Towards the use of Online Social Networks for Efficient Internet Content Distribution

Amit Ruhela*, Rudra M. Tripathy*, Sipat Triukose† , Sebastien Ardon†, Amitabha Bagchi*, Aaditeshwar Seth*
*CSE Department, IIT Delhi, India 110016
† NICTA, Australia

***Abstract*—A large contributor to the growing Internet traffic is user generated content shared via online social networking websites. Our insight is that these websites can reveal valuable information that can be used in content delivery networks for better caching and pre-fetching performance. In this paper, we combine five different datasets from Twitter and other sources, and make several observations that can lead to helpful heuristics for better content placement. In particular, we study the temporal growth and decay, the geographical spread, and the social spread, of topics on the social network. We also describe in detail our methodologies for data collection, that can be useful for other researchers working in this space as well. In the future, we will use these observations to design heuristics for improved CDN performance.**

I. Introduction

Internet traffic has increased manifold in the last few years. Consumer traffic largely composed of User generated Content (UGC), uploaded and accessed via online social networking websites, is growing at an impressive rate of 36% per year [1]. Access bandwidth however remains the bottleneck, and a key challenge for edge ISPs therefore is to efficiently manage and distribute the content with least investment in upgrading their infrastructure. Content Delivery Networks (CDNs) are seen as primary vehicles to help service providers in this regard, by pre-fetching and caching content likely to be demanded by their customers [2], [3]. Prior work on caching and pre-fetching in CDNs relies on local demand patterns to predict future content demand, but the changing nature of content from broadcast to UGC indicates that different mechanisms may be needed for more accurate placement prediction of user generated content [4]. In this paper, we aim to study three predictors, namely geographical, temporal, and social graph indicators of UGC, and determine their potential to design efficient content distribution strategies. We believe that our study will also assist in the design of better recommender systems, and reveal interesting patterns about the spread of topics discussed on online social networking websites.

Our key contributions in this paper are as follows:

1) We explain the challenges in collecting and mining large datasets, and describe our approach to address these challenges.
2) We combine information from 5 different datasets that we will eventually make public.
3) We study the geographical, temporal, and social characteristics of topic spread within the Twitter OSN, and show their potential in inventing better content placement algorithms.

The roadmap of this paper is as follows: We first describe several available datasets and their limitations. Second, we explain the challenges we faced in collecting missing information of the available datasets. Third, we describe three predictors of topic spread and their usefulness in predicting the popularity of topics for better content placement.

II. Related work

There are numerous studies on the analysis of large datasets from Twitter, Facebook and Youtube [5]–[10]. Our study differs in the extensiveness of the Twitter datasets we combine: We analyze the spread of topics across time, across the geographical location of the users, and across the social network graph of these users. To the best of our knowledge, no other prior work has examined topic spread across all these dimensions. Our work also differs in the technique we use to identify the topic of the tweets: We use OpenCalais on a large dataset – given the ease of use of the open service, our method can be used by other researchers for semantic analysis of tweets and other online content.

The use of online social networking information to aid content distribution further situates our work in its own niche. Earlier studies on content distribution mechanisms have focused on static [11], [12] and dynamic replication techniques [13]–[16]. These works formalized the content replica placement as a complex combinatorial problem and proved it to be NP-hard. However, none focused on the use of social network information to mine dynamic patterns of topic spread, and use it for content placement.

III. Datasets

We next describe various datasets we used, and their individual limitations.

A. Available datasets

Two large Twitter datasets exist in the public domain. The SNAP dataset [17] crawled between June 11, 2009 to September 1, 2009 has 196 million tweets. Along with the tweet content, the SNAP dataset also provides the screen_name of each Twitter user, and the creation times of tweets posted by the users. The second dataset is from KAIST [18], and provides complementary information to the SNAP dataset: it contains nearly the complete social network graph of Twitter

TABLE I
SUMMARY OF AVAILABLE DATASETS

Count of tweets	196,985,580
Count of users	9,801,062
Duration	June 11, 2009 To Sept 1, 2009
Count of re-tweets	15,126,588
Count of hashtags	1,341,733
Count of URLs	54,443,857
Count of follower relations	1468365182

TABLE II
GEOGRAPHICAL STATISTICS

Count of users whose account details were found on Twitter	7,394,244
Count of deleted/banned accounts	2,406,818
Count of users who submitted non-blank locations on Twitter	4,582,233
Count of users whose geographical location was mapped to a latitude-longitude pair	4,007,026
Count of users whose latitude-longitude pair was mapped to a triplet of city, state, and country	3,943,621

users crawled during July 6, 2009 to July 31, 2009, consisting of 1.47 billion follower relationships for 41.7 million Twitter users. Table I summarizes details of both the datasets.

B. Limitations

Since the SNAP and KAIST datasets overlap in time, they can potentially be combined together to yield both the social network as well as the tweets posted by the users. This is important for our analysis since we want to study how topics spread on the social network. Second, none of the datasets provide the geographical location of Twitter users. This is again important for us because we want to study how topics spread across geography, and eventually also use the geographical distribution of users to determine the density of CDN caches that service providers would need to deploy. Third, we need to identify the broad topic or event to which a tweet belongs, so that we can look at a granular spread of topics geographically and on the social network. We next describe how we overcame these limitations by combining the different datasets, and enhancing them with information collected from other sources.

IV. DATA EXTRACTION

Since both the SNAP and KAIST datasets lack geographic information of users and topics of the tweet, we queried the Twitter website for user information, used the Yahoo Geo API to obtain geographic information, and the OpenCalais web service to get topic information. Thus, we effectively made use of 5 datasets, by combining the SNAP and KAIST data with Twitter user information, Yahoo Geo API, and OpenCalais. We next outline the methodology adopted by us to extract this information.

A. Geographical Information

We wanted to find the geographical locations of Twitter users. We therefore queried Twitter to get user profile information, and then used the Yahoo Geo API to find the precise latitude and longitude information of the given location. Details of our data collection methodology are mentioned below.

1) *Querying for the location of Twitter users*:
The SNAP dataset provided us with the Screen_Name of each Twitter user. We then used the Twitter API to find the location given by users in their profile information. We obtained the details of ∼7.39M Twitter users, which comprise 75.4% of all Twitter user accounts in our dataset. The remaining users had either deleted their accounts, or had been banned by Twitter.
2) *Rate Limiting by Twitter*:
We used a white-listed Twitter account that was allowed to execute up to 20,000 queries per hour from Twitter. We ran multiple instances of our crawler to pipeline the requests and launch queries at the maximum allowed rate limit.
3) *Parsing of location information provided by Twitter users*:
We observed that only 61% of all ∼7.39M users in our dataset supplied their location information. No common format was used though. Therefore, we first converted all the extracted locations into a common format as a pair of latitude and longitude coordinates, and then we reverse-converted the coordinates to a triplet of city, state and country. We considered options of Yahoo! PlaceFinder [19], Google Geocoding [20] and the Mapquest Geocoding APIs [21] to geocode the extracted locations. Since Yahoo provided the maximum free rate limit of 50,000 requests per IP addresses/day, we used it to geocode all locations. After this pre-processing, the top 10 countries identified in our dataset in decreasing order were USA, Brazil, UK, Canada, Germany, Indonesia, Netherlands, Australia, India and France.

Table II summarizes the statistics regarding geographical location of Twitter users.

B. Topic Extraction

We wanted to determine the topics of all tweets to analyze how topics spread geographically, temporally, and via the Twitter social network. With ∼196M tweets in our dataset, it was imperative to use automated topic detection tools. Although the "hashtags" nested inside tweets do indicate the topic, but not all tweets contain hashtags. We first pruned our tweet dataset, and then ran a topic identification tool on the tweets. During the pruning stage, we discarded a few tweets of length less than 5 characters, non-English tweets, and those posted by users whose accounts were no longer available on Twitter. Statistics of this pruning exercise are given in Table III.

We then compared four publicly available online topic detection tools: OpenCalais [22], Alchemy [23], Yahoo Term Extractor [24] and Zemanta [25]. Since OpenCalais allowed

TABLE III
TWEETS STATISTICS

Number of tweets with length less than 6	167,814
Number of tweets that fall outside the date range	5,879
Number of non-English tweets	6,779,907
Number of tweets by unknown users	49,741,357
Number of discarded tweets	55,224,096
Number of remaining tweets	141,761,484

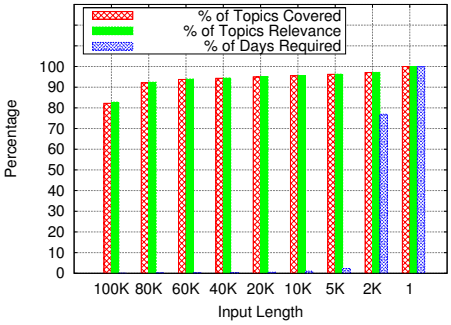


Fig. 1. Comparison of OpenCalais input length

the maximum number of queries per day, we decided to use it for topic identification.

OpenCalais is a Reuter's web service that takes unstructured text as input, and uses machine learning techniques to extract tags and entities automatically from the text. OpenCalais classifies the tags and entities into 38 general categories, and gives a description of each tag and its relevance score (0-1). In this paper, we assume that each distinct tag or entity returned by OpenCalais is a separate topic.

With even 50,000 requests per day, OpenCalais would require more than 3,000 days to obtain topics for the entire dataset of tweets. We handled this by bundling several tweets together when passing them on to OpenCalais. Although the bundling would reduce the time for extraction of topics, but bundling would also change the context and meaning of individual tweets. Therefore, topics returned by OpenCalais for bundled inputs could differ from topics returned for individual tweets. To determine the correct balance, we randomly selected a subset of 25,000 tweets and invoked the OpenCalais API for input lengths of sizes 100K, 80K, 60K, 40K, 20K, 10K, 5K, 3K, 2K, and 1K. This comparison of various input lengths is given in Fig. 1. We plot the following quantities:

1) Percentage of topics returned by OpenCalais when tweets are bundled, compared to topics returned with no bundling.
2) Percentage of the average relevance value of topics returned by OpenCalais when tweets are bundled, compared to the average relevance with no bundling.
3) Expected time to extract topics, as a percentage of the maximum time taken when topics are extracted for individual tweets (estimated as 2976 days to extract topics for the the complete dataset of tweets).

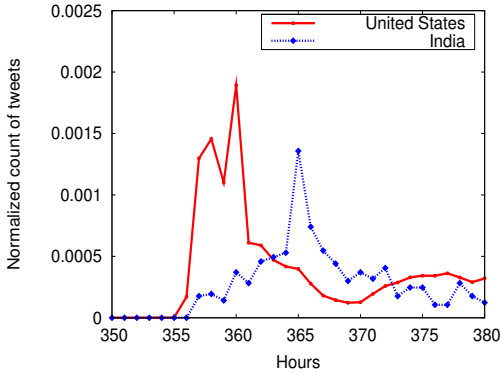


Fig. 3. Time difference of topic MichaelJackson

Fig. 1 shows that as the length of OpenCalais input decreases, the coverage and relevance of topics returned by OpenCalais increases, with a proportional increase in the time taken to extract the topics. We therefore had to choose an optimum value of input length that would minimize the topic identification time and maximize the topic coverage of the tweets in the bundle. On the basis of the above plot, we chose an input length of 40K, that would give a topic coverage of $\sim$94.31% and complete the extraction process in 2 weeks. Overall, we found 39 million URLs and 7.5 million unique topics from OpenCalais. Table IV shows the top topic entry of for each category as defined by OpenCalais.

V. TOPIC SPREAD

We next outline the geographical, temporal, and social spread of topics, and describe UGC placement strategies that could benefit from this information.

A. Spatial/Geographical

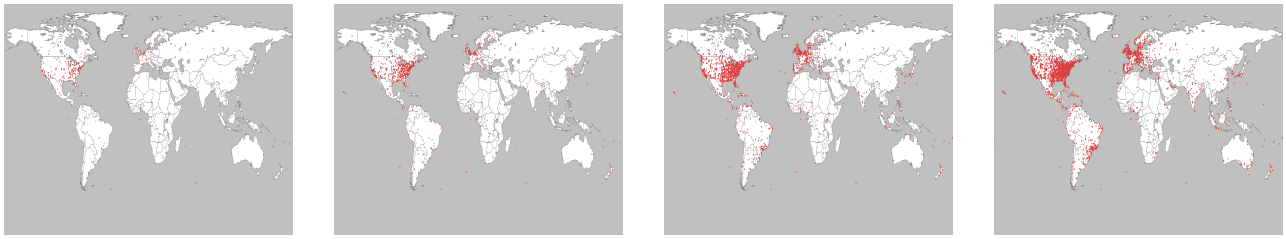
We first use the geographical location information of each twitter account to study how topics spread geographically over time. As an example, using tool given in [26], Fig. 2 plots the location of tweets relating to Michael Jackson's death, which occurred at 2:26 PM PDT. As can be seen, interest in the topic grows over time, but peaks at different times in different countries. This is likely to be related to the time difference across various zones, which we confirmed next.

Fig. 3 shows separately the normalized count of tweets sent by users from India and USA. The time lag of peaks in the two countries differs by five hours. 2:26 PM PDT when the event occurred, was past midnight in India. Indians began tweeting about the event when they woke up the next day. This heuristic can be easily used to predict content demand in different time zones.

Conclusion: The time lag of topic spread between the site of event occurrence and different time-zones of the world, can be used to replicate UGC pro-actively before the arrival of content demand from these locations.

TABLE IV
TOP ENTRY OF EACH OPENCALAIS CATEGORY

Class	Instance	Class	instance	Class	Instance
City	London	Company	Twitter Inc.	Continent	America
Country	United States	Currency	USD	EmailAddress	flag@whitehouse.gov
EntertainmentAwardEvent	Teen Choice Awards	Facility	App Store	FaxNumber	866-374-8858
Holiday	Christmas	IndustryTerm	food	MarketIndex	Top 40
MedicalCondition	swine flu	MedicalTreatment	surgery	Movie	Julie & Julia
MusicAlbum	Canvas	MusicGroup	Energy	NaturalFeature	Grand Canyon
OperatingSystem	Windows 7	Organization	National Football League	Person	Michael Jackson
PhoneNumber	3722911	PoliticalEvent	presidential election	Position	President
Product	iPhone	ProgrammingLanguage	php	ProvinceOrState	California,United States
PublishedMedium	New Moon	RadioProgram	This American Life	RadioStation	Wine
Region	Southwest	SportsEvent	NFL	SportsGame	football
SportsLeague	NFL	TVShow	G.I. Joe	TVStation	Kyle
Technology	http	URL	http://140mafia.com		



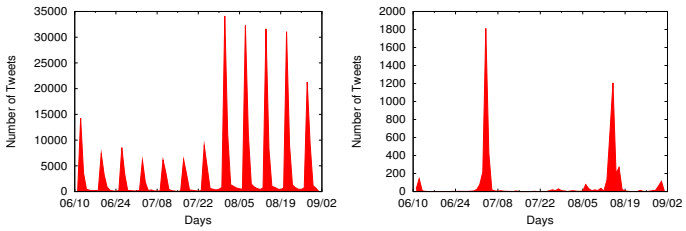
(a) 1:56 PM (b) 2:26 PM (c) 3:26 PM (d) 5:26 PM

Fig. 2. Geographical dispersion of topic Michael Jackson with time (PDT or GMT- 7Hrs)

B. Temporal Classification

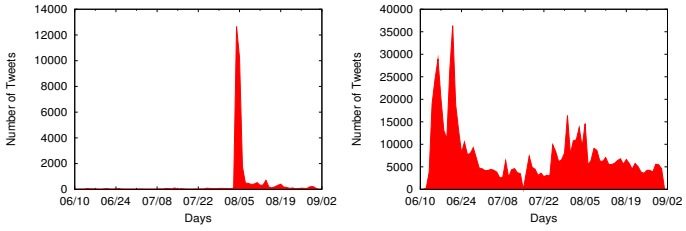
We next study how topic popularity changes over time. By studying 150 most popular topics, we show that topics can be broadly classified along 4 axis: periodicity, stability, growth rate, and decay rate. Each class would require different content placement strategies, and we believe that semantic analysis of hashtags and tweets can indicate which class a particular topic would belong to.

1) **Periodic vs Aperiodic**: A few topics are periodic and gain popularity at regular intervals. Fig. 4 shows instances of a few periodic topics. We further classify periodic topics as geography-independent and geography-dependent periodic topics. The topic (*#followfriday*) as shown in Fig. 4a repeats globally on weekends and is not confined to a particular geography, whereas the topic in Fig. 4b corresponding to the topic "Independence Day" repeats over time but confines itself to two different geographies (U.S. and India).
Conclusion: Once topics have been identified as periodic, content placement strategies can react accordingly and replicate content at the appropriate time.
2) **Ephemeral vs Stable Topics**: Topics across their lifetime can be classified under two categories: Ephemeral topics and Stable topics. Ephemeral topics have distinct peaks extending over a few days, whereas stable topics remains popular for a large number of days of the order of weeks and months. Fig. 5 shows examples of a few ephemeral and stable topics. The first topic corresponds to the event of "*Bill Clinton went to North Korea to Seek Release of U.S. Reporters*" and the second topic corresponds to the "*Iran Election*" whose popularity spans for more then 2 months in our dataset.
Conclusion: Content placement strategies should take into account the stability of topics to determine which content to evict from caches and which content to retain.
3) **Slow vs Sharp Growing Topics**: Based on the growth rate of topic popularity, we see that some topics become popular gradually. These topics typically include scheduled events that people already know about, so the rate of tweets rises gradually and the topics finally attain maximum popularity closer to the event dates. On the other hand, unscheduled events gain popularity sharply, and may even causes flash crowds. Fig. 6a shows an example of a gradually growing topic which corresponds to the release of a movie "*G.I. Joe: The Rise of Cobra*", whereas Fig. 6b shows an example of a sharply growing topic corresponding to the "*death of Michael Jackson*".
Conclusion: The growth rate of topics and the scheduled dates of events (if known), can assist in content placement strategies to predict times of maximum demand for the content, and hence assess cost-benefit tradeoffs of when to replicate the content.



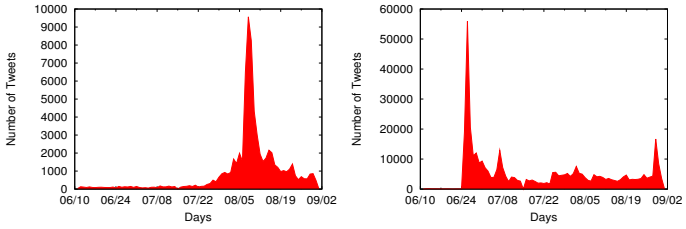
(a) FollowFriday (b) IndependenceDay

Fig. 4. Periodic topics



(a) Bill Clinton (b) Iran Election

Fig. 5. Ephemeral vs stable topics



(a) GiJoe (b) MichaelJackson

Fig. 6. Slow vs sharp growing popularity

4) **Slow vs Sharp Decaying Topics**: Similar to differences in growth rate, topics also exhibit different decay profiles. In case some events follow the main event, the decay is gradual. Fig. 6a shows an example of a slowly decaying topic about a movie; a number of spikes are also observed on the weekends. Some other topics on the other hand, as shown in Fig. 5a, decay sharply and die out soon after the event.
Conclusion: The decay rate of the topics can assist in content replacement to either uphold or delete content at CDN servers.

C. Social Network

We next wish to determine if the spread of topics on the social network indicates any patterns that can be used to predict content demand. We examine popular topics and niche topics separately, and use a metric for social cohesion defined as the ratio of the number of follower relations that exist between UGC producers of that topic and the maximum possible follower relations that could exist between them $= 2$ x n_2^C. We identify niche topics as those of interest to a small

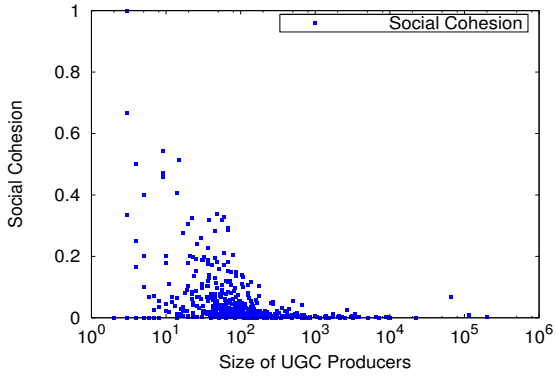


Fig. 7. Social cohesion strength of UGC producers

segment of users, ie. topics having a small number of users and comprising the long tail. We select 500 niche topics manually from our topics database.

Fig. 7 plots a graph between the count of users for a topic versus the social cohesion value of that topic. We observe that as the number of users talking about a certain topic increases, the number of social relations that exist between these users decreases. Thus we can say that as we move from popular to niche topics, the social cohesion values increase. This has direct implications on content placement strategies: The social network of niche content producers can be used to predict users who would be interested in the topic, and consequently the locations of these users can reveal the geographies where demand would arise in the future. Thus, the social network information of UGC producers can help decide UGC placement strategies.

VI. Conclusion

In this paper, we describe how geographical, temporal, and social characteristics of UGC spread can be obtained from online social networking websites, and potentially used for better content caching and placement strategies. In particular, we identified different classes of temporal growth patterns of UGC, we noticed that UGC popularity peaks at different times in different time zones, and that social cohesion among users interested in topics is greater for niche topics as compared to popular topics. These cues can be used to design content placement heuristics such as the following: use semantic information about the topic to assess what class the temporal growth of the topic would belong, use time-zone information to predict when a particular content would become popular in some other country, and use social network predictors for niche content in priority to geographical predictors. We also described several challenges we faced in collecting datasets from social networking websites, and described our methodology to address these challenges. We are currently running more extensive tests to confirm our hypotheses about geographical, temporal, and social predictors for content distribution. We will then subsequently design distributed algorithms that can coordinate the placement of UGC on content delivery networks.

ACKNOWLEDGMENT

The authors would like to thank Anirban Mahanti from NICTA for his continuous guidance in conducting this study; and Yogesh Kumar and Ravi Gupta for their valuable support and suggestions. This work was supported by the Commonwealth of Australia and the Department of Science and Technology, India, under the Australia-India Strategic Research Fund, and the Department of Information Technology, India under project ID RP02355.

REFERENCES

[1] "Cisco visual networking index: Forecast and methodology, 2010-2015 [visual networking index] - cisco systems," 2011. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360.pdf

[2] K. Stamos, G. Pallis, C. Thomos, and A. Vakali, "A similarity based approach for integrated web caching and content replication in cdns," in *Database Engineering and Applications Symposium, 2006. IDEAS '06. 10th International*, dec. 2006, pp. 239 –242.

[3] S. Bakiras and T. Loukopoulos, "Combining replica placement and caching techniques in content distribution networks," *Comput. Commun.*, vol. 28, pp. 1062–1073, June 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=1646674.1647256

[4] N. Sastry, E. Yoneki, and J. Crowcroft, "Buzztraq: predicting geographical access patterns of social cascades using social networks," in *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*. Nuremberg, Germany: ACM, 2009, pp. 39–45.

[5] M. C. Munmun De Choudhury, Scott Counts, "Find me the right content! diversity-based sampling of social media content for topic-centric search," in *In Proceedings of the 5th Int'l AAAI Conference on Weblogs and Social Media*. www.aaai.org, 2011, pp. 129–136.

[6] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *WWW '10: Proceedings of the 19th international conference on World wide web*. New York, NY, USA: ACM, 2010, pp. 591–600.

[7] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook," in *WOSN '09: Proceedings of the 2nd ACM workshop on Online social networks*. New York, NY, USA: ACM, 2009, pp. 37–42.

[8] B. Xu and L. Liu, "Information diffusion through online social networks," in *Emergency Management and Management Sciences (ICEMMS), 2010 IEEE International Conference on*, aug. 2010, pp. 53–56.

[9] 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*. San Diego, California, USA: ACM, 2007, pp. 15–28.

[10] 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*, vol. 17, no. 5, pp. 1357–1370, 2009. [Online]. Available: http://dx.doi.org/10.1109/TNET.2008.2011358

[11] J. Kangasharju, J. Roberts, and K. W. Ross, "Object replication strategies in content distribution networks," *Computer Communications*, vol. 25, no. 4, pp. 376 – 383, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140366401004091

[12] S. Sivasubramanian, G. Pierre, M. van Steen, and G. Alonso, "Analysis of caching and replication strategies for web applications," *IEEE Internet Computing*, vol. 11, no. 1, pp. 60–66, 2007. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4061123

[13] Y. Chen, R. H. Katz, and J. Kubiatowicz, "Dynamic replica placement for scalable content delivery," in *Revised Papers from the First International Workshop on Peer-to-Peer Systems*. Springer-Verlag, 2002, pp. 306–318.

[14] F. L. P. N. Bartolini, "Optimal dynamic replica placement in content delivery networks," in *Proceedings of ICON 2003*, Sydney, Australia, 2003, pp. 125—130.

[15] F. L. Presti, C. Petrioli, and C. Vicari, "Dynamic replica placement in content delivery networks," in *Proceedings of the 13th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*. IEEE Computer Society, 2005, pp. 357–360.

[16] C. Vicari, C. Petrioli, and F. L. Presti, "Dynamic replica placement and traffic redirection in content delivery networks," *SIGMETRICS Perform. Eval. Rev.*, vol. 35, no. 3, pp. 66–68, 2007.

[17] "Snap: Network datasets: 476 million twitter tweets," 2011. [Online]. Available: http://snap.stanford.edu/data/twitter7.html

[18] "What is twitter, a social network or a news media?" 2011. [Online]. Available: http://an.kaist.ac.kr/traces/WWW2010.html

[19] "Yahoo! placefinder guide," 2011. [Online]. Available: http://developer.yahoo.com/geo/placefinder/guide/

[20] "The google geocoding api," 2011. [Online]. Available: http://code.google.com/apis/maps/documentation/geocoding

[21] "Mapquest geocoding api service," 2011. [Online]. Available: http://developer.mapquest.com/web/products/dev-services/geocoding-ws

[22] "How does calais work? — opencalais," 2011. [Online]. Available: http://www.opencalais.com/about

[23] "Alchemy api," 2011. [Online]. Available: www.alchemyapi.com/api/entity/

[24] "Yahoo! term extraction," 2011. [Online]. Available: http://developer.yahoo.com/search/content/V1/termExtraction.html

[25] "Zemanta - contextual intelligence for everyone!" 2011. [Online]. Available: http://developer.zemanta.com/

[26] "Caida plot-latlong," 2011. [Online]. Available: http://www.caida.org/tools/visualization/plot-latlong/