Demo Abstract:

Free-form Text Summarization in Social Sensing

Hongzhao Huang, Sam Anzaroot, Heng Ji
City University of New York
{hengjicuny}@gmail.com

Hieu Khac Le, Dong Wang, Tarek Abdelzaher
University of Illinois at Urbana-Champaign
{zaher}@illinois.edu

ABSTRACT

This demonstration illustrates an information aggregation and summarization service for social sensing applications. Social sensing, using mobile phones and other networked devices in the possession of individuals, has gained significant popularity in recent years. In some cases, the information collected is structured, such as numeric data from temperature sensors, accelerometers, or GPS devices. Aggregate statistical properties, such as expected values, standard deviations, and outliers, can be easily computed, and can be used to summarize the data set. In other cases, however, the collection includes unstructured data types such as text or images with textual annotations. The concepts of expected values and outliers are harder to define, yet it is still important to be able to aggregate and summarize the data. We demonstrate a system which can automatically summarize real-time textual data common to social sensing applications. Specifically, we focus on text messages that describe events in the environment. The output of our service provides a reliable summary of observations that can be used in many contexts from military intelligence to participatory sensing campaigns.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Social Sensing, Information Summarization, Free-form Text

1. INTRODUCTION

Our work is motivated by the emerging needs of social sensing applications. It is clear that significant enhancements are possible to the capabilities of social sensing systems if they were able to process and summarize unstructured data types. Consider, for example, a participatory sensing campaign that allows individuals to geo-tag locations of relevance to the campaign (e.g., using their GPS phone), and add a short annotation to explain the reason for geo-tagging. For example, in a "clean campus" campaign, individuals might be asked to geotag locations where they observed a need for maintenance or upkeep, and add a text message that explains the nature of the problem. For another example, in a military or peace-keeping scenario, a friendly local population might be encouraged to report observations of relevance to the military or peace-keeping mission in the form of text.

In this demonstration, we use microblogs as a way of sharing observations. Twitter already offers an easy way to report short observations, as well as a way to include pointers to visual and location information. It has exhibited advantages over traditional news agencies in the success of reporting news more timely, for instance, in reporting the Chilean earthquake of 2010 [1]. A comparative study [2] shows that Twitter users tend to seek more temporally relevant information than those of general web search engines. This makes Twitter an ideal vehicle for real-time dissemination of observations in social sensing applications.

As with numeric data types (such as temperature), it is useful to summarize a large body of reported textual observations in order to convey the information more efficiently. This is especially important in large-scale applications, where in the absence of summarization, users of the data collection service might be overwhelmed by the sheer amount of collected observations. The main contribution of the demonstrated service lies in the implementation of novel (i) clustering and (ii) ranking techniques for summarizing unstructured *textual* observations in social sensing. The components of the service are described below.

2. A DATA SUMMARIZATION SERVICE

Our service addresses the need for summarization in the context of free-form textual sensing data. The service has three components:

Clustering: In the case of numeric data, one way to summarize a data set would be to report statistical properties such as average and standard deviation. Since textual data does not have an inherent order (unlike numeric data types), in the text domain, the corresponding solution is to (i) perform clustering of data, and (ii) report representatives of different clusters. For example, if several individuals describe the same event, one should be able to cluster those observations and report one or a few representative descriptions only. Our service uses clustering as a way to significantly reduce the size of input data and offer a quick representative view of the entire observation set. Our contribution lies not

IPSN'12, April 16–20, 2012, Beijing, China.
ACM 978-1-4503-1227-1/12/04.

in how clusters are formed (which is borrowed from natural language processing), but how their credibility is evaluated. Cluster size alone is not a good indication of authenticity of data. For example, it does not eliminate rumors where, a significant number of observers may report the same event because they heard it from others.

Ranking: When the collected observations are noisy, it is important to eliminate statistical outliers, since they are more likely to represent measurement noise. This is straightforward to do in the case of numeric data, but is needed when handling unstructured data as well. Outlier elimination, in the context of unstructured data, requires algorithms for data ranking that help sort the observations by statistical likelihood, such that less likely ones can be ignored. Our ranking techniques are novel in opportunistically exploiting heterogeneous *relations* between observations. For example, if we happened to observe (from reported textual annotations or from a map) that individuals reporting litter at some locations were frequently at a stop for Bus 13, other bus stops on the same route may be considered more likely candidates for having litter as well. Finally, if individuals reporting an observation are not at the location of the observed event, they are likely not observing the event first-hand. Hence, understanding relations between observations, sources, and locations can improve ranking. In general, if observation attributes (such as locations and reported events) were nodes, and their different semantic relations were (heterogeneous) links, we call the resulting network an *information network*. Our main contribution lies in information-network-assisted summarization of textual sensing data.

Query engine: Given the generated summary view, we make it possible to answer queries about the data. For example, much like a person can ask for all locations where the average daily temperature is above a threshold, we allow a user to ask about all locations where excessive litter was commonly observed. The former is a standard query on a numeric data type. The latter, in contrast, requires processing of textual annotations of geo-tagged locations, clustering, ranking, and returning the best representatives of the most statistically credible clusters.

3. TECHNICAL APPROACH

Given a set of textual observations (in our implementation, we use Twitter to share the data), we cluster them by linking together those pairs with similarity larger than a threshold, then assign initial credibility weights to each cluster using an expectation maximization approach (published in this IPSN [3]).

We then exploit linguistic lexical, syntactic and semantic analysis to extract pre-specified types of facts and entities from written text, and convert them into structured representations (i.e., information networks). An information network is a heterogeneous network that includes a set of "information graphs" $G = G_i(V_i, E_i)$, where V_i is the collection of entity nodes, and E_i is the collection of edges linking one entity to the other, labeled by relation or event attributes, such as "hometown", "employer", or "spouse". We adopt the taxonomy defined in the ACE2005[1] Information Extraction system [4, 5], which includes 7 types of entities (such as persons, locations, and organizations), 18 types of relations, and 33 distinct types of events.

One can think of the resulting information network as one that encodes inter-dependencies among various types of observations, from which certain regularities, called *association rules*, are mined. These regularities can be thought of as latent constraints. We then rank the observations based on the following hypothesis: an observation is more likely to be correct if it's more consistent with other extracted regularities [6]. Observations that are the most inconsistent are removed.

4. SCRIPT OUTLINE

In the demonstration, the audience will be presented with raw textual data sets collected from observations of recent events of interest, such as Hurricane Irene, and recent riots. We shall ask the users to summarize what happened in those events. It will be evident that the raw amount of information received makes it impossible to do so in a reasonable amount of time, and in fact makes it easy to miss important observations in the deluge of raw reported data.

We shall then demonstrate our summarization service. The service will automatically cluster tweets, extract information units from tweet clusters, and construct an information network to express relations between such clusters. It will then mine the constructed information network for regularities (the association rules), and use these to rank the individual tweet clusters based on conformance with these regularities. A summary consisting of individual tweets representing the highest ranked clusters will then be displayed. This will be compared to other information summarization techniques that do not exploit relations between clusters. It will be shown that the new summarization technique offers a more relevant picture of what transpired, making it a good candidate for summarizing unstructured (textual) data in social sensing.

5. REFERENCES

[1] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proc. 1st Workshop on Social Media Analytics*, 2010.

[2] Jaime Teevan, Daniel Ramage, and Meredith Ringel Morris. Twittersearch: A comparison of microblog search and web search. In *Proc. WSDM11*, 2011.

[3] Dong Wang, Lance Kaplan, Hieu Le, and Tarek Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *ACM/IEEE IPSN*, 2012.

[4] Heng Ji, Ralph Grishman, Zheng Chen, and Prashant Gupta. Cross-document event extraction and tracking: Task, evaluation, techniques and challenges. In *Proc. Recent Advances in Natural Language Processing 2009*, 2009.

[5] Zheng Chen, Suzanne Tamang, Adam Lee, Xiang L, Wen-Pin Lin, Matthew Snover, Javier Artiles, Marissa Passantino, and Heng Ji. Cuny-blender tac-kbp2010 entity linking and slot filling system description. In *Proc. TAC 2010 Workshop*, 2009.

[6] Qi Li, Sam Anzaroot, Wen-Pin Lin, Xiang Li, and Heng Ji. Joint inference for cross-document information extraction. In *Proc. 20th ACM Conference on Information and Knowledge Management (CIKM2011)*, 2011.

[1] http://www.itl.nist.gov/iad/mig/tests/ace/