

LITMUS: a Multi-Service Composition System for Landslide Detection

Aibek Musaev, *Student Member, IEEE*, De Wang, *Student Member, IEEE*,
and Calton Pu, *Senior Member, IEEE*

Abstract—Landslides are an illustrative example of multi-hazards, which can be caused by earthquakes, rainfalls and human activity among other reasons. Detection of landslides presents a significant challenge, since there are no physical sensors that would detect landslides directly. A more recent approach in detection of natural hazards, such as earthquakes, involves the use of social media. We propose a multi-service composition approach and describe LITMUS, which is a landslide detection service that combines data from both physical and social information services by filtering and then joining the information flow from those services based on their spatiotemporal features. Our results show that with such approach LITMUS detects 25 out of 27 landslides reported by USGS in December 2013 and 40 more landslide locations unreported by USGS during this period. LITMUS is a prototype tool that is used to investigate and implement research ideas in the area of disaster detection. We list some of the current work being done on refining the system that allows us to identify 137 landslide locations unreported by USGS during a more recent period of September 2014. Finally, we describe a live demonstration that displays landslide detection results on a web map in real-time.

Index Terms—Landslide detection service, multi-service composition, social media, physical sensors, event detection.

1 INTRODUCTION

TRADITIONAL method for natural disaster detection mostly relies on dedicated physical sensors. For instance, people use seismometers for earthquakes. However, few physical sensors exist for the detection of multi-hazards such as landslides, which may have multiple causes (earthquakes and rainfalls, among others) and happen in a chain of events. For multi-hazards like landslides, they are not like earthquakes which have strong signals that can be much easier caught by dedicated sensors. So, it is a real challenge to figure out where to allocate physical sensors and how many we need. Also, deploying physical sensors involves high costs for installation during early stage and maintenance afterwards.

As social media become real time information propagation platform, many researchers have studied the area of social media mining in natural disaster detection. Twitter data streams functioning as social sensors [1] are a good example of exploring the big data from social information services. High expectations have been placed on social sensors due to the fact that physical sensors (e.g., seismometers) are specialized for detection of specific disasters. But despite some initial successes, social sensors have met serious limitations due to the big noise in big data generated by social sensors such as all kinds of spam [2], [3] and other low quality information [4]. Also, one interesting example of the noise about “landslide” is the 70’s rock song “Landslide” [5]. Twitter filter for the word “landslide” gets more

tweets on this song than landslide disasters that involve soil movement. Therefore, an approach that relies on one kind of sensor is no longer able to fulfill the efficiency and accuracy requirements in the detection of multi-hazards.

By investigating existing approaches using physical and social sensors, we proposed a new landslide detection service — LITMUS [6], [7], which is based on a multi-service composition approach to the detection of landslides, a representative multi-hazard. It has the following characteristics compared with traditional or existing approaches for natural disaster detection.

Instead of trying to refine the precision and recall of event detection in each one of the physical and social information services¹, LITMUS composes information from a variety of sensor networks. The information services include both physical sensors (e.g., seismometers for earthquakes and weather satellites for rainfalls) and social sensors (e.g., Twitter and YouTube). More information services not only provide wider coverage than single source, but also improve the accuracy and reduce the latency overall.

Instead of trying to optimize the filtering process for each social sensor in isolation, LITMUS uses state-of-art filters for each social sensor and then adopts geo-tagging to integrate the reported events from all physical and social sensors that refer to the same geo-location. Our work shows that with such integration LITMUS achieves better landslide detection when compared to an authoritative source. Meanwhile, the geo-location information not only provides the base for the integration but also enables us to do real-time notification in the future.

LITMUS is one of the few systems that makes use of the composition of multiple heterogeneous information ser-

1. Throughout this paper the terms “information service” and “sensor” will be used interchangeably.

- A. Musaev is with the School of Computer Science, Georgia Institute of Technology, Atlanta, GA, 30332. E-mail: aibek.musaev@gatech.edu.
- D. Wang is with the School of Computer Science, Georgia Institute of Technology, Atlanta, GA, 30332. E-mail: wang6@gatech.edu.
- C. Pu is with the School of Computer Science, Georgia Institute of Technology, Atlanta, GA, 30332. E-mail: calton.pu@cc.gatech.edu.

vices. It is not tied to disaster detection and can be applied to other application areas involving service composition. This work presents a generic approach to the problem of composition of multiple heterogeneous information services and uses landslide detection as an illustrative example. Traditional approach to the composition of web services makes strong assumptions about services, which it then uses to select services when composing a new service, such as quality of service [8] or service license compatibility [9]. In practice, the real world services do not satisfy such assumptions. The claim we make in our work is that more information services should provide a more solid result and we demonstrate that it is the case with LITMUS.

The rest of the paper is organized as follows. Section 2 provides a set of requirements used to guide the design of the system followed by an overview of its implementation. Section 3 introduces the supported physical and social information services, and Section 4 describes implementation details of each system component. In Section 5 we present an evaluation of landslide detection using real data and compare the results generated by LITMUS with an authoritative source. Section 6 contains description of the ongoing work on refining the key components of the system. The web service demonstration is described in Section 7. We summarize related work in Section 8 and conclude the paper in Section 9.

2 FRAMEWORK OVERVIEW

2.1 System Requirements

Web service LITMUS was designed to detect landslides based on a multi-service composition approach that combines data from physical and social information services. Physical services should include earthquake and rainfall real-time feeds as possible causes of landslides. LITMUS should also support various social information services, which we expect to help detect landslides. However, the data from the social services must be filtered as they often contain a lot of noise. The system should also adopt geo-tagging to integrate the reported events from all physical and social sensors that refer to the same geo-location. Finally, a web client was designed to demonstrate LITMUS functionality by displaying detected landslides on a web map.

2.2 Implementation Overview

Based on the requirements described above, we implemented 3 independent components that perform filtering, integration and semantics-aware detection shown in Figure 1. The Filtering component downloads the data from social and physical sensors and filters out noise from social sensors. The Integration component combines the filtered data from social sensors with the data from physical sensors based on a Bayesian model integration strategy to generate a list of potential landslide locations. The last component performs semantics-aware detection of landslides by grouping locations related to the same event and excluding the results that are not current.

LITMUS provides access to its resources via a web service. The web service is implemented in a representational state transfer (REST) style [10]. The architectural

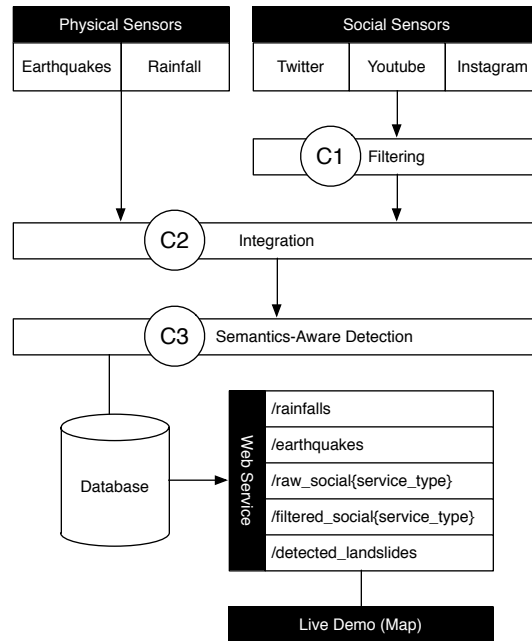


Fig. 1. Overview of LITMUS Framework

properties of this style are driven by several constraints, including client-server architecture and stateless communication. Client-server architecture ensures separation of concerns between clients and servers. This means, for example, that clients are not concerned with data storage, which is handled by servers. Servers are not concerned with the user interface or user state. The client-server interaction is further constrained by stateless communication, which means that no client context is stored on the server and that each client request contains all necessary information to be executed by the server. In addition to the described constraints, the central principle in REST is support for resources that are sources of information provided by the web service. Each resource is referenced with a global identifier, such as URI. In our landslide detection service the resources are the physical sensor feeds, the raw social feeds that are downloaded, the filtered social feeds that are processed by the system and the resulting feed of detected landslides as shown in Figure 1 as part of the Web Service component.

We implemented a web service demonstration consuming the resources provided by LITMUS, which is located in the GRAIT-DM portal accessible at [11]. GRAIT-DM is a SAVI project for Global Research on Applying Information Technology to support Effective Disaster Management.

3 PHYSICAL AND SOCIAL INFORMATION SERVICES

The physical information services supported by LITMUS do not provide information about landslides directly, but they provide data about other kinds of hazards, which may be possible causes of landslides. In particular, we use a real-time earthquake activity feed from the US Geological Survey (USGS) agency [12]. This feed is updated every minute and provides information about earthquakes of various magnitude. LITMUS collects data on earthquakes with a 2.5

magnitude and higher. The data is provided in a convenient GeoJSON format, which among other things provides time, magnitude, latitude, longitude, name of the place and ID, which is used to avoid duplicate records in the system.

Another physical information service supported by LITMUS is provided by the Tropical Rainfall Measuring Mission (TRMM) [13], which is a joint project between NASA and the Japan Aerospace Exploration Agency (JAXA). This project generates reports based on the satellite data of the areas on the planet that have experienced rainfalls within the past one, three and seven days. The reports are provided in multiple formats, including reports in a uniform manner on the project's web page, from which we parse and extract the rainfalls data.

LITMUS also supports various social information services, which we expect to help detect landslides, since there are no physical sensors that would detect landslides directly. Twitter is one of the supported information services that receives a lot of attention due to the volume and real-time nature of its information — 500 million tweets are posted daily². Tweets are often used to give a status update of what is happening in daily life of millions of people around the world. A tweet is even referred to as status in Twitter's documentation for developers. Users can follow status updates of other users, which enables them to read other people's tweets. Twitter is a microblogging service as the size of tweets is limited to 140 characters. One of the main reasons why Twitter is popular in research community is due to a rich set of public APIs provided by Twitter for accessing its data. Due to Twitter's popularity, the company introduced a number of API rate limits for accessing its public data to make sure that it can continue to provide uninterrupted service to its users. For example, the search functionality is currently rate limited at 180 queries per 15 minutes³.

In addition to Twitter, LITMUS also supports Instagram, which is an example of an image based social network, and YouTube, which is an example of a video based social network. Both of these social information services are among the leading social networks in their respective areas – 55 million photos are sent per day using Instagram⁴ and 100 hours of videos are uploaded per minute using YouTube⁵. The sliding window at Instagram for rate limiting purposes is larger than at Twitter and currently equals 1 hour. The rate limit is 5,000 requests per hour⁶.

The YouTube Data API uses a different structure of quota based on units. Different types of operations have different quota costs with the daily limit equal to 50 million units⁷. LITMUS operates within these limits and the data rates are sufficient for collecting landslide data.

All of the supported information services provide Search API based on keywords. This approach requires developers to implement a mechanism that avoids duplicate items in the system. LITMUS employs Search APIs provided by these

services and uses item IDs to avoid data duplication. A more efficient and complete approach is to use a Streaming API, which pushes items to its clients in real-time based on a list of keywords provided. However, not all of the services provide such API at the moment.

Next we present the implementation details of each system component.

4 SYSTEM COMPONENTS

C1. Filtering Component

The C1. Filtering Component applies only to social information services. This is due to the fact that all data provided by both earthquake and rainfall feeds is considered relevant for landslide detection purposes. Also, physical information services provide data with geo-coordinates, hence there is no need to apply the geo-tagging processing either.

To remove noise from social sensors, we process the downloaded data in a series of filtering steps. There are four stages in this process that filter out items, which are neither related to landslides (steps F1, F2 and F4) nor are useful to LITMUS due to lack of geo-location (step F3).

F1. Filter based on search terms

Each social information service provides search API based on keywords for software developers. The system periodically downloads the data from each social sensor based on "landslide", "mudslide", "rockslide", "rockfall" and "landslip" keywords. The period is currently set to 30 minutes, which is a configuration parameter and can be modified if necessary.

F2. Filter based on stop words & phrases

The social information services require additional filtering as they contain a lot of items unrelated to landslides and most of the time they are not geo-tagged either. The following is a set of frequent examples of unrelated items from the social information services:

- "Landslide" song by Stevie Nicks from Fleetwood Mac: "Climbed a mountain and I turned around, and I saw my reflection in the snow-covered hills, and the landslide brought me down. -FleetwoodMac"
- Used as an adjective describing an overwhelming majority of votes or victory: "Robert Mugabe's party claims landslide victory in Zimbabwe's key election as ... - Daily Mail <http://t.co/Hf4sVU3E8F>"
- Lyrics from "Bohemian Rhapsody" by Queen: "Caught in a landslide, no escape from reality..."

The first two items can be filtered out using a simple exclusion rule based on the presence of stop words "FleetwoodMac" and "election". The third item is filtered out using another exclusion rule based on the presence of stop phrases that currently include the lyrics of some popular songs, e.g. "no escape from reality".

But even after the stop words and stop phrases filters are applied, many unrelated to landslide items remain in the feeds from the social information services. Exclusion of those items requires a more sophisticated approach, including filtering based on geo-location and filtering based on penalized classification, described next.

2. <https://about.twitter.com/company>

3. <https://dev.twitter.com/rest/public/rate-limiting>

4. <http://instagram.com/press/>

5. <http://www.youtube.com/yt/press/>

6. <http://instagram.com/developer/limits/>

7. <https://developers.google.com/youtube/v3/getting-started#quota>

F3. Filter based on geo-location

To detect landslides within a particular period, we need to determine their locations. The data from the physical sensors already contains geo-coordinates. However, the data from the social sensors is usually not geo-tagged although each social network provides support for users to disclose their location. So, if an item has not been geo-tagged already, then we suggest to look for mentions of place names that refer to locations of landslides in the item's text.

For details of the geo-tagging algorithm used in LITMUS we refer the reader to [6]. Next we describe the changes that we made in this algorithm.

The geo-tagging algorithm extracts geographical terms (geo-terms) from incoming messages and assigns geo-coordinates to each geo-term using a gazetteer, which is a dictionary that maps places to geo-coordinates. In our system we use a public gazetteer data from the GeoNames database that covers all countries and contains over 10 million places [14]. Our geo-tagging algorithm uses a subset of this data, namely 448k places. This subset includes countries, administrative divisions of countries of first to fourth orders, cities with population greater than 1,000 people and islands. In the future we plan to increase this subset by including even more detailed places.

Some news sources mentioned in social media data, such as "Boston Globe" or "Jamaica Observer", contain valid geo-terms which must be removed from consideration by the geo-tagging algorithm; otherwise it would return incorrect results. That is why LITMUS maintains a list of major news sources, including "Boston Globe" and "Jamaica Observer". Consider the following tweet: "Boston Globe - Typhoon, mudslides kill 14 in Japan; 50 missing <http://t.co/nEUbk60Pzl>." "Boston" is positioned closer to the landslide keyword "mudslides" than "Japan", however it is a part of the "Boston Globe" news source that LITMUS automatically removes from consideration, such that the correctly extracted geo-term is "Japan".

F4. Filter based on classification

Majority of items returned by social sensors are not relevant to landslides, even though they contain landslide keywords and valid geo-terms. The following are examples of irrelevant items with respect to landslide disasters that contain valid geo-terms:

- Extracted geo-term "California": "Laying on a landslide with a bag of California to smoke."
- Extracted geo-term "New York City": "Lupica: As Bill de Blasio takes the mayoralty of New York City, let's not forget Chris Christie's landslide vi <http://t.co/0CoR1zQY55>"

To filter out such irrelevant items LITMUS employs machine learning binary classification, which automatically labels each item as either relevant or irrelevant based on a classifier model built from a training set containing labeled items.

Our training set contains labeled items from each social information service. There are two labels in our training set to indicate whether a particular item is relevant or irrelevant to landslide disasters. We use a list of landslide

events reported by the USGS agency. Every month it publishes a list of confirmed landslides collected from other reputable sources, including ABC News, Reuters, Xinhua News Agency, Latin Times and National Geographic. For each event in the USGS list we automatically identify the date of publication and a list of geo terms.

To find the actual items from social information services that were related to the confirmed landslides within each month under study, we first filtered out the data based on landslide locations. Then we manually went through each item in the filtered list to make sure they described corresponding landslides by comparing the contents of the items with the corresponding landslide articles. And whenever there were URLs in those items, we viewed them too to make sure they reported landslide events.

Here is an example of a landslide confirmed by the Latin Times news source, which was published on September 11, 2013:

- "Mexico Mudslide 2013: 13 Killed In Veracruz Following Heavy Rains" [15].

The geo tag term that we extracted from this news is "Veracruz, Mexico", whose latitude and longitude values are 19.4347 and -96.3831.

In order to create a list of items unrelated to landslides we randomly picked items in each social source and manually went through each item. This time we only had to make sure that the items did not describe landslide events.

For classification purposes we designed a set of features based on the textual description of items from each social information service, namely three sets of features that are applied to the items in each social sensor:

- Common statistical features including length of the text, number of uppercase/lowercase characters, position of the search term in the text, min/avg word length.
- Binary features indicating the presence of various elements in the text, including at sign, URL, percentage, exclamation/question marks, numbers.
- Vocabulary-based features: relevant and irrelevant vocabulary scores – see description below.

Vocabulary-based features are designed to mimic the decision making process during manual labeling. It was noticed that during manual labeling it was usually trivial to identify which items were relevant to landslide disasters and which ones were not. There are particular words like "victim", "rescue" or "killed" that are relevant to disasters and words like "election", "fleetwood" or "supernova" that are irrelevant. Hence, this set of features implements this idea using statistics. We compute the most frequently used words in the training set for relevant and irrelevant items. Then for each downloaded item we count the total number of words that are present in the list of relevant items, which we call a relevant vocabulary score. Similarly, we count the total number of words that are present in the list of irrelevant items, which we call an irrelevant vocabulary score.

Classification algorithm computes the values of the described set of features for each item in the training set, builds a model and uses it to predict whether unlabeled

items are relevant to landslide disaster or not. In practice this method may assign both correct and incorrect labels to data items. To improve the quality of classification we propose to convert the filtering problem of each item to the filtering problem of the aggregation of items assigned to each landslide location, i.e. geo-term. In most cases a particular geo-term is mentioned in multiple incoming items from social sensors. Each social item is labeled by a machine learning classifier, so there are multiple classification results for each geo-term. Although this method generates both correct and incorrect labels for a particular geo-term, but on average it provides more correct than incorrect labels. The idea of penalized classification uses this heuristics to improve the results of classification by accepting the label assigned to the majority of items for each geo-term and only considering locations whose majority label is positive.

C2. Integration Component

To generate a list of potential landslide locations, LITMUS combines the data from physical information services with filtered and geo-tagged data from social information services. This is a two stage process, where a grid-based landslide location estimation is followed by an integration of results from multiple services.

To estimate landslide locations, we propose to represent the surface of the Earth as a grid of cells. Each geo-tagged item is mapped to a cell in this grid based on the item's coordinates. After all items are mapped to cells in this grid, the items in each non-empty cell are used for computing an integrated landslide score. The size of the cells is equal to 2.5 minutes in both latitude and longitude, which corresponds to the resolution of the Global Landslide Hazard Distribution [16] that we plan to add as an additional physical sensor to the system. This is the maximum resolution supported in LITMUS. The actual resolution is driven by the precision of the geo-tagging algorithm described earlier.

After mapping the items from each sensor to cells in this grid, which represent potential landslide locations, we calculate the probability of landslide occurrence in location cells based on a Bayesian model integration strategy. Here is the description of this strategy. We use a subscript i to distinguish different sensors. Suppose we have a cell x and ω is the class associated with x , either being true or false. Then, assuming a hidden variable Z for an event to select one sensor, a probability for a class ω given x , $P(\omega|x)$, can be expressed as a marginal probability of a joint probability of Z and ω :

$$P(\omega|x) = \sum_i P(\omega, Z_i|x) = \sum_i P(\omega|Z_i, x) P(Z_i|x) \quad (1)$$

Here, we use external knowledge $P(Z_i|x)$ to express each sensor's confidence given x . For instance, if a certain sensor becomes unavailable, the corresponding $P(Z_i|x)$ will be zero. Also one could assign a large probability for the corresponding $P(Z_i|x)$ if one sensor dominates over other sensors.

In our experiment, we use prior F-measure R from the August data as the confidence for each sensor since F-measure provides a balance between precision and recall,

namely $F\text{-measure} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$. We use the August data in our experiments, because there were many landslides reported by USGS during this month, so the data collected during this month is representative. To generate the results in the range from 0 to 1, we normalize the values of F-measure into a scale between 0 and 1 first. Taking all items from each sensor into account, the formula will be further converted into the following format:

$$P(\omega|x) = \sum_i R_i \frac{\sum_j POS_{ij}^x - \sum_j NEG_{ij}^x - \sum_j STOP_{ij}^x}{\sum_i N_i^x} \quad (2)$$

Here, R_i denotes the normalized prior F-measure of sensor i from historic data. POS_{ij}^x denotes positively classified items from sensor i in cell x , NEG_{ij}^x denotes negatively classified items from sensor i in cell x , $STOP_{ij}^x$ denotes the items from sensor i in cell x that have been filtered out using stop words and stop phrases, and N_i^x is a total number of items from sensor i in cell x .

C3. Semantics-Aware Detection Component

The actual number of landslides is usually less than a total number of potential landslide locations returned by the Integration component. Some of the potential landslide locations may be referring to the same event. For example, this may happen when users refer to the same event by using geo-terms of different level of detail. Consider the following tweets that describe a landslide, which occurred in Italy in December:

- Extracted geo-term "italy": "Giant craters were ripped out of roads in Italy and homes and shops sank into the ground after a major landslide: <http://t.co/JZ6l63vXLH>"
- Extracted geo-term "montescaglioso": "VIDEO: Landslide rips apart Italy roads: Heavy rains and floods cause a powerful landslide in the southern Italian town of Montescaglioso."

Both geo-terms extracted from these tweets are valid, but they describe the same event, which should be correctly detected by the Semantics-Aware Detection component. Another possible scenario of having multiple geo-terms describing the same event can be described by the following tweets:

- Extracted geo-term "wadhurst": "Christmas come early thanks to #southeastern, the bad santa of train services. Landslide at Wadhurst has block line. Working from home."
- Extracted geo-term "hastings": "Train delayed due to landslide. That's a first for the Hastings line."

These two landslide locations are actually related as shown in the following tweet:

- Extracted geo-terms "wadhurst" and "hastings": "Avoid the trains on the Hastings line folks. Word is there's been a landslide on the line near #Wadhurst #UKstorm"

Social Sensors	F1.Filter based on search terms	F2.Filter based on stop words & phrases	F3.Filter based on geo-location	F4.Filter based on classification
Twitter	45391	24993	7879	2815
Instagram	1418	1263	308	15
YouTube	4890	3936	557	318

TABLE 1
Overview of Filtering Results

Metrics	F1.Filter based on search terms	F2.Filter based on stop words & phrases	F3.Filter based on geo-location	F4.Filter based on classification
Signal-to-Noise Ratio	0.05	0.10	0.44	1.90
Information Gain	-	0.028	0.121	0.197

TABLE 2
Evaluation Results of Filtering Component

Semantics-Aware Detection component must also handle temporal issues by excluding results that refer to the past or future events:

- “The Kedarnath disaster in Uttarakhand, India in June remains the worst landslide accident of 2013 to date <http://t.co/Mf31ztjwQ2>”

Even though the year matched the year of the evaluation period, the month did not, hence the landslide locations extracted from this message must be excluded from the final result.

The Semantics-Aware Detection component is currently semi-automated. LITMUS is able to group landslide locations that were referred to in the same message and also to exclude messages containing references to either past or future years. We plan to improve the performance of this component as part of future work.

5 EVALUATION USING REAL DATA

To evaluate the performance of our landslide detection service we designed three sets of experiments. We start with evaluating the performance of the filtering process of social information services. Next we compare the effectiveness of three multi-service composition strategies for landslide detection. The final experiment provides the detection comparison results between LITMUS and an authoritative source.

5.1 Evaluation of Filtering Component

In this experiment, we view different social media as our social sensors and process the data from these sensors in a series of filtering steps. For simplicity, we only focus on the textual description of each item during filtering. For Instagram, the textual description is an image’s caption text. For YouTube, the textual description is a concatenation of a video’s title and description. And for Twitter, the textual description is the text of a tweet itself. We use December 2013 as our evaluation period. Table 1 shows the total number of items downloaded during this month. For each filtering step, Signal and Total indicate the remaining items after the filtering steps, where Signal is a number of items from social sensors that are relevant to landslide detection and Total is a total number of items from social sensors.

There are two metrics that we use to evaluate the performance of the filtering steps in Table 2, namely Signal-to-Noise Ratio (SNR), and Information Gain (IG), where for each filtering step i :

$$SNR_i = \frac{Signal_i}{Noise_i} = \frac{Signal_i}{Total_i - Signal_i} \quad (3)$$

Here, $Signal_i$ is the number of relevant items remaining in step i , $Total_i$ is the number of all items remaining in step i .

$$IG_i(T_{i-1}, a_i) = H(T_{i-1}) - H(T_{i-1}|a_i) \quad (4)$$

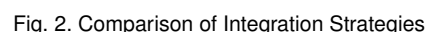
Here, we consider the filtering process as a binary classification problem which has two classes: relevant and irrelevant. T_{i-1} denotes a set of training examples before step i . a_i is the attribute (filtering condition) we used in step i . H denotes information entropy. The filtering conditions in the process are considered as attributes. For instance, the F2 filter based on stop words and phrases will be converted into a classifier based on a boolean attribute whether an item contains stop words and phrases. Information gain measures the relevance of attributes.

It can be seen from Table 2 that SNR is improving after each filtering step and eventually exceeds 1. Information gain ranks the relevance of the filtering conditions, which shows the rank of filters in decreasing order: filter based on classification, filter based on geo-location, and filter based on stop words and phrases. For F1, we cannot calculate the values of Information Gain, since we are not able to know how many items there are in filtered out data and what the prior information entropy is.

5.2 Evaluation of Integration Component

For comparison purposes, we introduce two major baselines - OR and AND integration strategies and compare them with the proposed Bayesian model strategy. We have five sensors in total, including social sensors (Twitter, Instagram, and YouTube) and physical sensors (earthquakes and rainfalls). Considering that each sensor has one vote to a particular cell, the cell will obtain one or zero votes from each sensor. For OR integration strategy, we grant equal weight to five sensors. And we obtain the decision (whether a landslide happened or not) by combining the

We present the results of comparison between integration strategies in Figure 2. This figure shows that the Bayesian model strategy has 71% precision, 82% recall and 77% F-measure. The OR strategy has the highest recall at 100%, but also the lowest precision at 2%. The AND strategy shows improvement compared to the OR strategy as its F-measure is higher, but the Bayesian strategy shows the best performance overall.



In addition to a real-time seismic feed, USGS also presents a continually updated list of landslides as reported by other reputable sources, including DailyMail.co.uk, [GlobalPost](http://GlobalPost.com), [HeraldNet](http://HeraldNet.com), and [Xinhuanet](http://Xinhuanet.com). In this experiment we compare the landslides provided by this authoritative source in December [17] versus the landslides detected by LITMUS during the same month — see Figure 3 for the results of this comparison. LITMUS was able to detect 25 out of 27 landslides reported by USGS. The 2 missed events were in Buncombe County, NC and Watchung, NJ. Both events affected very small areas, which is the reason why they did not attract significant public interest in social media and were missed by the system. In addition to the overlapping events detected by both LITMUS and USGS, our system managed to find 40 more landslide locations in December that were unreported by USGS. This is due to the fact that the USGS results are based on news sources, which can only report a limited amount of information. Whereas LITMUS employs multiple information services, plus the landslide

6 REFINEMENT OF THE SYSTEM

6.1 Evaluation of Geo-tagging Algorithms

One of the common approaches implementing this idea is based on the use of a gazetteer. An exact match of words in the item's text is performed against such gazetteer. For the list of places we can use the titles of the geo-tagged Wikipedia articles. For details of the geo-tagging algorithm based on Wikipedia articles as a gazetteer we refer the reader to [6]. An alternative gazetteer is the Geonames.org database that covers all countries and contains over 10 million places [14]. Both implementations using the gazetteer approach, however, have a poor quality as they often extract irrelevant geo locations. For example, the Wikipedia based algorithm produces many irrelevant locations for the social media items during the evaluation period, including "hill", "alot" and "uni" matches, whereas the Geonames.org based algorithm returns "most", "plan" and "cry" as candidate geo locations during the same period. That is why we need to apply various heuristics to remove irrelevant matches

when we use a gazetteer approach. We can remove common nouns from consideration by making use of a list of the most frequent words in English [18]. We can also exclude non-nouns from the list of geo locations using Part-Of-Speech tagging. Finally, in case there are multiple matches found in a single tweet, we can use the geo term that is the closest to the landslide keyword and ignore the rest.

An alternative approach for extracting geo locations from social media texts employs a natural language processing (NLP) technique called named entity recognition or NER. Among various entities, NER libraries seek to locate and classify elements in text into pre-defined categories, including names of persons, organizations, time and location. For the purpose of geo-tagging we are interested in the location entity. LITMUS employs Stanford CoreNLP library, which is a Java suite of NLP tools [19], to identify all location entities mentioned in social media. For each of the found locations the Geonames.org gazetteer is used to find the corresponding geographic coordinates of these locations. If there is no match in the gazetteer then LITMUS uses the Google Geocoding API [20] to convert locations into geographic coordinates. All locations that LITMUS finds geographic coordinates for are considered candidate landslide locations. Although the overall quality of the NER approach is superior to the gazetteer approach as can be observed in our experiments, it still has a number of issues. For example, the CoreNLP library extracts “China Landslides” as a location entity for this tweet: “11 Killed, 27 Missing in China Landslides”. It also fails to identify a location entity in this tweet: “At Least 19 Dead After Landslides and Flooding on the India/Pakistan Border”.

Next we provide details of the evaluation of the Wikipedia, Geonames.org and NER based geo-tagging algorithms using real data in September 2014.

In this section we evaluate the following geo-tagging algorithms: two algorithms based on gazetteer approach with Wikipedia articles and Geonames.org chosen as gazetteers and one algorithm based on named entity recognition (NER) approach for location entities.

In total, the Wikipedia based gazetteer approach identified 611 locations in September and out of 55 landslide locations reported by USGS during this period it found 49. The six missing landslide locations were Collbran Colorado, American Samoa, St. Lucia, Moravia, Panama Isthmus and Smolyan Bulgaria. Overall, there were 357 landslide locations in September and it found 128 of them.

The Geonames.org based gazetteer approach identified 1568 locations in September and out of 55 landslide locations reported by USGS it found 48. The 7 missing landslide locations were Collbran Colorado, American Samoa, St. Lucia, Mughal Kashmir, Moravia, Panama Isthmus, Smolyan Bulgaria. Overall, it was able to identify 153 landslide locations out of 357 during this month.

The NER based approach identified 811 landslide locations and it missed only 3 landslide locations reported by USGS, namely: St. Lucia, Moravia and Smolyan Bulgaria. During September, LITMUS did not find any items in English relevant to these events in Twitter, Instagram or YouTube. This may be due to a very local nature of those events, which did not attract the attention of users of the corresponding social networks. All of the landslide locations

discussed in corresponding social networks in September were successfully identified by this approach.

Overall, the NER based approach contained fewer irrelevant locations and produced the best precision and recall for geo-tagging purposes among the described approaches, which is why LITMUS now follows this approach for geo-tagging purposes.

6.2 Revision of Filtering Component

Majority of the data items from social information services collected using landslide keywords contain information, which is irrelevant to landslide events. This situation is compounded by the fact that even the items that describe landslide events may also contain unrelated information in addition to the relevant data. Consider the following description of a video from YouTube:

- “Thousands Still Waiting for Aid in Flood Ravaged Mexico. Flooding and mudslides have destroyed villages in Mexico’s southwest state of Guerrero, leaving many residents waiting for aid. . video,cnn news, fox news, abc news, world news, breaking news, us news, europe news, latino news, syria news, middle east news, russia news, china news, economic news, latest news, bbc news, pasific news, arab news, money news”

In addition to relevant information about flooding and mudslides in Guerrero, Mexico, the video description also contains unrelated keywords to generate more traffic. Such instances decrease the quality of the training set. Which is why in the training set we break the text of each data item into sentences and keep the ones that contain landslide keywords only. Similarly, we perform preprocessing before classification by removing sentences without landslide keywords from consideration. This strategy results in successful classification of data items as in the following video description from YouTube:

- “Floods, landslide kill 13 in Indonesia; 2 missing Breaking News MUST SEE. Enjoy the news, subscribe for more!”

6.3 Revision of Cell-based Integration

The first step that the integration component performs is it maps the items from each sensor to cells in a grid covering the surface of the Earth. Then it proceeds by considering only non-empty cells. Although this approach is easy to understand and its implementation is fast to compute, it has a few challenges. It is obvious that the size of cells can be either too coarse or too granular for detection purposes, for example big sized cells will include multiple landslides in them. Another challenge is that it ignores semantics of data items, such that unrelated items may be incorrectly considered as related to the same event and processed together. Let us consider the following items mapped to the same cell and treated in one batch in a cell based approach:

- 3 items on a landslide in Indonesia, including this tweet: “Floods, landslide kill 13 in Indonesia; 2 missing Breaking News MUST SEE. Enjoy the news, subscribe for more!”

- 4 unrelated items from social media mentioning Jakarta, including this Instagram image caption: “Enjoyed this much #creamycomfort #dessert #jakarta #brightspot #baileys #mudslide”

All of these items are mapped to the same cell, because the geo-tagging component returns the same geographic coordinates for both Indonesia and Jakarta. The filtering component classifies these items correctly, but the cell is not deemed a landslide location due to a low integrated landslide score. As this example shows, there are multiple topics connected to the same cell and they should be handled separately. The easiest approach that works in this particular case is to cluster data items based on a geo term within each cell. Such approach correctly detects a landslide in Indonesia. A more advanced approach is to use semantic clustering to group data items with similar content together. This is an area of research that we are currently investigating.

The refinements described in this section allowed us to significantly improve the quality of landslide detection. Based on the data collected during the month of September, LITMUS was able to detect 137 landslide locations unreported by USGS during the same period.


7 WEB SERVICE DEMONSTRATION

We developed a live demonstration [11] that consumes the resources provided by the web service. This web application shows live feeds from each resource described in the paper. The data from all feeds is displayed on a Google Map, which can be set to either Map or Satellite view by a user. A user can view detailed information about items from each feed — see Figure 4 for a detailed view of an item from a YouTube feed.



Fig. 4. Example of video from the YouTube feed

A separate feed shows a list of detected landslides that are a result of the multi-service analysis based on the Bayesian model integration strategy. A user can also view detailed information about all items that were used to make a decision regarding a landslide in each location — see Figure 5.


Detected Landslide Details

Twitter			
Location	Text	Created at	
colombia	Colombia hit by deadly landslide http://t.co/jVVq1sg87a	2013-12-02 01:14:59	
colombia	Colombia hit by deadly landslide http://t.co/RnKHOEPvTC	2013-12-02 17:31:07	
colombia	Colombia hit by deadly landslide http://t.co/gaoptzfHMQ	2013-12-03 22:13:40	


YouTube			
Location	Video	Text	Uploaded
colombia		Raw 20 Missing After Landslide in Colombia.	2013-12-29 06:35:27

Fig. 5. Example details of a detected landslide

8 RELATED WORK

Many researchers have proposed detection systems for disaster events, such as earthquakes and hurricanes, by physical sensor networks and real-time web monitoring. Cotofana et al. [21] described an SOA-based framework for instrument management for large-scale observing systems which are used as the dominant means of study for a variety of natural phenomena including natural disasters. Suzumura et al. [22] proposed a real-time Web monitoring system called “StreamWeb” on top of a stream computing system called System S developed by IBM Research, which provides a platform for developers to monitor streaming data such as Twitter streaming. Guy et al. [23] introduced TED (Twitter Earthquake Detector) that examines data from social networks and delivers hazard information to the public based on the amount of interest in a particular earthquake. Sakaki et al. [1] proposed an algorithm to monitor tweets and detect earthquake events by considering each Twitter user as a sensor. Kitsuregawa et al. [24] launched info-plosion project (IOT) to show how info-plosion analytics, especially sensor analytics, created disruptive services. Cameron et al. [25] developed platform and client tools called Emergency Situation Awareness - Automated Web Text Mining (ESA-AWTM) system by identifying tweets relevant to emergency incidents. Wang et al. [26] proposed a mixture Gaussian model for bursty word extraction in Twitter and then employed a novel time-dependent HDP model for new topic detection. Hua et al. [27] presented STED, a semi-supervised system that helps users to automatically detect and interactively visualize events of a targeted type from Twitter, such as crimes, civil unrests, and disease outbreaks. Twitter has been used in those studies as information source since it supports real-time propagation of information to a large group of users [28]. People can post tweets using a wide range of services: email, SMS text-messages, and smartphone apps. Besides Twitter, some studies targeted search engines, news, blogs, and time series data by analyzing spatiotemporal patterns [29], [30], [31], [32]. Our work continues development of LITMUS [6] by creating a landslide information service based on a multi-service composition approach that combines data from both physical and social information services.

Multi-service analysis requires a system to control and

overcome difficulties of several kinds such as data difficulties, development difficulties and analysis difficulties. The data from multiple services are complex, heterogeneous, dynamic, distributed and may be quite large. Some studies have been done in biostatistics and bioinformatics scenarios. Horton et al. [33] introduced a tutorial in biostatistics to perform regression analysis of multiple sources and multiple informant data from complex survey samples. Huopaniemi et al. [34] presented multivariate multi-way analysis of multi-source data in bioinformatics applications. Swain et al. [35] provided a technical report on multi-source data analysis in remote sensing and geographic information processing. They proposed a general approach for computer analysis using quantitative multivariate methods of remote sensing data combined with other sources of data in geographic information systems. Milanovic et al. [36] provided a survey of existing proposals for web service composition. Constantinescu et al. [37] presented an algorithm that supports dynamic service composition based on partial matches of input/output types. Truong et al. [38] proposed information quality metrics for identifying and reducing irrelevant information about web services. Service composition in LITMUS is static rather than dynamic, because the data from all of our sensor information services is downloaded at each cycle. As the number of information services supported by LITMUS grows, we plan to add support for dynamic service composition and execution.

Another important aspect for disaster detection systems is situational awareness. The challenge for social sensors is that users may use alias or location names in different granularities in messages resulting in inaccurate location information. Multiple studies have been done on location estimation for information from social networks based on content of tweets, e.g. [39]. [40] demonstrated a rapid unsupervised extraction of locations references from tweets using an indexed gazetteer. Our system also employs a public gazetteer and adopts a grid-based approach with customizable granularities in location estimation.

Also, social sensors involve many techniques in machine learning or data mining. For instance, researchers spent lots of efforts on text classification. Text classification (also known as text categorization, or topic spotting) is used to automatically sort a set of documents into classes (or categories, or topics) from a predefined set [41]. It has attracted a booming interest from researchers in information retrieval and machine learning areas in decades. Recently, several novel classification approaches have been proposed and implemented in text classification. Wang et al. [42] presented semantics-based algorithm for cross-domain text classification using Wikipedia based on co-clustering classification algorithm. Lex et al. [43] described a novel and efficient centroid-based algorithm Class-Feature-Centroid Classifier(CFC) for cross-domain classification of web-logs, also they have discussed the trade-off between complexity and accuracy. Pan et al. [44] proposed a spectral feature alignment (SFA) algorithm to align domain-specific words from different domains into unified clusters, with the help of domain independent words as a bridge. Zhen et al. [45] propose a two-stage algorithm which is based on semi-supervised classification to address the different distribution problem in text classification. Due to the noises and

evolution of those low quality information for those social sensors [46], [47], [48], our system also employs text classification and we propose a penalized classification technique to improve the results by accepting the label assigned to the majority of the items for each location.

9 CONCLUSION AND FUTURE WORK

Multi-hazards are disasters with causally chained events such as the 2011 Tohoku earthquake triggering tsunami, which caused the Fukushima nuclear disaster, and landslides, often caused by earthquakes or rainfalls. Detecting such multi-hazards is a significant challenge, since physical sensors designed for specific disasters are insufficient for multi-hazards. As promising alternatives [1], social information services have difficulties with filtering the big noise in the big data being generated. We show that multi-service composition approach, which combines data from both physical and social information services, can improve the precision and accuracy of multi-hazard detection when the participating sensors are relatively independent of each other.

Applying this approach, we built a landslide detection service called LITMUS, which composes physical information services (USGS seismometers and TRMM satellite) and social information services (Twitter, Instagram, and YouTube). LITMUS provides a REST API for obtaining its resources, including social and physical sensor feeds and a list of detected landslides. A live demonstration [11] is developed that consumes these resources to display the results on a Google Map.

The effectiveness of landslide detection is evaluated using real world data collected in December 2013. Individual filtering results for each social sensor are provided followed by the full integration of 5 sensors applying a modified Bayesian model integration strategy that achieved 71% in precision, 82% in recall and 77% in F-measure for landslide detection, which is significantly better than the baseline integration strategies. A comparison is performed against an authoritative list compiled by USGS, which shows that LITMUS detects 25 out of 27 reported events as well as 40 more events unreported by USGS in December.

The coverage of landslides detected by LITMUS can be improved by supporting other languages in addition to English, such as Chinese. Support for other languages can be implemented using two distinct approaches. One approach can be denoted as "native" and requires significant efforts, including preparation of a training set, a set of stop words and stop phrases in the new language as well as an NER library that would support recognition of a location entity in that language. An alternative approach requires minimal efforts where data from social information services in the new language is automatically translated into English, such that the existing LITMUS infrastructure can be used without modification for detection of landslides reported in other languages. We are implementing both approaches, so that we can compare their results. In addition to languages we are also working on adding more sources, such as social information services like Facebook as well as news sources like BBC and Weather Channel.

We are also interested in improving the precision of landslide detection. As shown above currently LITMUS uses feature selection based on the textual content of data items from social information services. We are adding additional features for classification purposes, such as user-based features in order to distinguish users based on their area of expertise, influence, past behavior as well as their information about their followers/friends among other factors. We are also analyzing a topic-based approach where the decision whether a particular data item is relevant to landslide events is based on its topics rather than words. For example, instead of using a set of stop words we are studying the use of stop topics or concepts.

Finally, we are interested in detecting other kinds of events using LITMUS infrastructure, for example epidemic events like ebola. This requires changes in the LITMUS architecture for performance reasons as the number of tweets containing keyword "ebola" is several orders of magnitude higher than the number of tweets containing keyword "landslide". One of those changes will require the use of the real-time Streaming API provided by Twitter instead of Search API to make sure we do not miss any tweet.

10 ACKNOWLEDGEMENTS

This research has been partially funded by National Science Foundation by CNS/SAVI (1250260, 1402266), IUCRC/FRP (1127904), CISE/CNS (1138666), NetSE (0905493) programs, and gifts, grants, or contracts from Singapore Government, and Georgia Tech Foundation through the John P. Imlay, Jr. Chair endowment. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding agencies and companies mentioned above.

REFERENCES

- [1] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in WWW, 2010.
- [2] D. Wang, D. Irani, and C. Pu, "A social-spam detection framework," in *Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, ser. CEAS '11, 2011, pp. 46–54.
- [3] D. Wang, D. Irani, and C. Pu, "A study on evolution of email spam over fifteen years," in *Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom)*, 2013 9th International Conference Conference on, Oct 2013, pp. 1–10.
- [4] D. Wang, "Analysis and detection of low quality information in social networks," in *IEEE 30th International Conference on Data Engineering Workshops (ICDEW)*, 2014, pp. 350–354.
- [5] S. Nicks, "Landslide," [http://en.wikipedia.org/wiki/Landslide_\(song\)](http://en.wikipedia.org/wiki/Landslide_(song)), accessed on 10/1/2014.
- [6] A. Musaev, D. Wang, and C. Pu, "LITMUS: Landslide detection by integrating multiple sources," in ISCRAM, 2014.
- [7] A. Musaev, D. Wang, C.-A. Cho, and C. Pu, "Landslide detection service based on composition of physical and social information services landslide detection service based on composition of physical and social information services," in *The 21st IEEE International Conference on Web Services (ICWS 2014)*, 2014.
- [8] S. Ran, "A model for web services discovery with QoS," in *ACM SIGecom Exchanges*, 2003, vol. 4, no. 1.
- [9] G. Gangadharan, M. Weiss, V. DAndrea, and R. Iannella, "Service license composition and compatibility analysis," in ICSOC, 2007.
- [10] R. T. Fielding and R. N. Taylor, "Principled design of the modern Web architecture," *ACM Transactions on Internet Technology*, vol. 2, no. 2, 2002.
- [11] GRAIT-DM, "Global Research on Applying Information Technology to support effective Disaster Management," <https://grait-dm.gatech.edu/demo-multi-source-integration/>, accessed on 10/1/2014.
- [12] USGS, "United States Geological Survey agency: Earthquake activity feed from the United States Geological Survey agency," <http://earthquake.usgs.gov/earthquakes/>, accessed on 10/1/2014.
- [13] TRMM, "Tropical Rainfall Measuring Mission: Satellite monitoring of the intensity of rainfalls in the tropical and subtropical regions," <http://trmm.gsfc.nasa.gov/>, accessed on 10/1/2014.
- [14] GeoNames, "GeoNames geographical database," <http://www.geonames.org/>, accessed on 10/1/2014.
- [15] LatinTimes.com, "Mexico mudslide 2013: 13 killed in veracruz following heavy rains," <http://www.latintimes.com/articles/8234/20130911/13-dead-veracruz-mudslides-landslides-mexico-rains.htm#UjDOD5LFVBk>, accessed on 10/1/2014.
- [16] CHRR et al., "Global Landslide Hazard Distribution," <http://sedac.ciesin.columbia.edu/data/set/ndh-landslide-hazard-distribution/>, accessed on 10/1/2014.
- [17] USGS, "United States Geological Survey agency: Recent landslide events," <http://landslides.usgs.gov/recent/index.php?year=2013&month=Dec>, accessed on 10/1/2014.
- [18] BYU, "Top 5,000 words in American English," <http://www.wordfrequency.info/>, accessed on 10/1/2014.
- [19] T. Stanford Natural Language Processing GGroup, "Stanford corenlp," <http://nlp.stanford.edu/software/corenlp.shtml>, accessed on 10/1/2014.
- [20] G. Inc., "The google geocoding api," <https://developers.google.com/maps/documentation/geocoding/>, accessed on 10/1/2014.
- [21] C. Cotofana, L. Ding, P. Shin, S. Tilak, T. Fountain, J. Eakins, and F. Vernon, "An soa-based framework for instrument management for large-scale observing systems (usarray case study)," in ICWS, 2006.
- [22] T. Suzumura and T. Oiki, "StreamWeb: Real-time web monitoring with stream computing," in ICWS, 2011.
- [23] M. Guy, P. Earle, C. Ostrum, K. Gruchalla, and S. Horvath, "Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies," vol. 6065, 2010.
- [24] M. Kitsuregawa and M. Toyoda, "Analytics for info-plosion including information diffusion studies for the 3.11 disaster," in *Web-Age Information Management*, 2011, vol. 6897.
- [25] M. A. Cameron, R. Power, B. Robinson, and J. Yin, "Emergency situation awareness from twitter for crisis management," in *WWW Companion*, 2012.
- [26] X. Wang, F. Zhu, J. Jiang, and S. Li, "Real time event detection in twitter," in *Web-Age Information Management*, 2013, vol. 7923.
- [27] T. Hua, F. Chen, L. Zhao, C.-T. Lu, and N. Ramakrishnan, "STED: semi-supervised targeted-interest event detection in twitter," in KDD, 2013.
- [28] S. Kumar, F. Morstatter, R. Zafarani, and H. Liu, "Whom should i follow?: identifying relevant users during crises," in HT, 2013.
- [29] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak, "Spatial variation in search engine queries," in WWW, 2008.
- [30] Q. Mei, C. Liu, H. Su, and C. Zhai, "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," in WWW, 2006.
- [31] K. Radinsky and E. Horvitz, "Mining the web to predict future events," in WSDM, 2013.
- [32] V. Guralnik and J. Srivastava, "Event detection from time series data," in KDD, 1999.
- [33] N. J. Horton and G. M. Fitzmaurice, "Regression analysis of multiple source and multiple informant data from complex survey samples," *Statistics in Medicine*, vol. 23, 2004.
- [34] I. Huopaniemi, T. Suviavaara, J. Nikkilä, M. Orešič, and S. Kaski, "Multivariate multi-way analysis of multi-source data," *Bioinformatics*, vol. 26, no. 12, 2010.
- [35] P. H. Swain, J. A. Richards, and T. Lee, "Multisource data analysis in remote sensing and geographic information processing," Purdue University, Tech. Rep., 1985. [Online]. Available: <http://docs.lib.purdue.edu/larstech/78/>
- [36] N. Milanovic and M. Malek, "Current solutions for web service composition," *IEEE Internet Computing*, vol. 8, no. 6, 2004.

- [37] I. Constantinescu, B. Faltings, and W. Binder, "Large scale, type-compatible service composition," in *ICWS*, 2004.
- [38] H. L. Truong, M. Comerio, A. Maurino, S. Dustdar, F. D. Paoli, and L. Panziera, "On identifying and reducing irrelevant information in service composition and execution," in *WISE*, 2010.
- [39] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *CIKM*, 2010.
- [40] E. A. Sultanik and C. Fink, "Rapid geotagging and disambiguation of social media text via an indexed gazetteer," in *ISCRAM*, 2012.
- [41] F. Sebastiani, "Text categorization," in *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, 2005.
- [42] P. Wang, C. Domeniconi, and J. Hu, "Cross-domain text classification using wikipedia," *IEEE Intelligent Informatics Bulletin*, vol. 9, no. 1, 2008.
- [43] E. Lex, C. Seifert, M. Granitzer, and A. Juffinger, "Efficient cross-domain classification of weblogs," *International Journal of Intelligent Computing Research*, vol. 1, no. 1, 2010.
- [44] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in *WWW*, 2010.
- [45] Y. Zhen and C. Li, "Cross-domain knowledge transfer using semi-supervised classification," in *AI 2008: Advances in Artificial Intelligence*, 2008, vol. 5360.
- [46] D. Wang, D. Irani, and C. Pu, "Evolutionary study of web spam: Webb spam corpus 2011 versus webb spam corpus 2006," *Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom)*, 2012 8th International Conference Conference on, pp. 40–49, 2012.
- [47] D. Wang, D. Irani, and C. Pu, "Spade: a social-spam analytics and detection framework," *Social Network Analysis and Mining*, vol. 4, no. 1, 2014.
- [48] D. Wang, D. Irani, and C. Pu, "A perspective of evolution after five years: A large-scale study of web spam evolution," *International Journal of Cooperative Information Systems*, vol. 23, no. 02, June 2014.



Calton Pu received his PhD from University of Washington in 1986 and served on the faculty of Columbia University and Oregon Graduate Institute. Currently, he is holding the position of Professor and John P. Imlay, Jr. Chair in Software in the College of Computing, Georgia Institute of Technology. He has worked on several projects in systems and database research. He has published more than 70 journal papers and book chapters, 200 conference and refereed workshop papers. He served on more than 120 program committees. His recent research has focused on big data in Internet of Things, automated N-tier application deployment and denial of information. He is a senior member of the IEEE and a member of the ACM.



Aibek Musaev received the BS and MS degrees in Computer Science from Georgia Institute of Technology in 1999 and 2000, respectively. He was an Application Engineer at Siebel Systems, Inc. until 2002. In 2002 he founded a software startup Akforta LLC that grew to over 30 people in staff. He is currently working toward the PhD degree in the College of Computing at Georgia Institute of Technology. His research interests are in the area of social media mining, with a current focus on natural disaster detection. He is

a student member of the IEEE.



De Wang De Wang received his bachelor's degree in Software Engineering with outstanding honor and his master's degree in Computer Software and Theory from Jinan University, Guangdong, China in 2007 and 2010 respectively. And he received his PhD from Georgia Institute of Technology in 2014. His dissertation research is in the area of cyber security with applications in systems and data analytics under the guidance of Prof. Calton Pu in the Distributed Data Intensive Systems Lab (DISL) and Center for

Experimental Research in Computer Systems (CERCS). He is currently working as a software engineer at Google Inc.