News Vertical Search using User-Generated Content

Richard McCreadie
School of Computing Science
University of Glasgow
richard.mccreadie@glasgow.ac.uk

September 2012

The online news landscape has been greatly affected by the emergence of user-generated content, as the general public summarise, discuss and comment upon news stories in real-time. Meanwhile, Web search engines serve millions of queries relating to news events each day, which could benefit from this new content. In this thesis, we investigate how user-generated content can enhance the news vertical aspect of a universal Web search engine, such that news-related queries can be satisfied more accurately, comprehensively and in a more timely manner. The aim is to provide end-users with an aggregate search result ranking that provides increased relevance and coverage of events for news-related queries. In particular, we focus on improving the search result ranking for those news-related queries that cannot be easily satisfied by newswire articles, either because the event that the user is searching for has just broken and hence no news articles have yet been published, or because current newswire articles are insufficiently detailed or out-of-date.

To do so, we propose a news search framework to describe the news vertical aspect of a universal web search engine, comprised of four components, each providing a different piece of functionality. The *top events identification* component identifies the most important events that are happening at any given moment using discussion in user-generated content streams. The *news query classification* component classifies incoming queries as news-related or not in real-time. The *ranking news-related content* component finds and ranks relevant content for news-related user queries from multiple streams of news and user-generated content. Finally, the *news-related content integration* component aggregates the previously ranked content for the user query into the Web search ranking.

For each component, we propose novel approaches that use user-generated content to enhance each. In particular, for top events identification, we propose a fully automatic real-time approach for the identification of currently important news-related events based upon voting theory. For news query classification, we propose a real-time classification approach that leverages recent discussions from multiple parallel news and user-generated content sources. To rank news-related content, we experiment with novel machine learned approaches that leverage the unique characteristics of the different user-generated sources used when ranking. Meanwhile, for news-related

content integration, we adapt resource selection techniques to select and integrate the content ranked from different news and user-generated sources into an enhanced ranking to display to the end-user.

We thoroughly evaluate these proposed approaches using standard TREC test collections (where possible) in addition to new evaluation datasets developed using crowdsourcing. Indeed, to facilitate our evaluation, we generated over 60,000 individual crowdsourced assessments. Moreover, we also performed a large-scale crowdsourced user-study examining whether users prefer the aggregate rankings produced by our news search framework in comparison to traditional Web search results. We show that user-generated content-enhanced approaches can effectively rank events by their current importance/newsworthiness; improve news-query classification accuracy; return additional relevant content for news-related queries in real-time (facilitating more comprehensive coverage of events); and enhance Web search rankings with additional content that end-users find useful. We conclude that user-generated content is indeed a very useful source of information to use in order to enhance a news vertical.

This thesis opens up multiple promising directions for future research in the fields of news/social search, real-time classification and vertical integration, while providing new viable crowdsourced methodologies to evaluate approaches in these fields.

Available Online at `http://theses.gla.ac.uk/3813`