Temporal Summarization of Event-Related Updates in Wikipedia

Mihai Georgescu, Dang Duc Pham, Nattiya Kanhabua
Sergej Zerr, Stefan Siersdorfer, Wolfgang Nejdl

L3S Research Center / University of Hannover, Germany
{georgescu, pham, kanhabua, zerr, siersdorfer, nejdl}@L3S.de

ABSTRACT

Wikipedia is a free multilingual online encyclopedia covering a wide range of general and specific knowledge. Its content is continuously maintained up-to-date and extended by a supporting community. In many cases, real-world events influence the collaborative editing of Wikipedia articles about the corresponding entities. In this paper, we present the *Wikipedia Event Reporter*, a web-based system that supports the entity-centric, temporal analytics of event-related information in Wikipedia by analyzing the whole history of article updates. For a given entity, the system first identifies peeks of update activities for the entity using bursts detection and automatically extracts event-related updates using a machine-learning approach. Further, the system summarizes event-related information through clustering of updates by exploiting different types of information such as update time, textual similarity, and the position of edits within an article. Finally, the system generates a meaningful temporal summarization of event-related updates and automatically annotates identified events in a timeline.

Categories and Subject Descriptors

Information systems [**Information systems applications**]: [Data Mining, Data stream mining]

General Terms

Algorithms, Design

Keywords

Wikipedia Updates, Event Detection, Entity Timeline, Temporal Summarization

1. INTRODUCTION

Wikipedia is a multilingual, web-based, free-content encyclopedia project based on an openly editable model. It is the most up-to-date encyclopedia available. Consequently, as new events take place all over the world, Wikipedia users will update the articles corresponding to the entities involved in these events, or influenced by them, causing an avalanche of edits on several articles, as more information regarding the event becomes available. Because Wikipedia has well defined standards for the quality of its articles, and it strives to provide a point of view as neutral as possible, in the course of the event, through a mutual agreement process the different points of view converge to a view accepted by the community of contributors.

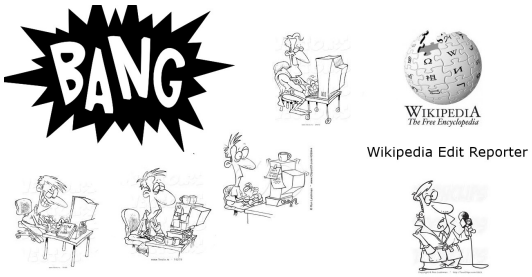

Figure 1: Events trigger the interest of Wikipedia users that update the pages of entities affected or involved. Our system uses this information for extracting and presenting events.

As an encyclopedia, Wikipedia covers much of what is of importance for a general reader. As it does not have periodical releases, and is a live encyclopedia, Wikipedia has to be constantly updated when events happen. The large community of users is always mobilized and takes care of keeping the accuracy and actuality of the information. To make sense of all the collaborative editing involved, all revisions of an article are kept, in an edit history. We use this as our news source.

The idea behind our system is visually explained in Figure 1. As an event happens, the Wikipedia community mobilizes itself to update the encyclopedia. Some information generated in a particular time period will no longer be available in a future version of the articles of the entities involved in the event. Thus, our main objective is to provide users (e.g., a journalist or a student studying about a history), an ability to visualize historical information, giving a comprehensive view of an event, and not only the socially accepted final interpretation of the event.

In this demonstration paper, we present *Wikipedia Event Reporter*, a system that automatically extracts events from the history of updates in Wikipedia and presents related information summary to the user in a meaningful way. To

WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.

this end, we first identify peeks of update activities for a specific entity using bursts detection and in the next step extract event-related updates using a classifier trained with manually labeled data. Further, we summarize event-related information through clustering of updates by exploiting different types of information such as update time, textual similarity, and the position of edits within an article. Finally, we visualize the obtained clusters as a colored histogram and present related sentences as well as affected section titles in form of a table.

The rest of this paper is organized as follows. In Section 2, we will present the description of dataset used in this demo as well as explain the event extraction methods underlying our system. In Section 3, we outline the system interface, and describe the proposed demonstration plan in Section 4. Finally, we present related work in Section 5 and conclude the paper in Section 6.

2. EVENT EXTRACTION METHODS

Our dataset used in this paper is the dump of the whole Wikipedia history (version from 30 January 2010). The history dump contains more than 300 million updates with the size of approximately 5.8 TB covering the time period between 21 January 2001 and 30 January 2010. We discarded updates made by *anonymous* users, resulting in a dataset containing 237 million updates belonging to 19 million articles. To store the revisions in a way that they are easily accessible for processing and information extraction we used the Wikipedia Revision Toolkit [5].

An *update* in Wikipedia represents the modifications present in one revision when compared to the previous revision of an article. It is accompanied by its creation time (timestamp), its author, and, possibly, comments provided by the updater. For a given update, we further consider the blocks of text added and removed, the title of the section where the modification occurred, and the relative and absolute positions of the blocks in their sections and in the article.

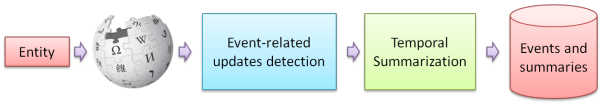

Figure 2: Pipeline for identifying and presenting the events related to an entity.

We extract event-related information from Wikipedia edits for a given entity and its corresponding article as follows. We first identify event-related updates, and in a second step we generate a temporal summarization by mapping the identified updates to their corresponding events and provide meaningful summaries. The pipeline for this process is depicted in Figure 2. In the following, we present the methods used for event-related update detection and summarization. The detailed description and experimental evaluation of the methods can be found in our previous work [6].

2.1 Detection of Event-Related Updates

For detecting event-related updates we make use of a combination of filters and classifiers based on burst detection, temporal information, and textual content.

Burst Detection Filter: Bursts of updates (peaks in the update activity) in a Wikipedia article are indicators for periods with an increased level of attention from the community of contributors. In order to detect bursts, we apply a simplified version of the burst detection algorithm presented in [10] on the temporal development of the update frequency of an article. The algorithm employs a sliding time window for which the number of updates is counted. The corresponding time intervals for which the update rate exceeds a certain threshold are considered *bursty*; our burst detection filter extracts the updates within those bursty periods.

Text Classification: Language and terms used in the update text can serve as an indicator whether an update is related to an event. We trained Support Vectors Machine classifiers [3] on manually labeled samples to distinguish between "event-related" and "not event-related" updates. To represent an update we constructed bag-of-word based tf*idf feature vectors (using stemming and stop word elimination) using the terms added in an update, terms removed, and terms from comments. As training data we used 10,680 article updates labeled, of which 2,616 as "event-related" and 8,064 as "not event-related".

2.2 Temporal Summarization of Events

The stream of event-related updates determined in the previous step serves as a starting point for identifying the events themselves and creating a meaningful summarization. In order to present event-related information in a understandable way, instead of using the detected event-related updates for summarization, we use the updated sentences. To this end, we start by identifying the sentences where the event-related updates were done, and assign to them a *weight*, corresponding to the number of times they were updated, and a list of positions at which the sentences appeared within the Wikipedia articles.

Temporal Clustering: In order to identify the distinct events, we first resort to a temporal clustering by identifying the bursts among the event-related updates. Each burst of event-related updates corresponds to a distinct event.

Text-Based Clustering: Within a burst of updates, in order to eliminate the duplicate sentences and group together the sentences treating the same topic we employ an incremental clustering based on the Jaccard similarity as a distance measure. Each *sentence cluster* is characterized by the aggregated *weight* of member sentences, and represented by the longest member sentence, that serves as a candidate for summarization.

Position-Based Clustering: Assuming that sentences on the same topic are located in spatial proximity of each other on the article page, by investigating the positions of all sentences modified in a burst we can identify *position clusters*. A cluster of positions is a contiguous succession of positions with no more than 10 positions gap in between; each cluster is assigned with a different color. Each sentence cluster belongs to the position cluster that has the maximum overlap of positions with member sentences.

Summarizing Events: Each detected event, corresponding to a burst of updates, is summarized using a ranked list of sentences. We rank the position clusters by how many sentence clusters are assigned to them and the sentence clusters by the aggregated *weight* of their member sentences. The proposed summarization for an individual event consists of displaying for each of the top-N position clusters, the representative sentences for the top-M clusters of sentences.

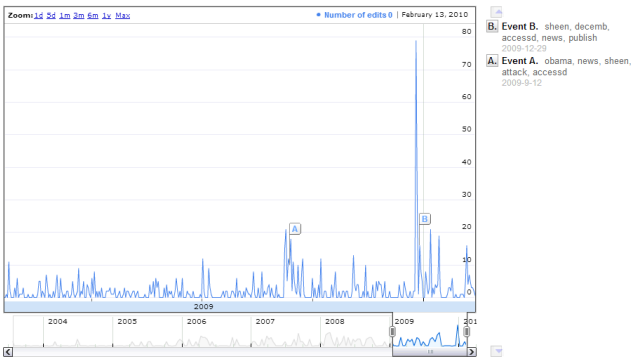

Figure 3: Entity timeline for Charlie Sheen annotated with the identified events.

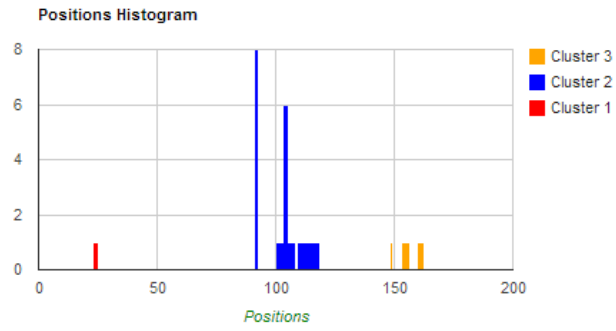

Figure 4: Histogram illustrating the positions where the event-related edits occurred.

3. SYSTEM INTERFACE

The information that we present when the user issues a query about one specific entity consists of a timeline of the updates, annotated with the detected events, a histogram depicting the positions of the edited sentences, and a list of sentences that characterize the event.

In Figure 3 we present an example for the timeline of the number of edits per day. Each of the peaks might be a candidate for an event, but because we employ a classifier to detect the event-related edits, only some of the peaks are actually be related to events. The detected events are marked on the timeline with the letter assigned to them, and are accompanied by a tag cloud description.

When clicking on the event from the timeline, the corresponding summary can be displayed for different time granularities: the *whole event history* and for each individual *day* that is a part of the event. All updates that were detected as being event related that belong to a common burst, characterize the same event. We cannot offer an intelligible description by using the updates themselves, because they are often just words or parts of sentence. Therefore, we use the sentences that were modified by the updates for summarization.

The histogram called *positions histogram* presented in Figure 4 represents the positions of all edited sentences that belong to the same event. The histogram is annotated using different colors for the identified positions clusters.

Because there most of the sentences are similar, they are clustered together in *sentence clusters*. As it can be seen in Figure 5, we rank the clusters based on their weight, and display the top 10 clusters presenting: the weight of the cluster, the representative sentence, the section name where most of the edits were made, and the positions cluster assignment. The color and number of the positions cluster assignment match to the positions histogram displayed above. When hovering over the positions cluster assignment, the user can see a histogram of all the positions of the sentences that are a part of the cluster. The positions clusters colors are easily identifiable to facilitate the understanding of the positions cluster assignment.

4. DEMONSTRATION OVERVIEW

In this demonstration, we will show how to generate and visualize temporal event summarization using our *Wikpedia Event Reporter* system. As one of the examples for an entity of interest, we exploit Wikipedia entry about Charlie Sheen's life and career. First, a user enters Charlie Sheen as an input query for the system. In this example, the system will display the timeline annotated with the events **A** and **B** on the dates *12 September 2009* and *29 December 2009* respectively as illustrated in Figure 3.

By checking the date of the detected Event **A** we find that it matches the following text in the current Wikipedia article of Charlie Sheen: *On September 8, 2009, he appealed to President Barack Obama to set up a new investigation into the attacks. Presenting his views as a transcript of a fictional encounter with Obama, he was characterized by the press as believing the 9/11 Commission was a whitewash and that the administration of former President George W. Bush may have been responsible for the attacks.* Figure 5 represents the temporal summary for Event A provided by the system, and Figure 4 depicts the corresponding positions histogram. It can be noticed that the generated summary that the system provides matches the current Wikipedia article. In addition our system presents links and content that are no longer available in the current version because as an encyclopedia, Wikipedia has to keep just the relevant, good quality content. This allows the user to discover details that were removed for the sake of brevity in the current version of Wikipedia such as: *Days before the eight anniversary of the 9/11 attacks, Sheen publicly requested a meeting with President Obama to discuss a list of 20 questions he had about the September 11th attacks which he says remain unanswered and is demanding an investigation into the attacks be reopened.*

For Event **B** about a domestic dispute, *On December 25, 2009, Sheen was arrested for assaulting his wife, Brooke Mueller in Aspen, Colorado*, most of the information we display is no longer available in Wikipedia: *Law enforcement sources cited by TMZ.com said Mueller initially told 911 dispatchers Sheen had assaulted her, alleging Sheen put a knife to her throat and made threats to kill. Mueller had a blood alcohol level of 0.13 that night (over the legal limit for driving), Sheen's BAC was 0.04 (well under the limit).* None of the links to the articles describing the incident are available in the current Wikipedia version but are discovered by our system. In this case the system uncovers details about an event that were not deemed as worthy of being kept in Wikipedia, but might be of interest to someone studying the entity in detail, that is not satisfied just with what is pro-

Weight	Sentence Representative	Section Title	Positions Cluster
11	= On March 20, 2006, Sheen stated during an Alex Jones (radio) Alex Jones interview that he questions the official story concerning the September 11 attacks of 2001.	September 11 attacks	2
6	On September 8, 2009, Sheen appealed to US President Barack Obama Obama to set up a new investigation into the attacks.	September 11 attacks	2
5	Presenting his views as a transcript of a fictional encounter with Obama, he characterized the 9/11 commission as "a whitewash" and alleged that the administration of former US President Bush was responsible for the attacks.	September 11 attacks	2
2	{{cite news last=Banerjee first=Subhajit journal=Daily Telegraph title=Charlie Sheen urges Barack Obama to reopen 9/11 investigation in video message date=September 12, 2009 url=http://www.telegraph.co.uk/news/newstopics/celebritynews/6177194/Charlie-Sheen-urges-Barack-Obama-to-reopen-911-investigation-in-video-message.html accessdate=September 13, 2009}}	September 11 attacks	2
1	Charlie Sheen has since become a prominent advocate of the 9/11 Truth movement .	September 11 attacks	2
1	Days before the eight anniversary of the 9/11 attacks, Sheen publicly requested a meeting with President Obama to discuss a list of 20 questions he had about the September 11th attacks which he says remain unanswered and is demanding an investigation into the attacks be reopened.	Charitable and political activities	3
1	Sheen stated that a friend of his died from breast cancer and he wanted to try to help find a cure for the disease.	September 11, 2001	3
1	ref name="CNN Showbiz March"	September 11 attacks	2
1	{{cite journal last1=Keating first1=Joshua last2=Downie first2=James title=The World's Most Persistent Conspiracy Theories journal=Foreign Policy date=September 10, 2009 url=http://www.foreignpolicy.com/articles/2009/09/10/the_worlds_most_popular_conspiracy_theories accessdate=September 13, 2009}}	September 11 attacks	2
1	{{cite news url=http://sports.espn.go.com/espn/page3/story?page=sheen/merron title=How Good Was Charlie Sheen? last=Merron first=Jeff date=2004-02-19 work=Page 3 publisher=ESPN accessdate=2009-03-21}}	September 11 attacks	1

Figure 5: Temporal summarization of the identified events for a given entity.

vided by the last version of the article, or does not want to spend too much effort searching the Web.

An online system and a video tutorial are published at `http://www.l3s.de/wiki-events`. More instructions on all the available summarization tools are provided in the Man page More instructions on all the available summarization tools are available in the *Man* page and a short description of the processes that take place in the background and the tools used can be found in the *Behind the Scenes* page of the online system.

5. RELATED WORK

Previous work has focused on detecting events from unstructured text like news, using features such as key words or named entities [1, 7, 9]. In this work, we employ Wikipedia article updates for event detection instead of using traditional news streams by leveraging a *Wisdom of the crowd* type of effect, instead of coming from a core group of users as in its early days.

Ciglan and Nørvåg [2] proposed to detect events by analyzing trends in page view statistics. In their recent work, Keegan et al. [8] studies the temporal dynamics of editorial patterns of news events using structural analysis, while Ferron and Massa [4] proposed different representations of events related to disasters by analyzing language usage.

To the best of our knowledge, we are the first to present an entity-centric, temporal analytics system for supporting a temporal analysis of event-related information in Wikipedia article updates.

6. CONCLUSIONS

We presented *Wikipedia Event Reporter*, a web-based system for generating temporal summarization of real-world events such as political conflicts, natural catastrophes, and new scientific findings that are mirrored by article updates in Wikipedia. Our system helps users explore the temporal development of events for entities of interest, by presenting an annotated timeline and a concise summarization. Moreover, it is able to find historical information about events that are no longer available in the current version of Wikipedia, giving the user a comprehensive view of the event, and not only the socially accepted final interpretation of the event and its implications. We demonstrated that our system is capable of automatic extracting and generating temporal summarization of events from Wikipedia updates enhancing real-world applications, such as, entity-specific, annotated timelines and news tickers.

Acknowledgments This work was partially funded by the European Commission FP7 under grant agreements No. 287704 and No. 600826 for the CUBRIK and ForgetIT projects respectively.

7. REFERENCES

[1] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of SIGIR '98*, 1998.

[2] M. Ciglan and K. Nørvåg. WikiPop: personalized event detection system based on Wikipedia page view statistics. In *Proceedings of CIKM '10*, 2010.

[3] C. Cortes and V. Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995.

[4] M. Ferron and P. Massa. Psychological processes underlying wikipedia representations of natural and manmade disasters. In *Proceedings of WikiSym '12*, 2012.

[5] O. Ferschke, T. Zesch, and I. Gurevych. Wikipedia revision toolkit: Efficiently accessing wikipedia's edit history. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. System Demonstrations*, pages 97–102, Portland, OR, USA, Jun 2011.

[6] M. Georgescu, N. Kanhabua, D. Krause, W. Nejdl, and S. Siersdorfer. Extracting event-related information from article updates in wikipedia. In *Proceedings of ECIR '13*, 2013.

[7] Q. He, K. Chang, and E.-P. Lim. Analyzing feature trajectories for event detection. In *Proceedings of SIGIR '07*, 2007.

[8] B. Keegan, D. Gergle, and N. Contractor. Staying in the loop: Structure and dynamics of wikipedia's breaking news collaborations. In *Proceedings of WikiSym '12*, 2012.

[9] Z. Li, B. Wang, M. Li, and W.-Y. Ma. A probabilistic model for retrospective news event detection. In *Proceedings of SIGIR '05*, 2005.

[10] Y. Zhu and D. Shasha. Efficient elastic burst detection in data streams. In *Proceedings of KDD '03*, 2003.