**Discovering Emerging Topics in Social Streams via**

**Link Anomaly Detection**

**Abstract:**

Detection of emerging topics is now receiving renewed interest motivated by the rapid growth of social networks. Conventional term-frequency-based approaches may not be appropriate in this context, because the information exchanged in social network posts include not only text but also images, URLs, and videos. We focus on emergence of topics signaled by social aspects of theses networks. Specifically, we focus on mentions of users – links between users that are generated dynamically (intentionally or unintentionally) through replies, mentions, and retweets. We propose a probability model of the mentioning behaviour of a social network user, and propose to detect the emergence of a new topic from the anomalies measured through the model. Aggregating anomaly scores from hundreds of users, we show that we can detect emerging topics only based on the reply/mention relationships in social network posts. We demonstrate our technique in several real data sets we gathered from Twitter. The experiments show that the proposed mention-anomaly-based approaches can detect new topics at least as early as text-anomaly-based approaches, and in some cases much earlier when the topic is poorly identified by the textual contents in posts.

**Architecture Diagram:**

****



**Existing System:**

A new (emerging ) topic is something people feel like discussing, commenting, forwarding the information further to their friends. Conventional approaches for topic detection have mainly been concerned with the frequencies of (textual) words.

**Disadvantages:**

A term frequency -based approach could suffer from the ambiguity caused by Synonyms ( or) homonyms. It may also require complicated preprocessing(ex. segmentation).Depending on the target language. Moreover, it cannot be applied when the messages are mostly non-textual information .On the Other hand the “words” formed by mentions are unique, require little preprocessing to obtain (the information is often separated from the contents )and are available regardless of the nature of the contents.

**Proposed System:**

1. In this paper , we have proposed a new approach to detect the emergence of topics in a social network Stream.
2. The basic idea of our approach is to focus on the social aspect of the posts reflected in the mentioning behavior of users instead of the textual contents.
3. We have proposed a probability model that captures both the number of mentions per post and the frequency of mentionee.

**Advantages of Proposed System:**

1.The proposed method does not rely on the textual contents of social network posts. It is robust to rephrasing and it can be applied to the case where topics are concerned the information other than texts, such as images, videos, audio and so on.

2.The proposed link anomaly based method performed even better than the keyword based methods on “NASA” and “BBC” dataset.

**Implementation Modules:**

1. Twitter Login

I . register ii. Followers

iii. Follow iv. Tweet

v. Re-tweet

2. Data sets

i. Job hunting ii. You tube

iii. NASA iv.BBC

3. Computing the link-anomaly score

4. Combining Anomaly Scores from Different Users

**Twitter Login**:

 Twitter has evolved from a micro blogging service into a popular social messaging platform, it has been instrumental in providing the "pulse" on news and events across the globe. In addition to its widespread acceptance among the news media and entertainment industry, Twitter has also become a popular social media marketing tool and a great way to communicate with both friends and co-workers. It is many different things to many different people. It can be used by a family to keep in touch, or a company to coordinate business, or the media to keep people informed or a writer to build up a fan base. Twitter is micro-blogging. It is social messaging. It is an event coordinator, a business tool, a news reporting service and a marketing utility.

**Follow:**

[Twitter](http://webtrends.about.com/od/twitter/u/twitter_guide.htm#s1) can be a great way to promote you, your brand or your company. Actors, writers, sports stars and others are turning to Twitter as a way to connect with fans and promote themselves to millions of people across the globe. Twitter is also a great way for bloggers to gain more traction in the [blogosphere](http://webtrends.about.com/od/glossary/g/Blogosphere_d.htm) and get noticed.

But how do you build up a Twitter following? These simple tips will help you get started.

**Follow People.** The number one way to gain followers is to follow people. Many Twitter users follow anyone that follows them, while others check to see if the profile is active before deciding to follow it. You can find interesting people to follow by doing a [Twitter search](http://webtrends.about.com/od/socialnetworking/ss/how_to_search_twitter_track_trends.htm).

**Be Active.** Don't just link your web feed to your Twitter profile. Post messages and make your Twitter page a place to connect with your readers.

**Engage People.** Ask questions. Conduct polls. Find ways to engage your followers in the discussion.

**Promote your Profile.** Add a link to your Twitter profile to your blog, website and as part of your signature in email messages and discussion forums.

**Respond to Messages.** Always pay attention to who mentions you. If you aren't following them, do so. And when you get a direct message, make sure you respond.

**Followers**:

A number of companies have sprung up offering to help people and companies grow their Twitter following the easy way by forking over a little cash. The usefulness of a large Twitter following isn't simply a number. Having 10,000 followers won't do you much good if only half of them are real people and most of them aren't interested in anything you [tweet](http://webtrends.about.com/od/glossary/g/what-is-a-tweet.htm). It may change the number of followers at the top of the screen, but it will have little effect on how many people actually follow what you say. The entire point of growing a large Twitter following is to communicate with people who are interested in your topic, your company or your brand. They want to read what you write, [retweet](http://webtrends.about.com/od/twitter/a/twitter_help_what_is_a_retweet.htm) what you tweet and check out the links you post. This is what makes a Twitter following valuable. And a Twitter following that simply ignores your tweets simply isn't worth it.

**Tweets:**

A tweet is a post or status update on [Twitter](http://webtrends.about.com/od/profiles/p/twitter-profile.htm), a micro-blogging service. Because Twitter only allows messages of 140 characters or less, "tweet" is as much a play on the size of the message as it is on the audible similarity to Twitter.

**Retweets**:

A "retweet" is a reply to a [tweet](http://webtrends.about.com/od/glossary/g/what-is-a-tweet.htm) that includes the original message or a tweet that includes a link to a news article or blog post that you find particularly interesting. Like [hash tags](http://webtrends.about.com/od/twitter/a/twitter_help_what_is_a_hashtag.htm), re-tweets are a recent community-driven phenomenon on [Twitter](http://webtrends.about.com/od/profiles/fr/Twitter-Profile.htm) with the aim of making the service better and allowing people to follow discussions easier.

**Registration:**

When you join Twitter from your phone, we collect all of your updates on the web. When you're ready to activate your web account, all you have to do is add an email address and a password to complete your profile and log in.

1. Enter your full details of the signup page.
2. Enter your **phone number** when prompted.
3. Twitter will send a **verification code** to your phone. When you get that text message, enter the code (shown below).
4. You'll then be asked to create an account by entering your **email address**, a **password**, and a **name** for your account. Your username will already be entered, since you created this when you signed up via SMS.
5. Click **Create my account** as shown below, and you're all done! Twitter will walk you through finding some friends you may know on Twitter, then will direct you to your home page

**Data Sets:**

**Job Hunting dataset:**

This data set is related to a controversial post by a famous person in Japan that “the reason students having difficulty finding jobs is, because they are stupid” and various replies to that post. The keyword used in the keyword-based methods was “Job hunting.” Figures 5a and 5b show the results of the proposed link-anomaly-based change detection and burst detection, respectively. Figures 5c and 5d show the results of the text-anomaly-based change detection and burst detection, respectively. Figures 5e and 5f show the results of the keyword-frequency-based change detection and burst detection, respectively.

**You tube dataset:**

This data set is related to the recent leakage of some confidential video by the Japan Coastal Guard officer. The keyword used in the keyword-based methods was “Senkaku.” Figures 6a and 6b show the results of link-anomaly-based change detection and burst detection, respectively. Figures 6c and 6d show the results of text-anomaly-based change detection and burst detection, respectively. Figures 6e and 6f show the results of keyword-frequency based change detection and burst detection, respectively.

**NASA dataset:**

This data set is related to the discussion among Twitter users interested in astronomy that preceded NASA’s press conference about discovery of an arsenic eating organism. The keyword used in the keyword-based models was “arsenic.” Figures 7a and 7b show the results of link anomaly- based change detection and burst detection, respectively. Figures 7c and 7d show the results of text anomaly- based change detection and burst detection, respectively. Figures 7e and 7f show the same results for the keyword-frequency-based methods.

**BBC dataset:**

This data set is related to angry reactions among Japanese Twitter users against a BBC comedy show that asked “who is the unluckiest person in the world” (the answer is a Japanese man who got hit by nuclear bombs in both Hiroshima and Nagasaki but survived).The keyword used in the keyword-based models was “British” (or “Britain”). Figures 8a and 8b show the results of link-anomaly-based change detection and burst detection, respectively. Figures 8c and 8d show the results of text-anomaly-based change detection and burstdetection, respectively. Figures 8e and 8f show the same results for the keyword-frequency-based methods.

**Computing the link-anomaly score:**

In this subsection, we describe how to compute the deviation of a user’s behaviour from the normal mentioning behavior modeled in the previous subsection; In order to compute the anomaly score of a new post ***x*** = (*t, u, k, V* ) by user *u* at time *t* containing *k* mentions to users V. The two terms in the above equation can be computed via the predictive distribution of the number of mentions , and the predictive distribution of the mentionee , respectively.

**Combining Anomaly Scores from Different Users:**

we describe how to combine the anomaly scores from different users; The anomaly score is computed for each user depending on the current post of user *u* and his/her past behaviour *T* (*t*) *u* . In order to measure the general trend of user behaviour, we propose to aggregate the anomaly scores obtained for posts ***x***1*, . . . ,* ***x****n* using a discretization of window size *τ >* 0 as follows:



**Algorithms:**

1.Change point detection

2.Burst detection.

**Algorithm definition:**

**Change point detection:**

In [statistical analysis](http://en.wikipedia.org/wiki/Statistical_analysis), change detection or change point detection tries to identify times when the [probability distribution](http://en.wikipedia.org/wiki/Probability_distribution) of a [stochastic process](http://en.wikipedia.org/wiki/Stochastic_process) or [time series](http://en.wikipedia.org/wiki/Time_series) changes. In general the problem concerns both detecting whether or not a change has occurred, or whether several changes might have occurred, and identifying the times of any such changes.The problem of change-point detection can be defined as finding the time of switching from state 1 to state 2 in this model. When inferring the hidden states$\mathbf{s}=s_1 s_2 \ldots s_t \ldots$ from the observed data $\mathbf{y}=y_1 y_2 \ldots y_t \ldots$, also ``missing" are the regression parameters for each segment. This can be addressed by the Expectation-Maximization (EM) algorithm which starts from some initial ``guess" of the regression parameters, and then iterates between the E-step and the M-step. In the E-step, the state probabilities are calculated assuming the regression parameters are fixed. The M-step uses weighted linear regression to estimate the regression parameters for each segment. It can be shown that EM converges to at least a local maximum of the likelihood function in the parameter space.After EM converges we have point estimates of the regression parameters. For any hypothesized state sequence $\bf {s}$, its decision on the change point $\hat{t}_c^{({\mathbf s})}$ is the time of switching from state 1 to 2. We pool together the decisions of all the state sequences ${\mathbf s}$, weighted by their posterior probabilities $p(\bf {s}\vert\bf {y})$. The estimated change time is the weighted average

|  |
| --- |
| \begin{displaymath} \hat{t}_c = \sum_{{\mathbf s}} p({\mathbf s}\vert{\mathbf y}) \hat{t}_c^{({\mathbf s})} \end{displaymath} |
|  |

**Burst Detection:**

The goal of this survey is to automatically identify a burst in a spike train. Bursts are considered as a unit of neural information since they denote a period of 'high activity' in a given spike train. A generally accepted, rough definition of a burst is "an occurrence of "many" spikes in a "small" time interval" (Palm, 1981). Most neurons can burst if stimulated or manipulated pharmacologically. Much research has been focused on the way that a neuron fires an individual spike or burst (for example see link 1). This study however will not focus on the *how* part of the process but will rather attempt to detect burst activity by examining the *output*ie the spike train. We will review some methods that are commonly used and also test them in a set of computer-generated data inspired from (Raos, 2006). The problem is set like this: given a spike train that may contain a number of bursts, find the same bursts that would be identified by an "expert" upon visual inspection.

# System Configuration:

**HARDWARE REQUIREMENTS:**

Hardware - Pentium

Speed - 1.1 GHz

RAM - 1GB

Hard Disk - 20 GB

Floppy Drive - 1.44 MB

Key Board - Standard Windows Keyboard

Mouse - Two or Three Button Mouse

Monitor - SVGA

**SOFTWARE REQUIREMENTS**:

Operating System : Windows

Technology : Java and J2EE

Web Technologies : Html, JavaScript, CSS

IDE : My Eclipse

Web Server : Tomcat

Tool kit : Android Phone

Database : My SQL

Java Version : J2SDK1.5