News Vertical Search: When and What to Display to Users

Richard McCreadie, Craig Macdonald, and Iadh Ounis
{firstname.lastname}@glasgow.ac.uk

University of Glasgow
G12 8QQ, Glasgow, UK

ABSTRACT

News reporting has seen a shift toward fast-paced online reporting in new sources such as social media. Web Search engines that support a news vertical have historically relied upon articles published by major newswire providers when serving news-related queries. In this paper, we investigate to what extent real-time content from newswire, blogs, Twitter and Wikipedia sources are useful to return to the user in the current fast-paced news search setting. In particular, we perform a detailed user study using the emerging medium of crowdsourcing to determine when and where integrating news-related content from these various sources can better serve the user's news need. We sampled approximately 300 news-related search queries using Google Trends and Bitly data in real-time for two time periods. For these queries, we have crowdsourced workers compare Web search rankings for each, with similar rankings integrating real-time news content from sources such as Twitter or the blogosphere. Our results show that users exhibited a preference for rankings integrating newswire articles for only half of our queries, indicating that relying solely on newswire providers for news-related content is now insufficient. Moreover, our results show that users preferred rankings that integrate tweets more often than those that integrate newswire articles, showing the potential of using social media to better serve news queries.

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

General Terms: Experimentation, Performance

Keywords: News Vertical, Web Search, User-generated Content

1. INTRODUCTION

Major universal Web search engines serve around a billion of user queries each day [25]. It has been reported that up to 11% of these Web search queries are related to current news [8]. We refer to these queries as *news-related queries*. When a news-related query is submitted to a Web search engine, a news-vertical is used to identify relevant news-related content from one or more providers, traditionally e-newspapers. If relevant content is found, then this content can be integrated into the Web search ranking, normally within a vertical result block, as illustrated in Figure 1.

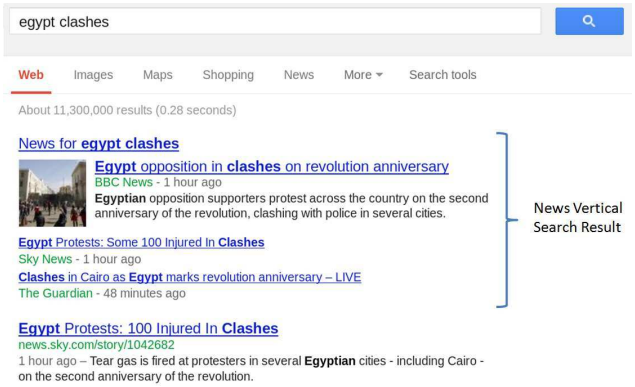


Figure 1: Example of vertical search result integration for the query 'egypt clashes' on the 25/01/2013.

The online news landscape has been greatly affected by the emergence of *user-generated content* sources such as Twitter. In particular, the role of the public in the online news space has shifted from static consumers to real-time reporters and commentators. Indeed, current events are now summarised and discussed in real-time [7, 27] using a variety of diverse media [21], driven by user interaction and user-generated content, e.g. news reporting in Twitter [26].

This shift has had important consequences for Web search engines when tackling news-related queries. In particular, users are now searching for information about events mere seconds after those events have occurred [14]. Moreover, new and potentially valuable content is being produced outside of the normal newswire providers that are used by news verticals. Indeed, we argue that relying solely on newswire providers is no longer sufficient to satisfy news-related queries.

In this paper, we investigate three questions with regard to real-time news vertical search, namely: Is newswire article integration still sufficient given the increasing pace of news-reporting? To what extent can content from user-generated sources also satisfy the end-user? How does the age of an event affect the types of content that users prefer? To answer these questions, we perform a large-scale user study. In particular, we first develop an evaluation interface for performing preference assessment between pairs of rankings. We then sample almost 300 queries relating to recent news events in real-time using the Google Trends and Bitly APIs.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

For each query and time, we also collect individual rankings of Web pages, newswire articles, blogs, tweets and Wikipedia pages for them. Using our proposed interface in conjunction with the crowdsourcing marketplace Amazon's Mechanical Turk, we employ workers (acting as surrogates for end-users) to compare Web page rankings with the same rankings enhanced with additional news content from newswire, the blogosphere, Twitter or Wikipedia. In this way, we evaluate the extent to which different sources can be used to better satisfy news-related queries. Indeed, this is the first user study examining the integration of different content types for news vertical search.

The remainder of this paper is structured as follows. In Section 2, we provide a background into prior works that have examined news vertical search and crowdsourcing. Section 3 describes our methodology, including the design of the crowdsourcing interface we use in our user study. In Section 4, we detail our experimental setup in terms of topic development and crowdsourcing/worker statistics. Section 5 describes the results of our crowdsourced user study with respect to our three research questions. Concluding remarks are provided in Section 6.

2. RELATED WORK

This paper builds upon two areas related to the field of information retrieval (IR), namely news vertical search and crowdsourcing. We discuss prior works that examined news vertical search in Section 2.1, while we provide a background on crowdsourcing for IR in Section 2.2.

2.1 News Vertical Search

The field of news vertical search has seen little investigation to date, with prior works focusing on how to predict when a user will click on a news-related document added into the Web search ranking. In particular, Diaz [12] proposed a machine learning approach for click prediction on newswire articles. This approach trains an initial model using features extracted from the news articles along with additional query features from past Web and news vertical query-logs. The approach then incorporates click-feedback, in that it allows a user's subsequent clicks to enhance the model over time. Arguello *et al.* [4] expanded upon the approach by Diaz for multiple verticals, extracting features from each vertical considered to build a classification model. Later, König *et al.* [20] also examined a learned approach for click prediction for news-related queries. They propose the use of additional features from the query, in addition to features describing the distribution of the query terms in blog, newswire and Wikipedia corpora, with the aim of better estimating the proportion of users that will click on a newswire article if displayed. Importantly, these works only examine the integration of newswire articles into the Web search results, not the integration of user-generated content (in König's case blogs and Wikipedia are used to better determine when newswire articles should be displayed, not as a source of content). In contrast, in this paper, we examine the integration of blogs, tweets and Wikipedia articles in addition to newswire articles. Moreover, rather than predicting document clicks to estimate when to integrate content (which requires access to large proprietary query logs), we perform a user study to ascertain from a user perspective when integrating additional content is useful. Furthermore, an added value of this work is that the assessments produced as part of our user study could be used to train machine learned models for news vertical search, like those described above.

Relatedly, in the context of search over multiple verticals (including the news vertical), Zhou *et al.* [31], proposed a general evaluation framework for aggregate search. However, their framework pre-supposes that relevance assessments for each document are available, which is not the case in our scenario. Hence, we take an alternative evaluation approach in this paper. Arguello *et al.* [5] performed a user study using 29 people from a nearby town to examine the relationship between task complexity and the use of integrated search results from multiple. However, their study focused on the user cognitive process for general tasks, rather than news-specific search tasks. Indeed, to the best of our knowledge, our user study is the first that examines news vertical search to date.

2.2 Crowdsourcing in IR

Crowdsourcing in general is the act of outsourcing tasks, traditionally performed by a specialist person or group, to a large undefined group of people or community (referred to as the "crowd"), through an open call [16]. There are many motivations for crowdsourcing tasks. For example, simple tasks can be completed at a relatively small cost, and often very quickly [3]. Moreover, by employing a crowd of 'users' to perform assessments as opposed to a few 'experts', a wider range of talent can be accessed and expert bias avoided [15]. However, crowdsourcing has also been the subject of much controversy as to its effectiveness. In particular, the work produced by crowdsourced workers is known to sometimes be of low quality [6] and results can be affected by random or malicious work [13].

In an IR setting, crowdsourcing has been used as an alternative method for evaluating search system performance. Alonso *et al.* [3] first suggested crowdsourcing as a cheap, fast, effective and flexible alternative to using specialist assessors when creating relevance assessments for ad-hoc test collections. Indeed, crowdsourcing was later used to generate relevance assessments for larger collections [2], in addition to being both the subject of one track[1] and used to support other tracks [24] at the Text REtrieval Conference (TREC). Relatedly, Zhu and Carterette [32] also used crowdsourcing to perform a successful pilot study examining the integration of images into the Web search results.

In this paper, we use crowdsourced workers from the Amazon's Mechanical Turk marketplace to compare Web search rankings to those enhanced with additional news-related content for news queries. Indeed, crowdsourcing is particularly suitable to use in our user study for three reasons. First, workers can act as surrogates of end-users on the assumption that those same workers also use Web search engines. Second, the worker-base is large and diverse [17] meaning that our results are less likely to suffer from demographic bias, in contrast to using students or co-workers. Third, assessments are quick, cheap and reproducible by other researchers.[2]

Prior works in the field of crowdsourcing have identified effective *validation* approaches for the identification of poor quality work, with the aim of increased accuracy. Snow *et*

[1] `https://sites.google.com/site/treccrowd/`

[2] Subject to implementation of the same assessment interface and access to the dataset described later in Section 4.1, which is available at `http://terrierteam.dcs.gla.ac.uk/vertical/`.

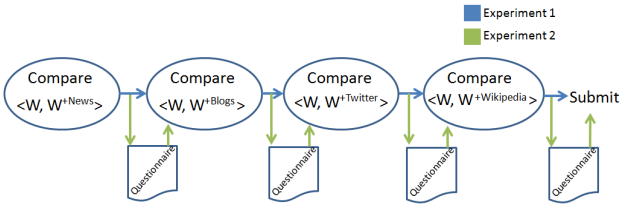


Figure 2: Workflow of a single assignment.

al. [29] and Callison-Burch [10] investigated the accuracy of crowdsourced labels generated for natural language processing tasks, concluding that 'expert' quality labels can be achieved by having three or five workers complete each crowdsourcing task and taking the majority (most commonly selected) label. Kittur *et al.* [19] proposed the introduction of questions with verifiable, quantitative answers, often referred to as 'gold judgements' or a 'honey-pot' to identify poorly performing workers. However, subsequent research by Ipeirotis [18] has indicated that within a Web page classification context, gold judgements are unnecessary when larger numbers of assessors work on each task. In particular, their results indicate that when 10 assessors work on each task and a majority result is taken, then no appreciable gains in accuracy are observed when further adding a gold judgement validation, i.e. crowdsourced work can be validated through redundancy. Following this prior work, as part of our user study, we use 10 redundant workers to compare each of our rankings. The majority vote is used to ensure quality.[3] Our assessment interface (described later in Section 3.1) also contains an in-built form of validation, to identify bots and users that randomly click on answers. In the next section, we describe our experimental methodology, including how we designed the interface we show to our crowdsourced workers.

3. METHODOLOGY

In an IR setting, rankings are usually evaluated using incomplete Cranfield-style relevance assessments [11] pooled from multiple systems [30]. However, such an approach is not suitable for evaluation of vertical search rankings because documents within a ranking cannot be judged independently of one another. In particular, whether the news-related results added by the news vertical improve the initial Web search ranking depends both upon the documents added and the extent to which the Web search ranking already satisfied the user's news need. Moreover, each ranking contains multiple types of documents, which from a pairwise perspective may not be directly comparable. For instance, how can the value of a tweet be directly compared to a newswire article? Furthermore, the real-time nature of news-related queries mean that the value that each news-related document brings to the ranking is dependant upon both its relevance and also its recency/timeliness with regard to the event underlying the query.

Instead, inspired by prior works by Thomas and Hawking [28] and Carterette [9], we propose a comparative evaluation approach, where the top ten documents from each ranking are considered as a single evaluation unit (rather than evaluating the individual documents). For this type of evaluation, multiple different rankings are displayed to the user, who then selects the one that he/she prefers. In our scenario, for a news-related query, the user is shown both the unmodified Web search ranking and a ranking with content from either newswire, the blogosphere, Twitter or Wikipedia integrated. In particular, for each news-related query, we have users compare the Web search ranking to rankings enhanced with content from each of the aforementioned sources in turn, such that they can identify those cases where integrating additional content is useful. The advantage of this approach is that it avoids the issues of pair-wise document comparability and redundancy by having end-users compare both document rankings as single units.

Formally, our approach takes as input a Web search ranking W and a set of news and user-generated content rankings $e \in E$ for a news-related query Q and a given point in time t. For our experiments, E is comprised of four rankings, namely rankings of newswire articles, blogs, tweets and Wikipedia pages. Each ranking is ranked by relevance to Q and with respect to the query time t, i.e. only documents from before t are ranked. For each query, the Web search ranking W is combined with each of the four news and user-generated content rankings in E to form four *enhanced rankings*. We denote an enhanced ranking as W^{+e}, e.g. W^{+Blogs}.

Our strategy for generating the enhanced rankings is as follows. For W^{+News}, W^{+Blogs} and $W^{+Twitter}$, we select the top three documents ranked for Q and add them to the top of W within a special result box. This simulates the look and feel of current news vertical search results, like those shown previously in Figure 1. For $W^{+Wikipedia}$, we include only the top ranked Wikipedia page for the query, based on the observation that Web search engines typically only return one page from Wikipedia in their top 10 results.

Finally, we combine the Web search ranking W with its four enhanced rankings to form four document ranking pairs, e.g. $<W, W^{+Blogs}>$. During our user study, for each query, a user is shown each of the four ranking pairs for that query in turn, selecting the one that they would prefer to see for Q and time t in each case. Figure 2 illustrates the workflow for a single assignment. As described later in Section 4.1, we evaluate using two datasets. For the second dataset, after the user selects a ranking, we also give the user a multiple-choice questionnaire about why they selected each ranking.

If the majority of users that compare a ranking pair for a query Q select the enhanced ranking W^{+e}, we say that W^{+e} *better satisfies* Q (the query topic) for time t. In Section 3.1, we describe the interface we show to the users, while Section 3.2 describes of the questionnaire that we use.

3.1 Crowdsourcing Interface

To facilitate our comparative evaluation approach, we develop an assessment interface to perform the presentation of document ranking pairs to the user, and record the results. Figure 3 provides an illustration of the assessment interface for the query 'mariangela melato' made on January 12th 2013. We enlarge the enhanced ranking for easy viewing in Figure 3. At the top of the interface, the title of the task is displayed along with a button that reveals the instructions for the task (hidden in Figure 3). The instructions are divided into two separate components, namely: the main instruction block that describes what the user is being asked to do; and the guidelines block that provides additional clarifications about the task. When a user first views the task

[3]Tie-breaking assessments are unneeded when using many (10 in our case) workers.

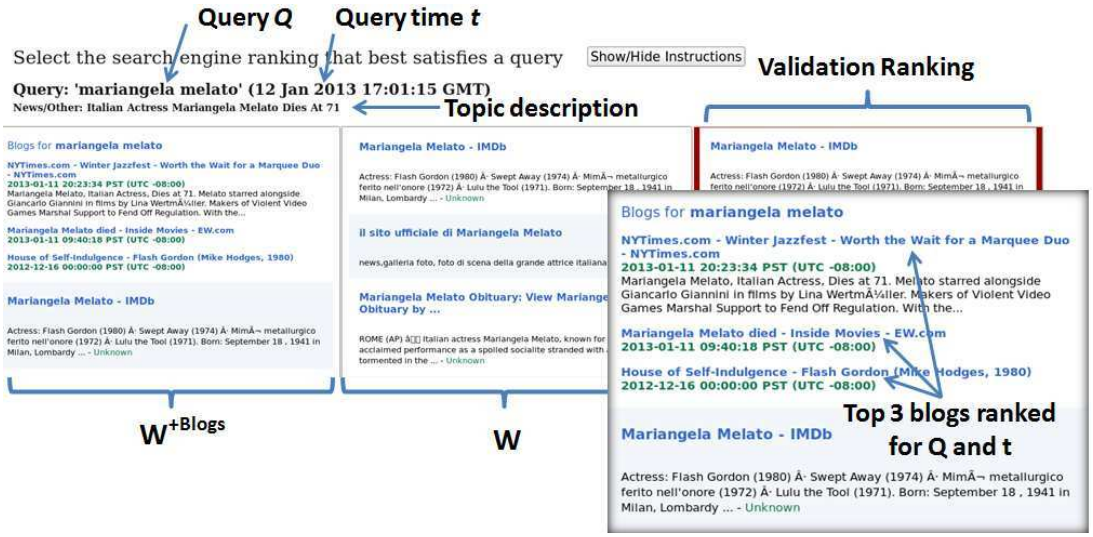


Figure 3: Illustration of our ranking interface produced for the query 'mariangela melato' on the 12/01/13 at 5:01pm. The first ranking is W^{+Blogs}, while the second ranking is W. The third (red bordered) ranking is a validation ranking designed to catch workers that are randomly clicking.

(before accepting it) the instructions are displayed to them. When actively attempting the task, the instructions are automatically closed by default to save screen real-estate.

Below the instructions, the news-related user query is displayed along with a short topic description. For the example in Figure 3, the user was searching for information about the death of Italian cinema and theatre actress Mariangela Melato. Below the query Q and topic description, the document rankings W and W^{+e} from the current pair are rendered (W^{+Blogs} in Figure 3). Users are instructed to select the ranking that best satisfies the query, while the guidelines instruct the user to focus on whether the additional results added to one of the rankings make it better or worse. To enable users to state their preference for one of the displayed rankings, each ranking is clickable. Clicking a ranking records the user's preference for that ranking. Recall that the rankings are produced with respect to the query time t, i.e. only documents published before the time of the query are ranked. Both the timestamp of the query and the publication time of each post (if available) are rendered as part of the interface. Users are instructed to consider recency as well as relevance when assessing using these timestamps.

The four ranking pairs for each query topic are assessed by a single user in sequence to increase efficiency [23]. After a user clicks on a ranking, if there are any ranking pairs remaining for the current query topic (of the four total), then the next ranking pair is rendered for assessment, else a link is displayed to enable the user to submit their assessments. If any of the news and/or user-generated content sources did not return documents for the query, e.g. because no relevant documents had yet been published, then both of the rankings in the associated pair would be identical. In these cases, the interface automatically skips to the next pair to be assessed.

As described in Section 2.2, one of the criticisms of crowdsourcing is a susceptibility to poor quality work. In addition to having 10 unique users assess each ranking pair, our assessment interface also integrates one method for validating user work. In particular, the interface renders the Web result ranking twice, creating three rankings. One of these rankings has a red border (as shown in Figure 3). The users are instructed never to select the red bordered ranking. We use this ranking to identify bots or malicious users that are randomly selecting rankings. Any set of assessments where a red-bordered ranking has been clicked is automatically rejected without payment. The three rankings are randomly ordered to avoid click bias for any one position.

3.2 Questionnaire

When assessing the topics in our second dataset (see Section 4.1), we show a questionnaire to the user after they select each ranking. Depending upon which ranking the user selected, a different questionnaire is displayed. In particular, for the enhanced ranking, the user is asked to select one or more statements that apply to the content added (either newswire articles, blogs, tweets or Wikipedia pages). The six statements are listed below:

```
> Relevant to the query
> Relate to a recent event (at the time of the query)
> Appear to be about breaking news
> Better satisfy the query by being included
> Are informative based upon the title/snippet shown
> Provide useful general information, unrelated to a
  recent event
```

In contrast, if the user selected the Web search ranking, they are asked to select one of the following reasons why the additional content did not better satisfy the query:

```
> Partially relevant, but I still preferred the basic
  ranking.
> Out-of-date, new information was added, but it was old.
> Redundant, no information was added that was not in the
  basic ranking.
> Completely irrelevant, the added documents were
  unrelated to the query.
> Spam, the added documents appear to be spam or
  advertising.
```

In this way, we first have users decide whether a ranking is improved by the integration of news or user-generated content and then have them elaborate as to why this is the case.

Table 1: Statistics for the two datasets used in this paper and the number of query topics belonging to each temporal class within each.

	2012_{Apr}	2013_{Jan}
Topic Source	Google Trends	Bitly Bursting Phrases
Time-Period	11/04/12 → 23/04/12	10/01/13 → 16/01/13
# Breaking	74	64
# Recent	38	7
# Long-Running	29	15
# Other	58	0
Total Topics	199	98

We use the interface and questionnaire described above in conjunction with Amazon's Mechanical Turk to perform our user study. In the next section, we describe our experimental setup for the user study.

4. EXPERIMENTAL SETUP

Due to a lack of publicly available datasets upon which we could evaluate ([12] and [20] described in Section 2.1 used private Yahoo! and Microsoft data for instance) we develop two new datasets for evaluation. In Section 4.1, we describe the creation of two new datasets that we use for evaluation. Section 4.2 summaries the structure of our user study. In Section 4.3, we detail the statistics of our crowdsourcing experiments, while Section 4.4 provides information about our MTurk workers.

4.1 Topic Development

We develop two new datasets to facilitate our evaluation using publicly available resources. Our first dataset, referred to as 2012_{Apr}, spans the period of the 11th to the 23rd of April 2012. The second dataset, referred to as 2013_{Jan} spans the period of the 10th to the 16th of January 2013. Each dataset contains a set of <query,timestamp> pairs and rankings of Web pages, newswire articles, blogs, tweets and Wikipedia pages for each.

To generate our news-related queries, we use the Google Trends[4] and the Bitly Data API[5]. In particular, for the 2012_{Apr} dataset, we sampled the queries that were reported trending by Google Trends on an hourly basis for that period. At the end of each day, we manually identified those queries that matched a reported news event, while all other queries were discarded. This left a set of 199 queries known to be news-related over the 12 day period. For the 2013_{Jan} dataset, we instead crawled the 'Bursting Phrases' service of Bitly's Data API at 30 minute intervals, since Bitly updates its bursting phrases more frequently than Google Trends. The bursting phrases service returns phrases that were receiving a consistently high volume of click traffic at the time. From this set of bursty phrases returned, we manually identified 98 that referred to news events, using these as our news-related queries for the second dataset. For both datasets, the timestamp of the query is the time when it was crawled using either API. Hence, this timestamp corresponds to a period when the query/phrase was either receiving high traffic as recorded by Google or Bitly. We also manually created topic descriptions for each query with reference to the main news stories around the time of each query.

To collect rankings for each topic, we again used publicly available API's. For the time of each query as recorded by that query's timestamp, we downloaded Bing Web search results[6], newswire articles and blogs from the Blekko API[7], Twitter tweets[8] and Wikipedia pages[9] for that query. These rankings are real-time in nature, i.e. they contain only documents published before the query timestamp t. The documents are ranked by relevance with respect to the query Q. Recall from Section 3 that the top three newswire articles, blog posts and tweets are selected for integration into the Web search ranking, while for Wikipedia, only the top ranked Wikipedia page is integrated.

[4]http://www.google.co.uk/trends/hottrends
[5]http://dev.bitly.com/

Table 2: Types of data collected from the two crowdsourcing tasks.

Data Subject	Data Type	2012_{Apr}	2013_{Jan}
Document Rankings	Ranking Preference	✔	✔
Document Rankings	Questionnaire	✗	✔
Worker	Work Submitted	✔	✔
Worker	Time Taken	✔	✔
Worker	Geo-Location	✗	✔

To aid our later analysis, we also divide our queries into four classes, representing the time elapsed between the event underpinning each query and that query's timestamp:

- **Breaking**: Queries related to events that have broken within the prior 12 hours.
- **Recent**: Queries related to a recent news event that occurred between 12 and 48 hours previously.
- **Long-Running**: Queries relating to older events that are still of interest.
- **Other**: The time of the originating event could not be accurately identified.

Table 1 summarises the statistics of our two datasets and the number of query topics belonging to each query class within them.

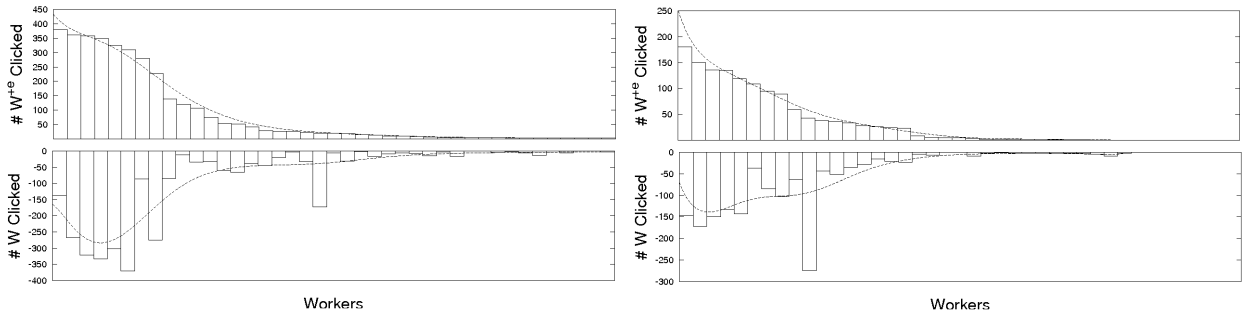
4.2 Experiment structure

Following best practices in crowdsourcing [1], we adopt an iterative design methodology. In particular, we separate our crowdsourced user study into two separate crowdsourcing experiments, one for each of the two datasets described in the previous section. Based upon worker feedback from the first experiment (2012_{Apr}), we improved the assessment interface, user behaviour logging and the provided instructions before crowdsourcing the second experiment (2013_{Jan}). Most notably, the questionnaires were added to enable users to better articulate why one ranking might be better than another, while IP-address logging was enabled to estimate the location of workers. Table 2 summaries the data that was collected from the two crowdsourcing experiments.

4.3 Crowdsourcing Configuration

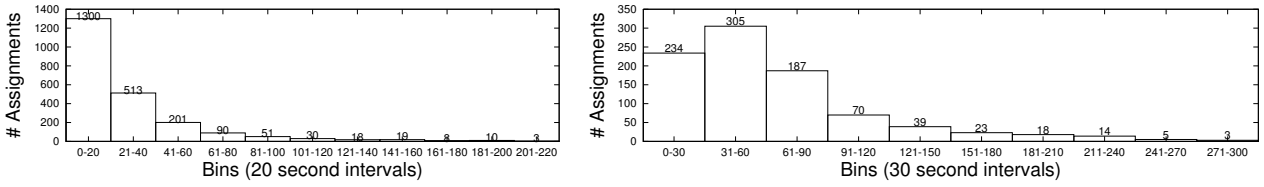
We use the crowdsourcing marketplace Amazon's Mechanical Turk (MTurk) to recruit workers for our evaluation. Following [18], as described in Section 2.2, we have ten individual workers assess each document ranking pair. Each MTurk

[6]http://www.bing.com/developers/
[7]http://blog.blekko.com/2012/10/15/powering-web-apps-with-the-blekko-api/
[8]https://dev.Twitter.com/
[9]http://www.mediawiki.org/wiki/API:Main_page



(a) 2012_{Apr} (b) 2013_{Jan}

Figure 4: The distribution of enhanced and Web search rankings selected by workers for each experiment.



(a) 2012_{Apr} (b) 2013_{Jan}

Figure 5: Assignments binned by the time they took to complete for each of the two experiments.

assignment involves the assessment of 4 document pairs (the Web search ranking in comparison to 4 different enhanced rankings). The second experiment also added the questionnaires after each assessment was made. We paid US $0.05 for each set of 4 document pairs assessed, since assessment requires that the user browses the two rankings and makes a single click. For the second experiment, an additional $0.05 was paid for completion of the questionnaires. We use the in-built worker validation described in Section 3.1 to detect bots and spammers. Any assignment where a red-bordered ranking was selected was rejected without payment and re-submitted for another user to complete. The total cost of running the two experiments was US $217.25, including MTurk's 10% fee.

In terms of the two experiments, the 199 topics from the first dataset translate into 1990 assignments (199x10), while the 98 topics from the second dataset translate into 980 assignments (98x10). Both experiments were completed individually in under 48 hours. Automatic validation (based on clicking of red-bordered rankings) resulted in the rejection of 9.5% and 3.05% of assignments for each of the two datasets respectively. Inter-worker agreement was 0.33 using Cohen κ for the first experiment and 0.23 using Cohen κ for the second experiment. Note that a low agreement is expected in this case due to both the large number of workers attempting each task and a greater scope for disagreement between workers due to individual preferences.

4.4 Worker Statistics

In total, 98 unique workers completed one or more assignments over our two experiments, 57 workers during the first experiment and 41 during the second experiment. Figures 4 (a) and (b) show the distribution of enhanced (W^{+e}) and Web search rankings (W) selected by workers. As is typical for crowdsourced tasks [2, 24], we see that the completed work follows a long-tail distribution, with a few workers completing many assignments and a long tail of workers that completed only a few assignments. Furthermore, we also observe that the majority of workers show no clear bias toward selecting either type of ranking. Note that because 10 individual workers assess each topic, the impact that a single worker can have on the final result is limited.

Table 3: Work statistics broken down by country for the second dataset.

Country	# of Page Requests	# Unique IP-Addresses	Proportion of Work Done	Proportion of IP-Addresses
India	6,479	41	68%	39%
United States	1,484	40	16%	38%
Macedonia	843	3	9%	3%
Philippines	242	1	3%	1%
Romania	210	3	2%	3%
United Kingdom	141	3	2%	3%
Pakistan	122	1	1%	1%
Bangladesh	10	1	0.1%	1%
Japan	8	5	0.1%	5%
Singapore	8	2	0.1%	2%
Hungary	6	1	0.1%	1%
Ireland	6	1	0.1%	1%
Albania	4	1	>0.1%	1%
Australia	3	1	>0.1%	1%
Canada	3	1	>0.1%	1%

The time taken by the workers to complete each assignment varies. Figures 5 (a) and (b) show the amount of time that assignments (4 assessments) took, separated into 20/30 second bins (for readabilty). From Figure 5, we observe that for the first experiment (2012_{Apr}), to assess four ranking pairs took the majority of the workers (57%) under 20 seconds, with a further 23% of assignments taking between 21 and 40 seconds. In contrast, for the second experiment (2013_{Jan}), assessment took longer on average, with 33% of workers taking between 31 and 60 seconds and a further 20% taking between 61 and 90 seconds. This is because the second experiment also required workers to fill in the questionnaire after each of the four ranking pairs per assignment.

Finally, for the second experiment (2013_{Jan}), we also logged the IP-addresses of the workers. From this information, we identified the county of origin that each worker comes from. Table 3 reports the number of page requests, unique IP-addresses and their proportions for each country made dur-

ing the second experiment. From Table 3, we observe that 39% of IP-addresses and hence workers came from India, followed by 38% from the U.S. However, while the number of unique workers is similar between these two countries, the volume of work done is much higher for Indian workers. This indicates that Indian workers completed more assignments on average than US workers.

4.5 Measures

In our subsequent experiments, we report the majority preference over the ten workers that compared each ranking pair. Recall from Section 3.1 that not all sources return content for all of the topics. In these cases, both rankings in a pair would be identical, hence we remove these cases from our dataset. For this reason, the click counts can be lower than the total number of assignments.

5. RESULTS

In this section, we investigate to what extent integrating newswire articles, blogs, tweets and Wikipedia pages into the search results for news-related queries can better satisfy the user. In particular, we examine the following three research questions, each in a separate subsection:

- Is newswire article integration still sufficient given the increasing pace of news-reporting? (Section 5.1)
- To what extent can content from user-generated sources better satisfy the end-user in comparison to returning the Web search ranking or integrating newswire articles? (Section 5.2)
- How does the age of a event affect the types of content that users prefer? (Section 5.3)

5.1 Integrating Newswire Articles

We begin by examining our first research question, i.e. is news article integration still sufficient to satisfy news-related queries in real-time. We report on the number of users that clicked (stated a preference for) rankings enhanced with newswire articles (W^{+News}) in comparison to the Web search ranking unaltered (W). If the users prefer the newswire-enhanced rankings, then this would indicate that newswire articles are still sufficient. On the other hand, if there are many topics where users do not prefer the newswire-enhanced ranking, then this would indicate that we need to look to other sources of content to satisfy the user for these queries.

Table 4 reports the raw number of users that clicked on the newswire-enhanced or Web search ranking, i.e. when $<W,W^{+News}>$ pairs were compared. Recall that we have ten users (workers) assess each ranking pair to counteract any noise introduced by poorly performing workers. Table 4 also reports the majority preference over the ten workers, i.e. the number of real-time news topics where the majority preferred either ranking. From Table 4, we observe the following. First, comparing the raw click counts for each of the newswire-enhanced and Web search rankings (rows 3 and 5), we observe that for 28.9% of topics the users preferred the Web search ranking, while 48.5% preferred the newswire enhanced ranking (aggregating both datasets). Second, when taking the majority vote (rows 4 and 6), we similarly see that the Web search ranking was preferred by the majority of users for 17.8% of topics, while for 44.4% of the topics, the news-enhanced rankings were preferred by the majority of users. These results show that only around half of our real-time news-related topics could be better satisfied by integrating newswire articles.

Table 4: Number of topics for which the majority of end-users preferred either the Web search ranking or the rankings enhanced with user-generated content.

Ranking Preferred	Majority Vote?	2012_{Apr}	2013_{Jan}	Total	Proportion
Web	✗	828	418	860	28.9%
Web	✔	24	29	53	17.8%
News	✗	1002	442	1444	48.5%
News	✔	98	34	132	44.4%

To investigate why newswire article integration was not able to satisfy many of our real-time news-related topics, we examine the reasons selected by our users in the questionnaire from our second experiment (see Section 3.2). Of the 418 Web search rankings clicked by our users in the second experiment, Figure 6 reports the percentage of those clicks that were attributed to the added documents in the newswire-enhanced ranking being either partially relevant, out-of-date, redundant, unrelated or spam. From Figure 6, we see that the main reason that users gave was that the results added were partially relevant, but did not make the enhanced ranking better. This result is interesting, since it indicates that the newswire sources were providing related content to the query, but that this content was not useful to the majority of our users. One example where the majority of users selected the Web search ranking was the query 'game industry', with the topic description 'News/Politics: Biden meeting with representatives from the video game industry'. For this query, the top two added newswire articles discussed violence relating to the video game industry, but did not specifically mention Joe Biden or the topic. The third integrated article was directly relevant to the topic. However, the original Web search ranking already contained two relevant Associated Press articles about the topic. Hence, the Joe Biden article was redundant, but the two other added articles were deemed by our users to be useful – even though they do not mention the topic directly.

Next, we examine the 442 instances from the second experiment where integrating newswire articles did improve the Web search ranking. Figure 7 illustrates the percentage of newswire-enhanced search rankings selected because they were more relevant, recent, reported breaking news, better satisfied the query, were particularly informative or provided general (background) information on the topic.[10] From Figure 7, we observe the following. First, as expected, we see that over 77.2% of the enhanced rankings were selected because they contained more relevant documents than the Web search ranking. We also see that over 63.3% of the enhanced rankings were selected because the added documents related to the recent event described in the topic description. Note the distinction between documents being relevant and documents relating to the event referred to by the query, i.e. some users may find that adding background articles for longer-running events is enough to improve over the Web search ranking. On the other hand, a much lower percentage of users (28.2%) indicated that they selected the enhanced ranking because the added documents were about breaking news. Also of note is that only 41.2% of users indicated that the introduction of newswire articles better satisfy the query

[10]Note that because a user can select multiple reasons for a single ranking, the percentages do not sum to one.

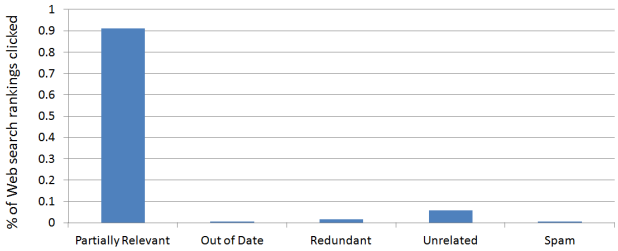


Figure 6: The percentage of Web search rankings clicked that were attributed to the added documents in the newswire-enhanced ranking being either partially relevant, out-of-date, redundant, unrelated or spam.

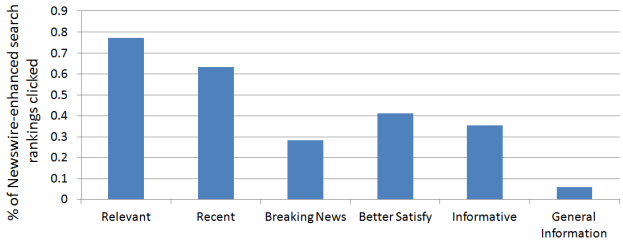


Figure 7: The percentage of newswire-enhanced search rankings selected because they were more relevant, recent, report breaking news, better satisfy the query, were particularly informative or provide general (background) information.

by being included, even though they stated a preference for the enhanced ranking. This indicates that either our users disagreed upon the meaning of 'satisfy' in the questionnaire, or that users can prefer the enhanced ranking for a topic, even if it does not better satisfy that topic. Indeed, almost all users selected the 'relevant' but not the 'better satisfied' choice for one or more topics.

To answer our first research question, newswire articles were not sufficient to satisfy the majority of topics in our two datasets, primarily because while some relevant content is added for most queries, users did not find that it improved the ranking. For the assignments where users preferred the newswire-enhanced ranking, users indicated that they preferred them because the added content was about the event underlying the query or was simply relevant to the query topic. In the next section, we investigate whether the integration of user-generated content can better satisfy users.

5.2 Integrating User-Generated Content

Next, we examine our second research question, i.e. to what extent can content from user-generated sources also satisfy the end-user? To this end, we begin by reporting on the number of users that clicked (stated a preference for) rankings enhanced with either blog posts, tweets or Wikipedia pages in comparison to the Web search rankings unaltered. If the number of topics where the majority of users selected the user-generated content-enhanced rankings is greater than for the number of topics that were similarly enhanced by integrating newswire articles (132 topics – see Table 4 row 6) then we can conclude that user-generated content sources may be more effective for news vertical integration than the traditional newswire sources used today.

Table 5 row 3 reports the number of topics where the majority of users preferred rankings enhanced with either blogs, tweets or Wikipedia pages to the Web search ranking unaltered. We report both the number of topics for each of the two datasets (2012_{Apr} and 2013_{Jan}), as well as the summation over the two. From Table 5 row 3, we observe that in total, the blog-enhanced rankings were preferred for 56.9% of the topics over the unaltered Web search rankings, while tweet-enhanced rankings were preferred for 71.2% of topics and Wikipedia pages for 52.1% of topics. Moreover, these proportions are consistent across both datasets. This result shows that all three of the user-generated content sources tested can enhance the Web search ranking for a substantial proportion of the 297 (see Table 1) real-time news-related topics tested. Most notably, the rankings including tweets were particularly favoured by our users, indicating that Twitter is a good source of real-time news content to integrate. Indeed, integrating tweets satisfied more query topics (168 – see Table 5 row 3, column 7) than integrating newswire articles (132 topics – see Table 4 row 6), indicating that Twitter is a better overall news source than traditional news providers.

However, if at the time t of each topic there exists relevant newswire articles, then user-generated content may not be required to satisfy the user, i.e. the extra content may be redundant. Indeed, in the previous section, we showed that approximately half of our topics could be better satisfied by integrating newswire articles than returning the Web search ranking unaltered. To examine this, we contrast the query topics that benefited from the integration of news articles to those that benefited by the integration of user-generated content. For a given query, if relevant user-generated content was found, integrated and the ranking was preferred by the majority of users, then user-generated content can be said to better satisfy that query than when integrating nothing. However, that same query may be equally or better satisfied by the integration of newswire articles. On the other hand, if there are news-related queries that user-generated content can satisfy, but newswire articles cannot (at the time those queries were made), then we can show that user-generated content is necessary to satisfy those queries.

Table 5 rows 4-6 report the number of topics for which integrating user-generated content aided but that integrating newswire did not (# Topics Better than Newswire). For comparison, we also report the reverse case (where integrating newswire articles aided but integrating user-generated content did not) and where integrating either newswire articles or user-generated content was equivalent (either both better satisfied the user or neither did). The total number of topics where integrating user-generated content aided but integrating newswire did not is highlighted in bold. From Table 5, we see the following two points of interest. First, each of our three user-generated content sources were able to satisfy topics which were not already satisfied by newswire articles. For instance, 22.2% of all topics could be better satisfied by integrating blogs but not newswire articles (row 4, column 4). Second, of the three user-generated content sources tested, we again see that Twitter is the most effective, improving 84 topics (34%), which were not better satisfied by integrating newswire articles. The query 'lionel richie' (issued on the 13th of April 2012) was one example where integrating tweets better satisfied our users but integrating newswire articles did not. For this query, the Twitter stream returned tweets reporting that music star

Table 5: The number of topics where users preferred rankings enhanced with either blogs, tweets or Wikipedia pages and comparison against those topics that could also be enhanced with newswire articles.

	Blogs			Twitter			Wikipedia		
	2012_{Apr}	2013_{Jan}	Total	2012_{Apr}	2013_{Jan}	Total	2012_{Apr}	2013_{Jan}	Total
# Topics Enhanced	86 (58.9%)	29 (51.8%)	115 (56.9%)	121 (73.8%)	47 (71.2%)	168 (71.3%)	66 (55.4%)	21 (43.8%)	87 (52.1%)
# Topics Better than Newswire	39 (22.2%)	15 (23.8%)	**54 (22.6%)**	53 (29.1%)	31 (47.0%)	**84 (33.9%)**	29 (17.3%)	12 (17.6%)	**41 (17.4%)**
# Topics Worse than Newswire	51 (29.0%)	20 (31.7%)	71 (29.7%)	30 (16.5%)	18 (27.3%)	48 (19.4%)	25 (36.8%)	25 (36.8%)	86 (36.4%)
# Topics Equal to Newswire	86 (48.9%)	28 (44.4%)	114 (47.7%)	99 (54.3%)	17 (25.7%)	116 (46.8%)	78 (46.4%)	31 (45.6%)	109 (46.2%)
# Topics Better than Blogs	—	—	—	66 (48.5%)	33 (48.5%)	99 (48.5%)	40 (26.8%)	16 (23.2%)	41 (20.2%)
# Topics Worse than Blogs	—	—	—	31 (22.8%)	15 (22.1%)	46 (22.5%)	60 (40.3%)	24 (44.4%)	84 (41.4%)
# Topics Equal to Blogs	—	—	—	39 (28.7%)	20 (29.4%)	59 (28.9%)	49 (32.9%)	29 (53.7%)	78 (38.4%)
# Topics Better than Twitter	31 (22.8%)	15 (22.1%)	46 (22.5%)	—	—	—	19 (14.1%)	10 (15.6%)	29 (14.6%)
# Topics Worse than Twitter	66 (48.5%)	33 (48.5%	99 (48.5%)	—	—	—	74 (55.2%)	36 (56.3%)	110 (55.6%)
# Topics Equal to Twitter	39 (28.7%)	20 (29.4%)	59 (28.9%)	—	—	—	41 (30.6%)	18 (28.1%)	59 (29.8%)
# Topics Better than Wikipedia	60 (40.3%)	24 (34.7%)	84 (38.5%)	74 (55.2%)	36 (56.2%)	110 (55.6%)	—	—	—
# Topics Worse than Wikipedia	40 (26.8%)	16 (23.2%)	56 (25.7%)	19 (14.2%)	10 (15.6%)	29 (14.6%)	—	—	—
# Topics Equal to Wikipedia	49 (32.9%)	29 (42.0%)	78 (35.8%)	41 (30.6%)	18 (28.1%)	59 (29.8%)	—	—	—

Lionel Richie was facing a large tax bill over allegations he owed more than $1 million to the U.S. government. On the other hand, our newswire stream returns no relevant articles for that query and time.

Having shown that integrating user-generated content can be more effective than integrating newswire articles, we now investigate whether the three user-generated content sources predominantly aid for the same or different topics. Indeed, even if Twitter is the most effective of the three user-generated sources overall, it may be that blogs and Wikipedia can aid for other topics. Table 5 rows 7-15 report the number of topics for which the rankings enhanced with content from each user-generated source were preferred over the rankings enhanced with another user-generated source. From Table 5, we observe the following. First, the three user-generated content sources appear to aid for different topics. For example, blog-enhanced rankings were preferred for 46 topics where rankings integrating tweets were not. However, there were similarly 99 topics for which the Twitter-enhanced rankings were preferred, where the blog-enhanced rankings were not. Second, the proportion of topics where integrating two different user-generated content sources were equivalent (both aided or neither did) is quite small, e.g. 22.5% of topics when considering tweets and blog posts. This indicates that user-generated content sources are useful for tackling different types of queries and that by effectively using multiple types of content, a larger proportion of news-related queries might better be satisfied. For instance, the topic 'james holmes' about a plea extension in the Aurora movie theatre shooting court case was only better satisfied by integrating tweets, while the topic 'baby lion' about a dog being mistaken for baby lion in Virginia was better covered by local newswire outlets.

To answer our second research question, integrating user-generated content into the Web search results can better satisfy many of our news-related topics. Moreover, many of these topics could not be similarly satisfied by integrating newswire articles published at the time. The most effective user-generated content source tested was Twitter, highlighting its usefulness for real-time news vertical search.

5.3 Content Integration and Event Age

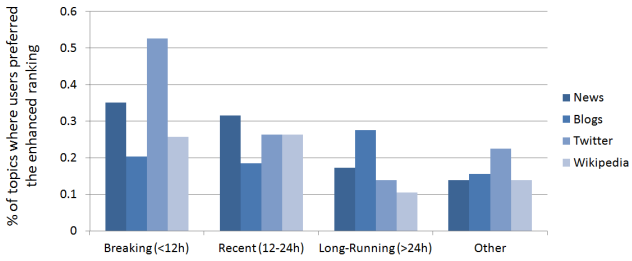
In this section, we investigate our final research question, i.e. how does the age of a event affect the types of content that users prefer. To this end, we examine the query topics for which the integration of user-generated content aided when we break down our topics by the time between the originating event and the topic time. The aim is to determine whether different sources are better to integrate at different points during an event's life cycle.

Figures 8 (a) and (b) report the proportion of query topics where users preferred the enhanced ranking in comparison to the Web search ranking divided into three time classes (see Section 4.1) for each dataset, respectively. From Figure 8 (a) and (b) we observe the following. First, the proportion of topics improved by integrating newswire articles shows a downward trend as the time elapsed from the event increases. This indicates that newswire articles are less useful to return for long-running events, possibly because newswire providers tend to publish a burst of articles around the time of the event, with little content published later. Secondly, in contrast, we see that integrating blogs satisfied a larger proportion of long-running queries than breaking or recent queries. This supports prior observations that for some types of events bloggers tend to lag behind newswire when publishing [22]. Next, we see that integrating Tweets was most effective for satisfying queries relating to breaking events. Indeed, on the first dataset, 52% of the breaking news queries could be better satisfied by integrating tweets into the Web search ranking. For example, one such query was 'levon helm', referring to American rock multi-instrumentalist and actor who was fighting cancer at the time. When this query was issued on the 19th of April 2012, Twitter returned tweets reporting his death, while our newswire stream only returned articles reporting that he was in the final stages of cancer. This illustrates the value that Twitter can bring from a content integration perspective for breaking news queries where no newswire articles have yet been published.

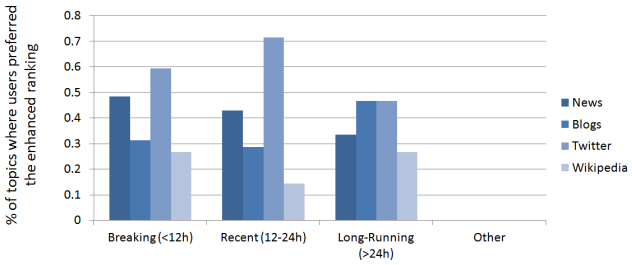
To answer our third research question, we find that users prefer rankings that integrate tweets or newswire articles soon after an event breaks. As an event matures, tweets and newswire articles become less effective, while integrating blogs becomes more effective.

6. CONCLUSIONS

In this paper, we examined to what extent integrating news and user-generated content for real-time news-related queries could better satisfy end-users. We performed a novel user study using the emerging medium of crowdsourcing, where we had ten different users compare Web search rankings for 297 news-related queries against rankings that integrate newswire articles, blog posts, tweets and Wikipedia pages. From this user study, we showed that newswire articles were not sufficient to satisfy the majority of news-related queries tested, highlighting the need to examine new



(a) 2012_{Apr}



(b) 2013_{Jan}

Figure 8: The percentage of topics where the enhanced rankings were selected divided by topic temporal class (the amount of time between the originating event and topic time).

sources of content. We also showed that integrating blog posts, tweets and Wikipedia pages instead can satisfy many of the queries where integrating newswire articles did not. Finally, our results indicate that users tend to prefer rankings that integrate tweets or newswire articles soon after an event breaks, with blogs becoming more useful over time.

From the results of this study, we believe that new approaches that effectively integrate both newswire articles and user-generated content based upon a combination of relatedness, novelty and timeliness will be able to better serve news-related queries in real-time than using newswire articles alone. For future work, we aim to investigate machine learned approaches to automatically select user-generated content to integrate for news-related queries, as well as further examining what documents are useful to return for such real-time news-related information needs.

7. REFERENCES

[1] O. Alonso. Crowdsourcing for relevance evaluation. In *Tutorial at ECIR'10*.

[2] O. Alonso and R. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. *Information Retrieval Journal* 2011.

[3] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 2008.

[4] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. In *Proc. of SIGIR'09*.

[5] J. Arguello, W.-C. Wu, D. Kelly and A. Edwards. Task Complexity, Vertical Display and User Interaction in Aggregated Search. In *Proc. of SIGIR'12*.

[6] J. Atwood. Is Amazon's Mechanical Turk a failure? `http://www.codinghorror.com/blog/2007/04/is-amazons-mechanical-turk-a-failure.html`, accessed on 28/01/2013.

[7] R. Bandari, S. Asur, and B. Huberman. The pulse of news in social media: Forecasting popularity. In *Proc. of ICWSM'12*.

[8] J. Bar-Ilan, Z. Zhu, and M. Levene. Topic-specific analysis of search queries. In *Proc. of WSDM'09 workshop on Web Search Click Data*.

[9] B. Carterette. *Low-Cost and Robust Evaluation of Information Retrieval Systems*. PhD thesis, University of Massachusetts Amherst, 2009.

[10] C. Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proc. of EMNLP'09*.

[11] C. W. Cleverdon. The significance of the Cranfield tests on index languages. In *Proc. of SIGIR'91*.

[12] F. Diaz. Integration of news content into Web results. In *Proc. of WSDM'09*.

[13] J. Downs, M. Holbrook, S. Sheng, and L. Cranor. Are your participants gaming the system? Screening Mechanical Turk workers. In *Proc. of CHI'10*.

[14] S. Hansell. Google keeps tweaking its search engine. `http://www.nytimes.com/2007/06/03/business/yourmoney/03google.html`, accessed on 28/01/2013.

[15] J. Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Three Rivers Pr, 2009.

[16] J. Howe. The rise of Crowdsourcing. `http://www.wired.com/wired/archive/14.06/crowds.html`, accessed on 28/01/2013.

[17] P. G. Ipeirotis. Demographics of Mechanical Turk. Tech Report, New York University, 2010.

[18] P. G. Ipeirotis. Crowdsourcing using Mechanical Turk: Quality Management and Scalability. In *Proc. of the WSDM'11 workshop on Crowdsourcing for Search and Data Mining*.

[19] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *Proc. of CHI'08*.

[20] A. C. König, M. Gamon, and Q. Wu. Click-through prediction for news queries. In *Proc. of SIGIR'09*.

[21] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media. In *Proc. of WWW'10*.

[22] L. Levon, P. Kaulgud, and S. Skiena. Newspapers vs. blogs: Who gets the scoop. In *Proc. of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[23] R. McCreadie, C. Macdonald, and I. Ounis. Crowdsourcing Blog Track Top News Judgments at TREC. In *Proc. of the WSDM'11 workshop on Crowdsourcing for Search and Data Mining*.

[24] R. McCreadie, C. Macdonald, and I. Ounis. Identifying Top News using Crowdsourcing. *Information Retrieval Journal*, 2013.

[25] J. Norman. Google processes 1,000,000,000 search queries per day. `http://www.historyofinformation.com/index.php?id=3276`, accessed on 28/01/2013.

[26] T. O'Brien. Twitter breaks news of plane crash in the Hudson. `http://www.switched.com/2009/01/15/twitter-breaks-news-of-plane-crash-in-the-hudson/`, accessed on 28/01/2013.

[27] T. O'Reilly and S. Milstein. *The Twitter Book*. O'Reilly Media, Inc., May 2009.

[28] P. Thomas and D. Hawking. Evaluation by comparing result sets in context. In *Proc. of CIKM'06*.

[29] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proc. of EMNLP'08*.

[30] K. Sparck-Jones and C. J. van Rijsbergen. Report on the need for and provision of an "ideal" judgements retrieval test collection. Technical Report, British Library Research and Development, 1975.

[31] K. Zhou, R.Cummins, M. Lalmas and J. M. Jose. Evaluating Aggregated Search Pages. In *Proc. of SIGIR'12*.

[32] D. Zhu and B. Carterette. An Analysis of Assessor Behavior in Crowdsourced Preference Judgments. In *Proc. of SIGIR'10*.