Int. J. Human-Computer Studies 74 (2015) 32–40

Contents lists available at ScienceDirect

Int. J. Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Human content filtering in Twitter: The influence of metadata ☆

Martin J. Chorley *, Gualtiero B. Colombo, Stuart M. Allen, Roger M. Whitaker

School of Computer Science & Informatics, Cardiff University, 5 The Parade, Cardiff CF24 3AA, UK

ARTICLE INFO

Article history:
Received 24 January 2014
Received in revised form
11 July 2014
Accepted 1 October 2014
Communicated by Dr. M. Zanker
Available online 14 October 2014

Keywords:
Decision-making
Twitter
Metadata
Cues
Recognition

ABSTRACT

Social micro-blogging systems such as Twitter are designed for rapid and informal communication from a large potential number of participants. Due to the volume of content received, human users must typically skim their timeline of received content and exercise judgement in selecting items for consumption, necessitating a selection process based on heuristics and content meta-data. This selection process is not well understood, yet is important due to its potential use in content management systems.

In this research we have conducted an open online experiment in which participants are shown quantitative and qualitative meta-data describing two pieces of Twitter content. Without revealing the text of the tweet, participants are asked to make a selection. We observe the decisions made from 239 surveys and discover insights into human behaviour on decision making for content selection. We find that for qualitative meta-data consumption decisions are driven by online friendship and for quantitative meta-data the largest numerical value presented influences choice. Overall, the 'number of retweets' is found to be the most influential quantitative meta-data, while displaying multiple cues about an author's identity provides the strongest qualitative meta-data. When both quantitative and qualitative meta-data is presented, it is the qualitative meta-data (friendship information) that drives selection. The results are consistent with application of the Recognition heuristic, which postulates that when faced with constrained decision-making, humans will tend to exercise judgement based on cues representing familiarity. These findings are useful for future interface design for content filtering and recommendation systems.

1. Introduction

Micro-blogging has become a significant channel of communication that is now widely used on a day-to-day basis around the world (Java et al., 2007), often through Twitter,[1] the current most dominant micro-blogging service. The defining characteristics of this medium are its informality, limited message size and the streamed nature of content provision where users opt-in to receive any content published by users of their choice. These online relationships provide a social network structure through which content is mediated, with users being able to republish or "retweet" received content as they wish (Webberley et al., 2011).

The ease with which content can be published results in a huge volume of potential content, much of this having limited relevance other than to a few users. A user's ability to choose whose content they receive counters to some degree the "long-tail" problem of social media content (Agichtein et al., 2008). However despite this, even following a small number of users can result in a significant real-time stream of tweets that becomes challenging to consume in its entirety. This necessitates the users exercising judgement and decision making to select content of higher priority or interest. This is a subconscious filtering process, deciding in real time whether content in the stream is worth consuming or ignoring. This filtering process may have many inputs, from parts of the tweet text itself (perhaps noticing keywords or hashtags) to metadata describing the tweet or tweet author.

In this paper we examine the role of metadata as cues for human decision making in content selection from the Twitter timeline. Our motivation is to develop findings that are useful in designing systems to support partial filtering of streamed content and reduce the cognitive burden for users. Systems to actively accomplish this are in their infancy (e.g., Yu et al., 2013). This task involves cognitive decision making under constrained conditions and in psychology, three main modelling approaches have emerged (Gigerenzer and Gaissmaier, 2011): logic, statistics or heuristics. Most recently, heuristics have grown in importance. Heuristics are "strategies that ignore information to make decisions faster, more frugally, and/or more accurately than more complex methods". Most significantly, the "less-is-more-effect"

☆This paper was recommended for publication by M. Zanker.

* Corresponding author.
E-mail address: m.j.chorley@cs.cardiff.ac.uk (M.J. Chorley).

[1] http://www.twitter.com

http://dx.doi.org/10.1016/j.ijhcs.2014.10.001

has been discovered (Gigerenzer and Gaissmaier, 2011), representing the existence of an inverted U shape relation between the accuracy of heuristic decision making and the resources (e.g., amount of information, level of accuracy, time) on which the decision is dependent. A range of different heuristics have been presented, including One-Reason Decision Making (e.g., Fishburn, 1974), Trade-off heuristics (e.g., Dawes, 1979), Fluency heuristics (e.g., Schooler and Hertwig, 2005) and Take the best heuristics (e.g., Martignon and Hoffrage, 2002). Arguably one of the most fundamental, and to which many others are related, is the Recognition heuristic (Goldstein and Gigerenzer, 2002) which states: "*If one of two objects is recognised and the other is not, then infer that the recognised object has the higher value with respect to the criterion.*". This is based on the assume that cues based on familiarity can drive human preference. As noted by Gigerenzer and Gaissmaier (2011), a sense of familiarity appears in human consciousness earlier than recollection (Ratcliff and McKoon, 1989).

We investigate to what extent recognition of cues dominate when readers are deciding which content to consume in Twitter. We have designed and built an open online experiment to observe human selection decisions. Participants are repeatedly shown metadata describing two pieces of content without revealing the content itself and are asked to choose which they would prefer to read. The content is taken from Twitter, and all participants are required to be Twitter users. One tweet is taken from the participants 'timeline', the stream of tweets they would usually see when browsing Twitter, so is a piece of content with which they have a relationship (typically they are already following the content author). The second piece of content is unknown to the participant, it is taken from a part of the Twitter network with which they have no relationship, so is a tweet they would not usually see or be exposed to. Without revealing the tweet content itself we control the information about both tweets on which the user is required to state her preference for reading. By varying the information presented on which the reader makes selection, we can ascertain the extent to which metadata cues, both quantitative and qualitative, have impact upon the decision making process.

The remainder of the paper is structured as follows: Section 2 gives an overview of work related to the experiment, Section 3 presents the experimental design and discusses details of the experiment operation and analysis, Section 4 presents and analyses the results of the experiment, while Section 5 summarises the conclusions of the work.

2. Related work

The growth of micro-blogging services (Twitter in particular) in recent years has been of much ongoing interest (Zhao and Rosson, 2009).

One of the reasons for success is the opportunity to post and receive small message updates in real time so as to draw attention to events while they are occurring (Java et al., 2007). The diversity and volume of content has led to much research on recommendation issues, in terms of those to follow Hannon et al. (2010), personalisation (Abel et al., 2011) and collaborative filtering (Cataldi et al., 2010).

However, independent of a user's following strategy, each user is faced with a real-time stream of content that has to be potentially consumed and which is difficult to do so in its entirety. Twitter updates in a user's timeline are visible for only a limited time, so that users are encouraged to frequently check and read what has been posted. This has motivated diverse research to identify ways in which the most relevant content can be filtered or ranked by machine for easy human consumption (e.g., Kapanipathi et al., 2011; Duan et al., 2010), and led to ways in which interactions with Twitter content can inform other online activity such as web search and browsing (e.g., Phelan et al., 2009).

The nature of the Twitter timeline and the need for human extraction of content from a real time stream has led to research concerning supportive technologies. Primarily work in this area has addressed techniques to extract implicit knowledge from Tweets and the social network, from which models for prioritisation can be built. The authors of Das Sarma et al. (2010) focus on the problem of formulating granular ratings, such as star ratings, and compare a range of mechanisms for doing so. Directly ranking micro-blogs is addressed in Nagmoti et al. (2010), proposing a number of different metrics that can be used in training targeted web applications to display Twitter updates in the modalities that best suit each specific user. These ranking methodologies are, however, mainly based on author's characteristics, such as their number of followers or updates posted.

The ranking technique in Weng et al. (2010) is different in applying a PageRank based model to follower/following relationships in Twitter. In Yamaguchi et al. (2010) it is pointed out that many approaches for ranking such as the PageRank algorithm are based on the numbers of relationships among users within the Twitter network while the importance of the content of the updates and differing strength of relationships are often ignored. Similar observations were also noted in Welch et al. (2011) and led to further variants of the PageRank algorithm. To counter this taking into account human activity and the social network, Webberley et al. (2011) focuses on the analysis of behavioural patterns of retweets by linking the user characteristics and physical properties of Twitter updates. Features such as the retweet chain length, time delay and retweet group size show how both tweet content and network structures are equally important to detect interest and define an implicit ranking of the most popular and frequently retweeted updates.

As a means to better provide content to individuals, broad behavioural characteristics of Twitter users have further been investigated in Quercia et al. (2011) through correlation between type of users and their personality, which governs their inherent disposition in many general situations. Users are profiled as belonging to three categories namely 'listeners' (following many users), 'popular' (followed by many others), and 'highly read' (listed in others reading lists). User personality is detected by crawling data from an existing Facebook application based on the OCEAN personality traits model (Goldberg, 1993) and the statistical correlations between the different traits and user types are then calculated. The work concludes by observing that user types are reflected in personality traits and this can be used to support recommendation systems. This work does not attempt to examine how or why users choose to consume content on Twitter as in this paper, but instead seeks to link their social behaviour in Twitter to their OCEAN personality.

A recent article (Morris et al., 2012) describes potential "credibility cues" such as author influence, topical expertise, history of on-topic tweeting and users' reputation as factors in guiding priorities for reading content. However, this work is focused on whether a tweet is a 'credible' source of information, and does not examine the psychological processes that happen when a user is deciding which content to pay attention to in Twitter. The research also examines the effect of user name, user image and tweet content on whether a tweet is perceived to be credible, but again does not look at whether a participant would make a decision to read or consume content based on these information cues.

Beyond this work we conclude from the literature that human interaction with Twitter for content filtering has received little attention yet it remains a fundamental issue for the human consumption of social media. To address this we investigate the

influence of metadata cues and the role that they play as a filtering mechanism for establishing a priority of interest for individual users. This is useful for further understanding human behaviour in information systems but also provides a basis for system design, search and automated filtering.

3. Experimental design

Our experiment, informally titled "TweetCues" is designed to provide evidence on the cues used by humans when deciding a priority in tweets for consumption. Decision making for content consumption is a heavily constrained problem, and this frames the human approach to asserting preference. In such scenarios the human is commonly thought to apply some heuristic based on a subset of variables (Sherman and Corty, 1984), which may be scant and simplistic, such as being led by a familiarity issue or strength of a variable. Our experiments are designed to address the following questions:

Q1 *In the absence of any further information, to what extent do participants prefer tweets that may be recognised as coming from their personal timeline?*
Q2 *Which quantitative metrics do participants use as cues to determine the quality of an unseen tweet?*
Q3 *Do the quantitative metrics or recognition of an existing relationship have the larger effect on the preferences of the participants?*

Q1 specifically addresses whether the Recognition heuristic (Goldstein and Gigerenzer, 2002) applies to such decisions on content. In addition to assessing the applicability of this heuristic Q2 and Q3 aim to discover which Twitter metadata have the greatest effect upon a participant's constrained decision making. The experiment allows us to discover the relative importance of friendship recognition as compared to quantitive measures that may be used to explicitly rank the importance of an unseen message.

The experiment has been implemented as a web application which requires participants to log in using their Twitter credentials (Fig. 1) after which they are asked a series of questions to assess their interpretation of cues when judging the value of unseen Twitter content. We pull content from outside an individual's Twitter feed so that we can sample alternative content and hence make relative comparisons concerning familiarity.

To ensure genuine Twitter users with sufficient tweets in their timeline, subjects are only permitted to continue to the experiment if they have at least 10 followers and are following 10 others. This requirement is intended to ensure that participants have at least some level of engagement within the Twitter social network, and will therefore be familiar with terms such as 'retweet', 'follower' and so on. Once a participant has been accepted they are subsequently presented with a series of questions which display partial information (i.e., selected metadata) about a pair of tweets. In each pair, one of the tweets is chosen from the user's timeline, while the other is chosen at random from a pool of tweets stored by the application, filtered to ensure the tweets have no direct connection to the user.

A subset of the associated metadata is displayed for each of the tweets, and they are asked the following question: 'Given the following information, which of these tweets would you prefer to read?'. The participant then selects the tweet that they would like to read, it is displayed to them, and the process is repeated in succession with different combinations of metadata being exposed. Fig. 2 shows the user being shown some information and being asked to choose which tweet they would like to read (Fig. 2(a)), then being shown the tweet after they have made their selection (Fig. 2(b)).

The design of the experiment is intended to comply with the recognised standards for Internet based psychological experiments (Reips, 2002a,b). To reduce the effect of 'automatic' selection of answers (for example, automatically clicking on all of the left items or answering without paying attention to the questions) the order of the tweets displayed is chosen randomly for each question. This selection bias is also limited by the user being required to click a centrally displayed 'continue' button between each question, so resetting the cursor position on screen for each question.

3.1. QuestionTypes and InfoTypes

In order to separate the influence of different cues on the user, we define 'QuestionTypes' which represent selected metadata concerning a tweet.

The information components from which QuestionTypes are built are called 'InfoTypes' which fall into two categories of cues:

Friendship The first subgroup contains metadata related to the issue of "recognising the tweet's author as a friend" either directly by providing the friendship status of the author, or indirectly by visualising their profile images or displaying their screen name for example. Metadata of this type includes: *Screen name*, *Name*, *Avatar*, and *Friendship* (if the authenticated user follows the author). Note that metadata about friendship is either directly or indirectly revealed. In addition we have a further entity defined by showing all the possible indirect metadata (i.e name, screen-name, and avatar image).

Quantitative The second subgroup contains metadata that indicates a quantitative measure of either some author characteristics within the Twitter network or about the tweet shown. These concern *Follower count*; *Following count* (giving the number of followers of the author or the number of others that he/she follows); *Tweets count* (the number of updates posted by the author); and *Number of Retweets* (for the given update).

Defining QuestionTypes as a combination of InfoTypes allows us to control the particular combinations that are presented as metadata and interpreted as cues by a user.

We have selected a fixed number of possible combinations (twenty five from all possible InfoType combinations) so that each user will receive the same number of questions and each of the considered combinations is shown a set number of times within the same survey. These Questions can be either 'single cue' (to explore the impact of each individual InfoType, or 'combined cue' questions to investigate the impact of multiple and potentially contradictory cues. As an example, Fig. 2 shows the InfoType 'Screen Name' being shown to the user. This meta-data can be seen displayed in the two tables, in the centre of the page in Fig. 2(a), and below the displayed Tweet and 'Next Question' button in Fig. 2(b). In this example, the QuestionType is a single cue QuestionType, as only one piece of meta-data is shown to the user, so the QuestionType is 'Single-Cue Screen Name'. Combined cue QuestionTypes use one (and one only) information entity from each of the 'friendship' and 'quantitative' categories.

In order to minimise the effect of drop-outs for each survey we display a permutation of the twenty five possible QuestionType combinations according to a uniform random distribution, i.e., the order in which the questions are presented is randomly chosen. In this way, we can still include participants who do not complete the full survey in our analysis, as their results do not skew our data

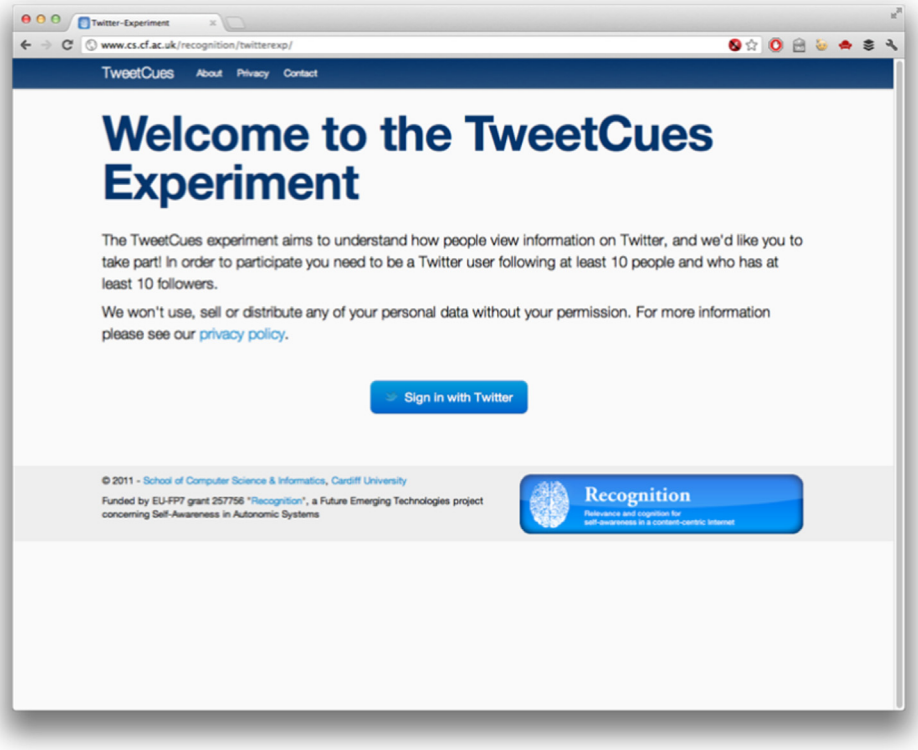

Fig. 1. TweetCues experiment website.

towards one particular QuestionType. Each participant who completes the entire survey is shown three questions for each QuestionType, for a total of 75 questions per survey.

3.2. Crowd-sourced participation

A trial was carried out using a similar experimental setup and results reported in Chorley et al. (2012). This analysis raised further questions surrounding the effects of the placement of tweets and the consistency of users between multiple questions of the same type.

The initial trial study (Chorley et al., 2012) relied on volunteer participants gathered from the local community and through the spread of the experiment on Twitter using tweets from initial participants. In order to gather a larger sample of users for this analysis, participants were paid to participate in the experiment through Amazon's Mechanical Turk[2] (AMT). Each participant was paid $0.50 for taking part through the AMT intermediary 'CrowdFlower'[3], with participants restricted to English speaking AMT users located in Europe, the United States or Australia. The survey was advertised on AMT through June and July 2012 during which time 239 surveys were started, and 227 surveys were completed.

Services such as Amazon's mechanical turk are now being used as a valid source of subjects for experimental and scientific research using this crowdsourcing paradigm, defined as a job outsourced to an anonymous crowd of online workers in the form of an open call (Hoßfeld et al., 2011; Chen et al., 2011). Jobs can be organised and submitted to securely paid online workers through AMT in the form of small tasks.

Other works have examined the use of AMT in a research setting, finding that it enables high quality behavioural online experiments to be conducted and is a useful research tool (Mason and Suri, 2012), and that worker demographics are arguably closer to the population as a whole than traditional university subject pools (Paolacci et al., 2010). In addition, result quality and subjects behaviour are comparable to those of laboratory subjects (Paolacci et al., 2010).

Other experiments achieving similar conclusions are summarised in Mason and Suri (2012), which also discusses a number of techniques to increase accuracy such as the introduction of multiple responses (Snow et al., 2008) and the use of 'reverse Turing test' questions as well as using AMT workers themselves to validate the responses of others' work. Of particular interest is the approach followed in Zhu and Carterette (2010) that suggests a correlation between low-quality responses and low entropy patterns of response, for example repeatedly choosing one option, or alternating between a small number of options in a regular pattern.

Finally, a number of works (Felstiner, 2011; Barchard and Williams, 2008) discuss the security and privacy aspects related to crowdsourcing as well as the novel legal grounds and ethical and issues related to this new paradigm. One of the main drawbacks of internet based experiments is a higher dropout rate than in normal lab experiments (Reips, 2002b). However, this obviously depends on experiment specific issues such as its duration and the cognitive effort required. The duration of our experiment is limited to a few minutes (we recorded an average completion time of around 11 min and 30 s), and in our experiment we observed a completion rate of 95% from 239 participants. The high completion rate is probably due to participants being paid to complete the survey and being required to complete the entire survey in order to receive payment. At the same time, because questions are displayed in random order the effect of dropouts has limited impact on the results.

As previously discussed, our experiment was limited to those Twitter users who are following at least 10 people and are followed

[2] https://www.mturk.com/mturk/welcome
[3] https://crowdflower.com/

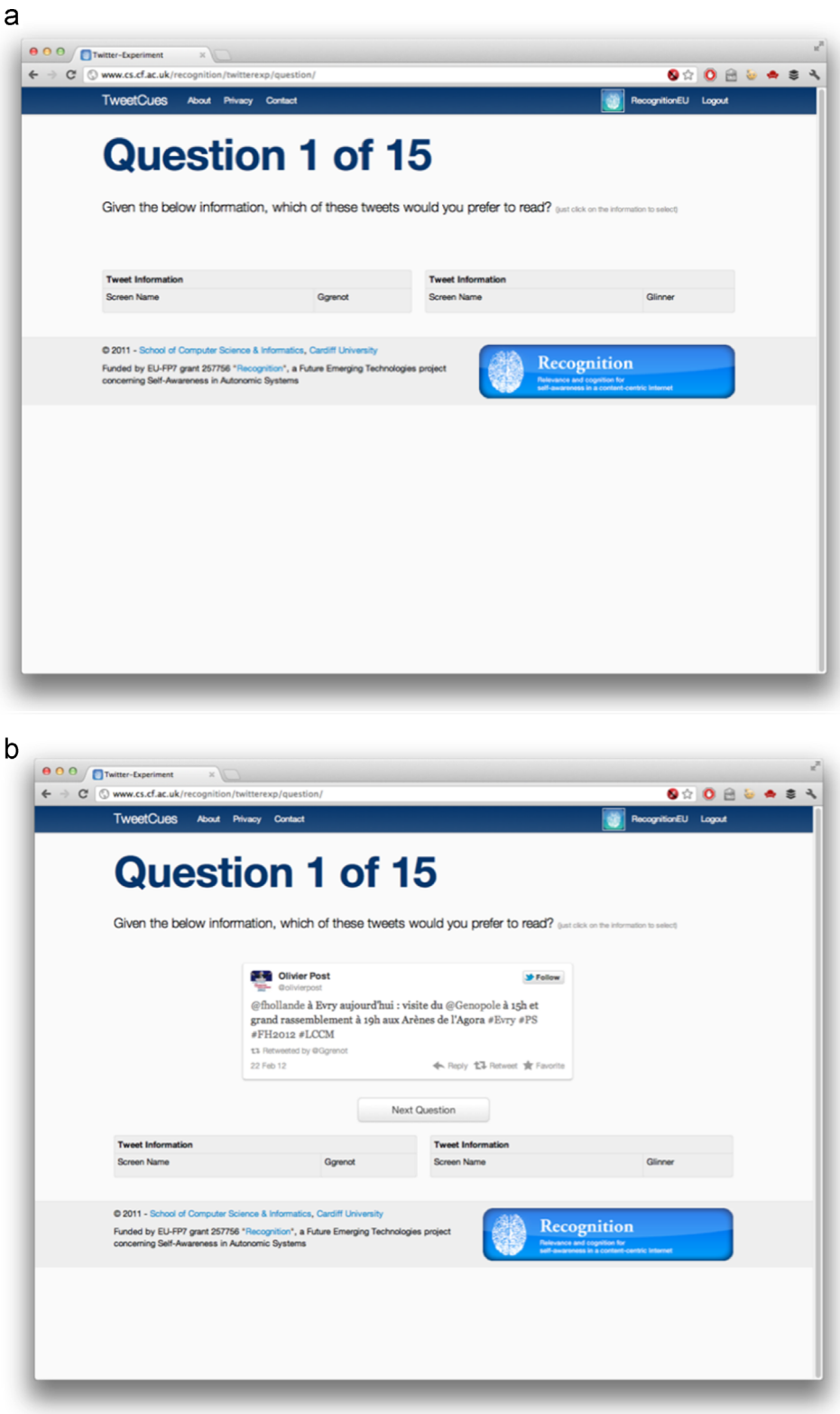

Fig. 2. Example question and selection, TweetCues experiment website. (a) Example Question (before selection) and (b) Example Question (after selection).

by at least 10 other users in order to ensure familiarity with Twitter, its operation and related terms. Average followers, following and tweet counts for users completing the survey are given in Table 1. The participants were drawn from several countries. The largest number of users came from the USA (80.54%), followed by the UK (8.05%), India (3.36%), Spain (2.68%), and Australia (2.01%). Poland, Estonia, Turkey, France and New Zealand each provided 0.67% of participants.

Table 1
Twitter user statistics for experiment participants.

Twitter data	Average	Std. dev
Number of followers	382.87	1585.94
Number following	412.98	1305.74
Number of tweets	3353.41	15,795.21

4. Results

When discussing results, we make reference to several phrases for describing tweets. 'Timeline' and 'Non-timeline' are used to differentiate between a tweet that is taken from a participant's timeline (i.e., the history of received tweets) and those selected from Twitter and published by users with whom they have no relationship. So, for example, if a user is presented with a Tweet written by a user they are following, this is a Tweet that they would normally see in their timeline when visiting the Twitter website, and so is termed a 'Timeline' Tweet. If the user is presented with a Tweet written by a user with whom they have no follower/following relationship, then this is a Tweet they would not normally see when visiting the Twitter website, and so is termed 'Non-timeline'.

When comparing two tweets on the basis of a particular quantitative InfoType, 'Greatest' is used to identify the tweet with the largest value, and the other tweet is identified as 'Smallest'. When considering the tweets selected by a user from the experiment, the proportion P_T refers to the proportion of tweets selected by the experiment subject that belong to her own timeline. P_G refers to the proportion of tweets selected by the experiment subject that are greatest with respect to a particular quantitative cue. In multiple cue questions, it is possible for a tweet to have metadata that provides both the 'Timeline' cue and the 'Greatest' cue. In these cases, the proportion of users selecting this Tweet is denoted as $P_{T,G}$.

This section begins by discussing overall statistics in Section 4.1, before examining whether users are more influenced by timeline metadata or quantitative metadata, looking over all questions (Section 4.2), over single cue questions (Section 4.3), and combined cue questions (Section 4.4). Finally, the results from Single cue and Combined cue questions are compared to discover how each type of cue (Friendship or Quantitative) influences the other (Section 4.5).

4.1. Overall statistics

239 users started the survey, and 227 completed all 75 questions. In total 17,497 questions were answered, with 472 answers coming from users who did not complete the entire survey. The average completion time for those users who finished all 75 questions was approximately 11 min and 30 s. The majority of users completed the survey in between 250 and 1500 s, suggesting response times for each question of between 3.3 s and 20 s per question, which is viable for parsing simple questions. Analysis using a dataset in which users who did not complete the survey are removed revealed no significant differences in results to the analysis using a dataset where non-completing users are included. The fuller dataset is therefore used for the analysis presented in this paper.

Over all surveys, the tweet on the left was selected 8901 times in a total of 17,497 answered questions (51% of the time), while the tweet on the right was selected 8596 times (49.1%) of the time. This difference is sufficient to reject the null hypothesis that *the tweet on the left was selected with equal probability to the tweet on the right*, ($\chi^2 = 5.317, p < 0.021$). It is not believed that this bias is caused by either the Timeline or Greatest tweet appearing on the left with any greater frequency than on the right. The Timeline tweet appears on the left in 50.024% of answered questions, while the Greatest value tweet appears on the left in 50.3% of answered questions. The slight left-hand bias may be introduced by users selecting the left tweet more often as most experiment participants were western and therefore read left to right. However, given the very small size of the bias and the random placement of tweets when displaying questions to participants we do not believe our results to be affected significantly.

4.2. Timeline vs. greatest tweet selection

Over all questions, *the Timeline tweet was selected with greater frequency than the Non-timeline tweet, and the Greatest tweet was selected with greater frequency than the Smallest tweet*. In both cases, Chi-square tests reject the null hypotheses that *the timeline tweet was selected with equal frequency to the non-timeline tweet* ($\chi^2 = 3310.789, p < 0.0001$) and that *the greatest tweet was selected with equal frequency to the smallest tweet* ($\chi^2 = 525.654, p < 0.0001$). The rejection of these hypotheses confirms that the metadata displayed is a factor that influences the decision to select one tweet over the other.

4.3. Selection of tweets based on single cues

Examining only those questions in which a single InfoType is displayed confirms similar results as those seen over all questions. *For questions displaying only a Friendship InfoType, the Timeline tweet is selected in a statistically significantly higher proportion of questions than the Non-timeline tweet. For questions displaying only a quantitative InfoType, the Greatest tweet is selected in a statistically significantly higher proportion of questions than the Smallest tweet.*

Table 2 reports the proportion of questions answered in which users selected the Timeline tweet (P_T) and the proportion selecting the Greatest value (P_G) for all the 'single cue' question types. All proportions are shown to be statistically significant (and so are highlighted in bold) under one sample χ^2 tests with $p < 0.001$.

Table 2
Proportion of tweet selections in which either the Greatest or Timeline tweet were chosen for each of the 'Single Cue' Question Types. Statistically significant ($p < 0.001$) values are highlighted in bold.

QuestionType		InfoType								Proportion	
		Friendship				Quantitative				P_T	P_T
		Screen name	Name	Avatar	Friendship	Follower count	Following count	Tweet count	Retweet count		
Single cue	1	X								**0.744**	–
	2		X							**0.750**	–
	3			X						**0.755**	–
	4				X					**0.804**	–
	5	X	X	X						**0.789**	–
	6					X				–	**0.741**
	7						X			–	**0.613**
	8							X		–	**0.655**
	9								X	–	**0.852**

4.3.1. Number of retweets is the strongest quantitative cue, but friendship is stronger overall

It is worth noting that the proportion of questions in which the greatest tweet was selected is highest with the 'number of Retweets' InfoType, highlighting that this metadata has a larger effect on decision making than the other numerical types. This is in line with the earlier findings on another dataset (Chorley et al., 2012). However, while this earlier work found no significant effect from the other numerical metadata, the analysis presented here based on a comprehensive sample finds that the effect from these information types *is* statistically significant. Also of note is that P_T is higher than P_G in all cases except for Retweets, indicating that *Friendship is a stronger cue in general than the other Quantitative cues*.

4.4. Selection of tweets based on combined cues

In general we cannot consider P_T and P_G to be independent in questions where both Friendship and Quantitative metadata are displayed concurrently. In some of our survey questions, one tweet may be represented by both the Timeline Friendship and a Greatest Quantitative metadata. In other survey questions one tweet may be represented by the Timeline Friendship metadata and the smallest Quantitative metadata, while the other tweet contains the opposite (the Non-timeline Friendship metadata and the Greatest Quantitative metadata). A relationship therefore exists between the Friendship and Quantitative InfoTypes in questions where both are presented at the same time. This is confirmed by performing cross tabulations with the χ^2 statistic to test the independence of the two variables *Timeline* and *Greatest*, which reveals that the majority of combinations (all those except the combinations involving the Following count) are statistically significant with $p < 0.05$, thus rejecting the null hypothesis that *there is no relationship between the Timeline and Greatest variables when presented together*.

4.4.1. Friendship cues are strongest when the timeline tweet and greatest tweet are not the same

Considering the two variables of Timeline and Greatest as dependent, our results show that *in cases where the Timeline tweet is also the Greatest tweet* ($T=G$), *users choose this tweet with significantly higher probability than the Non-timeline & Smallest tweet. In cases where the Timeline tweet is not also the Greatest tweet* ($T \neq G$), *users choose the Timeline tweet with significantly higher probability than the Greatest tweet*. Table 3 shows $P_{T,G}$ for the case where the Timeline tweet is also the Greatest, along with P_T and P_G for the case where the tweet displaying the timeline cue and the tweet displaying the quantitative cue are not combined. Statistically significant values ($p < 0.001$) are shown in bold.

The proportion of questions in which users select the Timeline and Greatest tweet in cases where these coincide ($P_{T,G}$) is above 77% for all combinations, indicating that the recognition of an existing relationship or the presence of a larger numerical cue (or the combination of both) is a significant driver in the selection of which tweet the user would prefer to read.

4.4.2. Number of retweets and follower count are the strongest quantitative cues

$P_{T,G}$ is highest in the case where the explicit Friendship cue is shown along with the number of Retweets. This might be expected, as both cues when shown individually also resulted in the highest values of P_T and P_G respectively (Table 2). This is strong evidence that these two cues are the most significant when users are deciding which content to consume.

In fact, for each of the Friendship cues, $P_{T,G}$ is highest when the cue is shown with the Number of Retweets, and in each case the Number of Followers produces the second highest proportion. This matches the results seen for the Single Cue questions, Retweets produced the highest value of P_G, and the Number of Followers produced the second highest value. This demonstrates the strength of both of these quantitative cues in the decision making process.

4.4.3. Explicit friendship cue is weakest of all friendship cues in combined questions

In cases where $T \neq G$ (the Timeline tweet is not the tweet with the Greatest numerical value), the proportion of users then selecting the Timeline tweet (P_T) is higher than the proportion selecting the Greatest tweet (P_G) in all cases, *except* in the case where the explicit Friendship cue is displayed with the number of Retweets. In this case, it is the number of Retweets that is the strongest cue, and so P_G is significantly higher than P_T. This indicates that when presented alongside other information, an explicit Friendship cue may be weaker than other Qualitative cues. Further evidence for this is given by the question combining Friendship and Follower Count. In this case, although P_T is larger than P_G, the difference is not significant. One possible reason for Friendship being weaker in the combined cue questions may be that it is the only cue that a user is not explicitly shown alongside a tweet when browsing Twitter outside of the experiment, and is therefore not recognised. In combined cue questions users are being presented with two pieces of information about a tweet, one which they recognise as seeing often (the number of retweets, followers, following or tweets), and one which they rarely see (an explicit declaration of a relationship between themselves and a Twitter user). They may then be choosing based on this recognised piece of information, rather than the explicit friendship cue.

Table 3 also shows evidence that in the case of combined questions, the Friendship cue revealing the Name, Screen Name and Avatar is the strongest Friendship cue, with higher values of P_T presented than with the other Friendship cues, no matter which Quantitative cue is shown alongside. So, no matter whether it is shown with a 'strong' numerical cue such as the number of retweets, or a 'weak' numerical cue such as the number following, the combined friendship cue has a larger value of P_T than any other friendship cue. This is in contrast to the Single cue questions, where the 'Name, Screen Name and Avatar' cue resulted in the second highest P_T value, beaten by the explicit Friendship cue.

4.5. Single vs. combined cues

Given that we know the effect a cue has when viewed on its own (Section 4.3), we can consider the results from the combined cue questions as examining *the additional effect on the decision-making process caused by adding a second cue to an existing single cue question*. We can examine this effect by comparing the combined cue results presented in Section 4.4 to the single cue results presented in Section 4.3. Contingency tables with chi-square analysis confirm many of the results from the previous two sections, as described in the following Sections below.

4.5.1. Friendship cues are always stronger

Examining the selection of the Timeline tweet (i.e. P_T, QuestionTypes 10–17 and 22–25 in Table 3), for all cases with the exceptions of the four in which the explicit Friendship cue is displayed, chi-square tests ($p > 0.05$) reveal that *the addition of the quantitative cue does not alter subject decision making from the selection of the tweet belonging to their timeline*. This means that when a user recognises a Tweet as coming from their timeline, the

Table 3
Proportions for the selection of the Greatest and Timeline tweet for the combined cue questions. Statistically significant ($p < 0.001$) values are highlighted in bold.

QuestionType		InfoType								Proportion		
		Friendship				Quantitative				$T=G$	$T \neq G$	
		Screen Name	Name	Avatar	Friendship	Follower Count	Following Count	Tweet Count	Retweet Count	$P_{T,G}$	P_T	P_G
Combined cue	10	X				X				**0.831**	**0.622**	**0.377**
	11	X					X			**0.777**	**0.703**	**0.296**
	12	X						X		**0.820**	**0.642**	**0.357**
	13	X							X	**0.868**	**0.599**	**0.410**
	14			X		X				**0.802**	**0.674**	**0.325**
	15			X			X			**0.779**	**0.726**	**0.273**
	16			X				X		**0.797**	**0.726**	**0.274**
	17			X					X	**0.843**	**0.620**	**0.388**
	18				X	X				**0.811**	0.512	0.488
	19				X		X			**0.779**	**0.622**	**0.377**
	20				X			X		**0.804**	**0.572**	**0.427**
	21				X				X	**0.901**	**0.413**	**0.586**
	22	X	X	X		X				**0.829**	**0.746**	**0.253**
	23	X	X	X			X			**0.804**	**0.801**	**0.199**
	24	X	X	X				X		**0.782**	**0.764**	**0.235**
	25	X	X	X					X	**0.883**	**0.699**	**0.300**

addition of a numerical value that may or may not be larger in the opposing tweet is not a powerful enough cue to result in more users selecting the opposing tweet. In particular, the questions displaying Name, Screen Name, and Avatar (i.e. QuestionTypes 22–25 in Table 3) do not produce significant results at a lower precision level ($p > 0.1$), confirming that this *combination of multiple friendship cues has the strongest influence on subjects for the selection of tweets inside their timeline*. However, for questions combining the explicit Friendship cue with a quantitative value (i.e. QuestionTypes 18–21 in Table 3), the cross table analysis produces statistically significant results at a higher precision level ($p > 0.0001$). This shows that the *addition of the quantitative cue is actually able to drive subjects against selection of their timeline tweets in a significant way*.

4.5.2. Quantitative cues are always weakest

From the point of view of the selection of the tweet having the Greatest quantitative value (i.e. P_G in Table 3) the differences in the selection are statistically significant in all but one case, meaning that *adding a Friendship cue always drives the selection towards the Timeline tweet in a significant way*. Furthermore this effect is more important when the Quantitative cue displayed is either the *Following Count* or the *Tweet Count*, which then both appear as having the weakest impact among the Quantitive cues considered. The only exception to this result is the question combining Friendship with the Follower Count (i.e. QuestionType 18 in Table 3), which does not produce a statistically significant result. This confirms the result observed in the previous subsection about this particular question and the effectiveness of Friendship as a cue in general. It also confirms the previous results describing Number of Retweets and Follower Count as the strongest of the quantitative cues, as each is affected less by the addition of a Friendship cue than the other two quantitative values.

5. Conclusions

The analysis of the experimental results presented in this paper reveals that metadata cues shown to users have a significant effect on the decision making process when those users are selecting which content to consume. These effects vary depending on the metadata presented, and the cues they represent may have a stronger or weaker effect dependent on whether they are presented individually or combined together. Overall, the results indicate that it is the relationship of the user to the content author that has the strongest effect on the decision making process, with metrics describing the quality of the content itself (the number of retweets in this case) also being important.

Considering the first of the original questions posed in Section 3 (**Q1**), it is clear that in the absence of any further information, participants prefer tweets that may be recognised as coming from their personal timeline. This illustrates that the Recognition Heuristic is a definite part of decision making within the content selection scenario, users are choosing the content from an author that they recognise as a signal of familiarity and reassurance. When single cues are provided, users prefer to read the tweet written by a user with whom they already have a relationship in at least 74% of all questions (see Table 2). This proportion is highest for the questions in which the existence of a 'Friendship' relationship is explicitly stated, by telling the user "You follow this author". When friendship cues are combined with numerical values these proportions are reduced, but are still statistically significant, although the explicit 'Friendship' cue becomes weaker than the other implicit friendship cues. Thus telling the user "You follow this author" becomes less important when additional information is also shown.

Where quantitative values are displayed (**Q2**), users select the greatest value in over 61% of questions answered for each cue type, with the 'Number of Retweets' providing the strongest cue (see Table 2).

This is the only quantitative cue displayed relating to the content itself rather than the author of the content. The second most important cue is the 'Follower Count', often seen as a quality indicator within the Twitter social network (the authors producing/sharing the best content will have more followers). The other quantitative cues still produce significant effects and should not be discounted. When quantitative cues are combined with friendship information, the proportion of users selecting the Greatest value are reduced, but remain significant except in the cases where the weaker quantitative cues (Tweet Count, Follower Count and Following Count) are combined with a strong friendship cue (Screen Name, Name and Avatar).

Considering the selection of Timeline or Greatest Tweet as dependent variables (**Q3**) shows that when the Timeline tweet also has the Greatest value it is selected in over 77% of cases (see Table 3). When Timeline and Greatest are in opposition (so the tweet displaying the timeline cue is not the same as the tweet

displaying the largest quantitative cue), it is the Timeline cue that drives selection of the tweet, with the Timeline tweet selected in a higher proportion of answers than the Greatest tweet in all but one case. Again, this strengthens the argument that recognition of an existing relationship with a content author is driving the selection of content. Considering the cases where two cues are presented in opposition, the 'Number of Retweets' cue is shown to be the strongest of the quantitative cues, as the proportion of users selecting against the quantitative cue is reduced in all cases where it is displayed. The cue combining Screen Name, Name and Avatar is shown to be the strongest Friendship cue in such cases, driving selection of the Timeline tweet in higher proportions than any other Friendship cue.

In general it is found that the addition of a quantitative cue to a friendship cue is does not deviate subjects from selecting the Timeline tweet. Additionally, the combination of friendship cues (Screen Name, Name and Avatar) has the strongest influence for the selection of Timeline Tweets. When combined with quantitative cues, the explicit 'Friendship' cue is shown to have the weakest influence. Similarly, when adding friendship cues to quantitative cues, the friendship cue drives selection towards the Timeline tweet and away from the Greatest tweet significantly, providing evidence that user recognition of existing relationships with content authors is a significant driver towards content selection.

These findings have a significant implication for future content recommendation and filtering systems. When presenting content to users within such systems, it may be necessary to present additional information or metadata describing the content alongside the content itself in order to attract the user to consume the content, or to convince the user that the presented content has some utility or relevance to them. In particular, content can potentially be presented to users based on prioritisation from metadata or through using selected metadata itself as a basis for cognitively efficient human decision making. This is an important contribution given the high volume of data that humans must navigate concerning social media. Alerting users to the high value of content based on information such as the number of retweets, or the social network graph linking the user to the content author could have a significant effect on the consumption of the content. Similarly, removing such cues from other content could reduce its importance within the content stream.

Acknowledgements

This research has been supported by RECOGNITION grant 257756, an EC-FP7 Future Emerging Technologies project concerning Self-Awareness in Autonomic Systems, and by an EPSRC Doctoral Award Fellowship, grant EP/L504749/1.

References

Abel, F., Gao, Q., Houben, G.-J., Tao, K., 2011. Analyzing user modeling on twitter for personalized news recommendations. In: User Modeling, Adaption and Personalization. Springer, pp. 1–12.

Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G., 2008. Finding high-quality content in social media. In: Proceedings of the International Conference on Web Search and Web Data Mining. ACM, pp. 183–194.

Barchard, K.A., Williams, J., 2008. Practical advice for conducting ethical online experiments and questionnaires for United States psychologists. Behav. Res. Methods 40, 1111–1128.

Cataldi, M., Di Caro, L., Schifanella, C., 2010. Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceedings of the Tenth International Workshop on Multimedia Data Minin. ACM, p. 4.

Chen, J.J., Menezes, N.J., Bradley, A.D., North, T., 2011. Opportunities for crowdsourcing research on amazon mechanical turk. Hum. Factors 5, 3.

Chorley, M.J., Colombo, G.B., Allen, S.M., Whitaker, R.M., 2012. Better the tweeter you know: social signals on twitter. In: 2012 ASE/IEEE International Conference on Social Computing, pp. 277–282.

Das Sarma, A., Das Sarma, A., Gollapudi, S., Panigrahy, R., 2010. Ranking mechanisms in twitter-like forums. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10, ACM, pp. 21–30.

Dawes, R.M., 1979. The robust beauty of improper linear models in decision making. Am. Psychol. 34, 571.

Duan, Y., Jiang, L., Qin, T., Zhou, M., Shum, H.-Y., 2010. An empirical study on learning to rank of tweets. In: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, pp. 295–303.

Felstiner, A.L., 2011. Working the crowd: employment and labor law in the crowdsourcing industry. Berkeley J. Employ. Labor Law 32 (2011).

Fishburn, P.C., 1974. Lexicographic orders, utilities and decision rules: a survey. Manag. Sci. 20, 1442–1471.

Gigerenzer, G., Gaissmaier, W., 2011. Heuristic decision making. Annu. Rev. Psychol. 62, 451–482.

Goldberg, L.R., 1993. The structure of phenotypic personality traits. Am. Psychol. 48, 26–34.

Goldstein, D.G., Gigerenzer, G., 2002. Models of ecological rationality: the recognition heuristic. Psychol. Rev. 109, 75–90.

Hannon, J., Bennett, M., Smyth, B., 2010. Recommending twitter users to follow using content and collaborative filtering approaches. In: Proceedings of the Fourth ACM Conference on Recommender Systems. ACM, pp. 199–206.

Hoßfeld, T., Hirth, M., Tran-Gia, P., 2011. Modeling of crowdsourcing platforms and granularity of work organization in future internet. In: Proceedings of the 23rd International Teletraffic Congress, ITCP, pp. 142–149.

Java, A., Song, X., Finin, T., Tseng, B., 2007. Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, ACM, pp. 56–65.

Kapanipathi, P., Orlandi, F., Sheth, A.P., Passant, A., 2011. Personalized filtering of the twitter stream. In: SPIM, pp. 6–13.

Martignon, L., Hoffrage, U., 2002. Fast, frugal, and fit: simple heuristics for paired comparison. Theory Decis. 52, 29–71.

Mason, W., Suri, S., 2012. Conducting behavioral research on amazons mechanical turk. Behav. Res. Methods 44, 1–23.

Morris, M.R., Counts, S., Roseway, A., Hoff, A., Schwarz, J., 2012. Tweeting is believing?: understanding microblog credibility perceptions. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work. ACM, pp. 441–450.

Nagmoti, R., Teredesai, A., De Cock, M., 2010. Ranking approaches for microblog search. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 01, WI-IAT '10. IEEE Computer Society, pp. 153–157.

Paolacci, G., Chandler, J., Ipeirotis, P., 2010. Running experiments on amazon mechanical turk. Judgm. Decis. Mak. 5, 411–419.

Phelan, O., McCarthy, K., Smyth, B., 2009. Using twitter to recommend real-time topical news. In: Proceedings of the Third ACM Conference on Recommender Systems, ACM. pp. 385–388.

Quercia, D., Kosinski, M., Stillwell, D., Crowcroft, J., 2011. Our twitter profiles, our selves: predicting personality with twitter. In: Proceedings of the Third IEEE International Conference on Social Computing, SocialCom '11, pp. 180–185.

Ratcliff, R., McKoon, G., 1989. Similarity information versus relational information-differences in the time course of retrieval. Cognit. Psychol. 21, 139–155.

Reips, U.-D., 2002a. Standards for internet-based experimenting. Exp. Psychol. 49, 243–256.

Reips, U., 2002b. Internet-based psychological experimenting: five dos and five don'ts. Soc. Sci. Comput. Rev. 20, 241–249.

Schooler, L.J., Hertwig, R., 2005. How forgetting aids heuristic inference. Psychol. Rev. 112, 610.

Sherman, S.J., Corty, E., 1984. Cognitive heuristics. Handbook of Social Cognition 1, 189–286.

Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y., 2008. Cheap and fast–but is it good?: evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 254–263.

Webberley, W., Allen, S., Whitaker, R., 2011. Retweeting: a study of message-forwarding in twitter. In: 2011 Workshop on Mobile and Online Social Networks (MOSN), pp. 13–18.

Welch, M.J., Schonfeld, U., He, D., Cho, J., 2011. Topical semantics of twitter links. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11, ACM, pp. 327–336.

Weng, J., Lim, E.P., Jiang, J., He, Q., 2010. TwitterRank: finding topic-sensitive influential twitterers. In: Proceedings of the Third ACM International Conference on Web Search and data Mining. ACM, pp. 261–270.

Yamaguchi, Y., Takahashi, T., Amagasa, T., Kitagawa, H., 2010. Turank: twitter user ranking based on user-tweet graph analysis. In: Proceedings of the 11th International Conference on Web Information Systems Engineering, WISE'10. Springer-Verlag, pp. 240–253.

Yu, J., Shen, Y., Xie, J., 2013. Mining user interest and its evolution for recommendation on the micro-blogging system. In: Web-Age Information Management, Springer, Berlin, Heidelberg, pp. 679–690.

Zhao, D., Rosson, M.B., 2009. How and why people twitter: the role that micro-blogging plays in informal communication at work. In: Proceedings of the ACM 2009 International Conference on Supporting Group Work. ACM, pp. 243–252.

Zhu, D., Carterette, B., 2010. An analysis of assessor behavior in crowdsourced preference judgments. In: SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation, pp. 17–20.