Insights on the Horizon of News Search

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

Craig Macdonald
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ
craigm@dcs.gla.ac.uk

Iadh Ounis
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ
ounis@dcs.gla.ac.uk

ABSTRACT

In recent years, news reporting and consumption has made the profound shift from paper-based media to free online publications, while the simultaneous emergence of Web 2.0 has fundamentally changed the way we react to news. In this paper, we argue that the rapid increase in volume of user-generated content now available presents new and exciting opportunities for the furtherment of news search. In particular, we discuss new applications for user-generated content when determining the stories of the moment, as well as adding value to the results returned.

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

General Terms: Ranking, Information Retrieval, Social networks

Keywords: News, Search, User Generated Content

1. INTRODUCTION

Over the last few years, newswire companies have been diversifying their content distribution onto the Internet. Indeed, although traditional (physical) news circulations are falling, the 'e-readership' of news articles has greatly increased [4]. This shift has motivated the need for real-time news search, satisfying the users need to find and access this content in a simple, convenient and familiar manner.

Thus far, news search has been satisfied through the integration of news article aggregation technologies into traditional search engines. For example, Google integrates stories from Google News[1] into their search ranking, while Microsoft's search engine Bing[2] provides a specific tab for the presentation of news related results to the user query.

However, while these systems are often effective, there are notable cases where they can fail the user. For example, when searching for a 'breaking' news story, i.e. a story which covers an event which has just occurred or is unfolding in real-time, the story might not be returned due to insufficient newswire articles having yet been published. Furthermore, when searching for a controversial story, the user may wish to see multiple viewpoints. However, search engines are limited by the newswire sources which they monitor, tending to favour a few highly trusted sources, which may share similar (neutral) viewpoints.

In this paper, we argue that the leverage of user-generated content can be applied to address issues such as these. Hence, in combination with the news article aggregation technologies currently employed, superior overall performance can be achieved, and more users can therefore be satisfied. In the next section, we examine why user-generated content might be a useful source of news related information. In Sections 3 and 4, we discuss the applicability of different user-generated content sources for detecting new news events and adding value to the returned results, respectively. Lastly, in Section 5, we provide concluding remarks.

2. NEWS PENETRATION WITHIN USER-GENERATED CONTENT

The advent of Web 2.0 has fundamentally changed the landscape of news content. The meteoric rise of blogging (and now micro-blogging) using web-sites like Blogger.com[3] and Twitter.com[4] has allowed users to freely copy, comment and discuss news stories in real-time like never before. Indeed, newswire content is constantly being replicated or re-written, opinions are given and supporting evidence is included or linked to.

Furthermore, due the ever-increasing availability of Internet access world-wide, the role of social media within a news context has changed. In particular, users now assume a dual role as both consumers and reporters of news. Indeed, there is a growing number of cases where news has been broken by social media, for example when a passenger plane landed in New York's Hudson river, the event was reported on Twitter a full 15 minutes before any journalists arrived on the scene.

However, news penetration is not limited to the publicly available user-generated content sources. Indeed, Jones & Diaz [2] showed how user search patterns extracted from Web search engine query-logs could be used to identify news events as they happened, confirming that users search in real-time for news events. Furthermore, it is doubtless that newsworthy information will exist within other forms of user-generated content. In point of fact, Wikipedia is well known to contain information about newsworthy events. Indeed, Michael Jackson's Wikipedia page was updated a total of 104 times on the day of his death, with a further 641 updates on the following day[5].

Moreover, we suspect that newsworthy information exists within other public user-generated content sources like forums, and 'private' user-generated content sources, i.e. where content generated by users has been logged by a third party, rather than published by the user. Notable examples of this are email correspondence and instant messaging logs, where users discuss the events of the moment with their contacts. Indeed, it should not be forgotten that large commercial organisations may collect vast quantities of user-generated content in these forms on a daily basis.

[1] news.google.com
[2] www.bing.com
[3] www.blogger.com
[4] www.twitter.com
[5] wikipedia.org/wiki/Michael_Jackson?action=history

SSM'2010, February 3rd, 2010, New York City, USA

Overall, we conclude that information pertaining to newsworthy events exists within many user-generated content sources. Not only might this make such sources useful for identifying different news stories, but we argue that they can also add value to known news stories. Indeed, in the following sections, we discuss how user-generated content might be useful for improving news search. In particular, Section 3 discusses some applications for detecting new news stories, while Section 4 explores the addition of user-generated content to search results.

3. NEW STORY DETECTION

Initially, the first task that any news search system must address, is whether the query a user has entered is news related or not. However, the unique temporal nature of news makes it paramount to know the current news context, i.e. what stories are important at that moment. As noted earlier, current systems rely on news articles published by one or more newswire companies. Hence, if a trusted newswire company publishes a story, then the topic of that story has been judged important. We refer to this as an editorial view of importance, as it is assigned by newswire editors. However, we argue that this is suboptimal for two reasons.

Firstly, the underlying assumption is that newswire editors will always report on the most important stories to the user, but this is not the case. Newswire editors select stories that they believe will interest their paper's readership, and there is no guarantee that the user is part of their target demographic. The counter-argument to this, is that newswire companies will always report on the most 'generically important' stories of the moment, no matter their inclination, and by selecting stories from a variety of newswire companies, sufficient coverage will be attained. However, we attest that this can never be sufficient to cover all users, as the definition of 'news' will vary from user to user according to their personal interests and current context. Indeed, from the authors personal experience when judging news stories during the TREC 2009 Blog track news task [5], there seemed to be little agreement on what constituted an important story among the judges.

Secondly, by relying on newswire editors to judge news, search systems implicitly restrict themselves to only the pre-defined categories that newswire reporting is built upon, e.g. politics, sport, etc. However, we argue that the user base for search engines is much broader, and hence there exists a need to adopt a less rigid definition of news.

To address these issues, we propose the real-time mining of user-generated content to determine the important stories of the moment from a user perspective. For example, we suggest that the mining of a micro-blogging stream like Twitter might be particularly effective for this task. Indeed, the extremely high post rate of such streams make them prime candidates as a source of evidence for detecting news as it happens. Specifically, we suspect that the distribution of discriminative terms over time can be used to detect when a new story emerges, for example through the application of burst detection [3].

However, the 'freshness' of such streams come at the cost of content. In the case of Twitter, 'tweets' are limited to 140 characters, while this increases the likelihood that posts will stay 'on-topic' for their duration, it severely limits the information contained. This negatively impacts the likelihood that posts discussing the same story use the same vocabulary, resulting in a decrease in detection performance. Indeed, we suspect that techniques to counter vocabulary mismatch like collection enrichment, or the application of Twitter tags might mitigate this effect.

Furthermore, we hypothesise that mining other streams with similarly high post rates concurrently may provide superior coverage, or at least collaborative evidence. In particular, we suggest that private instant messaging logs and user search engine queries are candidates for mining that might hold promise. Indeed, Diaz [1] explored the application of search engine click-through data in combination with user feedback for training a news query classifier. He showed that search engine query-logs when combined with user feedback were effective for determining whether a query is news related or not.

4. ADDING VALUE TO SEARCH RESULTS

In this section, we discuss how user-generated content can add value to the results returned to the user for a news query. In particular, we speculate upon the addition of such documents to the final ranking of documents returned to the user. Currently, Web search engines only inject news articles to their rankings. However, we argue that there are cases where this is insufficient.

In particular, we note the case where a user is searching for a controversial story. Here users may wish to see multiple viewpoints, i.e. in this case the information need of the user is not, "give me information about X", but rather "what do people think about X?". Indeed, it seems reasonable that people would want to know what others think before deciding where they stand on an issue. However, when integrating news results into a search ranking, search engines will often only show stories from neutral newswire providers to avoid being seen as biased, and as such cannot satisfy this type of query.

We believe that the inclusion of user-generated content into the search ranking can add value. In particular, certain user-generated content sources are opinionated by nature, e.g. blog posts, which make them ideal to satisfy the user's information need. Moreover, we hypothesise that user-generated content is superior to traditional newswire articles, as they lack any institutional bias. Furthermore, by 'taking the pulse of the blogosphere' in this manner, we can gauge the public's reaction to a story, rather than just what some subset of the newswire industry believes. Indeed, such was the motivation behind the Blog track top stories identification task at TREC in 2009 [5], where participants were asked to select the top diverse blog posts for a set of news headlines.

5. CONCLUSIONS

In this paper, we discussed how news related information permeates user-generated content on the Web. In particular, we argued that news now can be broken in social media, and that this makes a case for leveraging user-generated content to detect new events as they happen, as well as include such content into news search results. Furthermore, we discussed applications for two user-generated content sources, namely the blogosphere and Twitter, to these tasks. Overall, we suggest that user-generated content can be leveraged to further news search, and there are many directions for future research in this area, both in new techniques to mine user-generated content for information, but also to exploit new user-generated content sources as they arrive.

6. REFERENCES

[1] F. Diaz. Integration of news content into web results. In *Proceedings of WSDM 2009*, Barcelona, Spain.

[2] R. Jones and F. Diaz. Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25(3):14, 2007.

[3] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of KDD 2002*, Alberta, Canada.

[4] J. Leibowitz. "Creative destruction" or just "destruction", how will journalism survive the internet age? In *the FTC Public Workshop: From town crier to bloggers: how will journalism survive the Internet age?*, Washington, DC, USA, 2009.

[5] C. Macdonald, I. Ounis and Ian Soboroff. Overview of TREC-2009 Blog track. In *Proceedings TREC 2009*, Maryland, USA.