What's Trending? Mining Topical Trends in UGC Systems with YouTube as a Case Study

Colorado Reed
University of Iowa
Iowa City, IA
colorado-reed@uiowa.edu

Todd Elvers
University of Iowa
Iowa City, IA
todd-elvers@uiowa.edu

Padmini Srinivasan
University of Iowa
Iowa City, IA
padmini-srinivasan@uiowa.edu

ABSTRACT

User-generated content (UGC) systems such as Twitter, Facebook, and YouTube are quickly becoming the dominant form of information exchange on the web: shifting informational power from media conglomerates to individual users. Understanding the popularity trends in UGC content has proven problematic as traditional content popularity techniques (e.g. those developed for television) are not suited for the disparate origins and ephemeral lifecycle of UGC. Content-based trend detection with UGC systems has been an intensely growing field of research in recent years, yet surprisingly, there is no single method or approach that can be used to track and compare trends in user posts across multiple UGC sources. Therefore, in this work, we develop a standard system for detecting emerging trends in user posts for UGC that contains some form of textual data. We demonstrate the use and implementation of this system through a case study with approximately 2 million YouTube video posts. Furthermore, to help facilitate future comparative studies in UGC trend analysis, we have made this system open-source and straightforward to integrate with various UGC systems (Twitter, Facebook, Flickr, Digg, Blogger, etc.).

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering, Selection process*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and indexing

General Terms

Algorithm

1. INTRODUCTION

Over the past six years, web publishing of User Generated Content (UGC) has rapidly reshaped both the dissemination and accessibility of information—shifting informational power from traditional news sources to individuals[4]. UGC has become ubiquitous in web culture through applications including digital videos, forums, social networking, [micro] blogging, wikis, user-reviews, and non-commercial (open-source) software. This UGC boom has generated immense amounts of data[1]. Accordingly, UGC research has evolved into a vast and variegated field with areas of research ranging from the study of social networks and recommender systems to the investigation of the societal impact of microblogging [16, 17].

An important branch of UGC research is the detection of popularity trends in UGC activities (posting, viewing, downloading, etc.). Understanding these trends is necessary for the implementation of fast and accurate information retrieval and recommender systems, as well as for directed advertising and marketing. Consequently, significant research effort is being devoted to characterizing and analyzing popularity trends for UGC systems such as Twitter[19], YouTube[10], and Digg[15]. Not surprisingly, we also see various commercial applications for detecting UGC posting trends, for instance, Twitter posting trends: *Trendistic*[27], *Twitscoop*[28], *hashtags.org*[13], and *Twopular*[29].

Despite the fact that trend detection is a hot topic, to the best of our knowledge no single method or approach exists to track trends for different UGC sources. Typically, trend detection is tailored to the UGC it serves. For example, the methods designed for analyzing Twitter posting trends cannot be readily applied to a different UGC such as YouTube (we discuss this aspect in detail later). Thus our first goal is to develop a general method for detecting emergent topic trends in any UGC system, so long as the UGC consists of textual entries or objects accompanied by textual entries in the UGC. A general trend detection system will allow for a standard method of comparison across various UGC services. We may then investigate why/how emerging topics vary across different UGC systems. Our second goal in this paper is to demonstrate the application of our general UGC-emerging-topic-detection (ETD) through the detection of popularity trends in YouTube video posts, which is by far the preeminent source for UGC video content (as of March 2011, YouTube had an Alexa Rank[1] of 3, behind only Google and Facebook)—specific motivation for this choice of UGC is provided in the following paragraph.

Understanding emerging popularity trends of UGC videos

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MDMKDD'11, August 21, 2011, San Diego, CA, USA.
Copyright 2011 ACM 978-1-4503-0841-0 ...$10.00.

[1]In August 2010, Google CEO Eric Schmidt said that "every two days [humans] create as much information as we did from the dawn of civilization up until 2003." http://techcrunch.com/2010/08/04/schmidt-data.

has become an important topic to both the business and technology domain as the number of users of these services has skyrocketed in recent years (recently YouTube reported that their website has over 3 billion views per day, see footnote 7). Consequently, in recent years a substantial body of research has addressed this issue (see for example [4, 5, 6, 9, 10, 25]). In short, this collection of research seeks to answer the question: *what UGC videos are people watching and why?* However, the detection and analysis of emerging topics in UGC video posts remains virtually unexplored. The underlying question here is: *what UGC videos are people posting and why?*. We present an answer by analyzing the textual data in YouTube video posts using our general ETD system. To the best of our knowledge, this is both the first generalized UGC-ETD system and the first study of popularity trends in UGC video posts.

Our ETD system was inspired by the work of [3] (hereby referred to as Cataldi et. al.), who developed an ETD system for Twitter posts. We selected this system (described in detail in §2) because of its ability to recognize general emerging topics without requiring *a priori* information from the user —unlike the aforementioned commercial applications which require a query phrase from the user. Furthermore, Cataldi et. al.'s system is more general than competing methods such as [30], which focuses on large scale events (natural disasters, wars, etc.), or [24], which detects emerging Twitter topics through "retweets" (a feature specific to Twitter).

Despite the strengths of Cataldi et. al.'s approach, we have had to make significant changes and extensions to better suit our goals. These are explained in detail in the following section, here we provide a few key examples: their methods prioritize the contributions from influential users, while we treat all UGC contributors as equals; they determine the emergence of terms based on an absolute measure of emergence, while we use a ranking-based strategy that depends on the relative importance of a term in a given timeframe. Our novel contributions through this work can be summarized as follows:

- we create a general UGC emerging-topic recognition system that can be readily implemented for arbitrary UGC types[2]
- we use this system to uncover and analyze emerging topics in textual data from YouTube video posts

This paper is organized in the following way: in §2 we discuss research related to our work, in §3 we provide a detailed description of our ETD system, in §4 we address our generalized data collection technique, in §5 we discuss our empirical results from a case study with YouTube data, and in §6 we summarize this work and address future extensions.

2. RELATED WORK

A substantial body of research has investigated viewing trends of YouTube videos. In 2007, [4] studied the popularity lifecycle of videos and found that the power-laws and Pareto distributions that define the popularity of YouTube videos are similar to non-UGC video content (i.e. NetFlix), but that the video popularity lifetimes were much shorter. Also in 2007, [12] monitored YouTube usage of students, faculty, and staff at the University of Calgary (approximately 33,000 individuals) over a three month period and used the popularity characteristics present in the viewing patterns to develop a caching system that could potetially reduce the load of YouTube's servers. [12] also examined basic statistics of the $\approx$ 600k video posts collected during this period (average length, average video duration, and distribution of the video categories). In 2008, [2] used the popularity trends of videos as well as user preferences and viewing patterns to develop a video recommendation system. In the development of this system, the authors investigated the co-view relationship of videos and found that using a random walk model (treating the related videos as out-links) to weight the relationships between videos had predictive power in determining the popularity lifecycle of a given video.

Applications in the viewing trends of YouTube videos have also experienced significant gains in recent years in response to the massive amount of video-based UGC data available on the web. Notable applications include YouTube's official *Trends Dashboard*[3] and YouTube's official trends blog: *YouTube Trends*.[4] These two applications "leverage internal YouTube data to uncover viral videos as they become popular." Viral is the term used to describe a video that becomes widely popular in a short time period through UGC-based websites. Neither application explains precisely how they determine that a specific video is becoming "trendy," however both sites (which appear to use the same underlying system) frequently cite the number of views for a video in the past twenty-four hours. In our investigation we will therefore look to uncover emerging topics in YouTube video *posts* during twenty-four hour time periods as this should correspond, at least loosely, to the time period of interest for emerging topics in YouTube video *views*.

The above applications and research are a small sample of the field of research devoted to the characterization of popularity trends in YouTube video views (see [6, 10, 25] and their citations for further investigations). As preluded in §1, there is a dearth of information regarding the popularity trends of UGC video postings. Various works include basic statistics on their collected (usually crawled) YouTube datasets: for example [6] crawled the meta-data of nearly 38 million YouTube video posts and provided information about this dataset such as the most popular video posting categories (45% of the videos are classified in either the *Music* or *Entertainment* category), and the temporal distribution of video posts (daily video posting rates peak at 1 p.m., weekly video posting rates peak on Sunday). Nevertheless, a thorough investigation of related literature indicated that the topical characteristics of YouTube video posts has not been previously examined.

On the other hand, several applications exist for detecting emerging topics in Twitter posts. The methods developed by Cataldi et. al. form the basis for our work as these methods are the most generally applicable methods found in a survey of modern techniques. This method was shown to empirically reveal emerging topics in Twitter posts, such as the five-term emerging topic {*eyjafjallajökull, volcano, airports, iceland, close*} following the 2010 eruptions of Mt. Eyjafjallajökull.

Unfortunately, this system has several Twitter-specific features and constraints that inhibit its implementation in other

[2]http://code.google.com/p/emerging-youtube-topics/ contains the source code for this system

[3]http://www.youtube.com/trendsdashboard

[4]http://youtube-trends.blogspot.com/

UGC systems: the use of a streaming API (this feature is not available for a number of UGC APIs, including YouTube, Flickr, and Blogger), calculation of a PageRank weight term for a given user's post (recent API query limits make this calculation infeasible for both Twitter and YouTube), the provided life cycle model for a given term (the model is susceptible to statistical fluctuations and is not a true indicator of emerging terms for non-Twitter UGC), and the automated technique for determining the set of emerging terms (the automated clustering methods provided by Cataldi et. al. can often yield undesirable results). This system and its constraints are discussed further in the following section.

3. DETECTING EMERGING TOPICS

In this section we first provide a summarized account of Cataldi et. al.'s system for ETD in twitter posts, and then provide a detailed comparison to our general ETD system. For the rest of this work, we define an emerging topic as *a set of semantically related terms that experience a sudden increase in interest during a specific time period.*

3.1 Cataldi et. al.'s System for ETD

Cataldi et. al. formulated the following process for ETD with Twitter posts:

1. collect Twitter data via Twitter's streaming API[5]
2. represent the collected Twitter posts as vectors of terms weighted by the relative frequency of the terms
3. create a directed graph of the users where an edge from *node a* to *node b* indicates user *a* "follows" user *b*'s Twitter posts, and weight a given user's posts by his/her PageRank[20] score in this directed graph
4. model the life cycle of each term in the Twitter post by subtracting the relative combined weight of the term in previous time intervals from its combined weight in the given time interval
5. determine the emerging terms through a user-defined threshold or an automated clustering-based approach on the values obtained from the previous step
6. use a directed graph of the emerging terms to create a list of emerging topics by weighting the links between the terms via a co-occurance measure
7. select emerging topics by locating strongly connected components in the graph with edge weights above a given threshold

In the following section, this method will be thoroughly compared with the general ETD system we develop.

3.2 Generalized ETD

Since our goal is to formulate a system for detecting emerging topics in general UGC systems, we describe our system without specific reference to a given type of UGC. The only constraint is that the UGC must contain textual data. When appropriate, we include additional details and examples specific to textual data from YouTube video posts. For reference, here is a condensed overview of our ETD system given a time interval of interest (I):

1. use a large dictionary of broad terms to query a given UGC system and obtain recent UGC posts
2. represent each returned post, j, as a vector of terms, v_j^I of length N_t^I, where N_t^I is the number of terms in the period of interest, I
3. weight each entry in v_j^I by the max term frequency in the post multiplied by $1/N_u^I$, where N_u^I is the number of posts made by the posting user, u, in I
4. sum the weighted term vectors in I
5. assign a rank to each term in I, where a rank of 1 is given to the term with the largest combined weight
6. model the emergence of each term in I by performing a weighted linear regression using the rank of each term in the previous s time periods and then calculate the fraction of error between the predicted (P) and actual (A) rank value in I via $(P - A)/P$
7. terms with the fraction of error close to 1 are considered emerging terms for time period I
8. create a navigable directed graph, where terms represent nodes and weighted links represent the semantic relationship between term pairs
9. extract emerging topics by locating strongly connected components in the graph such that all of the edge weights are above a given threshold and the graph contains at least one emerging term

Choosing a time interval, I, in which emerging topics are of interest is the first component in our system. Cataldi et. al. studied emerging topics in Twitter posts in fifteen minute intervals as they were seeking to uncover breaking news before it was reported by news sources. As discussed in §2, emerging YouTube viewing trends likely happen in approximately 24-hour intervals, so this seems an appropriate interval for posting trends as well. The large difference in intervals of interest between these types of UGC can be attributed to the frequency of user posts: 140 million Twitter posts per day[6], and an estimated 100 thousand YouTube posts per day.[7] Therefore, it is not possible to choose a rigid time interval that is applicable to all UGC systems, and so we introduce this time-interval as the first parameter in our system.

A UGC post, in the context of our system, is defined as the textual information provided by the user when posting his/her content. We represent each post as a vector of weighted terms and make no distinction between the various fields in a given post. For example, YouTube and Flickr posts contain three text fields (title, tags, and description), while Twitter posts only contain one text field.

For our system, we combine all fields into a single weighted term vector v_j^I, where j denotes the j^{th} post in the time period I. The length of all post vectors in I is N_t^I: the total number of terms from all posts in I. Each term, denoted t_x, in post j of time-period I, denoted p_j^I, is weighted via:

$$W(p_j^I, t_x) = \frac{tf(p_j^I, t_x)}{\arg\max_i tf(p_j^I, t_i) \times N_u^I} \quad (1)$$

where the numerator is the term frequency of t_x and the denominator is the maximum term frequency in p_j^I multiplied

[5]http://dev.twitter.com/pages/streaming_api

[6]http://blog.twitter.com/2011/03/numbers.html

[7]http://youtube-global.blogspot.com/2011/05/thanks-youtube-community-for-two-big.html

by the number of posts the posting user, u, made in I. If t_x does not appear in p_j^I then $W(p_j^I, t_x) = 0$. Each post can then be represented as a vector of weighted term-scores in the following way:

$$v_j^I = \left\{ W(p_j^I, t_1), W(p_j^I, t_2), \ldots, W(p_j^I, t_{N_t^I}) \right\} \quad (2)$$

In order to justify the above term weights, which differ from the term-weights used by Cataldi et. al., we must first define the *nutrition* of t_x^I in period I as the sum of the term's weights over all posts in I, or formally:

$$nutr_x^I = \sum_{j=1}^{N_p^I} W(p_j^I, t_x) \quad (3)$$

where N_p^I is the number of posts in I. The nutrition for a term is a biological metaphor that represents the importance of a term in a given time interval—the larger the nutrition, the more prevalent the term is in the given period. This biological metaphor was coined in [7] and was adopted in the work of Cataldi et. al. Our particular term-weighting serves to mitigate the influence of individual users on determining the nutrition value of a given term. Specifically, through this implementation, a user can contribute a maximum of 1 to the total nutrition of a term. Also, the user's influence is diminished if the user posts several times in I, which serves primarily as an anti-spam measure. In brief, the provided term-weighting ensures that the interests of the crowd, not individuals, is apparent in the relative nutrition scores of I.

Cataldi et. al. also used the sum of term weights to determine the nutrition scores in a given time period. Using the same notation as above, their term weights were defined as follows:

$$W(p_j^I, t_x) = PR(u) \left(0.5 + 0.5 \frac{tf(p_j^I, t_x)}{\arg\max_i tf(p_j^I, t_i)} \right) \quad (4)$$

where $PR(u)$ is the damped PageRank (PR) of the user that made the Twitter post p_j^I. The PR algorithm is a well-known method used to measure the importance of a node in a directed network[20]. Cataldi et. al. used a directed graph of the Twitter user's followers/followings network to calculate the PR of the posting users.

For our purposes, this weighting scheme can often create undesirable nutrition scores in a given time period. For example, say a highly-influential user (super-user) had a PR value that was 1000 times the value of an average user. If the super-user made a one-word post consisting of the term "cat," then this would be equivalent to 1000 average users posting the term "cat"—while in our system the collective posts of 1000 average users is 1000 times more influential than a single post from one super-user. In other words, Cataldi et. al.'s weighting scheme creates an oligarchy, while we assume a democracy for ETD. Furthermore, Cataldi et. al.'s weighting is susceptible to spamming users. To continue the above example, if the aforementioned super-user made 500 posts in I consisting of the term "cat" then this would contribute to the nutrition of "cat" 500 times more than 1000 users making single posts of the term "cat." While in our system the super-user would still only contribute 1 to the nutrition of "cat" and the 1000 users would contribute 1000. Therefore, a term in our system can only be considered important if it is popular with a large number of posters.

We recognize that particular studies could find counter-examples to our weighting system, and therefore provide one further justification as to why we avoided a user-authority-based weighting scheme. Namely, that it is becoming impractical to form complete graphical representations of user communities in various UGC services. In the past year, Twitter and YouTube have decreased the number of queries allowed to their authenticated APIs, which provide information such as the number of followers or subscribers for a user. Twitter allows 350 requests to this service per hour, and given the 170 million (and quickly growing) Twitter accounts, it would take around fifty years to obtain the complete user-network needed to calculate a true PageRank value (approximately five years for YouTube). Various work-arounds are possible, but as our focus is on usability across multiple UGC services, we have explicitly avoided this technique.

Before finding emerging *topics* in I, we must first uncover emerging *terms* that will be used to form the root of the topics. A term is considered emergent if it experienced a significant, unexpected increase in nutrition in a given time period. We define this measure of emergence as the *energy* of the term in I—this biological metaphor was also taken from [7] and adopted in the work of Cataldi et. al. In our approach, the energy of a given term in I, $energy_{t_x}^I$, is determined by first obtaining a predicted nutrition rank (NR) of t_x in I and calculating the fraction of error between the predicted and actual NRs via:

$$energy_{t_x}^I = \frac{P^I(s, t_x) - A^I(t_x)}{P^I(s, t_x)} \quad (5)$$

where $P(s, t_x)$ is the predicted NR of t_x in I, s is the number of previous time intervals to take into account for the prediction, and $A^I(t_x)$ is the actual NR of t_x in I. The NR of a term is its relative rank in nutrition in the given timeframe—where the term with the greatest nutrition in I will have an NR of 1, the term with second greatest nutrition will have a NR of 2, etc. $P^I(s, t_x)$ is then calculated by performing a weighted linear regression on the normalized NRs for the previous s time periods, obtaining a normalized predicted NR for I, and multiplying this normalized NR by the number of terms in I to determine the actual predicted NR. The value of s, our second parameter, depends on the type of UGC under investigation. It is important to choose an s value that allows for an accurate prediction of the NR while not including too large of a timeframe that could blend distinct phases of emergence for a given term.

We used a weighted least-squared linear regression to predict the NR in I as it is a non-parametric model that can be efficiently applied to a wide range of s values. In the linear regression fit, we apply a $1/h$ weight-factor to the fit-distance of the h^{th} previous NR value so as to bias our prediction towards recent NRs of the given term. The justification for this bias is that the energy of a term in I should be more heavily influenced by the deviation of the term's nutrition from recent nutrition scores, rather than earlier nutrition scores. We recognize that this model is limited through the assumption that the NR of a term follows a linear model. However, we find this assumption to yield reasonable results in our YouTube case study and plan to explore the necessity of non-linear models in future work.

The fraction of error between $P^I(s, t_x)$ and $A^I(t_x)$ was chosen to represent the energy of t_x in I because this def-

inition favors predicted NR deviations towards the higher ranked nutrition scores. For example a 10 NR deviation between a predicted NR of 100 and an actual NR of 90 yields an energy of $(100-90)/100 = 0.1$, while a 10 NR deviation between a predicted NR of 20 and an actual NR of 10 yields an energy of $(20-10)/20 = 0.5$. As desired, the second example has a greater level of emergence than the first. The energy has a practical range of $(-N_p^I,\ 1)$, where positive (negative) values indicate the term is more (less) popular in I than predicted.

Using the same notation, Cataldi et. al. determined the energy of a given term in I via:

$$Cenergy_{t_x}^I = \sum_{h=I-s}^{I-1} \left(\frac{(nutr_{t_x}^I)^2 - (nutr_{t_x}^h)^2}{I-j} \right) \qquad (6)$$

where the C indicates that this is the energy formulated by **C**ataldi et. al. This relies on the weighted sum of the difference of the squared nutrition values for the previous s time intervals. A key difference between the $Cenergy$ and $energy$ is that the $Cenergy$ uses the absolute nutrition values of a term, not the NR values. We transform nutrition values to NR so that terms with consistently large nutritional values would not be considered emergent due to small statistical variations in their nutrition.

An example we found in our YouTube post analysis was that ubiquitous terms, such as "video," tend to have nutrition values that are an order of magnitude greater than their neighbors. As a result of using the $Cenergy$, "video" was considered emerging every Sunday simply due to the larger number of video posts that take place on this day[6], even though its relative importance across different time periods did not change (it was consistently the most popular term). The rankspace transformation of our energy metric alleviates this problem as uniform fluctuations in the total number of posts do not change the normalized rankings across time intervals. Thus terms such as "video" will not have a large energy, and consequently, are not identified as emergent by our system. One could argue that the use of an appropriate stop-word list could rectify the $Cenergy$; however, this presents two problems: (1) removing the consistently-highest-nutrition terms simply creates new high-nutrition terms that have periodic emergence, (2) this list would be unique for each UGC service, and as a result, make our system less general.

Cataldi et. al. used both a user-defined threshold-technique and an automated threshold-technique for determining the emerging terms in a given time interval. As we wish to keep the number of parameters in our system as small as possible, we have chosen not to implement a user-defined threshold. The automated technique determines the emerging-term threshold in the following way:

1. Rank all terms in descending order by energy
2. Let $\max(drop^I) \equiv$ the maximum change in energy (drop) between adjacent terms in the sorted list
3. For all terms that are ranked before $\max(drop^I)$, compute the average drop between adjacent terms
4. The first drop which is higher than the computed average drop defines the threshold for emerging terms

This technique essentially separates a figure-of-merit (FOM) ordered list into two clusters, where the maximum change in a given FOM (in our case, the energy of a keyword) defines the boundary between the two clusters. We applied this procedure to our YouTube dataset and found that the maximum FOM change is nearly always between the highest and second-highest energy-ranked terms: there is usually one anomalous term with an energy significantly larger than the energy values of the other terms. This creates a cluster of size one in which we are to find the mean FOM *difference*—an undefined scenario since a difference requires at least two terms. In addition, using a relative FOM-difference can produce undesirable results in the selection of emerging terms. For example, the last and penultimate energy-sorted terms may have a large relative-FOM-difference, and from the definition provided by Cataldi et. al., we should include the term with the lowest energy in our emerging terms list: clearly something that should be avoided.

Consequently, we have implemented the following automated method to determine the emerging terms:

1. Rank all terms in I in descending order by energy
2. Remove all terms that have a non-positive energy
3. Compute the mean and standard deviation of the positive distribution
4. Label all terms that are greater than two standard deviations larger than the mean as emerging terms

Chebyshev's inequality places an upper bound of 25% on the number of terms that can have energy values larger than two standard deviations from the mean value [14]; however, we find that in practice the actual number of terms is usually around 10-50 (0.003-0.01% of the total terms in I).

Once a list of emerging terms is formed for I, the final step is to extract a set of emerging *topics*. To do this, we explore semantic relationships among the terms in I using directional co-occurrence as a metric, e.g. if "Obama" only appears when "Barack" appears then we can say "Barack" has a strong semantic relationship with "Obama," (although the converse can not directly be inferred, "Barack" may appear quite frequently without "Obama"). Extracting the emerging topics is important because they provide context for the emerging terms. Cataldi et. al. used a weighted correlation vector to express a semantic relationship between terms. Specifically, they used the following probabilistic feedback mechanism developed in [22] to express the semantic correlation, $c_{k,z}^I$, from term k to term z:

$$c_{k,z}^I = log\left(\frac{n_{k,z} / \left(n_k - n_{k,z}\right)}{\left(n_z - n_{k,z}\right) / \left(N - n_z - n_k + n_{k,z}\right)} \right) \times \left| \frac{n_{k,z}}{n_k} - \frac{n_z - n_{k,z}}{N - n_k} \right|$$

where:

- n_k is the number of posts that contain k in I
- n_z is the number of posts that contain z in I
- $n_{k,z}$ is the number of posts that contain k and z in I
- N is the number of total posts in I

Cataldi et. al. used this probabilistic metric to imitate a typical information retrieval problem of locating documents that are correlated with a user's query. Instead of a query and relevant documents, however, Cataldi et. al. used a pair of keywords where one keyword can be viewed as the query

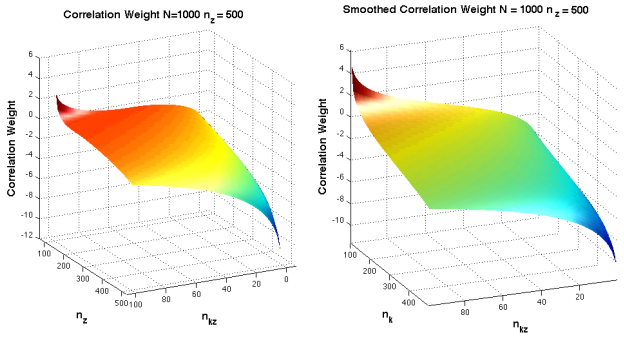

Figure 1: Standard (left) and smoothed (right) correlation weight function ($N = 1000$, $n_z = 500$).

and the other as a document. The correlation weight then represents the directed correlation between two terms rather than the correlation between a query and a document. Note that in the above equation the first term increases as the number of posts that contain both k and z increase, and the second term decreases as the number of posts that contain k but not z increases. Given a term, k, it is then possible to form a *correlation vector*:

$$\mathbf{cv_k^I} = \langle c_{k,1}, c_{k,2}, ..., c_{k,v} \rangle \tag{7}$$

which is used to represent the relationship between k and all of the other $|v|$ keywords in I. In our YouTube data analysis we found this to be a reliable method of inferring semantic correlations of the terms in a given period; however, we encountered a problem with this specific implementation as the correlation vector is not defined if $n_{k,z} = n_k$ or $n_{k,z} = n_z$. A solution to this problem was found through the smoothed correlation metric also introduced in [22]:

$$c^I_{k,z} = log\left(\frac{\left(n_{k,z} + n_k/N\right) / \left(n_z - n_{k,z} + 1\right)}{\left(n_k - n_{k,z} + n_k/N\right) / \left(N - n_k - nz + nk, z + 1\right)}\right)$$

As can be seen in Fig. 1, the two forms of correlation have very similar shapes for a given N and n_z. As is discussed in [22], these correlation metrics tend to yield virtually indistinguishable relative rankings, and to make our system as general as possible, we have used the smoothed correlation metric to determine the directed semantic relationships between terms.

After the formation of the correlation vectors, we created a directed, edge-weighted graph, G, of the emerging terms and all of the first and second order co-occuring terms. A first-order co-occuring term appears in the same post as an emerging term, and a second-order co-occuring term appears in the same post as one of the first-order co-occuring terms. The emerging topics are then found by extracting the strongly-connected-components (SCCs) from G for various threshold values of the correlation weights, where the threshold values are chosen iteratively so that each emerging term appears in at least one SCC. We then rank the SCC's by the average energy of the emerging terms in the SCC and return an interactive graph to the user (see §5). Cataldi et. al. included techniques for determining the minimal set of terms in an emerging topic. Instead, we opt to keep all strongly correlated terms and produce a highly interactive graph that allows the user to explore the different relationships in the SCCs and interactively determine whether to increase the minimal edge-weights.

4. DATA ACQUISITION

In this section we describe a data acquisition process that may be applied to arbitrary UGC systems, allowing for future comparative studies. We then present the specific process used to build the dataset for our YouTube case study. For this study, we retrieved the text contents of approximately 2.2 million YouTube videos. Acquiring all of the YouTube video posts for a given timeframe is a challenging, if not impossible, task for a non-YouTube affliate. Therefore, instead of attempting to gather all of the videos for a given timeframe, our aim was to uniformly sample all of the YouTube video posts within a given timeframe—our justification being that general background noise (non-emerging topics, spam, etc.) will be sampled at the same rate as emerging topics, and since classification of emerging topics is a relative (not absolute) measure, the same emerging topics should be prevalent in a scaled dataset. We did not use stop-lists, but removed all punctuation symbols, web addresses, and single-letter words from our dataset.

We used the YouTube Data API[8] (YDA) to collect our corpus of video posts. The YDA does not have the ability to return a large, uniformly random sample of YouTube video posts. Furthermore, unlike the Twitter API, YouTube does not have a streaming-data API that returns a mixture of realtime results. Twitter, YouTube, and several other UGC APIs (Digg, Facebook, Blogger, Flickr) have a query based functionality that lets users query a particular phrase and obtain relevant results. We exploit this similarity to develop a data collection method that should work consistently across diverse UGC APIs. Specifically, we used the individual terms from the Enron corpus[18] as query phrases in the YouTube API. The Enron corpus contains 28,101 terms from the collection of publicly released Enron emails.

The Enron corpus was chosen as it consists of a large number of broadly defined terms that return a diverse range of UGC content. Using a particular corpus, as opposed to random words chosen from a dictionary, allows for consistency across data collection periods and UGC types, an important requirement for comparative studies. We justify the use of the Enron corpus in several ways. First, we found that only 8.8% of the 28101 terms in the Enron corpus did not return any video results when used as a query term, and the median number of videos returned for each term was 166 (mean 879). Using the same number of query terms randomly sampled from the PubMed abstracts corpus (a corpus heavily biased towards scientific terminology) [11], we found that 45.2% of the terms did not return any video results and the median number of videos returned for each term was 1 (mean 335). This test serves to show that the Enron corpus is capable of returning a large number of videos, as opposed to a selective subset as would be expected for a heavily biased corpus such as PubMed.

We also examined the categorical distribution of the returned videos to address the possibility that the collected videos may be biased towards a particular category (e.g. our dataset could contain a disproportionate amount of *Music* videos). Each video post is assigned a predefined category from the submitting user. The predefined categories for videos have frequently changed in recent years, as of May

[8]http://code.google.com/apis/youtube/overview.html

2011 there are 18 predefined categories, in 2007 there were 14 predefined categories. The two most popular video categories, *Music* and *Entertainment*, have remained unchanged. Using the Enron and PubMed corpuses, we collected data on three distinct occasions and recorded the categorical distributions of the collected videos. In Table 1 we compare the percentage of collected videos in the *Music* and *Entertainment* category with the results from four previous studies that used breadth-first crawlers to collect YouTube posts. Our categorical distribution is in relative agreement with the other studies, and we accept this as evidence that a broad query based approach (supplied with terms from the Enron corpus) can sample data at least as uniformly-random as a breadth-first crawler.

Table 1: A categorical comparison of previous crawler-based studies and our query-based study.

Data	% (Ent. + Music)	Size [Millions]	Collect. Year
[9]	40.7	2.68	2007
[8]	48.1	5.14	2008
[23]	43.04	0.81	2008
[6]	45.0	37.9	2010
enron	41.9	2.71	2011
pubmed	33.2	1.42	2011

The YDA does not allow for the query of videos within a specific timeframe, e.g., it is not possible to request videos from 9/20/08 to 9/22/08. Instead, the user may restrict the video query to the following (inexact) timeframes: *today, this_week, this_month, all_time*. We collected two weeks of data using the *this_week* parameter on 3/12/11 and 3/19/11. Staying within the query confines of the YouTube API, we were able to obtain approximately 2.2 million (unique) video posts during this time frame from an estimated 40 million (non-unique) available video posts.

5. CASE STUDY: YOUTUBE VIDEO POSTS

In this section we discuss the implementation of the UGC-ETD system formulated in §3 for the detection of emerging topics in YouTube video posts. We examine emerging topics in 24-hour time periods (I=24 hours) from 2.2 million YouTube video posts with posting dates between 3/5/2011 and 3/19/2011. We used a historical time range of 5 solar days for predicting the nutrition-rank of a given term, $s = 5$ days. This s value was chosen for two reason: the predicted rank of a term from the regression analysis seldom changed for $s > 5$ (because of the least-squared weighting term), and because an s value of 5 enabled us to place the necessary analysis data into the main memory of our workstations, greatly increasing the speed of the analysis. Furthermore, each day has an average of over 150,000 posts, which attests to the statistical stability of regression analysis: even though the analysis is based on a small number of data points, each data point is drawn from a large amount of data.

In table 2 we report sample emerging topics found by our system that display both the strengths and weaknesses of our generalized implementation. We first note the two emerging topics (on March 11th and 12th) related to the magnitude 9.0 earthquake and resulting tsunami that spawned off the Japanese coast on March 11th 2011, which then caused malfunctions at the Fukushima nuclear power plant in the following days[26]. Video posts relating to this catastrophic event ranged from news reports and video-diary entries to first-hand recordings and music tributes.

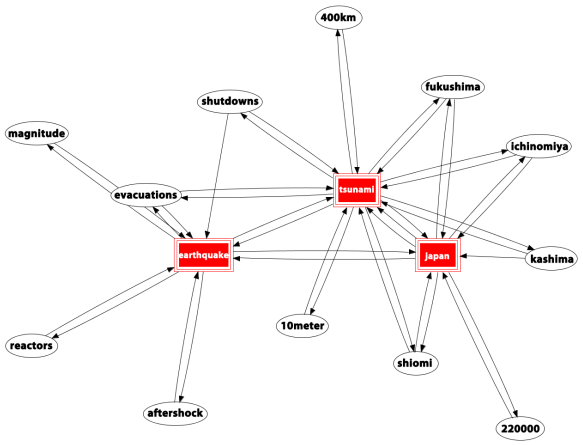

Figure 2: Top emerging topic for March 11 2011. The red squares indicate emerging terms, and the directed edges indicate a high rate of co-occurance.

Examining only the top emerging terms for March 11th, 12th, and 16th essentially provides a summary of the emerging topics: {tsunami, earthquake, japan, magnitude} indicates that an earthquake and tsunami took place near Japan, {fukushima, explosion, nuclear, sendai} indicates that a nuclear explosion took place near Fukushima or Sendai, and {sadiq, batcha, suicide, death} indicates that someone with the name Sadiq or Batcha was involved in either a suicide or death. These summarized topics show that our ETD system can recognize news events, even though the UGC medium (YouTube) was not explicitly designed for this purpose.

Table 2: Top emerging topics for March (9th, 11th, 12th, 16th) detected using our UGC-ETD system.

Date	Top Emerging Topic
9	{**momsen**, beuty, gossip, badgley}
11	{**tsunami**, earthquake, japan, magnitude}
12	{**fukushima**, explosion, nuclear, sendai}
16	{**sadiq**, batcha, suicide, death}

In Figure 2 we display the complete strongly connected component (SCC) for the foremost emerging topic for March 11th. In addition to the most emergent (summarizing) terms, the March 11th SCC also displays non-intuitive correlations with terms that have smaller energies. For example, 220000 and *japan* were found to be highly co-occuring. As it turns out, 220000 was the frequently-cited estimated-number of casualties from the U.S.'s nuclear attack on Japan during World War II. After the earthquake and tsunami, many YouTube video posts described the tsunami as the *worst Japanese tragedy since the 220000 Japanese civilian deaths in WWII*. Other interesting correlations can be found in this SCC such as the correlation between "10meter" and "tsunami" (stemming from posts claiming the tsunami was generating 10-meter high tidal waves), and the correlation between "shutdown" and "earthquake," which is not bidirectional ("earthquake" nearly always accompanied the term "shutdown" but not vice versa).

The foremost emerging topic for March 12th stems from

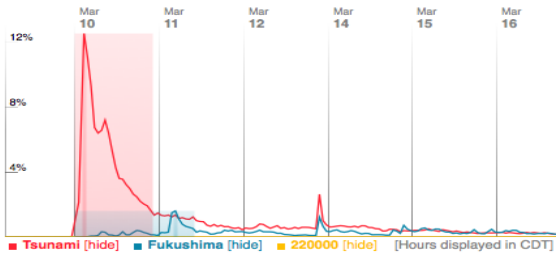

Figure 3: Temporal volume of Twitter posts for *tsunami*, *fukushima*, and *220000* [27].

the March 11th catastrophe, and it is interesting to note that according to Trendistic[27], the same delayed emergence between the terms *tsunami* and *fukushima* was also present in Twitter posts [Fig. 3]. On the other hand, Fig. 3 shows that the term "220000" was unique to YouTube postings. This observation supports the claim that future comparative studies will be able to use our system to study the unique qualities in posting trends from various UGC systems. We also note that smaller-scale news events can be perceived as emerging YouTube topics. A prime example is the foremost emerging topic for March 16th, {sadiq, batcha, suicide, death}, which was related to the suicide of Sadiq Batcha, a key aide to the allegedly corrupt telecommunications minister Andimuthu Raja[21]. This topic became emerging through the large number of video-posts that discussed suspicions of foul-play underlying his suicide.

The SCC in Fig. 2 is a small piece of the full March 11th interactive graph displayed in Fig. 4—note that this graph includes both emerging and non-emerging SCCs that can be explored by the users of our system, where emerging topics contain red dots (emerging terms). These large generated topic graphs are an integral part of our system as they allow the user to interactively explore large SCCs and uncover non-summary terms that are unique to a UGC-system: 220000 was the 13th ranked emerging term in Fig. 2 and would probably not have appeared in the small, typically 4 or 5 term, minimum-spanning SCCs implemented by Cataldi et. al. These graphs are created using a hybrid implementation of the open-source JGraphT Java library[9] in combination with the open-source graph visualization software ZGRViewer.[10]

Our system is not perfect; however, and we find that certain kinds of spammed posts can be perceived as emerging topics. For example, the foremost emerging topic for March 9th [Fig. 5] was formed due to the posting of approximately 600 YouTube videos from different users that contained nearly the exact same textual data (a good indication that these videos were actually from a spammer that created multiple user accounts to avoid YouTube's spam filters). These posts lured YouTube users to a website that allegedly contained pirated episodes of the tv-show Gilmore-Girls (featuring actress Taylor *Momsen*). This type of spam presents a difficult challenge for our general system, and other systems as well, because these spammed posts originate from multiple users. One way we could address this problem is by removing posts with similar text contents. This technique may be successful for YouTube posts, but

[9]http://www.jgrapht.org/
[10]http://zvtm.sourceforge.net/zgrviewer.html

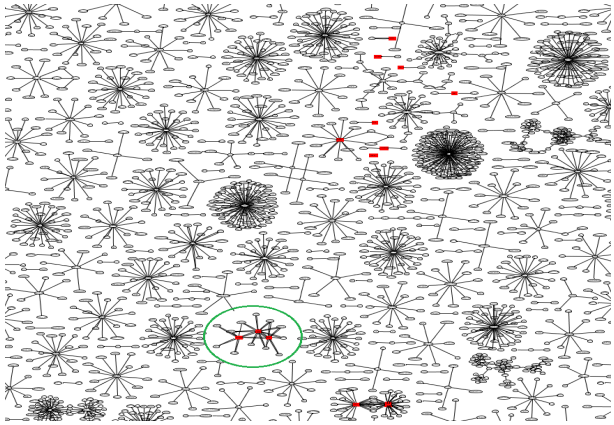

Figure 4: A slice of the March 11th, 2011 emerging topic graph after all emerging terms are included in at least one SCC. Red squares indicate emerging terms (and consequently emerging topics) and a green ring is drawn around the top emerging topic.

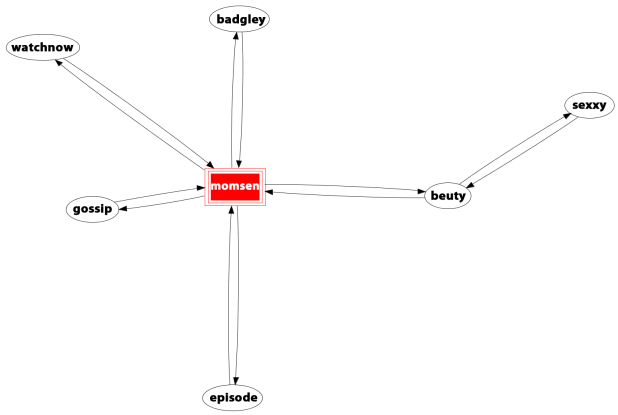

Figure 5: Top emerging topic for March 16 2011: this provides an example of a spammed topic.

at the same time, it could incorrectly bias Twitter posts as Twitter has a feature that allows a user to directly copy another user's posts. Therefore, removing these posts could be detrimental to the detection of emerging Twitter topics, and as a result, negatively impact the generality of our system. In future work, we will study these spam characteristics in the context of multiple UGC systems.

6. CONCLUSION

In this paper we presented a general UGC-ETD system and showed its implementation for the textual data of 2.2 million YouTube video posts published between 3/5/2011 and 3/19/2011. To the best of our knowledge, this is the first general UGC-ETD system created to date. We are currently in the process of including more types of UGC into our system and exploring nonlinear trending models, and we plan to display these results in a follow-up publication. In addition, we plan to use human-based classification of emerging topics to statistically characterize the reliability of our system.

7. REFERENCES

[1] Alexa. http://www.alexa.com/.

[2] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly. Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceeding of the 17th international conference on World Wide Web*, p.895–904. ACM, 2008.

[3] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on Twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, p.1–10. ACM, 2010.

[4] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. I tube, you tube, everybody tubes. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement IMC 07*. ACM Press, 2007.

[5] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Networking (TON)*, 17(5):1357–1370, 2009.

[6] G. Chatzopoulou, C. Sheng, and M. Faloutsos. A first step towards understanding popularity in YouTube. In *2010 INFOCOM IEEE Conference on Computer Communications Workshops*, p.1–6. IEEE, Mar. 2010.

[7] C. Chen, Y. Chen, Y. Sun, and M. Chen. Life cycle modeling of news events using aging theory. *Machine Learning: ECML 2003*, p.47–59, 2003.

[8] X. Cheng, K. Lai, D. Wang, and J. Liu. Ugc video sharing: Measurement and analysis. *Intelligent Multimedia Communication: Techniques and Applications*, p.367–402, 2010.

[9] X. Cheng, J. Liu, and C. Dale. Understanding the characteristics of internet short video sharing: A youtube-based measurement study. *IEEE Transactions on Multimedia*, 2010.

[10] F. Figueiredo, F. Benevenuto, and J. Almeida. The tube over time: characterizing popularity growth of youtube videos. In *Proceedings of the fourth ACM international conference on Web search and data mining*, p.745–754. ACM, 2011.

[11] A. Frank and A. Asuncion. UCI machine learning repository [http://archive.ics.uci.edu/ml]. University of California, Irvine, School of Information and Computer Sciences, 2010.

[12] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: a view from the edge. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, p.15–28. ACM, 2007.

[13] hashtags.org. http://hashtags.org/.

[14] R. Hogg and A. Craig. Introduction to mathematical statistics. *Prentice Hall*, 1994.

[15] S. Jamali and H. Rangwala. Digging digg: comment mining, popularity prediction, and social network analysis. In *Web Information Systems and Mining, 2009. WISM 2009. International Conference on*, p.32–38. IEEE, 2009.

[16] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, p.56–65. ACM, 2007.

[17] H. Kautz, B. Selman, and M. Shah. Referral Web: combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65, 1997.

[18] B. Klimt and Y. Yang. Introducing the Enron corpus. In *First conference on email and anti-spam (CEAS)*, 2004.

[19] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 international conference on Management of data*, p.1155–1158. ACM, 2010.

[20] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. 1999.

[21] L. Polgreen. India Scandal Has Andimuthu Raja, Ex-Minister, at Heart. *The New York Times*, Nov. 2010.

[22] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(02):95–145, 2003.

[23] A. Sharma and M. Elidrisi. Classification of multi-media content (videos on youtube) using tags and focal points. *Unpublished manuscript. Retrieved from http://www-users.cs.umn.edu/~ankur/FinalReport_PR-1.pdf*, 2008.

[24] J. Story and J. Wickstra. Discovering trending topics on twitter via retweets. *Unpublished manuscript. Retrieved from http://cs.uiowa.edu/~jwikstr/finalPaper.pdf*, 2011.

[25] G. Szabó and B. Huberman. Predicting the popularity of online content. *CoRR, abs/0811.0405*, 2008.

[26] H. Tabuchi. Company believes 3 reactors melted down in japan. *New York Times*, p.12–13, May 24, 2011Âą 2011.

[27] Trendistic. http://trendistic.com/.

[28] Twitscoop. http://www.twitscoop.com/.

[29] Twopular. http://twopular.com/.

[30] Y. Wu, Y. Ding, X. Wang, and J. Xu. On-line hot topic recommendation using tolerance rough set based topic clustering. *Journal of Computers*, 5(4):549–556, 2010.