Leveraging User-Generated Content for News Search

Richard M. C. McCreadie
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ
richardm@dcs.gla.ac.uk

Categories and Subject Descriptors
H.3.3 [**Information Storage & Retrieval**]: Information Search & Retrieval

General Terms - Experimentation

Keywords - News, Blogs, Social Media

ABSTRACT

Over the last few years both availability and accessibility of current news stories on the Web have dramatically improved [3]. In particular, users can now access news from a variety of sources hosted on the Web, from newswire presences such as the New York Times, to integrated news search within Web search engines. However, of central interest is the emerging impact that user-generated content (UGC) is having on this online news landscape. Indeed, the emergence of Web 2.0 has turned a static news consumer base into a dynamic news machine, where news stories are summarised and commented upon. In summary, value is being added to each news story in terms of additional content.

Importantly, however, while there has been movement in commercial circles to exploit this extra value to enrich online news [5], there has been little research from the academic community on how can be achieved. Indeed, the main purpose of this thesis is to research practical techniques for the integration of UGC to improve the news search component of the most ubiquitous of Web tools, i.e the Web search engine. Importantly, we identify the following three key aspects of news search which might be improved through the application of UGC.

Intuitively, the first task that the news vertical search aspect of a Web search engine needs to accomplish when confronted with a user query is to decide whether the query is in fact news-related, and hence requires news content to be included. However, queries themselves are sparse in nature, being often comprised of one of two tokens only. This presents issues when performing query classification, as there are few features to distinguish the news related queries. We attest that UGC can help alleviate this ambiguity. Indeed, we hypothesise that there is a strong link between the volume of UGC content being posted mentioning a query and the likelihood of that query being news-related within a specific timeframe.

Secondly, we consider the task of real-time event detection. It is imperative for search engines to maintain knowledge of the events of the moment, such that the results displayed are updated. Traditionally, systems have detected new events through the clustering of newswire articles [1]. However, in the current fast-paced news search environment where users begin querying for events within a couple of minutes of their occurrence [4], relying on slow newswire reporting is unacceptable. On the other-hand, UGC sources such as Twitter provide a natural alternative, as the high post rate and popularity of news topics makes a site such as this an ideal medium from which to monitor emerging events. Indeed, many paid journalists maintain personal blogs and other social media accounts for the reporting of fast-breaking news stories [2].

Lastly, we examine the presentation of results to the user. The presentation of news articles to satisfy news-searches is generally accepted. However, with the ever-increasing pace of news reporting world-wide, there is now no guarantee that a trusted news source will have yet published upon the story. In these cases, one must look else-where for content to satisfy the user. We hypothesise that UGC is ideal for presentation in these cases as the delay between an event occurring and commentary appearing in UGC sources like Twitter or the Blogosphere is mear minutes. Moreover some information needs cannot be easily solved using newswire articles alone. For example, the correct result for the query 'current news' would be a list of news stories ranked by their importance for the day in question. This is a difficult ranking problem, as 'importance' is greatly dependent upon the perspective of the user. In this case, one solution might be to leverage 'public opinion' as represented in UGC, for example by taking 'the pulse of the Blogosphere'. Indeed, we have examined such during TREC 2009.

In conclusion, we have identified multiple areas of the news-search process which cannot be satisfied by traditional newswire articles. We hypothesise that the application of user-generated content can be leveraged to improve the field of news-search in relation to the rich and timely information that UGC provides.

1. REFERENCES

[1] James Allan. *Topic detection and tracking*. Springer, 2002.

[2] D. Matheson. Weblogs and the epistemology of the news: some trends in online journalism. *New Media and Society*, 6(4):443–468, 2004.

[3] Newspaper Association of America (NAA). Newspaper Web sites attract more than 70 million visitors in June; over one-third of all Internet users visit newspaper Web sites, 2010. `http://www.naa.org/PressCenter/SearchPressReleases/2009/NEWSPAPER-WEB-SITES-ATTRACT-MORE-THAN-70-MILLION-VISITORS.aspx`, accessed on 25/01/2010.

[4] J. Pedersen. Keynote speech. In the *Third Annual Workshop on Search in Social Media*, 2010.

[5] Amit Singhal. Relevance meets the real-time Web, 2010. `http://googleblog.blogspot.com/2009/12/relevance-meets-real-time-web.html`, accessed on 25/01/2010.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.