Towards Personalized Learning to Rank for Epidemic Intelligence Based on Social Media Streams

Ernesto Diaz-Aviles[1]
diaz@L3S.de

Avaré Stewart[1]
stewart@L3S.de

Edward Velasco[2]
velascoe@rki.de

Kerstin Denecke[1]
denecke@L3S.de

Wolfgang Nejdl[1]
nejdl@L3S.de

[1]L3S Research Center / University of Hannover. Hannover, Germany
[2]Robert Koch Institute (RKI). Berlin, Germany

ABSTRACT

In the presence of sudden outbreaks, how can social media streams be used to strengthen surveillance capabilities? In May 2011, Germany reported one of the largest described outbreaks of *Enterohemorrhagic Escherichia coli* (EHEC). By end of June, 47 persons had died. After the detection of the outbreak, authorities investigating the cause and the impact in the population were interested in the analysis of micro-blog data related to the event. Since Thousands of tweets related to this outbreak were produced every day, this task was overwhelming for experts participating in the investigation. In this work, we propose a Personalized Tweet Ranking algorithm for Epidemic Intelligence (PTR4EI), that provides users a personalized, short list of tweets based on the user's context. PTR4EI is based on a learning to rank framework and exploits as features, complementary context information extracted from the social hash-tagging behavior in Twitter. Our experimental evaluation on a dataset, collected in real-time during the EHEC outbreak, shows the superior ranking performance of PTR4EI. We believe our work can serve as a building block for an open early warning system based on Twitter, helping to realize the vision of *Epidemic Intelligence for the Crowd, by the Crowd*.

Categories and Subject Descriptors: H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–*Information Filtering*; K.4 [**Computer and Society**]: General

General Terms: Algorithms, Performance, Experimentation

Keywords: Learning to Rank, Recommender Systems, Twitter

1. INTRODUCTION

Epidemic Intelligence (EI) encompasses activities related to early warning functions, signal assessments and outbreak investigation. Only the early detection of disease activity, followed by a rapid response, can reduce the impact of epidemics. Recent works have shown the potential of using Twitter for public health, and its real-time nature makes it even more attractive for public health surveillance. These works have either focused on: the text classification and filtering of tweets, e.g., [8]; or finding predictors for diseases that exhibit a seasonal pattern (i.e., influenza-like illnesses) by correlating selected keywords with official influenza statistics and rates, e.g., [7]. Still others have focused on mining Twitter content for topic and aspect modeling [4]. Furthermore, these existing approaches have all focused on countries where the tweet density is known to be high (e.g., the UK, or U.S.). In contrast these studies, ours focuses on a sudden outbreak of a disease that does not involve any seasonal pattern. Moreover, our work shows the potential of Twitter in countries where the tweet density is significantly lower, such as Germany [6].

This paper contributes an innovative personalized ranking approach that offers decision makers the most relevant and attractive tweets for *outbreak analysis and control*, by exploiting the social hash-tagging behavior in Twitter. A complementary analysis of this work, exploring the potential of Twitter for outbreak detection, is discussed in [2].

2. PERSONALIZED RANKING FOR EI

We will rank and derive a short list of tweets based on limited user context information. The user context C_u is defined as a triple $C_u = (t, MC_u, L_u)$, where t, MC_u, L_u are, respectively, the discrete time interval, set of medical conditions (e.g., disease or symptom), and set of locations of user interest during the investigation of an outbreak.

Our learning approach, PTR4EI, is shown in Algorithm 1. We build upon a learning to rank framework by considering a personalized setting that exploits user's individual context [3].

More precisely, we consider the context of the user, C_u, and prepare a set of queries, Q, for a target event (e.g., a disease outbreak). We extract the hash-tags that co-occur with the user context, by considering the medical conditions and locations in C_u as hash-tags themselves, and finding which other hash-tags co-occur with them within a tweet. We use an indexed collection of tweets for epidemic intelligence ($\mathcal{T}$), where not all tweets are necessarily interesting for the target event. The set Q is constructed by *expanding*[1] the original terms in C_u with the co-occurring hash-tags, which are previously classified as medical condition, location or complementary context entities.

Then, we build a set D of tweets by querying index $\mathcal{T}$ using $q \in Q$ as query terms. Next, we elicit judgments

[1]This phase of the algorithm can be considered a particular case of the *query expansion* task in information retrieval, where search terms are named entities (i.e., medical conditions, locations, name of people, organizations, etc.), whose implicit correlations are discovered in the reduced dimensional space induced by the top co-occurring hash-tags.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

Algorithm 1 Personalized Tweet Ranking algorithm for Epidemic Intelligence (PTR4EI)

Input: User Context C_u, and an index $\mathcal{T}$ of tweets
Output: Ranking Function f_{C_u} for user context

1: Consider each $mc \in MC_u$ as a hash-tag, and extract from $\mathcal{T}$ all co-occurring hash-tags: $coHashTags$
2: Classify the hash-tags in $coHashTags$ as Medical Condition MC_x, Location L_x or Complementary Context[2] CC_x
3: Build a set of queries as follows: $Q = \{q \mid q \in MC_u \times \mathcal{P}(\{L_u \cup MC_x \cup L_x \cup CC_x\})\}$
4: For each query $q_i \in Q$ obtain tweets D from the collection $\mathcal{T}$
5: Elicit relevance judgments Y on a subset $D_y \subset D$
6: For each tweet $d_j \in D$, obtain the feature vector $\phi(q_i, d_j)$ w.r.t. $(q_i, d_j) \in Q \times D$
7: Apply learning to rank to obtain a ranking function for the user context C_u: $f_{C_u}(q, d) = \vec{w} \cdot \phi(q, d)$
8: **return** $f_{C_u}(q, d)$

from experts on a subset of the tweets retrieved, in order to construct $D_y \subset D$. We then obtain for each tweet $d_j \in D$ its features vector $\phi(q_i, d_j)$ with respect to the pair $(q_i, d_j) \in Q \times D$. Finally, with these elements, we apply a learning to rank algorithm to obtain the ranking function for the given user context. The ranking function is applied to rank existing and new incoming tweets.

In the rest of the section, we evaluate our approach considering as event of interest the EHEC outbreak in Germany, 2011.

Experiments and Evaluation

To support users in the assessment and analysis during the EHEC outbreak, we set the user context as $C_u = (t, MC_u, L_u) = ([\{2011\text{-}05\text{-}23; 2011\text{-}06\text{-}19\}], \{\text{“EHEC”}\}, \{\text{“Lower Saxony”}\})$, in this way, we are taking into account the main period of the outbreak, the disease of interest, and the German state having the most reported cases.

Following Algorithm 1, we computed the co-occurring hash-tags using an indexed collection $\mathcal{T}$ of 7,710,231 tweets collected during May and June, 2011. Table 1 presents an example of hash-tags co-occurring with *#EHEC*.

Three experts, one from the Robert Koch Institute, Germany's federal institution responsible for disease control and prevention, and two from the Lower Saxony State Health Department, provided their individual judgment on a subset D_y of 150 tweets, evaluating for each tweet, if it was relevant or not to support their analysis of the outbreak. Any disagreement in the assigned relevance scores were resolved by majority voting.

For each tweet, we prepared five binary features whose corresponding value was set to *true* if a medical condition, location, hash-tag, complementary context term, or URL were present in the tweet, and *false* otherwise. For learning the ranking function, we used Stochastic Pairwise Descent algorithm [5].

We compared our method against a vector space model based on TF-IDF scores. For evaluation, we used *Precision at Position n* (*P@n*), that measures the relevance of the top n documents in the ranking list with respect to a given query, e.g., “EHEC in Lower Saxony” [1].

[2]**Complementary Context** CC is defined as the set of nouns, which are neither Locations nor Medical Conditions, e.g., names of persons, organizations or affected organisms. $CC \cap (L \cup MC) = \emptyset$

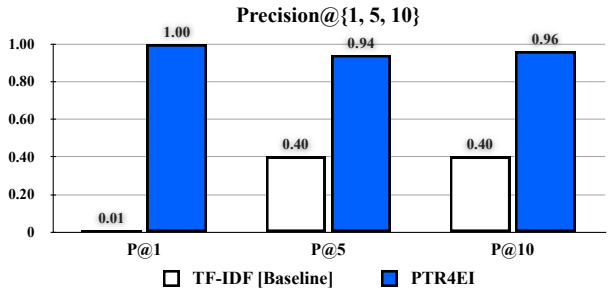

Figure 1: Ranking Performance in terms of Precision

Week 21: May 23 – 29, 2011			
Medical Condition	**Location**	**Complementary Context**	
bacteria	bremen	cucumber_salad	cdu
diarrhea	cuxhaven	cucumbers	edeka
ehec_victim	hamburg	ehec_vegetable	fdp
hus	münster	tomatoes	merkel
intestinal_infection	northern_germany	vegetables	rki

Table 1: Example of hash-tags co-occurring with *#EHEC* during the first week of the outbreak.

We randomly split the dataset into 80% training tweets, which will be used to compute the ranking function, and 20% testing tweets. To reduce variability, we performed the experiment using ten different 80/20 partitions. The reported performance is the average over the ten rounds on the test set. The ranking results are shown in Figure 1, where we can clearly see the superior performance of PTR4EI.

3. CONCLUSION

We have shown the potential of Twitter to support the task of outbreak analysis and control, and empirically demonstrated how personalized ranking for epidemic intelligence can be achieved. Currently we are working closely with German and international public health institutions to help them integrate monitored social media into their existing surveillance systems. We hope that this paper provides some insights into the future of epidemic intelligence based on social media streams.

Acknowledgments This work was funded, in part, by the European Commission FP7/2007-2013 under grant agreement No.247829 for the M-Eco Medical Ecosystem Project.

4. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 2nd edition, 2011.
[2] E. Diaz-Aviles, A. Stewart, E. Velasco, K. Denecke, and W. Nejdl. Epidemic Intelligence for the Crowd, by the Crowd (full version). http://arxiv.org/, 2012.
[3] T.-Y. Liu. Learning to Rank for Information Retrieval. *Found. Trends Inf. Retr.*, 3:225–331, March 2009.
[4] M. J. Paul and M. Dredze. You Are What You Tweet: Analyzing Twitter for Public Health. In *ICWSM'11*, 2011.
[5] D. Sculley. Large Scale Learning to Rank. In *NIPS 2009 Workshop on Advances in Ranking*, Dec. 2009.
[6] Semiocast. Countries on Twitter. http://goo.gl/RfxZw, 2012.
[7] A. Signorini, A. M. Segre, and P. M. Polgreen. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE*, 2011.
[8] M. Sofean, A. Stewart, K. Denecke, and M. Smith. Medical Case-Driven Classification of Microblogs: Characteristics and Annotation. In *ACM IHI 2012*, 2012.