Find Me the Right Content!
Diversity-Based Sampling of Social Media Spaces for Topic-Centric Search

Munmun De Choudhury[†*] **Scott Counts**[‡] **Mary Czerwinski**[‡]
[†]Arizona State University, Tempe, AZ 85281, USA
[‡]Microsoft Research, Redmond, WA 98052, USA
[†]munmun@asu.edu, [‡]{counts, marycz}@microsoft.com

Abstract

Social media and networking websites, such as Twitter and Facebook, generate large quantities of information and have become mechanisms for real-time content dissipation to users. An important question that arises is: *how do we sample such social media information spaces in order to deliver relevant content on a topic to end users?* Notice that these large-scale information spaces are inherently 'diverse', featuring a wide array of attributes such as location, recency, degree of diffusion effects in the network and so on. Naturally, for the end user, different levels of diversity in social media content can significantly impact the information consumption experience: low diversity can provide focused content that may be simpler to understand, while high diversity can increase breadth in the exposure to multiple opinions and perspectives.

Hence to address our research question, we turn to diversity as a core concept in our proposed sampling methodology. Here we are motivated by ideas in the "compressive sensing" literature and utilize the notion of sparsity in social media information to represent such large spaces via a small number of basis components. Thereafter we use a greedy iterative clustering technique on this transformed space to construct *samples matching a desired level of diversity*. Based on Twitter Firehose data, we demonstrate quantitatively that our method is robust, and performs better than other baseline techniques over a variety of trending topics. In a user study, we further show that users find samples generated by our method to be more interesting and subjectively engaging compared to techniques inspired by state-of-the-art systems, with improvements in the range of 15–45%.

1 Introduction

The advent of the Web 2.0 technology has given considerable leeway to the creation of vast quantities of user-generated information content online. Such information often manifests itself in social media spaces, via status updates on Facebook, tweets on Twitter, and news items on Digg. In almost all of these websites, while end users can 'broadcast' information that interests them, they can also 'listen' to their peers by subscribing to their respective content streams. Consequently, these avenues have emerged as means of *real-time* content dissemination to users for timely happenings like the BP oil spill, the elections in Iran, the earthquake in Haiti, or the release of the Windows Phone.

However, with the content from half a billion Facebook users or with more than 60 million tweets generated every day, the domain of topic-centric search of social media content faces tremendous challenges. How do we identify the *right* content from these spaces, that can best satisfy an end user in the context of real-time search on a topic?

Our answer to this question lies in devising methods geared towards effective *sampling* of social media spaces. The problem of sampling information signals has been studied extensively in the information theory literature (Cover and Thomas 1991); a noted method being the celebrated Nyquist-Shannon theorem that provides a technique for sampling bandlimited signals. However, this sampling technique does not apply to social media information spaces, because (1) they do not have a notion of bandwidth, and (2) they inherently feature a wide ensemble of attributes. For example, tweets (on Twitter) can be 'rich' in themes (e.g., political and economic perspectives on the same topic), can be posted by individuals in disparate geographic locations, can be updates from a celebrity, or can be conversational between two or more individuals with conflicting opinions. In essence, social media information spaces are of high dimensionality: a characteristic property we refer to as *"diversity"*.

In order to leverage rather than be limited by this diversity when sampling, we first consider the wide variety of ways an end user can best use this diversity property, when searching for topic-centric social media content. To take an example, a user searching for content on Twitter after the release of the Windows Phone in November 2010 might intend to find *homogenous* samples (*low diversity*)—say, tweets posted by the technical experts. In another situation, if she is interested in learning about the oil spill in the Gulf of Mexico back in 2010, an appropriate sample would be *heterogenous* in terms of the "mixing" of its attributes (*high diversity*). It will therefore span over attributes like author, geography and themes like Politics or Finance.

Generalizing, we contend that generating *samples that align to a desired diversity level* can have practical utility implications to the end user in a search context (Brehm 1956; Ziegler et al. 2005; Radlinski and Dumais 2006). Content of low diversity, or homogenous samples can cater to scenarios where the user seeks focused information qualifying certain pre-requisites (knowledge *depth*). Highly diverse con-

[*]Part of this work was performed while the author was an intern at Microsoft Research, Redmond.

tent, being heterogeneous in its representation of various attributes, is likely to benefit the user in terms of information gain along multiple facets (knowledge *breadth*).

Thus we describe diversity as a core property of social media content, and we quantify it via its measure of entropy in a conceptual structure called the diversity spectrum (discussed in more detail in section 3). Our central goal is therefore *to determine social media information samples on a topic that match a desired degree of diversity*.

Our Contributions. We have developed a weighted dimensional representation of the information units (e.g., tweets) characterizing large-scale social media spaces. Next we propose a sampling methodology to reduce such large social media spaces. The sampling method borrows ideas from the compressive sensing literature that emphasizes the notion of representing an information stream via a small set of basis functions, assuming the stream is fairly sparse. We thereafter deploy an iterative clustering framework on the reduced space for the purpose of sample generation. The algorithm utilizes a greedy approach-based entropy minimization technique to generate samples of a particular sampling ratio and matching a desired level of diversity.

Main Results. We perform quantitative evaluation of the proposed sampling method over a Twitter dataset (*Firehose* comprising 1.4 Billion tweets in June 2010). There are several key insights in our results. (1) We find that the compressive sensing based reduction step can prune the social media space by as much as 50–60%, and still yield robust samples that are very close to a given desired information diversity level, compared to other baseline techniques. (2) Overall, we observe that information diversity appears to be a useful attribute to sample social information spaces consistently over multiple thematic categories (Politics, Sports etc.). (3) Nevertheless, depending on the thematic category of content, the choice of the dimensional type (e.g., tweet features like recency; nodal features like the social graph topology of the tweet creator) can make a notable difference to the samples generated.

We also address the issue of evaluation of sample results sets in the absence of ground truth data. Ultimately it is the end users who decide the *"goodness" of samples* in a topic-centric search context. Hence our evaluation criterion of judging sample goodness relies on the end user's perception of the sample quality. This is accomplished via a user study involving 67 active Twitter users at a large corporation. We evaluated how the samples generated by different methods are perceived by users via two metrics: interestingness and subjective engagement, which has been found to be of utility in information comprehension tasks (Czerwinski, Horvitz, and Cutrell 2001). From the participant responses in the user study, we observe that our technique yields samples of better quality, with improvements in the range of 15–45% over state-of-the-art techniques.

The rest of this paper is organized as follows. We discuss related work in the next section. Section 3 presents our problem definition. In sections 4 and 5, we present dimensional importance learning of social media, followed by the sampling methodology. Section 6 presents our evaluation strategy. We discuss quantitative and subjective experiments in sections 7 and 8. Section 9 and 10 present a discussion of our results followed by the conclusions.

2 Related Work

Although the burst of informational content on the Web due to the emergence of social media sites is relatively new, there is a rich body of statistical, data mining and social sciences literature that investigates efficient methods for sampling large data spaces (Kellogg 1967; Frank 1978; Das et al. 2008). Sociologists have studied the impact of snowball sampling and random-walk based sampling of nodes in a social network on graph attributes and other network phenomena (Frank 1978). Recently, sampling of large online networks (e.g., the Internet and social networks) has gained much attention (Rusmevichientong et al. 2001; Achlioptas et al. 2005; Leskovec and Faloutsos 2006; Stutzbach 2006; De Choudhury et al. 2010; Maiya and Berger-Wolf 2010) in terms of how different techniques impact the recovery of overall network metrics, like degree of distribution, path lengths, etc., as well as dynamic phenomena over networks such as diffusion and community evolution.

Most of the above mentioned work on social media sampling focused on how the sampling process impacts graph structure and graph dynamics. Thus, the focus of the sampling strategies was to prune the space of nodes or edges. However, this does not provide insights into the various characteristics (e.g., degree of diffusion, topical content, level of diversity, etc.) of social media spaces in general.

Moreover, while these works addressed the issue of how to sample relevant entities in dynamic graphs, no principled way to sample or prune large social media spaces has been proposed. These spaces are unique, because of the nature of user generated content, including its high dimensionality and diversity. To the best of our knowledge, in this paper, generalized sampling methods for large social media spaces are being proposed for the first time. Our focus includes how the generated samples can improve the information consumption experience of end users.

3 Problem Definition

We begin by formalizing our problem definition.

Diversity Spectrum. It is a conceptual structure that quantifies the measure of diversity in a social media information sample. Any point on the spectrum can be specified in the form of a *diversity parameter* (referred to as ω), which is any real value in the range $[0, 1]$. In this work we utilize the information-theoretic metric "entropy" (Cover and Thomas 1991) to represent the diversity parameter that matches a generated sample. Samples with near zero entropy (or a diversity parameter value of near zero) will therefore be highly homogenous, while those with entropy nearing one, at the other end of the spectrum, will be highly heterogenous.

Social Media Dimensions. We utilize several attributes (referred to as "dimensions") along which we can sample social media information content on a certain topic. In this paper, we define the different dimensions in the context of Twitter, where the information space comprises the tweets posted by users in any given time period. Our motivation was to choose a wide range of dimensions that characterize tweets based on

Table 1: Dimensions of social media content, e.g. tweets.

1.	Diffusion property of the tweet—measured via whether the given tweet is a "re-tweet".
2.	Responsivity nature of the tweet—measured via whether a given tweet is a "reply".
3.	Presence of external information reference in the tweet, i.e., a URL.
4.	Temporal information i.e. time-stamp of the tweet.
5.	Location attribute of the tweet—given by the time-zone information on the profile of the tweet author.
6.	The thematic association of the tweet within a set of broadly define categories—such as "Business, Finance", "Politics", "Sports" or "Technology, Internet". This association is derived using the natural language toolkit, OpenCalais (http://www.opencalais.com/).
7.	Structural features of the tweet author—number of followers and number of followings / friends.
8.	Degree of activity of the tweet author—given by her number of status updates.

their content, their temporal attributes and dynamics as well as the structural properties of their creators in the social network as a whole. A description of the dimensions is given in Table 1. We assigned the dimensions into three categories: social characteristics of tweets, S (1–5), content characteristics, C (6) and nodal characteristics, N (7–8).

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 sampling ratio ρ, our goal is to determine a (sub-optimal) tweet sample, $\hat{\mathcal{T}}^*_\omega(\rho)$, 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 sample in terms of the entropy measure. This involves the following steps: (a) *Dimensional importance learning* (section 4), and (b) *Social media content sampling,* developing an approach for sample generation that matches a desired value of the diversity parameter (section 5).

4 Dimensional Importance

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 representation of t_i, based on its values for the different dimensions (section 3). Let $t_i \in \mathbb{R}^{1\times K}$ be the dimensional representation of a tweet for the set of K dimensions.

Our goal is now to determine the mutual concentrations (in other words, "importance") of the various dimensions K in the occurrence of any tweet t_i. In several text mining tasks, the observed distribution in documents is often described by multivariate mixture densities. Assuming the same density for a tweet $t_i \in \mathcal{T}$ we have:

$$P(t_i) = \sum_{\mathcal{T}_\ell} P(\mathcal{T}_\ell)P(t_i|\mathcal{T}_\ell), \tag{1}$$

where we assume that the tweet t_i is associated with a latent result set $\mathcal{T}_\ell$, that is to be shown to the user. Hence based on a K component mixture model, probability of occurrence of a tweet t_i can now be written as:

$$P(t_i) = \sum_{\mathcal{T}_\ell} P(\mathcal{T}_\ell)P(t_i|\mathcal{T}_\ell) = \sum_{k=1}^{K} \pi_k \cdot P(t_i|\lambda_k), \tag{2}$$

where π_k is the concentration parameter for the k-th dimension and $P(t_i|\lambda_k)$ is the probability distribution corresponding to the k-th dimension, with parameters λ_k. Hence the likelihood function over the entire collection $\mathcal{T}$ is given as:

$$P(\mathcal{T}|\pi,\Lambda) = \Pi_{t_i\in\mathcal{T}} \sum_{k=1}^{K} \pi_k \cdot P(t_i|\lambda_k), \tag{3}$$

where $\Lambda = [\gamma_1, \gamma_2, \cdots, \gamma_{K-1}, \mu_K, \Sigma_K]$ is the vector of the model parameters of the different distributions on each dimension. The log likelihood function is therefore given by:

$$L(\pi,\Lambda) = \ln P(\mathcal{T}|\pi,\Lambda) = \sum_{t_i\in\mathcal{T}} \ln \left\{ \sum_{k=1}^{K} \pi_k \cdot P(t_i|\lambda_k) \right\}. \tag{4}$$

Hence our goal is to maximize the above log likelihood function. We use the EM algorithm as an iterative procedure for maximizing $L(\pi,\Lambda)$. This gives us the optimal estimates of the concentration parameters π_k (and also Λ) for each dimension $1 \le k \le K$ in our collection. Thus each tweet t_i is given as: $t_i = [\pi_1 \cdot t_{i1}, \pi_2 \cdot t_{i2}, \cdots, \pi_K \cdot t_{iK}]$, where t_{ij} is the value of the j-th dimension for the tweet t_i. We now discuss how this weighted information space can be utilized in our sampling methodology to generate a sub-optimal sample $\hat{\mathcal{T}}^*_\omega(\rho)$, of a certain sampling ratio ρ and diversity ω.

5 Sampling Methodology

Our proposed sampling methodology is described in the following subsections. It has three major steps: sample space reduction, sample generation and ordering of information units in the generated sample.

5.1 Sample Space Reduction

The basic approach to any sampling methodology involves systemic pruning of the information space to disregard redundant or less relevant information and then construct samples satisfying a given pre-condition, such as minimizing a loss function (Cover and Thomas 1991). This is because, typically, the information space is very large to start with (say, in the order of millions). In our problem context, this is strongly valid—the social media information space as given by $\mathcal{T}$ can, therefore, benefit from an efficient reduction step. In this light, our solution is motivated by the work in the signal processing literature on "compressive sensing" that emphasizes that images or signals can be reconstructed reasonably accurately and sometimes even exactly from a number of samples that are far smaller in number than the actual resolution of the image or the signal (Candes and Wakin 2008). This idea is based on the observation that real signals often bear the property of being highly "sparse" (Romberg 2008). Compressive sampling exploits the sparsity notion in signals to describe it as a linear combination of a very small number of basis components.

We utilize this notion of sparsity defined in compressive sensing to reduce the social media space. That is, in the context of Twitter, we assume the sparsity property holds true for most tweets, when each tweet is described by the set of K sampling dimensions. This assumption is reasonable because we have observed that many of the dimensions, such as thematic associations of a tweet, as well as diffusion property or re-tweet (RT), are non-zero only for a handful percentage of the tweets. Hence, the assumption is that the information space in general is actually *compressible*, meaning that it essentially depends on a number of degrees of freedom which is smaller than the total number of instances N. Hence it can be written exactly or accurately as a superposition of a small number of vectors in some fixed basis. Given $\mathcal{T} \in \mathbb{R}^{N \times K}$, we are interested in the "underdetermined" case $M \ll N$, where we intend to have fewer measurements than actual information unit instances. Formally our goal is to find a smaller (transformed) matrix $\hat{\mathcal{T}} \in \mathbb{R}^{M \times K}$, that allows us to reconstruct $\mathcal{T} \in \mathbb{R}^{N \times K}$ from linear measurements $\hat{\mathcal{T}}$ about $\mathcal{T}$ of the form:

$$\hat{\mathcal{T}} = \Phi \mathcal{T}. \tag{5}$$

Here M is the number of basis functions whose coefficients can reconstruct $\mathcal{T}$ (as $\hat{\mathcal{T}}$) via linear measurements about $\mathcal{T}$. Typically, M is chosen based on the non-zero coefficients in the linear expansion of $\mathcal{T}$.

There are several standardized techniques proposed in prior literature that provide approximations to computing the transformation matrix Φ (Romberg 2008). Here we utilize the popular wavelet transform, called "Haar wavelet". The details of computing the transform based on the Haar wavelet can be referred to in (Nason and von Sachs 1999).

5.2 Sample Generation

We now present an iterative clustering technique to generate a sub-optimal sample of a certain sampling ratio ρ, such that it corresponds to a chosen value (or pre-specified measure) of the diversity parameter on the diversity spectrum, given as ω. The clustering framework, that utilizes the transformed (and reduced) information space $\hat{\mathcal{T}}$, follows a greedy approach and attempts to minimize distortion of entropy measures between the generated sample and the desired diversity parameter ω. Specifically, in order to construct the sample $\hat{\mathcal{T}}^*_\omega(\rho)$, we start with an empty sample, and pick any tweet from $\hat{\mathcal{T}}$ at random. Let us refer to this tweet as t_1. We iteratively keep on adding tweets from $\hat{\mathcal{T}}$, say t_i, such that the distortion (in terms of ℓ_1 norm) of entropy of the sample (say, $\hat{\mathcal{T}}^i_\omega$) on addition of the tweet t_i is *least* with respect to the specified diversity measure ω. That is, we iteratively choose tweet $t_i \in \hat{\mathcal{T}}$, whose addition gives the minimum distortion of entropy of $\hat{\mathcal{T}}^i_\omega$ with respect to ω, where ω is simply the pre-specified diversity parameter, as specified on the diversity spectrum.

Note that we continue the iterative process of adding one tweet at a time to the sample, until we satisfy the sampling ratio ρ. Finally, we get the optimal sample: $\hat{\mathcal{T}}^*_\omega(\rho)$.

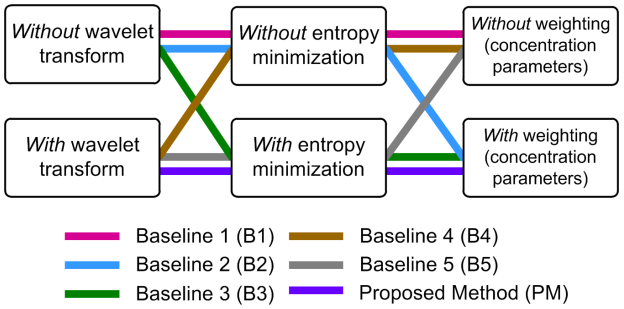

Figure 1: Evaluation strategy showing the description of different baseline techniques (B_1 through B_5).

5.3 Sample Ordering

In the last step, we present a simple entropy distortion based ordering technique of the tweets in the sub-optimal sample $\hat{\mathcal{T}}^*_\omega(\rho)$. Our central intuition is that the ordering should be based on how close the entropy of a particular tweet in $\hat{\mathcal{T}}^*_\omega(\rho)$ is with respect to the specified diversity parameter ω. Hence we compute the distortion (ℓ_1 norm) of entropy of tweet t_i, given as $H_O(t_i)$, with respect to ω. The lower the distortion, the higher is the "rank" or position of the tweet t_i in the final sample.

6 Evaluation Strategy

Data. To evaluate our proposed sampling methodology, we utilized the Firehose of tweets from the social media site Twitter, 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.

We constructed samples of various sizes (e.g., 10, 20, 30, ...) based on tweet sets segmented over 24-hour periods,[1] using our proposed method. Note that each of these samples was defined over a given "trending topic" (such as "oil spill"), and a desired diversity parameter value.

Baseline Techniques. Apart from the tweet samples constructed using our method as discussed above, we also developed a set of different baseline techniques for comparative evaluation of our proposed sampling strategy (referred to as "Proposed Method" or PM in the rest of the paper). The baseline techniques are variants of our method in terms of: use of the wavelet transform for reducing the space of tweets, use of the entropy minimization technique to achieve a desired diversity parameter level, and weighting of the tweet dimensions in terms of the learned concentration parameters. A description of the different baseline variants (B_1

[1]The motivation for segmenting over day-long periods was to preserve a reasonable recency of the information presented in the samples.

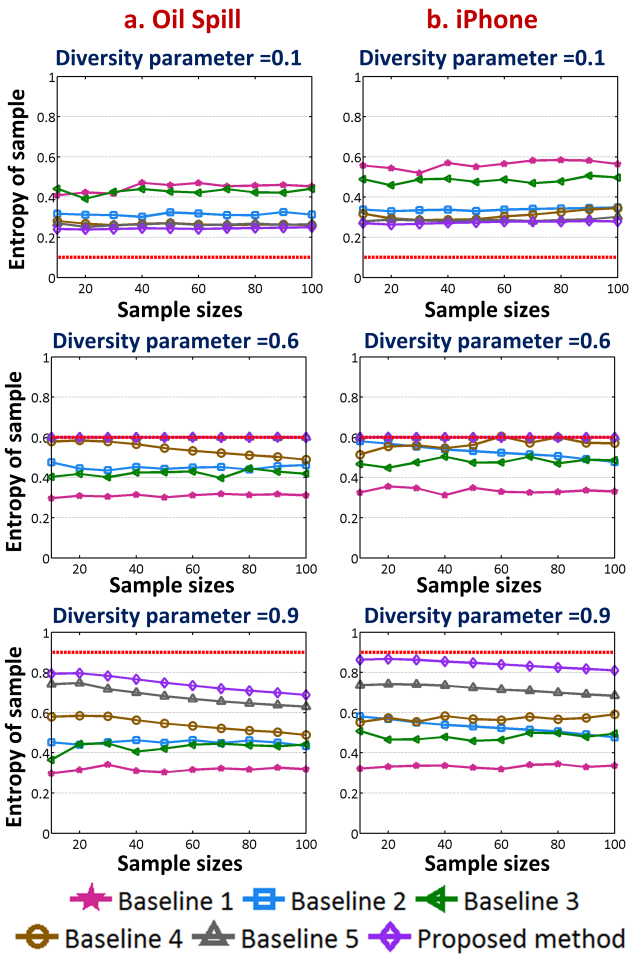

Figure 2: Comparison of proposed sampling method against baseline techniques. Results are shown for two topics and corresponding to three diversity levels. The topics are: (a) "oil spill", and (b) "iphone".

through B_5) are shown in Figure 1.

We also use two versions of current state-of-the-art tweet sampling methods. In the Most Recent Tweets (or MR) technique, we generate a sample (on a given topic) of a pre-specified size, based on reverse chronological ordering of their timestamps of posting. The final baseline method is called the Most Tweeted URL-based tweets (or MTU) where we determine all the URLs that were shared (via tweets) on the particular topic and on the given day. Thereafter we sort them by the number of times they were mentioned in different tweets throughout the day. We yield the result sample given by the "first" tweet that mentioned each of these highly shared URLs.

7 Quantitative Evaluation

7.1 Comparison against Baseline Techniques

Figure 2 gives the performance of our proposed method against baseline techniques, over two topics ("Oil Spill" and "iPhone"). Samples of sizes between 10 and 100 are generated using these methods corresponding to three values of the diversity parameter: 0.1, 0.6 and 0.9. Each figure reports the entropy of the generated sample (by each method) and shows how well it matches with the corresponding diversity parameter value (shown in a red dotted line).

The best performance, we observe, is given by our proposed sample technique, PM (as it is the closest in all cases, to the associated diversity parameter value). Among the baseline techniques, note that Baseline 5 ($B5$) yields samples of entropy which are very close to PM. Our conjecture is that $B5$ being the unweighted version of the proposed sampling method, is able to construct samples that match the desired diversity reasonably well, but is worse in performance with respect to PM, because it is not able to learn the significance (concentration parameters) of the different tweet dimensions in terms of their natural distributions in the data. On the other hand, the worst performance is given by Baseline 1 ($B1$) because it is not able to generate samples fitting the specified diversity level, nor does it consider the concentration parameters of the different tweet dimensions.

7.2 Robustness of Proposed Method

We now study the robustness of our proposed method from three different aspects: (1) effect of performing the space reduction using compressive sensing, (2) multiple iterations of the algorithm that subsumes choice of different seeds (i.e., tweets), and (3) generalizability across topics.

Effect of Sample Space Reduction. Our proposed sampling method utilizes the compressive sensing concept of pruning the information space using a wavelet transform. We studied the effect of this reduction step. We generated samples on both topics "Oil Spill" and "iPhone" for diversity values 0.1, 0.6 and 0.9; first without using the reduction step (i.e., we generate samples directly from the entire space), and second using the reduction as proposed. Results indicated that we obtained a 50–60% reduction in size of the tweet space, and as much as ∼95% overlap in content between the samples generated with the two cases. This indicates that the reduction phase helps to prune down the information space by a significant margin; while also extracting almost the same sample of tweets and preserving the requisite diversity in the sample. This further substantiates that our pruning step

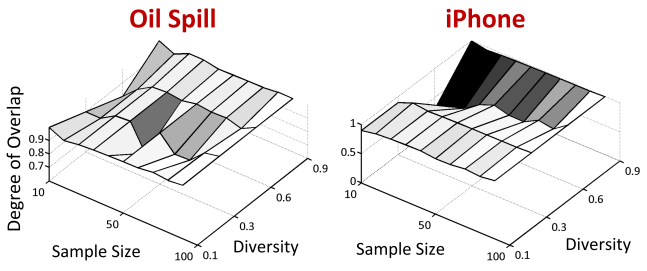

Figure 3: Robustness of proposed sampling method across multiple iterations, i.e., choice of different seeds. We show the degree of overlap of tweets across samples generated across iterations (Z-axis). These values are shown for various sample sizes (Y-axis) and diversity parameter levels (X-axis).

is approximately a lossless reduction.

Robustness across Choice of Seeds. Note that our sampling technique utilizes a greedy iterative clustering method for sample generation—every time we draw a tweet for the sample from the information space, we ensure that the entropy is as close as possible to the desired diversity level. This requires our method to start with a random tweet as the seed. In that light, how robust is our algorithm when different seeds are chosen? To this end, we study the degree of overlap of tweets in the samples across several iterations of our technique for topics "Oil Spill" and "iPhone" (Figure 3).

The results reveal that our algorithm is *indeed* consistent, with mean sample content overlap 83% (min: 68%, max: 95%). Further, they are consistent regardless of the size of sample over different diversity values. We conjecture that such high consistency in the appearance of the same tweets across different iterations occurs because the greedy strategy is *indeed* able to identify tweets whose addition minimizes the distortion between the sample entropy and the desired diversity. Consequently, it appears that the Twitter information space is likely to possess "entropy signatures" or regularities that enables the discovery of the set of tweets featuring entropy in the close neighborhood of the desired diversity.

Robustness across Topics. Next we study the robustness of our algorithm across different choices of topics. We choose a set of 30 trending topics, spanning several broad thematic categories, as shown in Table 2. Samples of sizes 10 through 100 (in increments of 10) were drawn for all topics using our sampling method, and over various diversity parameter values (0.1-0.9, in increments of 0.1).

We now quantify the performance of our sampling technique over the 30 topics based on the mean absolute difference between each sample's entropy and the diversity parameter for each category. Based on the results in Table 2, we observe the absolute differences of entropies for different categories are consistently low, providing evidence that our proposed method performs consistently across topics.

7.3 Impact of Dimensions

Finally, we analyze the impact of choosing different dimensional categories of the information space in the sampling process. The goal is to be able to study how the choice of different types of dimensions (as introduced in section 3) performs with respect to generating samples on a given thematic category and matching a certain diversity level. For this purpose we conducted an experiment to generate samples of sizes between 10 and 100, for all the 30 topics listed in Table 3. These samples were generated matching two values of ω, 0.1 and 0.9. We report the mean entropies of the samples for each thematic category in Figure 4.

The results show that different dimensional types perform differently as we focus on different thematic categories. Overall we still observe that using all features and weighting them using our proposed method (AW–All features, Weighted; Figure 4) yields samples with entropies closest to the desired value. However, in certain cases, choosing dimensional categories judiciously, instead of all features, can indeed yield samples of better quality.

Content features C seem to be effective in the case of "Celebrities". Likely this is because celebrity related news are often event-based (e.g. Lindsay Lohan's sentence to jail in June 2010), and hence the content of the tweets are often very important in judging the sample quality. For the nodal features N, the best performance is observed for the themes "Technology" and "Politics". Since these topics often comprise real-world news items, tweets from certain topical authorities (such as @cnn or @mashable) are likely to be judged more relevant in the samples than those based on other dimension types.

Table 2: List of 30 (randomly chosen) trending topics from Twitter that were used for studying the robustness of our proposed sampling method. Broad thematic categories (hand-labeled) are indicated to indicate a wide span of topics. Mean of the absolute difference between sample entropy and diversity parameter values are also shown for each category (minimum absolute difference is zero; so values close to zero are good).

TYPE	TRENDING TOPICS	MEAN
Sports	NBA, Vuvuzela, #worldcup, Lakers, Suns	0.0927
Entertainment	Star Trek, Harry Potter, New Moon, Twilight, American Idol, Inception	0.1284
Celebrities	Lady Gaga, Michael Jackson, Justin Bieber, Lindsay Lohan	0.1073
Technology	Tweetdeck, iPad, Snow Leopard, iPhone, Apple, At&t, Google wave, Motorola, Steve Jobs	0.1318
Politics	Barack Obama, McCain, Afghanistan	0.0727
Global Affairs	H1N1, Haiti, Oil Spill	0.0933

8 Subjective Evaluation

To reinforce our quantitative studies, we conducted a user study to evaluate the "goodness" of our proposed method.

8.1 Method

Participants. Participants were 67 employees of a large technology company who were required to be Twitter users at least two times per week. Median age of participants was 26 years (range: 19–47 years).

Stimuli and Procedure. A web-based application was developed for the purpose of our study. 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. Each participant was presented with 12 samples of tweets spanning a topic, either "Oil Spill' or "iPhone", generated by the different baseline techniques and proposed method. Participants saw tweets for only one topic, and topic assignment was random—out of the 67 participants, 32 were shown "Oil Spill" and remaining 35 "iPhone". Also the ordering of the samples generated by different techniques was randomized. All partici-

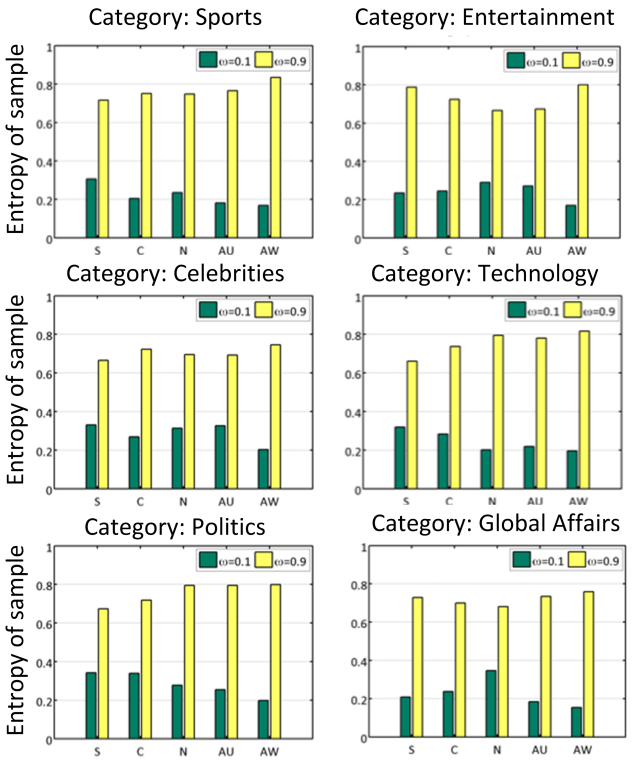

Figure 4: Evaluating impact of different dimensional types. The different dimensions on X-axis are represented as the following: S=social features, C=content features, N=nodal features, AU=all features (unweighted) and AW=all features (weighted).

pants saw samples at three levels of diversity: 0.1, 0.6, 0.9. Each sample contained 10 tweets, along with their corresponding usernames and the time of creation. After each sample, the participant was asked to, (a) estimate the length of time spent reading the tweets, and (b) rate the interestingness of the tweets (on a Likert scale of 1 to 7).

Metrics. We included two metrics to evaluate user performance with the different content sampling techniques. The first one was an explicit metric consisting of a 7-point Likert scale rating question, corresponding to the "interestingness" of each sample shown to the participants. We also used an implicit metric for evaluation, which is a normalized version of subjective duration assessment (Czerwinski, Horvitz, and Cutrell 2001). We refer to it as "subjective engagement". It is given by the ratio of the difference between the actual and perceived durations involved in going through a sample. Ideally, if the information presented is very engaging, the participant would underestimate the time taken to go through it and subjective engagement would be a positive value. In our case, we saw negative values for subjective engagement, though what is important to observe here are the relative differences in engagement when reading tweet samples generated by the different methods.

Table 3: Performance of different techniques over the metrics as follows: Interestingness (M1) and Subjective engagement (M2). Note that higher values are better.

	$B1$	$B2$	$B3$	$B4$	$B5$	PM	MR	MTU
M1	0.34	0.37	0.38	0.39	0.41	0.42	0.30	0.39
M2	-11.2	-9.6	-8.5	-6.1	-5.2	-4.4	-15.5	-6.9

Table 4: p-values against significance level of 0.05 for different baseline techniques against our proposed method. Only the comparisons which yielded significance for at least one of the two measures are shown.

	Interestingness	Subjective engagement
$B1 \times PM$	*0.0093*	*0.0126*
$B2 \times PM$	0.0735	*0.0436*
$B4 \times PM$	0.2927	*0.0216*
$MR \times PM$	*0.0004*	*0.0138*

8.2 Results

Now we comapare the overall performance of our proposed method against the different baseline sampling techniques using the qualitative responses obtained in the user study. In Table 3 we show the measures of the two dependent metrics for all the sampling techniques, averaged over topics and the values of diversity. We observe that the best ratings are obtained for our proposed sampling method PM and the worse for $B1$ (i.e., on an average 15–45% improvement for our proposed technique).

In the light of these observed differences, we study the statistical significance of our proposed sampling technique with respect to the baseline techniques. To this end, we perform a one-tail paired t-test on the participant ratings obtained from the user study. Our experimental design comprises 2 topics: "Oil Spill", "iPhone" $\times$ 3 levels of diversity: 0.1, 0.6, 0.9 $\times$ 8 sampling techniques: $B1$-$B5$, PM, MR, MTU. Additionally, our null hypothesis is that participant ratings on samples generated by different techniques come from the same distribution and hence have the same mean. We observe from Table 4 that for both the metrics, the comparisons of PM to most baseline techniques yield low p-values. This indicates that the improvement in performance of PM is statistically significant, particularly for the engagement metric. Exceptions are observed for $B5 \times PM$ and $MTU \times PM$, and to a lesser extent $B3 \times PM$ which are not reported in Table 4. This is in conformity with our observations from the quantitative experiments, supporting the ability of our proposed method to generate samples not only fitting a desired diversity level, but also ones that are more conducive to users' content consumption process.

9 Discussion

Social media and diversity. A central observation in this work has been that social media spaces engender a diverse set of attributes and controlling the diversity in samples can benefit end users. Hence we devised a methodology that generates samples based on a desired level of diversity. However, we acknowledge that diversity is not the only core property of social media content; there can be other properties that one might intend to optimize in the sampling context of social media—e.g., controlling for novelty of information in the samples, or optimizing samples for degree of past familiarity of the end user to the broader topic.

Entropy signatures. Our experimental observations indicate that the Twitter information space has *structure* and regularity to it. Recall that, on applying the transformation during the reduction phase (based on the sparsity assumption of the dimensions), we were still able to retain information that yields samples of high quality, as estimated by the participant rating in the user study. Besides, regardless of the initial seed tweet, our samples contained significant overlap in the final set of tweets and were able to preserve the requisite diversity level (see Figure 3). Consequently, one question that arises is whether these regularities reflect "entropy signatures" of the information space. If so, then how can the sampling methodology benefit from these signatures?

Evaluation of sample goodness based on loss functions. Note that many typical sampling algorithms are evaluated based on some loss function—the lower the functional error with respect to actual data, the better is the method. For example, in graph sampling, the loss function could be the error in the sample in recreating the degree distribution of the actual graph (Leskovec and Faloutsos 2006), or the error in predicting a related time series variable (De Choudhury et al. 2010). However, in our context, if the loss function was, e.g., geared towards recreating the re-tweet distribution on Twitter, it might create samples which contain noisy or redundant information, and therefore not satisfy an end user's real-time information consumption experience. Hence our subjective evaluation provides greater utility in this context.

Personalization of samples. Note that we acknowledge that in a practical social media content sampling scenario, we would like to present tweets to the end user that are personalized with respect to her activity patterns, demographics like location, the structural properties of her egocentric network and so on. Although our paper caters to an average user, the proposed sampling method can easily be applied to different personalized contexts.

10 Conclusions

The Internet is a *big* place, bustling with rapidly growing user-generated content. Making sense of current happenings on a topic from such large-scale repositories of information therefore involves selecting (i.e., sampling) the "right" set of items for the end user, so that the user finds the presented information (a) to be suitably diverse, reflective of the high dimensional social media space, as well as (b) to be interesting and engaging. We have presented a sampling framework to cater to this issue, using ideas from compressive sensing. Along with a host of baseline techniques, we evaluated the samples generated by our method, quantitatively as well as qualitatively through a user study, to get 15–45% improvement over some of the state-of-the-art social media tools.

Sampling of social media spaces is of paramount significance in web-scale data management, social data analytics, as well as user interface design. Through this work, we have observed: (a) what kind of sampling methods perform better than others; and (b) what kind of (cognitive) metrics quantify the sample quality as perceived by the end user. In all, information diversity turned out to be a useful attribute in generating samples from social media spaces, that users found interesting and engaging. However we noted that the choice of the dimensional type (e.g., social versus nodal features) can make a notable difference to the quality of the samples generated. In future work, we are interested in extending our observations to other social media information spaces (e.g., Facebook); determining when to apply which sampling methods and dimensions, as well as to study in more detail the statistical characteristics of these spaces in the hopes of gathering useful insights about the sampling process in general.

References

Achlioptas, D.; Clauset, A.; Kempe, D.; and Moore, C. 2005. On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. In *STOC '05*.

Brehm, J. 1956. Post-decision changes in desirability of alternatives. *J. of Abnormal and Social Psych.* 52:384–389.

Candes, E., and Wakin, M. 2008. An introduction to compressive sampling. *Signal Processing Magazine, IEEE* 25(2):21 –30.

Cover, T. M., and Thomas, J. A. 1991. *Elements of information theory*. New York, NY, USA: Wiley-Interscience.

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

Das, G.; Koudas, N.; Papagelis, M.; and Puttaswamy, S. 2008. Efficient sampling of information in social networks. In *SSM*, 67–74.

De Choudhury, M.; Lin, Y.-R.; Sundaram, H.; Candan, K.; Xie, L.; and Kelliher, A. 2010. How does the data sampling strategy impact the discovery of information diffusion in social media? In *ICWSM '10*.

Frank, O. 1978. Sampling and estimation in large social networks. *Social Networks* 1(91):101.

Kellogg, W. 1967. Information rates in sampling and quantization. *IEEE Trans. on Info. Theory* 13(3):506 – 511.

Leskovec, J., and Faloutsos, C. 2006. Sampling from large graphs. In *KDD '06*, 631–636. ACM.

Maiya, A. S., and Berger-Wolf, T. Y. 2010. Sampling community structure. In *WWW '10*, 701–710. ACM.

Nason, G. P., and von Sachs, R. 1999. Wavelets in time series analysis. *Phil. Trans. R. Soc. Lond. A* 357:2511–2526.

Radlinski, F., and Dumais, S. 2006. Improving personalized web search using result diversification. In *SIGIR '06*.

Romberg, J. 2008. Imaging via compressive sampling. *IEEE Signal Processing Magazine* 25(2):14–20.

Rusmevichientong, P.; Pennock, D.; Lawrence, S.; and Giles, C. 2001. Methods for sampling pages uniformly from the world wide web. In *AAAI Fall Symposium on Using Uncertainty Within Computation*, 121–128.

Stutzbach, D.; Rejaie, R. D. N. S. S. W. W. 2006. Sampling techniques for large, dynamic graphs. In *INFOCOM 2006*.

Zaichkowsky, J. L. 1985. Measuring the involvement construct. *Journal of Consumer Research: An Interdisciplinary Quarterly* 12(3):341–52.

Ziegler, C.-N.; McNee, S. M.; Konstan, J. A.; and Lausen, G. 2005. Improving recommendation lists through topic diversification. In *WWW '05*, 22–32. ACM.