Identifying Relevant Social Media Content: Leveraging Information Diversity and User Cognition

Munmun De Choudhury*
Arizona State University
Tempe, AZ 85281, USA
munmun@asu.edu

Scott Counts
Microsoft Research
Redmond, WA 98052, USA
counts@microsoft.com

Mary Czerwinski
Microsoft Research
Redmond, WA 98052, USA
marycz@microsoft.com

ABSTRACT

As users turn to large scale social media systems like Twitter for topic-based content exploration, they quickly face the issue that there may be hundreds of thousands of items matching any given topic they might query. Given the scale of the potential result sets, how does one identify the "best" or "right" set of items? We explore a solution that aligns characteristics of the information space, including specific content attributes and the information diversity of the results set, with measurements of human information processing, including engagement and recognition memory. Using Twitter as a test bed, we propose a greedy iterative clustering technique for selecting a set of items on a given topic that matches a specified level of diversity.

In a user study, we show that our proposed method yields sets of items that were, on balance, more engaging, better remembered, and rated as more interesting and informative compared to baseline techniques. Additionally, diversity indeed seemed to be important to participants in the study in the consumption of content. However as a rather surprising result, we also observe that content was perceived to be more relevant when it was highly homogeneous or highly heterogeneous. In this light, implications for the selection and evaluation of topic-centric item sets in social media contexts are discussed.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems; J.4 [**Social and Behavioral Sciences**]: Sociology

General Terms

Algorithms, Experimentation, Human Factors.

Keywords

Cognition, Information Seeking, Information Selection, Real-time Search, Social Media, Twitter, User Interfaces.

*This work was performed while the author was an intern at Microsoft Research, Redmond.

1. INTRODUCTION

Social media sites like Twitter continue to expand rapidly, garnering users who are forming more connections, sharing more information, and finding new ways to appropriate these communication media spaces. Apart from being conducive avenues for users to express their thoughts and opinions, as well as share their everyday experiences, these platforms have begun to evolve as mechanisms to reflect and reveal news related information on a range of real-world events [20, 22]. To take examples, Twitter played an instrumental role during Barack Obama's 2008 presidential campaign, the 2009 demonstrations in Iran, and in Moldova's "Twitter Revolution" in Eastern Europe in 2009 (ref. CNN). Consequently, today Twitter serves tremendous potential to cater to information seeking and exploration on timely happenings [18, 20]. This is also supported by the statistics that by April 2010, Twitter was receiving over 600 million search queries per day[1].

Our Goal. The central question addressed in this paper is: *how do we identify the most "relevant" or "'best" set of items on a given topic, from millions and even billions of units of user generated content on social websites?*

Challenges. Retrieving relevant social media content for the end user given a certain topic is a challenging task not only because the social media information space exhibits profuse scale and exceedingly high rate of growth, but also because it features *a rich set of attributes* (e.g., network properties of the content author, geographic location, timestamp of post, presence of multiple themes and so on). Affected by these attributes, the information authorities of relevant social media content are, therefore, likely to be emergent and temporally dynamic. This, in turn, renders the content deemed relevant over a given topic, to be temporally changing as well. Hence approaches utilizing static structural metrics (such as HITS, PageRank) might not suffice in this context because they are likely to point to the celebrities, journalists, A-list bloggers or government bodies whose posted content might not be deemed relevant to the end-user at all points in time. Consequently, it appears that traditional search engines, such as Google and Bing are not well equipped with the capability of searching for social media content (also see [1]).

However, there have been recent attempts to tackle the problem of retrieval of social media content in a commercial setting. This motivated us to undertake a short background survey in an organization to understand these current state-of-the-art retrieval techniques. The survey (discussed in greater detail in section 4) involved asking a set of individuals about the different tools they use for

[1]Huffington Post. Twitter User Statistics Revealed: http://www.huffingtonpost.com/2010/04/14/twitter-user-statistics-r_n_537992.html, Apr. 2010

Table 1: Usage of current social content search / exploration tools, based on an organizational survey.

INTERFACES/TOOLS	#RESPONSES
Twitter website	50
Search engines, such as Bing Social	25
Twitter clients, such as Tweetdeck, Twitterific etc.	19
Third party apps, such as Twitter plug-in for Google	9

exploring and searching Twitter for tweets on a given topic. The responses from the survey are given in Table 1. We observe that the two highly used tools are the native search engine by Twitter, and second, the social search tool developed by Bing (Bing Social).

However, we note that the retrieval mechanisms on both of these tools do not adequately address the challenges discussed in the previous paragraphs, because they rely on content presentation based on a *fixed attribute*, ignoring the rich span of attributes that the Twitter information space features. For example, while Twitter search gives a list of tweets on a topical query that are ordered reverse chronologically (i.e. most recent tweets), there is no scope for the end user to seek content that might be posted by authors in geographically disparate locations, or content that includes pointers to external information sources via URLs. Although Bing Social goes one step beyond the temporal recency attribute, and yields URLs that have been shared widely among users on Twitter, the end user might still intend to seek content that have been conversational on Twitter (to know about conflicting or agreed upon opinions), or wish to see tweets spanning a variety of themes on a topic (e.g., political versus economic perspectives).

Hence it is intuitive that while exploring or searching for social media content on a given topic, an end user might like information filtered by only a specific attribute (i.e. information that is *homogeneous*), or can be interested in content that features a "mixing" over a wide array of attributes (i.e. information that is *heterogeneous*). We take an example for each case. Suppose an end user is looking for relevant Twitter content after the release of the Windows Phone in November 2010. It would be natural to display tweets that are homogeneous in terms of authorship, i.e. tweets posted primarily by the technical experts. On the other hand, if the user wanted to learn about the oil spill in the Gulf of Mexico that took place in summer of 2010, a good set of social media items for the user to explore would span over a range of attributes like author, geography and themes such as Politics or Finance.

1.1 Diversity in Social Media Content

Given the above challenges and observations, we are motivated to utilize the rich attribute-based characteristics of the user-generated social media content in identifying relevant information. Because social media information spaces feature a wide variety of attributes and since end users might want to seek content on any combination of these attributes, we refer to this characteristic property of social media content as its "diversity". That is, the diversity of a set of social media items characterizes the range of attributes considered (e.g., geography, author characteristics, etc.) and the values of those attributes (e.g., for geography, all content from the same region versus from around the globe). This notion of diversity in the context of social media is supported by extensive prior literature in different areas ranging from economics, ecology and statistics [10], where the diversity index of a sample population has been widely used to measure the differences among members of a population consisting of various types of objects.

Our hypothesis is that the diversity property needs to be incorporated into the framework of identifying relevant social media content in order to regulate the degree of desired homogeneity or heterogeneity of the information presented. This is because prior research in consumer markets [28] suggests that individuals' involvement and perception of items differs significantly depending on the attributes of the item presented. Additionally, in the context of social media research, it has been observed that the information consumption process is often affected by a variety of attributes of the content author apart from his/her identity, ranging from relationships between identity presentation of the author and perception of the reader to the interpretation of temporality to the reader [4]. In other words, in our specific context, there can be different sets of attributes, or variable degrees of *information diversity* across those attributes in the social media information space, that an end user is likely to find useful while seeking relevant information on a topic.

1.2 Cognitive Measures & Content Relevance

Note that an outstanding challenge in this problem is the subjective notion of relevance; and hence how to assess the quality of topic-centric sets of social media content, especially in the face of absence of any ground truth knowledge. Relevance performance has traditionally been addressed objectively in information retrieval contexts using metrics such as precision/recall [3], new ranking mechanisms [13, 19], relevance feedback [3], eye gazing patterns [6], quantifying expected contribution of the retrieved information in accomplishing the end user task [26] and so on. However except for a very few pieces of prior research that has considered user perception based metrics in the context of information retrieval [12, 15, 27], evaluation of the subjective notion of relevance remains fairly under-investigated.

Our perspective on this issue stems from the observation that relevant social media content is likely to streamline the end user's cognitive information comprehension experience. Hence in this work, we propose to evaluate the quality of topic-centric results sets by measuring aspects of human information processing when end users are engaged with the social media content. Because there may not be a clear best result in the same way that there is a best web page result for many web queries, we assume that the best information will be interpreted as *interesting* and *informative*, and will be more *engaging* to the user during reading [11] and *better remembered* later (i.e. better encoded in the human long-term memory) [24, 25].

1.3 Our Contributions

In this light, the following are the major contributions of the work presented in our paper:

1. We characterize social media information spaces through an entropy-based measure known as *diversity* that captures the relative representation of different attributes featured in the information. We further identify the importance of these different informational attributes in the social media space, based on feedback from users at a large corporation.

2. We propose a methodology to identify relevant social media content. The proposed framework is motivated by information theoretic concepts and is based on a greedy iterative clustering technique. It uses the attribute representation of the social media space developed in (1) to construct relevant item sets on a given topic, matching a desired level of diversity.

Note that we do not make apriori assumptions about what degree of diversity of the information space is more desirable for the content selection task. Instead, diversity is considered a parameter

in our experimental design, and we provide discussions on how the choice of its value affects the end user's perception of the information consumed.

We performed an elaborate user study involving 67 participants at a large corporation to evaluate our proposed method on Twitter Firehose data. The study entailed showing the participants tweet sets on a range of topics and thereafter seeking their feedback along the lines of different cognitive measures. There are two key observations in the results of our user study:

1. First, our proposed method outperformed baseline techniques in yielding tweet sets that were perceived as interesting, informative, engaging and memorable (by a margin of $\sim$25-30%): validating the utility of incorporating the diversity aspect of social media in the content selection framework.

2. Second, somewhat surprisingly, we found that participants found content to be of better quality (in terms of the four cognitive measures) when they were of very low diversity (i.e. highly homogeneous) or when they were extremely diverse (i.e. highly heterogeneous).

The rest of this paper is organized as follows. We review related literature in the next section, and then formalize our problem definition in section 3. In section 4 we discuss a short survey that was conducted to evaluate the importance of the user-derived attributes, followed by the proposed content selection methodology in section 5. Section 6 gives our result set generation methodology on Twitter data along with several baseline techniques for comparison. Sections 7 through 9 present empirical observations on our proposed method, and present results of our method's performance based on a user study. Finally we discuss some open questions and conclude with our major contributions in sections 10 and 11.

2. RELATED WORK

There has been considerable prior research on recommending, filtering and searching social media content on the web [1, 5, 7, 13, 20, 23]. More recently, to tackle the issue of the availability of large scale social media content, Bernstein et al. [5] proposed a Twitter application called "Eddi" that allows users to quickly find popular discussions in their Twitter feed by searching, or navigating a tag cloud, timeline or categories. In other work, Chen et al. [7] explored three dimensions for designing a recommender of social media content: content sources, topic interest models for users, and social voting.

While this prior work has attempted to address the issue of how to manage and present relevant content for large repositories of social media content, no principled way of selecting or pruning such large spaces has been proposed. To the best of our knowledge, this is the first time methods to select topic-centric information from large social media information spaces are being investigated, particularly with regard to human information processing.

We discuss some literature in the light of using diversity in recommendation and information retrieval tasks. There has been a significant amount of work focused on diversification of recommendation items or search results [2, 8, 9, 16, 19, 21, 29]. Ziegler et al. [29] proposed a similarity metric between items in a recommended list to assess its topical diversity and thereafter a topic diversification approach for decreasing the intra-list similarity. Clarke et al. [9] address the problem of ambiguity and redundancy in information retrieval systems with the help of a cumulative gain based evaluation measure to increase novelty and diversity in search results. In the context of social tagging, Chi et al. [8] used a mutual information based measure to infer that diverse tags better describe shared items in a social bookmarking setting than popular tags.

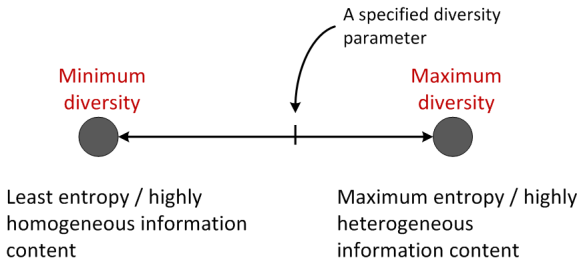

Figure 1: Visual representation of diversity spectrum featuring entropy of the social media information space.

To summarize, research in this direction has primarily focused on maximizing information gain in retrieval paradigms by presenting the end user with diverse content. However we do not have a clear insight into how the user perception of information relevance changes with high and low diversity, or whether highly homogenous or highly heterogenous content is more desirable—developing this understanding is a major motivation in this work.

3. PROBLEM DEFINITION

We begin by formalizing our problem definition.

Diversity Spectrum. Recall that social media content today can be viewed as having a wide array of attributes, ranging from numerous geographic locations, the extent of diffusion of the topic in the associated social network, and so on. As a consequence, social media information spaces are inherently *diverse*. Therefore, we conjecture that the content presented to an end user should match a certain level of diversity, that is cognitively conducive to his or her information consumption process.

In this light, we define a conceptual structure that characterizes the nature of the social media information space in terms of "entropy" [10, 17]. Entropy quantifies the degree of "randomness" or "uncertainty" in the data featured by its attributes in an information theoretic sense. We call this structure the "diversity spectrum" (schematic representation shown in Figure 1). In the diversity spectrum, the two ends of the continuum represent content that is *homogeneous* (i.e., information homophily) and content that is *heterogeneous* (i.e., information heterophily). Any point on the spectrum can be specified in the form of a *diversity parameter* (referred to as ω), which is any real value on the spectrum, in the range $[0, 1]$.

We identify that although there are a host of measures to estimate such diversity (e.g., species richness, concentration ratio, etc.), the most popular and robust measure by far is Shannon's entropy based quantification [17]. Note that entropy based measures have also been used in the past to characterize other forms of social content, such as email traffic [14].

Social Media Content Attributes. We define the attributes along which we characterize social media content on topic. A description of the different attributes used in our work is given in Table 2. We note here that because we use Twitter as our test social media platform, some of our content attributes are Twitter-specific. For example, we noted that content shared on Twitter has distinct attribute along which information is typically dissipated, such as retweets, @-replies, presence of URLs and thematic distribution of the tweets. Moreover, the authorship of the content is also likely to play a significant role in consumption of information to an end user: hence we include attributes such as author location, as well as

Table 2: Description of different social media content attributes (posts on Twitter, or tweets, in this context).

1.	Diffusion property of the tweet—measured via whether the given tweet is a "retweet" (RT tag).
2.	Responsivity nature of the tweet—measured via whether a given tweet is a "reply" from one user to another.
3.	Presence of external information reference in the tweet—whether the tweet has a URL in it.
4.	Temporal relevance of the information, i.e., time-stamp of posting of the tweet.
5.	The thematic association of the tweet within a set of broadly defined categories—such as "Business, Finance", "Politics", "Sports" or "Technology, Internet". This association is derived using the natural language toolkit, OpenCalais (www.opencalais.com) that utilizes the content of the tweet, as well as the information about any URL that it might contain, to return a thematic distribution over the tweet.[2]
6.	Geographic dimension of the tweet—measured via the time-zone information on the profile of the tweet creator.
7.	Authority dimension of the creator of the tweet—measured via the number of followers of the user who posts the particular tweet.
8.	Hub dimension of the creator of the tweet—measured via the number of followings / friends of the user who posts the particular tweet.
9.	Degree of activity of the creator of the tweet—measured via the number of statuses of the user who posts the particular tweet; i.e., the number of tweets the creator had posted up to that point in time.

network structure related attributes like #followers and #friends.

Additionally we acknowledge that we have focused on a finite set of attributes to characterize the tweets. However, a host of additional attributes might be relevant. Potential attributes include sentiment or linguistic style of the content, relationship strength between the creator of content and the consuming end user, community attrition of the creator and the consumer, sophisticated network metrics of the consumer, such as clustering coefficient or embeddedness, and so on. Incorporating such attributes might prove to be useful especially while personalizing the recommendation of social media content to users.

Problem Statement. Given, (1) a stream of tweets from all users in a time span, and filtered over a certain topic θ, say, $\mathcal{T}_\theta$; (2) a diversity parameter ω; and (3) a set size s, our goal is to determine a (sub-optimal) tweet set, $\mathcal{T}^*_\omega(s)$, such that its diversity level (or entropy) is as close as possible to the desired ω and also has a suitable ordering of tweets in the set in terms of the entropy measure. This involves the following steps: (a) Estimating the importance of the different attributes that characterize the entire tweet information space (section 4); (b) Developing a greedy optimization technique to construct a tweet set that matches the desired diversity ω, and finally organizing the tweets in it based on the relative distances of their entropies from diversity ω (section 5).

4. USER RATINGS OF ATTRIBUTES

We begin by discussing the aforementioned background survey. The focus of the survey was to have users rate, for importance, each of the different content attributes (ref. section 3), while assessing the quality of Twitter content. The survey solicited responses from 11 active Twitter users[3]. These users were employees at a large technology corporation (8 male, 3 female; median age 25). Each participant was requested to rate each of the tweet attributes on a scale of 1 through 7, where 1 implied "not important at all", and 7 meant "highly important". The survey also allowed them to identify other attributes that they might think to be significant in aiding in the exploration/search of Twitter content.

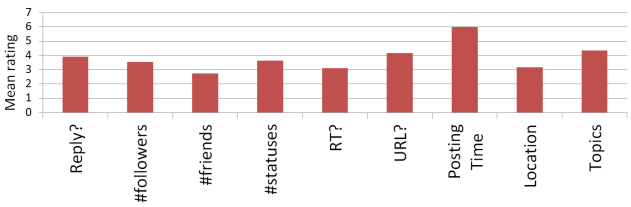

Figure 2: Ratings (on a scale of [1–7]) of different attributes characterizing tweets. These ratings correspond to the weighting of the attributes in the content selection methodology.

As can be seen in Figure 2, the importance of the attributes varied, with posting time rated notably higher than the others. However, we were not interested in comparing these values statistically. Instead, as described below, we utilized this user-generated configuration of importance ratings of the attributes by incorporating them as weights for content selection.

5. CONTENT SELECTION METHOD

We now present our proposed social media content selection methodology. The goal is to identify sets of topically relevant content by leveraging the notion of information diversity and the rich attribute structure of social media information spaces. We start with a filtered set of tweets $\mathcal{T}_\theta$, or simply $\mathcal{T}$ corresponding to the topic θ. For each tweet $t_i \in \mathcal{T}$, we develop a "vectored" (and weighted) representation of t_i, based on its values for the different attributes (ref. previous section). Let $t_i \in \mathbb{R}^{1\times K}$ be the attribute representation of a tweet for a set of K attributes.

Thereafter, the following are the two major steps in our content selection methodology. First, we determine a set of tweets of a certain size s, such that it corresponds to a pre-specified measure of the diversity parameter on the diversity spectrum, given as ω. We refer to this step as entropy distortion minimization. Second, we develop an organizational framework for the selected set of tweets in the set, such that it enforces ordering on the nature of the content in terms of entropy. These two steps are described as below.

5.1 Entropy Distortion Minimization

At the heart of our content selection methodology is a greedy iterative clustering technique that yields a set of tweets of size s on a given topic corresponding to a pre-specified diversity. To construct the set $\mathcal{T}^*_\omega(s)$ for a topic with diversity ω, we start with an empty set, and pick any tweet from $\mathcal{T}$ at random. We iteratively keep on adding tweets from $\mathcal{T}$, say t_i, such that the distortion (in terms of ℓ_1 norm) of entropy of the sample (say, $\mathcal{T}^i_\omega$) on addition of the tweet t_i is *least* with respect to the specified diversity measure ω. That is, we iteratively choose tweet $t_i \in \mathcal{T}$, whose addition gives the minimum distortion of normalized entropy[4] of

[2]Note that the set of topics is pre-defined by the OpenCalais domain; hence making the topical associations of tweets to be semantically meaningful.

[3]By "active" users, we filtered participants based on their frequency of activity on Twitter. We restricted our survey to users who use Twitter at least twice a week.

[4]Normalized entropy of a distribution is given as the ratio of the entropy of the distribution to the maximum entropy of its given dimensions.

$\mathcal{T}_\omega^i$ with respect to ω, where ω is simply the pre-specified diversity parameter, as specified on the diversity spectrum. This can be formalized as follows: $t_i \in \mathcal{T}_\omega^i$ if and only if, $\|H_O(\mathcal{T}_\omega^i) - \omega\|_{\ell_1} < \|H_O(\mathcal{T}_\omega^j) - \omega\|_{\ell_1}, \forall t_j \in \mathcal{T}$, where $H_O(\mathcal{T}_\omega^i)$ is the normalized entropy given as, $H_O(\mathcal{T}_\omega^i) = -\sum_{k=1}^{K} P(t_{ik}) \cdot \log P(t_{ik})/H_{\max}$, and $H_{\max}$ being given as $\ln K$.

We continue the iterative process of adding a tweet t_i to the sample $H_O(\mathcal{T}_\omega^i)$ until we achieve the requisite size s. Finally, we get the optimal tweet set as: $\mathcal{T}_\omega^*(s)$.

5.2 Content Organization

We now present a simple entropy distortion based organization technique of the tweets in the set $\mathcal{T}_\omega^*(s)$. Our central intuition is that the ordering should be based on how close a particular tweet $t_i \in \mathcal{T}_\omega^*(s)$, in terms of its different attributes K, is with respect to the specified diversity parameter ω. Hence we compute the distortion of the normalized entropy of tweet t_i, given as $H_O(t_i)$, with respect to ω. The lower the ℓ_1-norm distortion, the higher is the "rank" or position of the tweet t_i in the final set presented to a user.

6. GENERATING TWEET SETS

We now discuss the generation of tweet sets for content exploration based on Twitter data. We utilized the "*full Firehose*" of tweets and their associated user information over the month of June 2010. This dataset was made available to our company through an agreement with Twitter. The different pieces of information we used in this paper (in anonymized format) were: tweet id, tweet text, tweet creator's id, tweet creator's username, reply id, reply username, posting time, tweet creator's demographics, such as number of followers, number of followings, count of status updates, time-zone and location information. The entire dataset comprised approximately 1.4 Billion tweets, with an average of 55 Million tweets per day.

The data were segmented into 24-hour long logs, a vectored and weighted representation of the tweets was generated based on the content attributes and user ratings discussed in sections 3 and 4. This tweet space was filtered for tweets on a certain topic (e.g. "oil spill", "iphone") based on string matching. Finally the proposed content selection method was run on each of them, given a pre-specified diversity parameter value. This process generated tweet sets with three pieces of information for each tweet: the tweet content, the username of its creator and its posting time. The size of the tweet sets was set to pre-specified "sizes", such as a 10-item sized tweet set. Note that although our proposed content selection technique can generate tweet sets of any given size, we considered sets of a reasonably small size (10 items) in our user study. The goal was to ensure that while going through the user study and evaluating different sets, the end-user participant was not overwhelmed by the quantity of information presented. However empirical evaluations were conducted for a range of tweet set sizes between 10 and 100.

Baseline Techniques. Using the same data as above, we also generate tweet sets for several baseline techniques that can enable us compare the effective of our proposed method. We first propose three baseline techniques that were simplified variations of our proposed technique. We primarily consider two aspects of our algorithm: entropy minimization and attribute weighting, and craft baseline techniques along these variables. Note that in cases when entropy minimization is not used, tweets are selected based on a random range of entropies. Relationship of these baseline techniques with our proposed method (henceforth referred to as PM) is shown in Table 3. Tweet sets of size 10 as above were generated for each of these techniques for the user study purpose.

We also use two versions of current state-of-the-art methods (i.e. similar to Twitter search and Bing Social). One of them is called the "Most Recent" or (MR) method, where we generate a set of tweets of a pre-specified size, based on their timestamps of posting. Filtered by a topic, the tweet set comprises the tweets with the 'most recent' timestamp on the particular day under consideration. The last baseline technique is called "Most Tweeted URL" (or MTU), where we determine all the URLs that were shared (via tweets) on the particular topic and on a given day. Thereafter we sort them by the number of times they were mentioned in different tweets throughout the day. We generate the tweet set of a certain size s, by selecting the top s most-tweeted URLs from the sorting process; and then yielding the "first" tweet on the same day that mentioned each of the s URLs.

Table 3: Different content selection techniques used in the experiments, based on variants of our proposed method.

Baseline 1 (or $B1$)	×	entropy minimization
	×	attribute weighting
Baseline 2 (or $B2$)	×	entropy minimization
	✓	attribute weighting
Baseline 3 (or $B3$)	✓	entropy minimization
	×	attribute weighting
Proposed Method (or PM)	✓	entropy minimization
	✓	attribute weighting

Table 4: Example tweet-sets generated using various content exploration techniques.

Most Recent (MR)
Some oil spill events from Monday, June 7, 2010: A summary of events on Monday, June 7, Day 48 of the Gulf of Mexico http://bit.ly/9HNG9Z
RT @DAYLEE that! Broken pipe is not NATURAL! RT @RayBeckerman FreedomWorks CEO, Calls Oil Spill Natural Disaster http://bit.ly/coUY4l
Is there a way to help save the wildlife affected by the oil spill?
Most Tweeted URL (MTU)
RT @TEDchris: A Gulf oil spill picture I will never forget. http://twitpic.com/1toz8a
Citizen Speaks The Truth ON BP Gulf Oil Spill–the Govt, BP Are Doing Nothing, There Are No Leaders Here http://bit.ly/BP-Gulf-Oil-Spill
RT @OliBarrett: Visualizing the BP Oil Spill http://www.ifitwasmyhome.com/
Proposed Method (PM)
Oil spill cap catching about 10,000 barrels a day: LONDON? BP's oil spill cap, designed to stop a huge leak http://oohja.com/xeWhD
How to Help with Oil Spill Aftermath. Links to sites to donate, volunteer, and support. http://ow.ly/1UKso
Looking for Liability in BP's Gulf Oil Spill: White Collar Watch examines the potential criminal and civil liability. http://nyti.ms/9lUMaT

7. EMPIRICAL OBSERVATIONS

We first present empirical observations to understand the effectiveness of our proposed method in generating high quality relevant tweet sets on a topic. In Table 4 we present sets of tweets constructed by two of the baseline content selection techniques (MR and MTU) investigated in this paper. Note that the content generated by PM (using a medium diversity parameter value of 0.5) is overall better in a qualitative sense compared to the two others: it seems to have pointers to external information sources, via URLs, regarding the business and political developments around oil spill, reveals some socio-economic aspect of the oil spill incident as well as information on the state of the spill.

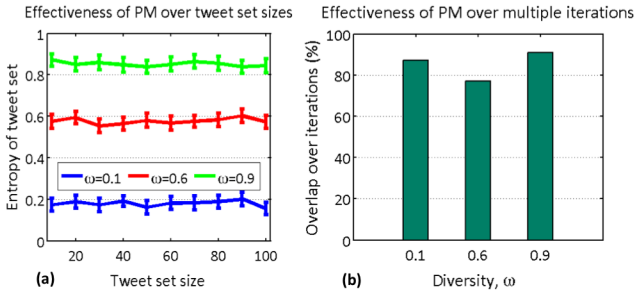

Figure 3: Experiments to illustrate the effectiveness of our proposed method PM. We show the mean entropy of the samples generated by PM for different tweet set sizes (between 10 and 100) in (a). In (b) we show the mean overlap of content in the generated samples across multiple iterations of PM, i.e. choice of different seeds across iterations. Both (a) and (b) are reported averaged over two topics "oil spill" and "iphone" and over 100 repeated iterations. The standard error bars in (a) report the deviation over repeated runs of the algorithm.

We also conduct some experiments to investigate the effective of PM in the light of scalability (i.e. choosing different sizes of tweet sets on a topic), as well as robustness over multiple iterations (i.e. choosing different tweets to seed the greedy iterative clustering during the entropy minimization step). From the results in Figure 3 we observe that for three different levels of diversity, we are able to generate tweet sets, (a) whose entropies are sufficiently close to the corresponding ω (note that all the three plots are relatively flat over tweet set size); and (b) whose percent overlap across consecutive iterations are fairly high ($\sim$78-91%). Both of these experiments show that our proposed method PM is consistent across the size of the set chosen as well as across choice of different seed tweets in repeated iterations.

8. USER STUDY

In the absence of ground truth to validate the tweet sets generated by different baseline techniques and our proposed method, we conducted a user study. Sets of tweets generated by the different methods were shown to participants in order to determine for which method the presented content was considered most interesting, most informative, the most engaging and memorable.

8.1 Method

Participants. Participants were 67 employees of a large technology corporation who were compensated for their time with a $10 lunch coupon. Participants were 'active' Twitter users who used Twitter at least two times per week. The median age was 26 years, although the age range was fairly wide spanning between 19 to 47 years; whereas Male/Female ratio was approximately 60/40%.

Stimuli and Procedure. A web-based application was developed for the purposes of our study (Figure 4). Using the site, participants were presented with a task of conducting a "real-time search" on a topic on Twitter over one of two topics, "Oil Spill' or "iPhone", that had been found to be of temporal relevance during the month of June 2010. The duration of the study was 20-30 minutes.

1. *Part I:* Each participant was presented with 12 sets of tweets spanning a topic, either "Oil Spill' or "iPhone". Participants saw tweets for only one topic, and topic assignment was random[5]. Each set contained 10 tweets, along with their corresponding usernames and the time of creation. After each sample, the participant was asked, (i) to estimate the length of time spent reading the tweets, (ii) the interestingness of the tweets (on a scale of 1 to 7), (iii) how diverse the tweets were in terms of their content, and (iv) how informative the content of the tweets was.
2. *Part II:* The participants were asked to respond to a survey on demographics and general activities on Twitter, such as frequency of tweet posting and searching behavior in Twitter. In addition to allowing us to collect demographics, this served as a "filler task" before a tweet recognition test (given in Part III).
3. *Part III:* Each participant was presented with a list of 72 tweets (randomly chosen $3 \times 12 = 36$ tweets from the 12 sets in Part I, while the remaining 36 tweets on the same topic but did not appear in the sets shown in Part I). Participants were asked to recognize, via "Yes"/"No"questions, if they had seen the tweets in any of the sets presented earlier.

8.2 Measures

We included four dependent measures to evaluate user performance with the different content selection techniques. Our measures fell into two categories that we refer to as explicit and implicit:

- *Explicit Measures.* Explicit measures consisted of two 7-point Likert scale ratings made after reading each tweet set (see middle section of Figure 4). The ratings corresponded to the following three aspects of tweet set quality as perceived by the participant: interestingness and informativeness.
- *Implicit Measures.* We used two measures considered to be implicit, because they were not based on direct, explicit evaluation by participants. The first implicit measure is motivated from prior literature on subjective duration assessment [11], and we refer to it as "cognitive engagement". It is computed using the function: $[(D_i - \hat{D}_i)/D_i]$, where D_i and $\hat{D}_i$ are respectively the actual and perceived time taken to go through the i-th tweet set by a participant. Note that ideally, if the information presented in the i-th tweet set is very engaging, the participant would underestimate the time taken to go through the tweet and the cognitive engagement measure would be a positive value [11]. In less engaging scenarios, engagement has been shown to be negative; hence, relative comparison across engagement measures of different techniques seems reasonable. Our second implicit measure, collected in Part III, was recognition memory for tweets seen in Part I verses unseen tweets. It is derived as: $[n_i(\text{'yes'})/n_i]$, where $n_i(\text{'yes'})$ is the number of tweets from the i-th set that a participant correctly recognized as having seen in Part I and n_i is the total number of tweets from the same set that appear in the recognition test. More memorable content read in Part I should generate better scores on this recognition task.

8.3 Design and Predictions

Our study was a 2 (topic: oil spill, iphone) $\times$ 3 (level of diversity: 0.1, 0.6, 0.9)[6] $\times$ 6 (content production technique: $B1$-$B3$, PM,

[5]Out of the 67 participants, 32 were shown "Oil Spill" and remaining 35 "iPhone" by random assignment.

[6]Again note that since we do not make a priori assumptions of what is a "best" level of diversity, we included diversity as a variable in the experimental design and chose values spanning the range between $[0, 1]$.

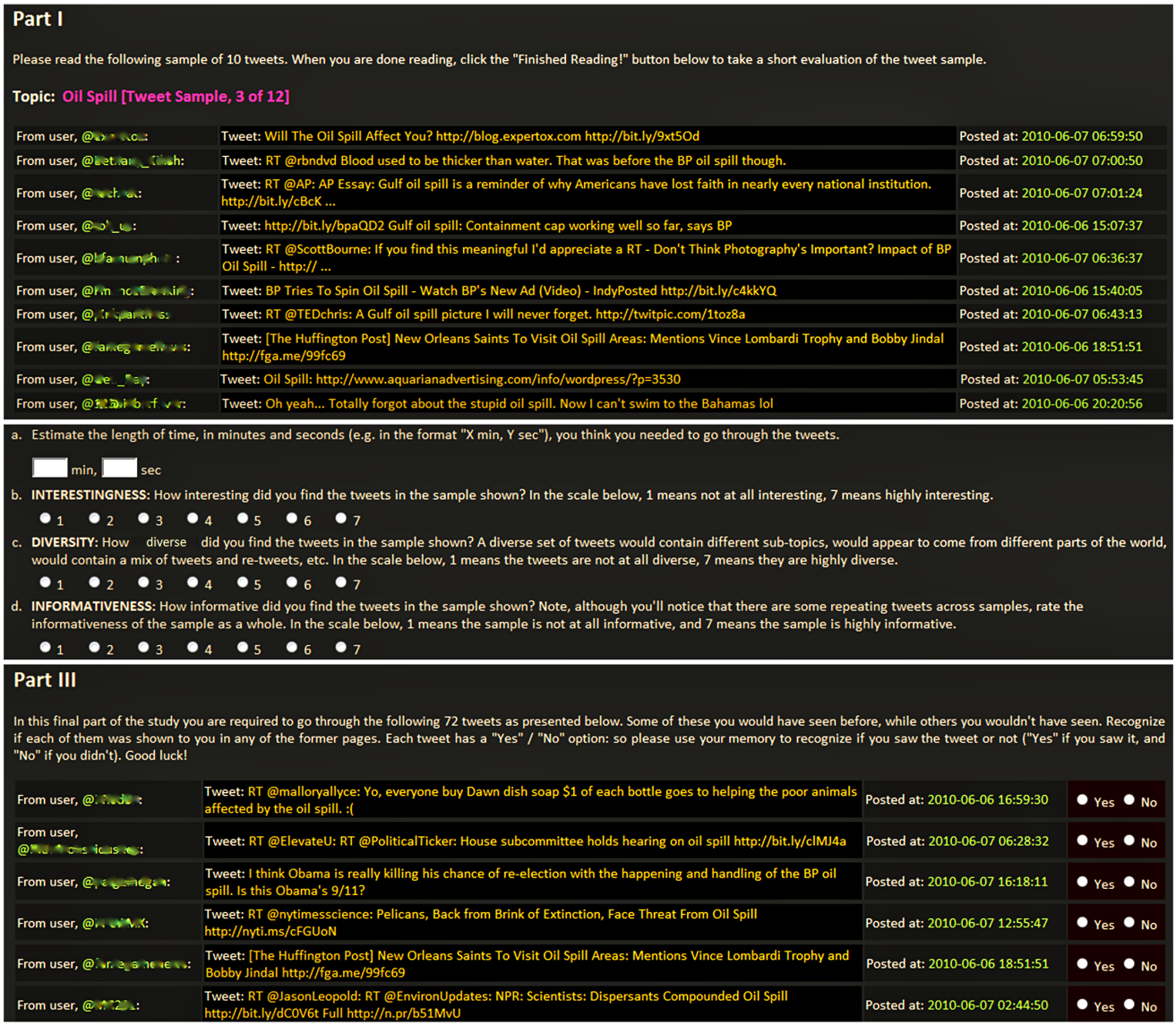

Figure 4: Screen-shots of Part I and III of the user study. The usernames have been blurred out for privacy concerns.

MR, MTU) experimental design. The content selection technique and topic variables were within subjects, while level of diversity was between subjects. Thus we propose testing the following hypotheses:

HYPOTHESIS 1. *(Performance of Proposed Method) Tweet sets generated by proposed method (PM) will be rated more interesting, informative, engaging and better recognized than those from baseline methods.*

HYPOTHESIS 2. *(Perception of Content Diversity) Participants will be able to perceive the diversity of tweet sets generated by the proposed method (PM) more accurately than those generated by baseline methods.*

HYPOTHESIS 3. *(Cognitive Measures and Content Diversity) Participants responses in the lines of interestingness, informativeness, engagement and recognition memory will be affected by the level of diversity in the various tweet sets shown.*

Note that we made no predictions about differences among the different implicit and explicit measures. In terms of the various baseline techniques, we did anticipate $B1$ would perform worse than the other baseline techniques and the proposed method. We also did not predict any differences across the topics in terms of the participants' cognitive perception—the two topics were included in the design for generalization purposes.

9. EXPERIMENTAL RESULTS

We present the experimental results based on our user study in this section. We organize our experiments in the lines of testing the three hypotheses in the previous section, that form the core of the findings in this work.

9.1 Performance of Proposed Method

In order to observe the performance of our proposed method across the baseline content selection techniques, we need to first examine interactions between our different aspects defining the various techniques (baselines and proposed). Using a repeated measu-

Table 5: Performance of the different content selection techniques using the four different cognitive measures (averaged over topics, diversity). Here, $M1$: interestingness, $M2$: informativeness, $M3$: cognitive engagement and $M4$: recognition memory. We show the mean participant ratings (r) and the standard error (se) for each case.

	$M1$		$M2$		$M3$		$M4$	
	r	**se**	**r**	**se**	**r**	**se**	**r**	**se**
$B1$	2.1	0.97	2.2	0.88	-9.1	3.02	0.2	0.32
$B2$	2.7	0.79	2.8	0.94	-5.7	3.96	0.2	0.29
$B3$	3.5	0.78	3.4	0.75	-3.3	2.88	0.3	0.44
PM	**4.3**	0.81	**4.5**	0.67	**-1.7**	3.12	**0.4**	0.43
MR	1.8	0.65	1.7	0.69	-15.4	10.8	0.1	0.51
MTU	3.7	0.65	3.8	0.71	-4.1	4.63	0.3	0.46

res ANOVA, we first tested for interactions between pairs of these aspects (use of entropy minimization, use of attribute weighting, ref. Table 3) in the participant responses on the four cognitive measures. The results (not shown due to space constraints) indicate that the interactions were not significant for the topics "Oil Spill" and "iPhone" (high p-value, therefore we accept the null hypothesis that the participants responses are generated from distributions with similar means). Hence we can proceed with main effect testing.

We now analyze and compare the performance of the various content selection techniques (our proposed method against the other baselines) across all measures (Table 5), in order to observe support for our HYPOTHESIS 1 (ref. section 8.3). In this table, the results shown are averaged across the three values of diversity ω and the two topics, "Oil Spill" and "iPhone". We see that our proposed method (again, that utilizes the entropy distortion minimization technique and uses user rating-based weighting of tweet attributes) generally yields the best performance for these measures by ∼25–30% over the baseline techniques.

To substantiate the above claim better, we present results of statistical comparisons of PM with others based on a one-tail paired t-test (Table 6). In comparing our selection technique (PM) to the other methods, we observe that the most significant difference was for the $MR \times PM$ comparison. This indicates that the approach of showing the most recent tweets on a topic (a commonly used technique) yields results sets that are less interesting, less informative, less engaging to read, and less recognized later. Baseline 1, effectively a random sample of on topic tweets, also performed poorly, though the improvement of our method for degree of recognition was only trend level significant (p<0.1).

Baseline 2 ($B2$), which incorporates the user weightings on the tweet attributes, but not the entropy minimization technique, also generated tweet sets that were less interesting and less engaging to read than those from our proposed method. Tweet sets containing the most tweeted URLs (MTU) performed fairly well, though were less engaging to read and marginally less interesting (p<0.1) than those generated by our proposed method. Finally, Baseline 3 ($B3$), which is similar to the proposed method except that it does not utilize the weightings on the tweet attributes, did not perform significantly differently from our proposed method.

9.2 Perception of Content Diversity

Since our goal in this work is to present the end user with content that matches a certain level of diversity, we were interested to investigate to what extent the participants were able "perceive" the diversity in the content generated by our method, against the baseline techniques (ref. HYPOTHESIS 2). Note we cannot compare the MR and MTU techniques in this context, because these methods do not have a notion of diversity in the tweet set generation process.

Table 6: p-values investigating statistical significance of our proposed content exploration method against other baseline techniques using one-tail paired t tests. Again, $M1$: interestingness, $M2$: informativeness, $M3$: cognitive engagement and $M4$: recognition memory.

	$M1$	$M2$	$M3$	$M4$
$B1 \times PM$	***0.002***	***0.009***	***0.007***	*0.097*
$B2 \times PM$	***0.027***	0.117	***0.011***	0.105
$B3 \times PM$	0.241	0.351	0.138	0.411
$MR \times PM$	***0.0003***	***<0.0001***	***0.003***	***0.005***
$MTU \times PM$	*0.061*	0.171	***0.004***	0.214

Table 7: Perceived level of diversity of participants, r (on the Likert scale). Percentage errors, e with respect to the corresponding actual diversities (0.1, 0.6 and 0.9) are also indicated.

	$B1$		$B2$		$B3$		PM	
	r	**e(%)**	**r**	**e(%)**	**r**	**e(%)**	**r**	**e(%)**
$\omega = 0.1$	2.8	20.6	2.2	11.1	2.1	8.8	1.1	*7.8*
$\omega = 0.6$	1.7	47.5	2.9	28.1	3.3	20.8	5.4	*13.6*
$\omega = 0.9$	5.1	20.6	5.5	14.6	6.1	9.5	6.8	*7.3*

From Table 7 we observe that our proposed method yields ratings that are in fact closest (minimum error) to the corresponding actual diversity levels. Additionally, we note that in general, the ratings on the Likert scale are better (or the errors with respect to the actual diversity are lower) and nearly symmetrical at the ends of the continuum (i.e., for diversity levels 0.1 and 0.9, in comparison to 0.6). This seems to be consistently true across the different content selection methods. Although this observation supports HYPOTHESIS 2, it is still a surprising result to us. One would conjecture that participants' perception of diversity will monotonically increase or decrease with diversity, but in the context of Twitter, we do not observe it to be true. Our explanation is that it is related to the overall characteristics of the information space, and that users appear to decipher the diversity levels better when the information is highly homogeneous or highly heterogeneous.

It appears that to the extent that people's perceptions of diversity in a result set are less accurate at medium levels of diversity. Therefore, one possibility worth further investigation is how systems raise or lower the diversity level of results in response to user input. For example, if a user could adjust a slider control to request more diverse results, perhaps the actual diversity of the results needs to increase non-linearly with the user specification in order to match the users' phenomenology.

9.3 Cognitive Measures & Content Diversity

Based on the participant responses in the user study, we now present some results to validate HYPOTHESIS 3. Recall that, HYPOTHESIS 3 states that participants responses in the lines of the cognitive measures, interestingness, informativeness, engagement and recognition memory will be affected by the level of diversity in the various tweet sets shown. Hence we report the average participant ratings for each of these measures corresponding to the three diversity levels on which the user study was conducted: 0.1, 0.6 and 0.9 (Figure 5). The figures reveal some interesting insights. We observe that the ratings, for all the four measures, are significantly better (∼30-45%) for very low (0.1) and very high (0.9) diversity values, compared to that in the middle of the spectrum (0.6). That is, it appears that users are able to comprehend the information better (and thereby find it more relevant) when it is highly homogeneous or highly heterogeneous, from a cognitive perspective.

9.4 Summary of Findings

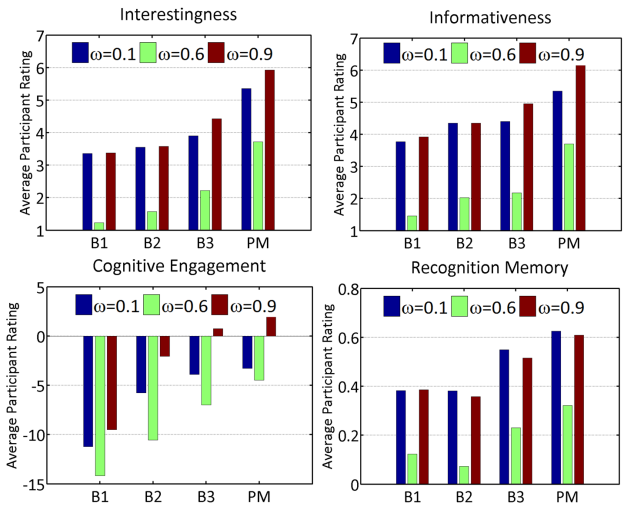

Figure 5: Cognitive measures over different values of the diversity parameter, ω.

To summarize the results, our predictions were largely confirmed for most comparisons, with our proposed method PM generally faring better than $B1$, $B2$, and MR. In particular, simple but common techniques (MR, MTU) like showing a set of recent tweets on a topic, were considerably worse than the proposed technique. In terms of the two components of our proposed method (entropy minimization and incorporating user-generated weights on tweet attributes), the two approaches in conjunction with one another seem to have had the strongest effect (PM performed best overall), though the entropy minimization component may be more helpful ($B3$ was closer in performance to PM than was $B2$). Moreover, diversity of the tweet sets indeed seemed to make a difference to the participants: they found information to be more interesting, informative, engaging and memorable if it consisted of highly homogeneous or highly heterogeneous content.

10. DISCUSSION

The key observation in our experiments is that our proposed method performed the best on the whole. Given the improvements over Baseline 2 (e.g., in Baseline 3 and the Proposed Method), this suggests that the entropy distortion minimization technique benefits the user when selecting sets of tweets about a given topic. Conversely, while the effects were in the expected direction, we did not see significant improvements of our method over Baseline 3, which used the entropy technique, but not the user-generated weights of the tweet attributes. This implies that there is room for improvement in finding weightings to place on the different tweet attributes. For example, in a future application, users could be given additional control in order to specify which attributes are most important to them in the context of their current task.

As these different tweet attributes are leveraged in the content selection process, the relatively high dimensionality of the social media information space begs the question of whether users are able to discern differences in levels of diversity. For example, a user may ask for a less diverse set of items with respect to geography and topic in the hopes of finding content specific to economic issues surrounding a local election. As the system reduces the level of diversity in the result set to accommodate this request, will this be noticeable to the user? Our results suggest that it would be, but that the answer may be complicated in that users may have greater difficulty discerning variations in levels of diversity that are closer to the middle of the spectrum—recall, error was higher for $\omega = 0.6$ in Table 7. Understanding the potentially nonlinear relationship between actual and perceived diversity will help us better design interfaces that allow users to scale sets of social media items along different attributes.

Moreover, the participants found the presented information to be more relevant (in terms of the different cognitive measures) for lower and higher diversity levels for almost all content selection methods (Figure 5). It therefore appears that there is a complex interaction between the perception of what nature of information is considered to be relevant to users, and the inherent diversity of the information space. At the first pass, highly heterogeneous content might seem to be of high utility, because of the higher information gain that can be obtained by a user: a greater mixing of attributes is likely to reveal information on a topic from a variety of perspectives. However the reason behind users finding homogeneous tweet sets to be of better quality necessitates more investigation. Our conjecture is that Twitter being a noisy social environment, it is possible that to certain users a great degree of diversity can create cognitive dissonance. Hence they might prefer homogeneous content, spanning only a limited combination of attributes, to be of better relevance. However, in the future, it will be worthwhile to understand and infer *empirical bounds*, if any, on what ranges of diversity levels are cognitively of better quality to users in the context of social media content consumption.

Finally, we comment on our choice of measures. As indicated in the introduction, evaluating topic-based sets of social media items may require different measures than traditional web search results. For many results (though certainly not all) in web search, there is a clear best result. If the user searches for "New York Times", the newspaper's home page should be the top result. In contrast, there may never be a single 'best' tweet for any given topic. Therefore, we focused on the perception of the user, both explicit and implicit, in evaluating the goodness of our proposed method. Exploring additional measures is an important area for future work.

11. CONCLUSIONS

Topic-based exploration appears to be gaining traction as a use case for social media. In this work, we addressed two problems related to this scenario: 1) what is the best technique for selecting social media content, and 2) how should we measure the effectiveness of these techniques? This paper compared several methods for selecting social media information content, with an eye towards the notion that the best results are those perceived to be interesting and informative, are engaging to read and are memorable.

A significant challenge and opportunity lies in the fact that information generated over social media sites like Twitter features a very high degree of diversity, due to the presence of a wide range of attributes. Our proposed method for content selection took advantage of this diversity by weighting content attributes according to ratings given by users in a survey. Thereafter we quantified diversity using the information theoretic measure entropy. We proposed a greedy approach of generating a result set which gives minimum distortion of its entropy compared to a desired diversity level. Based on a user study using a dataset from Twitter, our method fared better than the baseline techniques, particularly better than the recency-driven approach that is commonly used. However, in understanding the role of information diversity on human cognition, we interestingly observed that users found content on a topic to be of better quality if it featured very low or very high diversity. Our results bear on measurement techniques for social media content selection and on interface design in these high dimensionality information spaces.

12. REFERENCES

[1] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 183–194, New York, NY, USA, 2008. ACM.

[2] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA, 2009. ACM.

[3] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[4] Eric Baumer, Mark Sueyoshi, and Bill Tomlinson. Exploring the role of the reader in the activity of blogging. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 1111–1120, New York, NY, USA, 2008. ACM.

[5] Suh B. Hong L. Chen J. Kairam S. Bernstein, M. and E.H. Chi. Eddi: Interactive topic-based browsing of social status streams. In *ACM User Interface Software and Technology (UIST) conference*, 2010. To appear.

[6] Georg Buscher, Andreas Dengel, and Ludger van Elst. Query expansion using gaze-based feedback on the subdocument level. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 387–394, New York, NY, USA, 2008. ACM.

[7] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and tweet: experiments on recommending content from information streams. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pages 1185–1194. ACM.

[8] Ed H. Chi and Todd Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, HT '08, pages 81–88, New York, NY, USA, 2008. ACM.

[9] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 659–666, New York, NY, USA, 2008. ACM.

[10] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

[11] M. Czerwinski, E. Horvitz, and E. Cutrell. Subjective duration assessment: An implicit probe for software usability. In *Proceedings of IHM-HCI*, pages 167–170, September 2001.

[12] P J Daniels. Cognitive models in information retrieval—an evaluative review. *J. Doc.*, 42:272–304, December 1986.

[13] Anish Das Sarma, Atish Das Sarma, Sreenivas Gollapudi, and Rina Panigrahy. Ranking mechanisms in twitter-like forums. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 21–30, New York, NY, USA, 2010. ACM.

[14] Jean-Pierre Eckmann, Elisha Moses, and Danilo Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40):14333–14337, October 2004.

[15] Nigel Ford. Modeling cognitive processes in information seeking: from popper to pask. *J. Am. Soc. Inf. Sci. Technol.*, 55:769–782, July 2004.

[16] Sreenivas Gollapudi and Aneesh Sharma. An axiomatic approach for result diversification. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 381–390, New York, NY, USA, 2009. ACM.

[17] Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, May 2006.

[18] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600. ACM.

[19] Qiaozhu Mei, Jian Guo, and Dragomir Radev. Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1009–1018, New York, NY, USA, 2010. ACM.

[20] Owen Phelan, Kevin McCarthy, and Barry Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, pages 385–388, New York, NY, USA, 2009. ACM.

[21] Filip Radlinski and Susan Dumais. Improving personalized web search using result diversification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 691–692, New York, NY, USA, 2006. ACM.

[22] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[23] Marc Smith, Vladimir Barash, Lise Getoor, and Hady W. Lauw. Leveraging social context for searching social media. In *Proceeding of the 2008 ACM workshop on Search in social media*, SSM '08, pages 91–94, New York, NY, USA, 2008. ACM.

[24] S. M. Smith. Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory*, 5(5):460–471, 1979.

[25] G. Sperling. A model for visual memory tasks. *Human Factors*, 5:19–31, 1963.

[26] Pertti Vakkari. Relevance and contributing information types of searched documents in task performance. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 2–9, New York, NY, USA, 2000. ACM.

[27] Yunjie (Calvin) Xu and Zhiwei Chen. Relevance judgment: What do information users consider beyond topicality? *J. Am. Soc. Inf. Sci. Technol.*, 57:961–973, May 2006.

[28] Judith Lynne Zaichkowsky. Measuring the involvement construct. *Journal of Consumer Research: An Interdisciplinary Quarterly*, 12(3):341–52, 1985.

[29] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 22–32, New York, NY, USA, 2005. ACM.