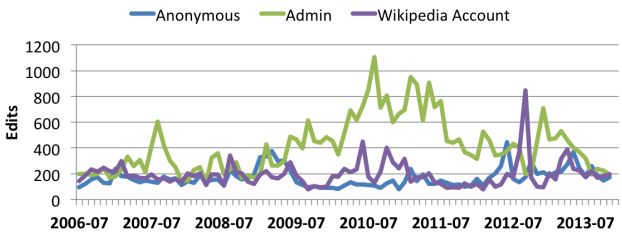

Figure 17: Number of "adding" edits by account type

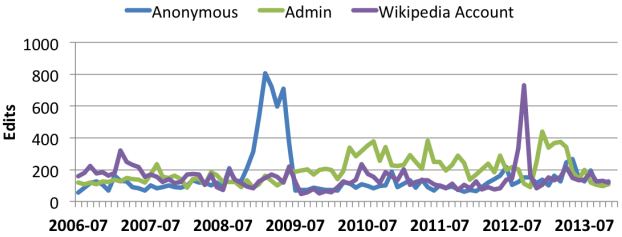

Figure 18: Number of "revising" edits by account type

5.4 Expert Profiling

We investigate the chance if there exists an group of editors who are experts; by experts, we mean the editors who actually focused on contributing in very few categories among others. Having few experts among a big number editors would be a good thing in en-

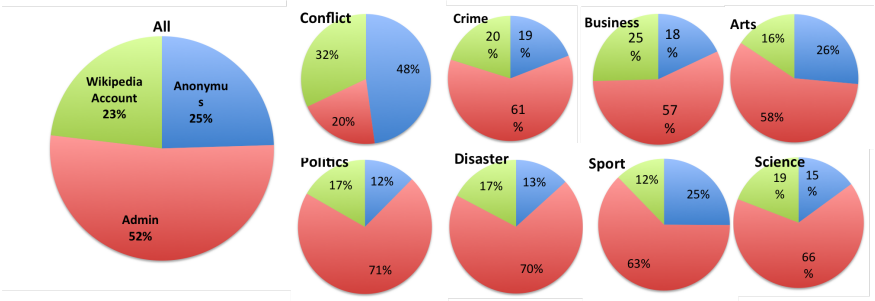

Figure 19: News content contribution of editors by account type

suring quality of WCEP. We set the criteria for an expert as the following: (i) must be an active editor (i.e., at least 50 contributions per year) and (ii) focus on maximum 2 categories of events, which means the proportion of their contributions on each of these domains is from 30%-100% of total contributions she made. The Figure 20 shows the distribution of experts over categories with total 75 experts. In general, the number of experts found is higher in major categories, such as, politics, crime, disaster. However, it is surprising to us that there has been no expert in conflict category. To make it clearer, we took a deep look into this category and found that major of editors are in anonymous users (see also in Figure 19), and neither they are not active editors nor also contributed to other categories with approximately same amount of edits.

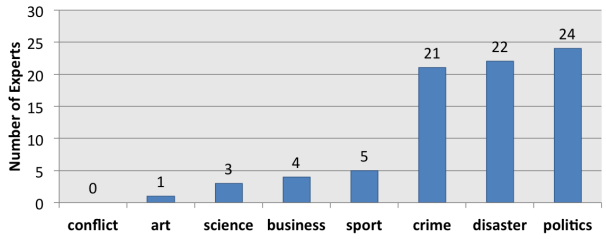

Figure 20: Number of experts over categories

6. RELATED WORK

To the best of our knowledge, there has been no study published on analyzing Wikipedia's Current Events except the work of [3] and [4]. In their initial work [3] the authors indexed all historical events from Wikipedia pages of years since 11 in different languages. More recently, in [4] monthly pages from Current Events were also indexed. However, the two papers focus mainly on describing the extraction and indexing process with little analysis on the collected data.

Our study is also related to research on mining Wikipedia revisions and views history, for example, [9], [1]. However, the work of Kittur et al. (2007) [6] is more related to our current study. Kittur et al. did a comprehensive analysis on Wikipedia users in general. In this work, the authors showed that there is a fall in the influence of admins to Wikipedia while other Wikipedians play major role in contributing to Wikipedia articles. Here, our focus is on the quality of Current Event portal, a special subset of Wikipedia where the content is about (important) daily events. It is worth mentioning that different to Wikipedia in general, WCEP accepts edits from *any* editors without revising procedure.

7. CONCLUSION

In this paper we presented the results of a deep analysis of Wikiepedia's Current Events Portal (WCEP) as a platform for crowdsourced daily summarization of news events. This study we tried to understand the characteristics of its data and community in order to better estimate its quality and reliability. We find that WCEP contains a significant amount of news events data enriched with annotations on categories, related stories and involved entities as well as links to external news articles. We also observe that the content of WCEP is growing and the community behind it is at a stable state with a core group making sure that the platform continues and a bigger more dynamic group of occasional contributors.

Our findings show that WCEP can be an invaluable resource both for public and research community. We also presented WikiTimes, our system for automatic extraction and indexing of news events from WCEP into a MySQL database, Lucene index and provides access with GUI and Restful interfaces for querying the index.

8. REFERENCES

[1] M. Ciglan and K. Norvag. Wikipop: personalized event detection system based on wikipedia page view statistics. In *CIKM 2010*, pages 1931–1932. ACM.

[2] M. Georgescu, N. Kanhabua, D. Krause, W. Nejdl, and S. Siersdorfer. Extracting event-related information from article updates in wikipedia. In *ECIR*, pages 254–266, 2013.

[3] D. Hienert and F. Luciano. Extraction of historical events from wikipedia. *CoRR*, abs/1205.4138, 2012.

[4] D. Hienert, D. Wegener, and H. Paulheim. Automatic classification and relationship extraction for multi-lingual and multi-granular events from wikipedia. In *Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2012)*, volume 902 of *CEUR-WS*, pages 1–10, 2012.

[5] B. Keegan, D. Gergle, and N. Contractor. Staying in the loop: Structure and dynamics of wikipedia's breaking news collaborations. WikiSym '12, pages 1:1–1:10, New York, NY, USA, 2012. ACM.

[6] 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. *CHI*, 1(2):19, 2007.

[7] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of NAACL'03 - Volume 1*, pages 71–78, 2003.

[8] S. Niederer and J. van Dijck. Wisdom of the crowd or technicity of content? wikipedia as a sociotechnical system. *New Media and Society*, 12(8):1368–1387, 2010.

[9] S. Nunes, C. Ribeiro, and G. David. Wikichanges: Exposing wikipedia revision activity. In *Proceedings of the 4th WikiSym*, WikiSym '08, pages 25:1–25:4. ACM, 2008.

[10] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the CoNLL '09*, pages 147–155, 2009.