Exploring Long Running News Stories using Wikipedia

Jaspreet Singh[†], Abhijit Anand[†], Vinay Setty[*], Avishek Anand[†]

[†]Forschungszentrum L3S Hannover, Germany
[*]Max-Planck-Institut für Informatik Saarbrücken, Germany
{singh,aanand}@l3s.de, vsetty@mpi-inf.de, anand@l3s.de

ABSTRACT

A significant portion of today's news articles are part of long running stories. To better understand the context of these stories journalists, social scientists and other scholars use news collections to find temporal and topical insights. However these insights are devoid of *user impressions*, derived from click-through data and query logs, and are only reliable if the collection is complete and consistent. In this work we introduce the notion of combining user impressions from Wikipedia with news collection based insights for long running news story exploration and outline promising new research directions. We also demonstrate our initial attempts with a prototype system called NEWSEX.

Categories and Subject Descriptors

H.3 [**Information Search and Retrieval**]: Search Process; H.4.3 [**Information Systems Applications**]: Information browsers

General Terms

Algorithms

Keywords

News, Exploration, Wikipedia

1. INTRODUCTION

News articles published today can be broadly classified into standalone pieces or contributions to long running news stories like "Syrian War" for instance. Articles from the latter unsurprisingly often require contextual information to fully comprehend. Upon analyzing Wikipedia's current events portal[1] we found that nearly 30% of recent news was linked to pages about long running stories (we consider pages as "long running" if edit activity spans 30 days or more). These pages are structured into various sections and subsections that describe the background and events relevant to the story. Each section is then typically linked to another page that gives more details on associated events. Wikipedia pages serve as a good starting point for exploration but (a) users need to read them in full to get a better understanding of the news story (b) such pages do not immediately indicate important time periods, important topics and entities and (c) do not provide relevant news articles to contextualize the content. To overcome these drawbacks, special user groups like historians and journalists turn to news archive exploration systems.

Given keywords describing the news story as input, news exploration systems return a ranked list of news articles augmented with exploratory aids like timelines and topic filters. These aids visualize insights to help users quickly understand the results. Timelines are useful for identifying important time periods by aggregating documents based on temporal features such as publication date and temporal expressions. Topical insights based on linguistic features like the most frequently occurring terms or entities help provide more context. Previous approaches like [2, 1, 5, 6] have relied solely on the underlying collection to rank documents and produce insights. Relying solely on the collection however has the following drawbacks: (a) insights, especially those based on aggregates, are only reliable if the collection is not missing a significant portion of documents; (b) a lack of *user impressions*.

"User impressions" is the term we use to refer to recorded user activity such as query logs and click-through data. Web search engines leverage user impressions to vastly improve retrieval but search engines designed for news archives are not privy to a large user base needed to generate vast query logs and clicks. However the collective impressions of users regarding popular topics and events resides indirectly in Wikipedia's rich edit history, authorship patterns and link structure. Recently, with a stricter editor policy and emphasis on citations, the reliability of popular pages like the ones related to major news stories has improved [4]. Wikipedia editors create, modify or redact news story articles whenever a major incident occurs or to reflect changing opinions in a timely manner - all of which is recorded in the edit history, including the edited text. The textual content of the page is also a rich source of impressions since it is constructed and organized collectively. Most notably, important aspects and events are given separate sections, certain sections and entities are linked to other pages providing more detail and pages are classified according to a defined taxonomy enabling us to determine domain context.

To address the above issues we introduce the following problem: Given a long running news story specified by its Wikipedia page and an external news collection as input, how to (i) rank relevant news articles (Section 2.1) (ii) identify important time periods (Section 2.2) and (iii) identify important topics and events (Section 2.2), by leveraging Wikipedia's construction dynamics in conjunction with collection based insights. In this paper we present our first attempts to tackle these challenges in a prototype system called NEWSEX (interface shown in Figure 1) and elaborate on the research challenges posed.

[1]http://en.wikipedia.org/wiki/Portal:Current_events

WebSci '15, June 28 - July 01, 2015, Oxford, United Kingdom
 ISBN 978-1-4503-3672-7/15/06…$15.00
DOI: http://dx.doi.org/10.1145/2786451.2786489

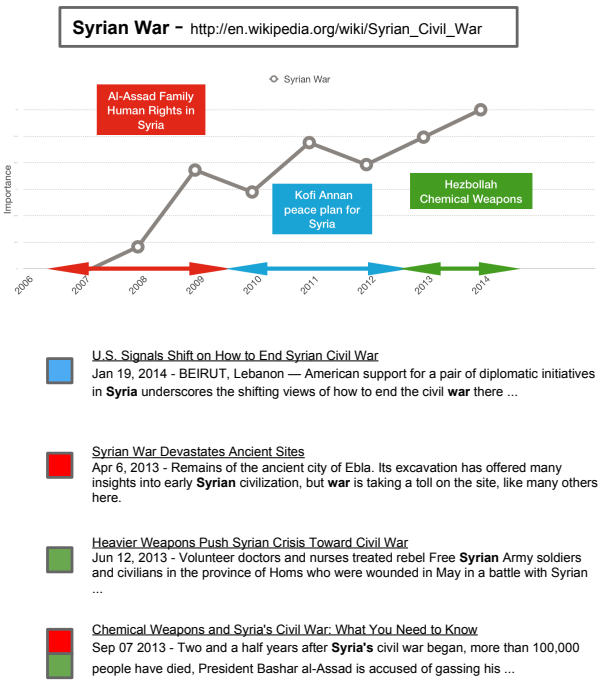

Figure 1: The NEWSEX prototype displaying an annotated timeline and list of results for the news story "Syrian Civil War"

2. METHODOLOGY

In this section we detail our methodology used to generate the elements shown in Figure 1 and highlight important research problems when exploring long running news stories. NEWSEX's interface contains the primitives common to most news exploration systems: temporal insights plotted on a timeline, topical insights and a list of news articles providing context. The timeline can be used to respecify (broaden or focus) the intent with a qualifying time period. The topical insights, visualized as timeline annotations in our interface, can be used to filter documents. Below the timeline is a list of results ranked by relevance to the news story and the period of interest. To make news story selection easier we first show the user a list of current events that have links to corresponding Wikipedia news story pages.

2.1 Ranking News Articles

The first question we pose for long running news story exploration is how do we identify and rank relevant news articles from the document collection using Wikipedia? In our implementation we rank news articles by computing the unigram language model score between the document and the title of the the news story (the query) in the Wikipedia page. We believe that more relevant results will arise when using more principled approaches that combine features extracted from Wikipedia's article structure, infobox, entity annotations, main article links and edit history. In addition to these features one can also consider various temporal features such as temporal expressions, publications date and document focus time due to the longitudinal nature of news corpora. When a time period is specified, the retrieval model can either filter out documents not published during that time or use new terms for query expansion. These new terms can stem from the edited text in that period or from other Wikipedia articles more relevant in that time period.

2.2 Timeline Construction and Annotation

The timeline of a long running news story should highlight important intervals and major events with annotations that best describe them. We construct a timeline for new story exploration that incorporates both: user impressions and collection statistics. In our prototype we first build the temporal profile of the query akin to [3] and then smooth the temporal profile with the fraction of edits made to the input Wikipedia page using Dirichlet smoothing. The temporal profile scores are then plotted to create the timeline in the interface. Finally, we annotate important time periods with the title of the subsection where the most edits were made. This approach however does not take into account pages related to the input article which can improve accuracy and give more context. Furthermore, to effectively estimate the temporal profile of a long running news story we can also consider the edit history of related pages, the temporal expressions found in both Wikipedia pages and the result set as well as dates mentioned in infoboxes and subsection headings. Similarly for more contextually relevant topical annotations we can exploit the edited text, subsection headings, links to other pages added during the time and key phrases mined from the news articles.

3. SYSTEM & CONCLUSION

NEWSEX utilizes the current events portal from Wikipedia to suggest input stories to the user. Once users select a story, they are presented with the exploration interface consisting of an annotated timeline and results from a specified document collection. The underlying document collection we use is the GDELT dataset[2] consisting of nearly 7 million news articles crawled from April 2013 to April 2014. For the Wikipedia based insights we use a Wikipedia dump from August 2014.

In conclusion, the intention of this paper is to introduce a novel framework for exploring long running news stories using user impressions from Wikipedia and an external news collection. To showcase our preliminary attempts to solve these problems, we built a prototype called NEWSEX. Further research is needed to find more sophisticated retrieval models, better insight mining techniques and robust evaluation procedures for such exploration systems. We believe that the framework described in this paper is important for news archive exploration by journalists, social scientists and various other special user groups.

4. ACKNOWLEDGMENT

This work is funded by the European Research Council under ALEXANDRIA (ERC 339233).

5. REFERENCES

[1] O. Alonso, K. Berberich, S. Bedathur, and G. Weikum. Neat: News exploration along time. Proc. of ECIR'2010.

[2] O. Alonso, M. Gertz, and R. Baeza-Yates. Clustering and exploring search results using timeline constructions. In *Proc. of the 18th ACM CIKM*, 2009.

[3] R. Jones and F. Diaz. Temporal profiles of queries. *ACM Transactions on Information Systems*, 2007.

[4] A. Lih. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. *Nature*, 2004.

[5] M. Matthews, P. Tolchinsky, R. Blanco, J. Atserias, P. Mika, and H. Zaragoza. Searching through time in the new york times. In *Proc. of the 4th Workshop on HCIIR*. Citeseer, 2010.

[6] V. Setty, S. Bedathur, K. Berberich, and G. Weikum. Inzeit: efficiently identifying insightful time points. *Proc. of the VLDB Endowment*, 2010.

[2]http://gdeltproject.org/data.html