Structural Analysis of the Emerging Event-web

Vivek K. Singh
University of California, Irvine
singhv@uci.edu

Ramesh Jain
University of California, Irvine
jain@ics.uci.edu

ABSTRACT

Events are the fundamental abstractions to study the dynamic world. We believe that the next generation of web (i.e. *event-web*), will focus on interconnections between events as they occur across space and time [3]. In fact we argue that the real value of large volumes of microblog data being created daily lies in its inherent spatio-temporality, and its correlation with the real-world events. In this context, we studied the structural properties of a corpus of 5,835,237 Twitter microblogs, and found it to exhibit Power laws across space and time, much like those exhibited by events in multiple domains. The properties studied over microblogs on different topics can be applied to study relationships between related events, as well as data organization for event-based, real-time, and location-aware applications.

Categories and Subject Descriptors

H.4.m [Information Systems]: Miscellaneous

General Terms

Experimentation, Human factors, Design

Keywords

Event-web, Microblogs, Power Law, Zipfs law, Pareto law

1. INTRODUCTION

The current Web is largely document-centric hypertext. Unlike events, hypertext has no notion of time, space, or semantic structures other than often ad hoc hyperlinks [3]. Going forward, we anticipate multiple applications organizing data around *events* in terms of their interconnections, geo-locations, and temporal bounds. The applications will range from real-time event search, to automated trip planning, to spatio-temporal situation assessment. We feel that microblogs like Twitter will have an important role to play in the realization of such an Event-web. Twitter posts are inherently spatio- temporal, and can be easily encoded with semantic structures like hash-tags (e.g. #iran-elections) to support data organization. Thus we anticipate the information organization to soon occur, not via 'URLs' but rather via 'event tags' and across 'geo-locations'. Early signs of such trends are visible with Google and Microsoft providing Twitter based search results for real-time events, and exponential growth of tools like Yelp and Foursquare.

In such a context, we undertook a structural analysis of the evolving micro-blogosphere with a specific focus on 'hash-tags'. We argue that people make a conscious decision to hash-tag their post, *because they want to relate it to an event which is relevant to others in the same spatio-temporal volume*. For example *#iran-election* was related to the real world Iran elections event, and *#iphone* tag in the data set, was related to the iphone 3G release event -- even though the name *#iphone* may not directly look like an event tag. In fact a recent study which monitored 'earthquake' related posts [5] has modeled Twitter as a distributed event sensor, and [6] has reported the direct impact of real world events on the popularity of topics in blogosphere. Hence, we consider each topic (i.e. hash-tag) in Twitter to signify a real world event.

Copyright is held by the author/owner(s).
WWW 2010, April 20–24, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

Power law and structural properties of web were first studied in 1999 by works like [1] and [2]. Recent studies have reported similar power law properties (e.g. in terms of number of posts per user [6], or interconnections between users [7]) in the micro-blogosphere. However, the impact of recent occurrences like Haiti earthquake, Michael Jackson's death, Iran elections etc. on the web, has underscored the importance of *events*, and their real world *location* and *time,* for data organization in practical applications. This motivates our study into the spatio-temporal structure exhibited by event related data in micro-blogosphere.

We report our findings on 5,835,237 Twitter microblogs collected from across the globe for one month. We found specific patterns in event-tag occurrences across the globe. In fact, while the author of each micro-blog has complete freedom in choosing which (and how many) event-tags to use for it, we found that the overall system obeys scaling and structural laws characteristic only of highly interactive self-organized systems and critical phenomena [2]. Organization of tweets based on event-tags exhibited patterns in terms of Power laws (Pareto's law, Zipf's law, and Richter-Gutenberg law). Significantly, we found the patterns to be surprisingly robust and not varying significantly from region to region or over time durations considered. Hence, these findings can indeed be used by application designers (e.g. using Pareto law, to focus only on top 20% events for real time indexing) in spatio-temporal event based applications.

2. METHODS AND FINDINGS

The findings presented here are based on 5.8 million tweets downloaded between Aug 22, 2009 and Sep 21, 2009 using Twitter's public 'spritzer' stream. The geo-location used is based on the 'home' location of the user, as converted by the Geonames server (*www.geonames.org*). Only tweets with locations successfully geocoded, were used for location-based experiments.

We define *magnitude* of each *event* as the frequency of occurrence of that event-tag in the corpus. In the data set studied we found a total of 2,310,104 (2.3 million) event-tag occurrences, corresponding to 292,981 unique event-tags.

2.1 Pareto's law

The Pareto's law states that, for many domains, roughly 80% of the effects come from 20% of the causes. We found that most popular 20% of event tags corresponded to more than 80% of all tag occurrences. In fact as shown in Table 1, we noticed that such a Pareto-law distribution holds even for the spatial and temporal subsets of the corpus, with the subsets becoming increasingly top-heavy as the size of the corpus increased.

Corpus Space	Corpus Time	Num. occurrences	Top 20%	Percentage
Whole world	30 mins	2,218	1,476	66.5%
Whole world	1 day	76,256	61,521	80.67%
Whole world	1 week	508,389	427,539	84.1%
Whole world	2 weeks	1,175,076	1,019,113	86.7%
Whole world	3 weeks	1,924,768	1,690,459	87.8%
Whole world	1 month	2,310,104	2,036,552	88.1%
US only	1 month	388,291	321,579	82.8%

Table 1: Distribution of event tags

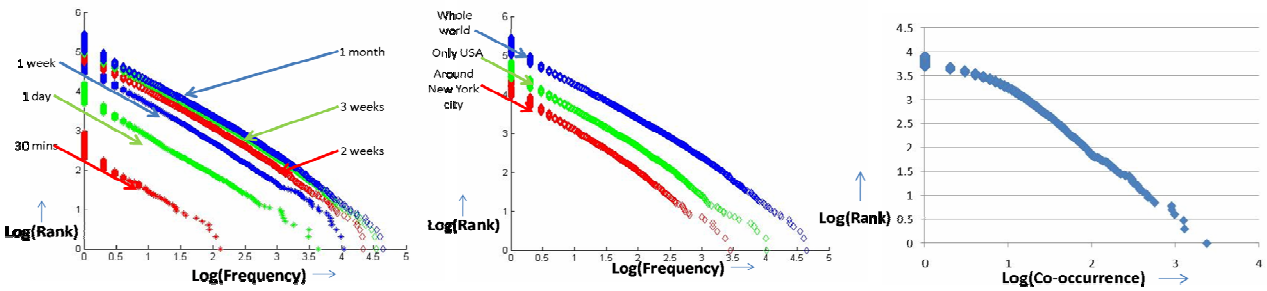

Fig. 1: Variation of frequency of event tags and their ranks for different (a) time durations. (b) geo-regions. (c) Co-occurrence pattern across event tags.

2.2 Zipf's law and Gutenberg-Richter law

Zipf's law, is the observation that in many domains the frequency of occurrence of some event ***P***, as a function of the rank ***i*** when the rank is determined by the frequency of occurrence, is a *power-law* function $P_i \approx 1/i^{\alpha}$. This pattern has been found for the corpus of English words (most frequent word occurs approximately twice as often as the 2nd most frequent word), population ranks of cities in various countries, income rankings etc.

Zipf's law (or more generally the Zeta distribution [4]) is most easily observed by plotting the data on a log-log graph, with the axes being **log(rank order)** and **log(frequency)**. A dataset that conforms to Zipf's law will lead to a linear plot.

For micro-blogosphere, we wanted to verify *if there is a fixed relative ratio for the occurrence of events of different magnitude*. Further still we wanted to see *if this ratio remains constant, if we vary the space or the time-duration being considered?*

In effect, we also wanted to verify another Power law, known as the Gutenberg–Richter law which (in seismology) expresses the relationship between the magnitude and total number of earthquakes in any given *region* and *time* i.e. **log N = a - b M** where: ***N*** is the number of events in a given magnitude range, ***M*** is a magnitude minimum, ***a*** and ***b*** are constants. While Gutenberg–Richter law can be interpreted as a variation of the Zipf's Law [4], the interesting aspect is that the relationship is robust, and does not vary significantly from region to region or over time.

In our data set, we ranked different event-tags based on their frequency of occurrence, and plotting the rank of the tags against the frequency of occurrence on a Log-Log scale gave a linear plot. More interestingly (See Fig 1(a)), when we took subsets of the data corpus for different time durations (30 mins, 1 day, 1 week, 2 weeks, 3 weeks, and 1 month), we observed separate but still linear (and with very similar slope), plots for each one of them . Further still (see Fig. 1(b)) when we took subsets of data based on location (whole world, just United States, and just a 10 latitude by 10 longitude block around New York city), each collection independently exhibited the Zipf's law. Hence we found that Zipf's law (and Gutenberg-Richter law) holds for each reasonably sized spatial and temporal collection of event data. This finding may be useful for localized indexing and data assimilation, which will be a vital thrust of the evolving event-web.

2.3 Linking across events

What made WWW a 'web', was the interlinking across different websites. The in-links and out-links from websites, and their structure (e.g. 80% links point to 20% websites [1]) were used in Page-rank and many related technologies to decide on relative importance of the websites. We believe that evolving dynamic web would be a 'web' of interconnected events across time and space [3]. Thus, the interlinking across event-tags would be an important indication of relative importance of events. Hence, we proceeded to study the co-occurrence patterns of event tags in the microblog data. For example a tweet like *"Show trials resume, Montazeri speaks out - The Majlis - http://shar.es/1bAel #iranelection #neda"*, shows a possible connection between event tags *#iranelection* and *#neda*. We found that such co-occurrence patterns also followed the Power law over several orders of magnitude. While event tags like *#jobs* and *#tcot*, were linked to thousands of other tags, there were thousands others which had no co-occurrences. Hence, just like Page-rank, such event link structures can be used to assign importance to events in future.

3. APPLICATIONS AND FUTURE WORK

Just like the findings about the structure of traditional web [1,2] were used to design better caching mechanisms (e.g. caching only top 20% of websites), our findings will allow the event web architects to make similar decisions e.g. focusing only on top 20% of events for real-time indexing, or customized advertisement creation, or real time marketing outreach. Further still, the spatio-temporal properties of the distributions studied, might be critical for resource planning in evolving location based services. The finding that microblog data exhibits 'Zeta distribution', which is significantly different from the traditional 'Poisson distribution' assumption, can be used for better/ correct modeling of web traffic growth in near future. The Zeta distribution pattern can also be used to extrapolate and calculate the impact of various events or campaigns across Geo-locations, by just using their relative ranking with other well studied events.

While the focus of studied Power laws and our current finding has been on *occurrence* and *co-occurrence* patterns; in near future, we also intend to study *causality* across such events.

4. REFERENCES

[1] L. Adamic and B. Huberman. Growth dynamics of the World Wide Web. Nature, 401(6749):131, 1999.
[2] R. Albert, H. Jeong, and A. Barabasi. Diameter of the world-wide web. Nature, 401(6749):130-131, 1999.
[3] R. Jain. Eventweb: Developing a human-centered computing system. IEEE Computer, 41(2), 2008.
[4] A. Clauset, C.R. Shalizi, M.E.J. Newman, "Power-law distributions in empirical data" SIAM Review 51(4), 2008.
[5] Sakaki et al., "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors". WWW conference 2010. (Accepted)
[6] Gruhl et al., "Information diffusion through blogspace", WWW conference, 2004.
[7] Java et al., "Why we twitter: understanding microblogging usage and communities". In WebKDD/SNA-KDD, 2007.